Técnicas de Aprendizaje Profundo para el Análisis de la Percepción de la Seguridad Urbana

Felipe Adrian Moreno Vera

Orientador: Dr. Jorge Luis Poco Medina

Jurado:
Dr. José Eduardo Ochoa Luna – Universidad Católica San Pablo – Perú
Dr. Guillermo Cámara Chavez – Universidade Federal de Ouro Preto – Brasil
Dr. Rensso Victor Hugo Mora Colque – Universidad Católica San Pablo – Perú

Tesis presentada al
Departamento de Ciencia de la Computación
como parte de los requisitos para obtener el grado de
Maestro en Ciencia de la Computación.

Universidad Católica San Pablo – UCSP
Mayo de 2024 – Arequipa – Perú

Este trabajo esta dedicado a mis padres, hermanos y demás familiares que siempre han estado apoyando y depositando su confianza en mi, siempre estarán conmigo.

Agradecimientos

Agradezco el apoyo y orientación que recibí durante todo este tiempo, padres, hermanos, orientador y profesores en general. También al **Consejo Nacional de Ciencia, Tecnología e Innovación Tecnológica** (CONCYTEC) y al **Fondo Nacional de Desarrollo Científico, Tecnológico e Innovación Tecnológica** (FONDECYT), que mediante convenio de Gestión 234-2015-FONDECYT, han permitido la subvención y financiamiento de mis estudios de Maestría en Ciencia de la Computación en la **Universidad Católica San Pablo** (UCSP). La cual pudo brindarme un espacio tan variado en conocimiento y personas con las cuales interactuar y poder plantear, desarrollar e implementar el presente trabajo. Así como también a la **Fundación Getúlio Vargas**, Rio de Janeiro, Brasil; en la cual desempeñé una labor de investigador utilizando el ambiente computacional proporcionado durante mi estadía en Brasil.

Abreviaturas

API *Application Programming Interfaces*

AUC *Area Under Curve*

CAM *Class Activation Map*

CNN *Convolutional Neural Network*

DCNN *Deep Convolutional Neural Network*

DeCAF *Deep Convolutional Activation Feature*

DSVR *Direction based Street View Retrieval*

FOV *Field of View*

GAP *Global Average Pooling*

GAN *Generative Adversarial Networks*

GBP *Guided Back-Propagation*

GEH *Global Edge Histogram*

GSV *Google Street View*

GVI *Green View Index*

HDI *Human Development Index*

HDMiR *Hierarchical Deep Multi-instance Regression*

HOG *Histogram of Oriented Gradients*

HPM *Human Perception Mapping*

KNN *K-Nearest Neighbors*

MLR *Multi Linear Regressor*

MTDRALN *MultiTask Deep Relative Attribute Learning Network*

PRN *Perception Rank Network*

RBF *Radial Basis Function*

RGB *Red-Green-Blue*

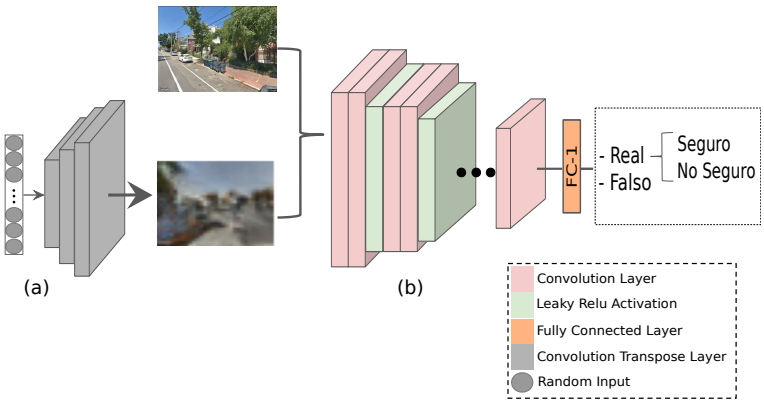
SG *SmoothGrad*

SAPN *Semantic-Aware Perception Network*

SVM *Support Vector Machine*

SVR *Support Vector Regressor*

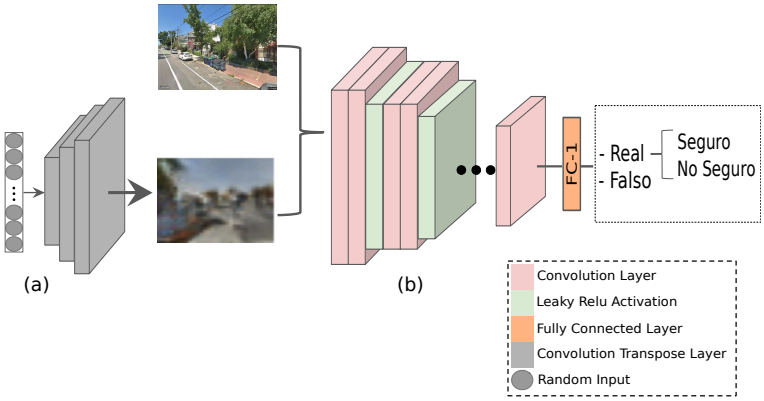
Abstract



Perception is how humans interpret and understand information from some environment. This information is captured after interacting with the environment that surrounds them, learning new experiences, or reinforcing others already lived. The perception of urban security can be described in how humans present a reaction to a particular stimulus from the visual appearance or prior knowledge of a specific place (streets, urban areas, etc.). Based on the previous idea, various studies sought to describe this phenomenon. A very notable example is the theory called *"The Broken Window"* which studied the behavior of people in environments whose visual appearance was chaotic. Likewise, recently this study has been implemented using various types of data, not only limited to surveys or social experiments, to determine the relationship between urban perception and intrinsic characteristics of cities. Which is one of the most noteworthy data sets is *Place Pulse*. In this work, we propose a methodology that allows the analysis and data exploration of *Place Pulse 2.0*. As the main results, we present an exploratory data set analysis, highlighting behavior and outliers. Besides, we show the comparison and training results of supervised and semi-supervised GAN-based models against other techniques. We are showing that our Semi-Supervised GAN approach presents better results in metrics and stability in dealing with this kind of data.

Keywords: *Deep Learning*, *Convolutional Neural Networks*, *GAN*, *features extraction*, *urban perception*.

Resumen



La percepción es la forma en que los humanos interpretan y comprenden la información captada después de la interacción con el entorno que les rodea, aprendiendo nuevas experiencias o reforzando otras ya vividas. La percepción de la seguridad urbana se puede describir en cómo los humanos presentan una reacción ante un determinado estímulo proveniente de la apariencia visual o conocimiento previo sobre un cierto lugar (calles, zonas urbanas, etc). A partir de esta idea, diversos estudios buscaron describir dicho fenómeno teniendo como ejemplo más notable la teoría denominada *"The Broken Window"*, la cual estudiaba el comportamiento de las personas frente a ambientes cuya apariencia visual era caótica. Así mismo, recientemente este estudio está siendo implementado utilizando diversos tipos de datos, no solo limitándose a encuestas o experimentos sociales, con el objetivo de determinar la relación entre la percepción urbana y características intrínsecas de los ciudades; de los cuales, uno de los conjuntos de datos más resaltables es *Place Pulse*. En este trabajo, se propone una metodología que permita analizar y explorar los datos de *Place Pulse 2.0*. Como resultados principales, presentamos un análisis exploratorio de los datos, resaltando la organización y comportamiento de los datos. Además, presentamos una comparación entre diferentes técnicas de aprendizaje supervisado y semi-supervisado. Mostrando que un modelo *Generative Adversarial Networks* (GAN) presenta mejores resultados que técnicas convencionales.

Keywords: *Deep Learning*, *Convolutional Neural Networks*, *GAN*, *features extraction*, *urban perception*.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

En este capítulo describimos las principales motivaciones que hemos tenido para el desarrollo de nuestro trabajo. Además también definiremos el problema que se pretende abordar y los objetivos que se pretende alcanzar en este trabajo.

1.1. Contexto y Motivación

"Las ciudades son diseñadas para moldear e influir en la vida de sus habitantes" (Lindal y Hartig, 2013). Diversos estudios han demostrado que la apariencia visual de las ciudades juega un papel central en la percepción humana y la reacción ante dicho entorno. Un ejemplo notable es la teoría de la ventana rota ("*The Broken Window*") (Wilson y Kelling, 1982) que sugiere que los signos visuales (apariencia) de un trastorno ambiental, tales como ventanas rotas, automóviles abandonados, basura y grafitis, pueden inducir resultados sociales negativos y aumentar los niveles de delincuencia. Esta teoría ha tenido una gran influencia en las estrategias de política pública que conducen a tácticas policiales agresivas para controlar las manifestaciones del desorden social y físico. Por ejemplo, en el estudio realizado en la ciudad de *New York* (Keizer et al., 2008), se realizaron experimentos sociales sobre la calidad de vida percibida en las calles. Estos experimentos comparaban lugares "impecables" como centros comerciales (paredes limpias, ordenado y tranquilo) con otros lugares en los cuales se contaba con la presencia de grafitis, calles antiguas o descuidadas y basura en las calles; concluyendo que en lugares en donde "se vulneran las reglas" conlleva a que a largo plazo las normas sociales no sean respetadas. Como conclusión, se determina que la apariencia visual descuidada influencia negativamente al entorno (p.ej. grafitis, basura esparcida por las calles, poca limpieza del entorno, etc.).

Así mismo, otros estudios han demostrado que el aspecto visual de los espacios de una ciudad afectan el estado psicológico de sus habitantes (Lindal y Hartig, 2013). Por ejemplo, a través de una evaluación psicológica se demostró que la presencia de áreas verdes en ciudades tienden a producir sensaciones positivos en los habitantes de

una ciudad por ejemplo seguridad, relajación, tranquilidad, etc. (Ulrich, 1979). Por el lado contrario, a través del estudio de 40 reportes psicológicos basados en encuestas y estudios de los estados mentales de sus habitantes, también se logró deducir que el desorden urbano induce angustia psicológica, estrés y miedo constante (Sampson et al., 2002). Además, también ha quedado demostrado que los grafitis y edificios en mal estado o abandonados están directamente relacionados a la percepción de inseguridad (Schroeder y Anderson, 1984).

Por lo tanto, a través de diversos estudios sobre el impacto de la apariencia visual de una ciudad en sus habitantes, se vuelve de vital importancia comprender las percepciones y evaluaciones de los espacios urbanos de las personas. En ese sentido, se han realizado diversos estudios acerca de cómo la ciudad y su apariencia visual influye en el comportamiento de la ciudad. Por ejemplo, en "*The image of the city*" (Lynch, 1984) se dividieron ciudades (p.ej. Boston, Jersey y Los Ángeles) en regiones de importancia (basados en datos de crímenes, sociedad, secciones urbanas o no urbanas, etc.) generando mapas mentales acerca de las características en común de estas ciudades; indicando que los elementos de cada ciudad se distinguen entre cientos, miles o millones de otros artefactos debido a sus formas únicas, tamaños, colores, etc. A partir de este conjunto de estudios, en el aspecto psicológico, comenzó una tendencia a estudiar y evaluar la percepción de los habitantes con respecto a los elementos visuales de la ciudad. En trabajo realizado por Nasar (1998), el cual estaba fuertemente relacionado a encontrar aquellas regiones/áreas de mayor agrado para los ciudadanos; demostró que en la mayoría de las evaluaciones siempre predominaban las áreas verdes, calles temáticas, espacios abiertos, centros comerciales y aeropuertos. Además, las áreas que fueron calificadas como "no agradable" para los habitantes fueron edificios con estilos poco atractivos, presencia de grafitis, parques sin una forma establecida y lugares abandonados. Además, otros estudios relacionados al desorden de la ciudad (Skogan, 1992) enfocados a la presencia de basura en las calles, edificios y carros abandonados o estacionados en esquinas desoladas; contribuyen a la percepción de descontrol, miedo e inseguridad ciudadana.

Por este motivo, para comprender el comportamiento de la criminalidad y la sensación de inseguridad se han realizado estudios acerca de la influencia de los delitos y crímenes en las calles, los cuales han ido incrementando en lugares muy concurridos (p.ej. turísticos). Estos delitos a largo plazo tienen un impacto negativo en cómo los posibles turistas perciben el nivel de seguridad de dichos lugares (Mawby, 2014; Mohammed y Sookram, 2015; Glaeser et al., 2018). Adicionalmente, con el paso de los años, se ha ido recopilando información acerca de los índices de criminalidad de diversas ciudades y sus tendencias, tal como el sitio web Numbeo (Numbeo, 2019), que nos informa acerca de que los índices de criminalidad en todos los países. Como dato curioso nos muestra que América del Sur se mantiene por encima de Asia y Europa en niveles de criminalidad (p.ej. Caracas-Venezuela está primero en el listado). Entonces, partiendo del hecho que contamos con un índice de criminalidad por ciudad se desarrollan diversas aplicaciones como mapas de criminalidad (USA, 2012; Google-Motorolla, 2019), estadísticas de datos (EuroStat, 2016) o aplicaciones que predicen la tendencia criminal de las zonas (Stalidis et al., 2018).

Todos estos estudios previos acerca del impacto de la apariencia visual y el índice de criminalidad en ciudades; han generado diversos enfoques para identificar qué elementos de la apariencia visual de una determinada calle influyen en la percepción urbana (Andersson et al., 2017). Por ejemplo, calidad de vida, áreas verdes, seguridad, entre otros. En los últimos años, con el avance de diversas técnicas para analizar información (p.ej. imágenes) y la evolución de técnicas como aprendizaje profundo (a partir de ahora *deep learning*); se ha evidenciado la creación de no solo reportes, sino también la creación de conjuntos de datos y la tendencia a la predicción de la percepción urbana. Un caso de ejemplo es el trabajo "*Which looks more safe*?" realizado por el *MIT-Media Lab*; creando el conjunto de datos *Place Pulse* (MIT-Media-Lab, 2013). Los datos registrados en *Place Pulse* son acerca de la percepción urbana de las personas a partir de una encuesta en línea realizada; en dicha encuesta un voluntario debe escoger entre dos imágenes de calles la más segura. Basado en dicho conjunto de datos, Li et al. (2015b); MIT-Media-Lab (2015) analizaron las áreas verdes y su influencia en la percepción urbana. Además, a través de técnicas como detección de objetos (p.eje. grafitis) y adición de otros conjuntos de datos, tales como tasas de criminalidad, niveles de violencia, presencia de árboles, índice de desarrollo humano, entre otros; fueron analizados en diversos trabajos tales como Porzi et al. (2015); Tokuda et al. (2019); Arietta et al. (2014); Li et al. (2015b) que detallaremos más adelante.

Así mismo, estudios acerca de la presencia de objetos y su correlación con la percepción de seguridad urbana también fueron realizados, mostrando que es posible dividir las ciudades a través de los tipos de objetos más frecuentes (p.ej. árboles, basura, edificios, cercas, grafitis, etc.) y la respectiva percepción de seguridad (Zhang et al., 2018; Min et al., 2019). Como se presenta brevemente en esta sección, existe una gran motivación por el estudio de la percepción urbana basado en las características y apariencia visual de imágenes de calles. Dichos estudios están basados en estadísticas recolectadas en un período de tiempo, así como también, están basados en modelos que utilizan estos datos estadísticos para realizar predicciones respecto a áreas influyentes, presencia de objetos (p.ej. grafitis, basura) o relaciones entre lugares y estadísticas de crímenes. Hemos identificado que en este tipo de estudios el **problema computacional radica en cómo identificar, diferenciar y relacionar las características de imágenes de las calles con la idea de percepción urbana**, debido a la similitud entre imágenes, la poca cantidad de muestras, etc. En nuestro trabajo nos enfocaremos en primer lugar explorar y analizar los datos de *Place Pulse 2.0* (el cual está compuesto de imágenes de calles). De esta forma, propondremos un modelo basado en *Deep Convolutional Neural Network* (DCNN) que nos permita solucionar las dificultades mencionadas en la tarea de predecir la percepción urbana de manera eficaz.

1.2. Objetivos

En esta sección presentaremos de manera concisa los objetivos del presente trabajo. La motivación principal está basado en la investigación sobre cómo predecir la percepción de seguridad urbana. A través del conjunto de datos *Place Pulse 2.0* pro-

ponemos una metodología que nos permita realizar esta tarea. Por lo cual presentamos a continuación los objetivos planteados en el presente trabajo:

1.2.1. Objetivo general

El objetivo general de este trabajo es describir, explorar, analizar y presentar las evaluaciones de diferentes modelos basados en DCNN para realizar una predicción de la percepción de seguridad ciudadana (p.ej. seguro y no seguro) utilizando el conjunto de datos *Place Pulse 2.0* previamente mencionado.

1.2.2. Objetivos específicos

De manera particular, podemos listar los objetivos específicos mencionados brevemente en el objetivo principal:

- Proponer una metodología que nos permita explorar y analizar el conjunto de datos de *Place Pulse 2.0* (Dubey et al., 2016), el cual se compone de 1.22 millones de comparaciones de 111 390 imágenes de 56 ciudades diferentes, conteniendo seis categorías diferentes de comparación *safety*, *lively*, *beautiful*, *wealthy*, *boring*, *depressing* (correspondientes a seguro, bueno para vivir, bonito, opulento, aburrido, aburrido y depresivo). Así mismo, de analizar y presentar las características y comportamiento de los datos. Con el objetivo de buscar posibles limitaciones que puedan estar presentes en el conjunto de datos (se discutirá a detalle más adelante en el presente documento), la categoría principal de estudio será la de *safety* (seguridad).
- Proponer y presentar un modelo basado en *Convolutional Neural Network* (CNN) para la tarea de clasificación. Este modelo nos permitirá diferenciar de manera eficiente la percepción urbana de una calle, teniendo en cuenta el comportamiento y distribución de los datos previamente analizados. Para las evaluaciones utilizaremos diferentes enfoques, tales como *transfer-learning*, *fine-tuning* y *generative adversarial networks*.

1.3. Contribuciones del Trabajo

El presente trabajo presenta 2 contribuciones principales. Siendo la primera contribución el estudio y análisis de *Place Pulse 2.0* cuyos datos son imágenes de 56 ciudades asociadas a una puntuación de percepción urbana (p.ej. seguridad). El objetivo de este estudio es explorar y analizar todas las características y distribución de los datos. El análisis expondrá los criterios a utilizar para dividir nuestros datos entre las categorías segura y no segura; así como también, se estudiará si es posible realizar

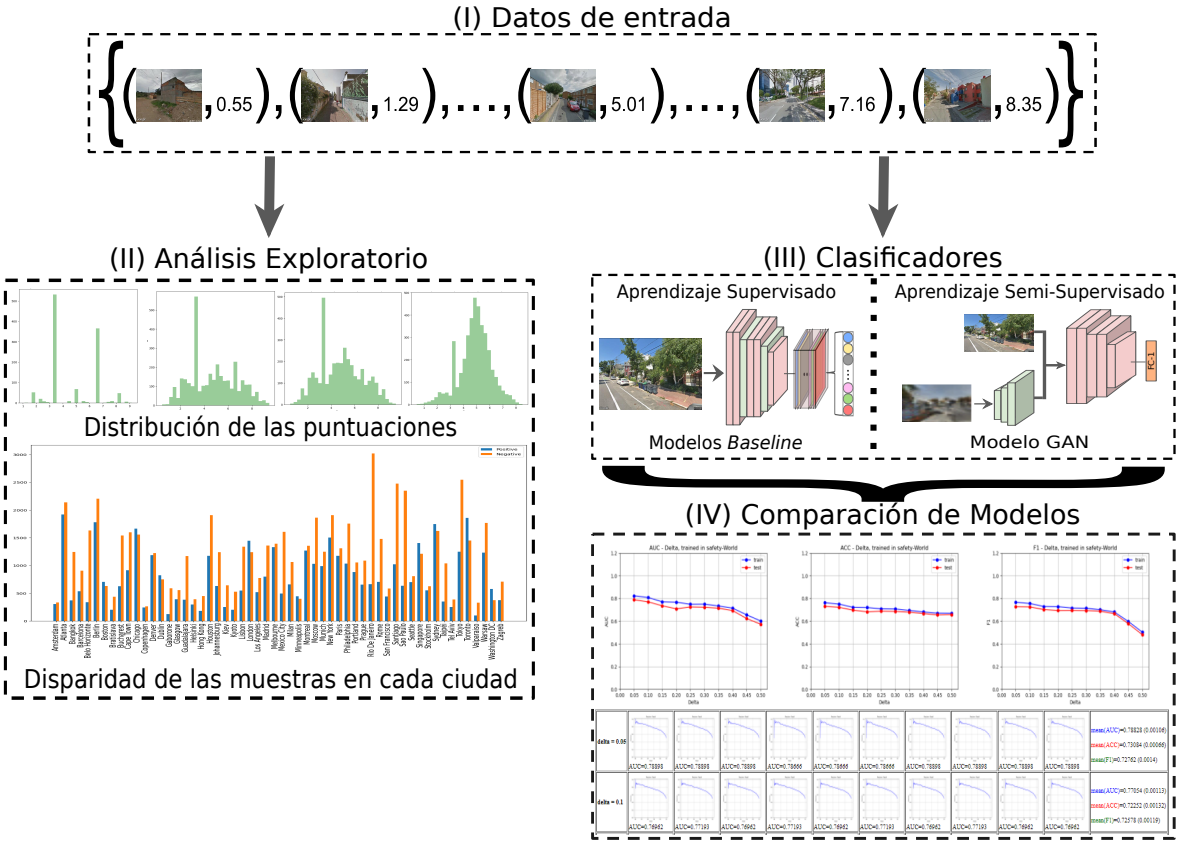


Figura 1.1: Metodología del trabajo: Se presenta de manera general los pasos a seguir para obtener los objetivos específicos descritos. Comenzando desde el conjunto de datos correspondientes a las imágenes y las puntuaciones asociadas a cada una, dicho conjunto se encuentra en (I); en (II) mostramos parte del análisis exploratorio realizado sobre los datos, resaltando resultados relevantes en nuestro análisis, tales como la distribución de las puntuaciones y la disparidad de los datos; en (III) se realiza el entrenamiento y seguimiento de los modelos a través de los enfoques supervisado y semi-supervisado; finalmente en (IV) reportamos y comparamos los resultados de las evaluaciones obtenidas, representándolos a través de gráficas respecto a los valores de las métricas evaluadas. Este esbozo tiene como objetivo darle al lector la idea general del presente trabajo, los cuales se detallarán más adelante en los próximos capítulos. Fuente: El autor.

una división en regiones a través de la percepción en diferentes "nivel de generalización geográfica" tales como ciudad, país, continente y global.

La segunda contribución corresponde a un modelo basado en DCNN, el cual será evaluado utilizando diversas técnicas y enfoques, tales como Aprendizaje Supervisado y Aprendizaje Semi-Supervisado. Este modelo será capaz de diferenciar, relacionar e identificar las características de las imágenes y realizar la predicción de la percepción urbana. Esto se expondrá en mayor detalle en el Capítulo 4 y 5, en los cuales se presentarán la descripción de las técnicas y modelos; además de los resultados obtenidos.

En la Figura 1.1 se muestra la metodología implícita que abarca todo el trabajo

realizado y presentado en este documento. Esta metodología nos permite realizar ambas contribuciones previamente descritas, comenzando desde el cálculo y agrupación de las imágenes entre las dos categorías estudiadas (segura y no segura) a través de las puntuaciones asociadas a cada una, en (I) observamos dicho conjunto de datos con cada imagen y su puntuación asociada. A partir de esas puntuaciones, en (II) realizamos un análisis exploratorio de los datos con la intención de entender el comportamiento de los datos estudiados, mostrando como resultado la disparidad de datos y distribuciones de las puntuaciones asociadas obtenidas. Este resultado corresponde al artículo publicado en Felipe Moreno-Vera (2021b). En (III) se realizan los respectivos entrenamiento y validación de los datos en los diferentes enfoques basados en modelos CNN, para finalmente en (IV) reportar, comparar y mostrar las evaluaciones y métricas obtenidas en cada modelo. Cada uno de estos pasos serán explicados y discutidos en detalle en los próximos capítulos.

Finalmente, el presente trabajo cuenta con tres publicaciones realizadas: La primera publicación fue realizada en *IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT) '21*, la cual abarca todo el análisis de las limitaciones presente en este conjunto de datos, las cuales explicaremos en el Capítulo 3 en detalle. La segunda publicación fue realizada en la conferencia *IEEE Mexican International Conference on Artificial Intelligence (MICAI '21)*, en la cual se presenta un estudio enfocado al análisis de la correlación de la presencia de objetos y la percepción de seguridad urbana de imágenes de calles. La tercera publicación fue realizada en la conferencia *International Conference on Intelligent Computing (ICIC '21)*, donde se presentó algunos resultados preliminares del entrenamiento de los modelos supervisados y una comparación entre dos métodos de explicación de modelos (buscando las regiones de relevancia para la predicción). Pueden encontrarlas siguiendo las referencias:

- **Moreno-Vera, Felipe**, Bahram Lavi, and Jorge Poco. "Quantifying Urban Safety Perception on Street View Images". *In IEEE/WIC/ACM International Conference on Web Intelligence* (WI-IAT '21), December 14–17, 2021, Essedon, Australia. (Felipe Moreno-Vera, 2021b).
- **Moreno-Vera, Felipe**, Bahram Lavi, and Jorge Poco. "Urban Perception: Can We Understand Why a Street Is Safe?". *In Mexican International Conference on Artificial Intelligence* (MICAI '21), October 25-30, 2021, Mexico City, Mexico. (Felipe Moreno-Vera, 2021a).
- **Moreno-Vera, Felipe**. "Understanding Safety based on Urban Perception". *In International Conference on Intelligent Computing* (ICIC '21), August 12-15, 2021, Shenzhen, China. (Moreno-Vera, 2021).

1.4. Organización del Documento

En el Capítulo 2 presentamos los trabajos relacionados sobre (a) análisis de la percepción urbana y (b) extracción de características y componentes visuales y (c)

interpretación y visualización de características extraídas. En el Capítulo 3 presentamos el análisis exploratorio del conjunto de datos *Place Pulse 2.0*. En el Capítulo 4 describiremos las arquitecturas de los modelos que utilizaremos para los experimentos sobre los datos. En el Capítulo 5 presentamos y describimos los resultados de los entrenamientos y evaluaciones realizadas en base a nuestras hipótesis. En el Capítulo 6 se presentan las discusiones y limitaciones del presente trabajo. Finalmente en el Capítulo 7 se presentan las conclusiones obtenidas de nuestros resultados.

Capítulo 2

Trabajos Relacionados

Los estudios sobre la percepción urbana se han ido incrementando debido a que cada vez se cuenta con mayor información de referencia geográfica de calles y ciudades (p.ej. *Google Street View* (GSV)) y la búsqueda de una manera para poder determinar el nivel de percepción (p.ej. de seguridad) de las calles. Los trabajos relacionados se podrían agrupar en tres grandes bloques: (a) análisis de la percepción urbana y (b) extracción de características y componentes visuales; y (c) interpretación y visualización de características extraídas. A manera de introducción al lector, primero tendremos una sección de conceptos previos en donde abordamos los conocimientos mínimos requeridos para poder entender el contenido del documento.

2.1. Conceptos Previos

En esta sección exponemos algunos conceptos básicos y necesarios para entender este documento, Algunos conceptos a tratar son las técnicas de aprendizaje supervisado, aprendizaje no supervisado y aprendizaje semi-supervisado. Así como también, Interpretación de modelos y algunas las tareas de clasificación de imágenes, detección y segmentación de objetos.

2.1.1. Técnicas de Aprendizaje Automático

Explicaremos brevemente las técnicas de aprendizaje que serán mencionadas en el presente documento, las cuales son Aprendizaje Supervisado, Aprendizaje No Supervisado y Aprendizaje Semi-supervisado.

2.1.1.1. Aprendizaje Supervisado

Es un método de aprendizaje automático en el cual se tiene un conjunto de datos conformado por "datos de entrada" y "etiquetas" asociados, al cual 'podemos llamar "datos de entrenamiento" (Abu-Mostafa et al., 2012). Podemos definir al conjunto de datos de "n" muestras como $I_{supervisado} = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$ donde cada $x_i \in \mathbb{R}^d$ es el i-ésimo "vector de características" y le corresponde su respectiva etiqueta (clase) y_i (Wikipedia, b). Algunas tareas realizadas con esta técnica son clasificación binaria, clasificación de imágenes, análisis de sentimiento, regresiones lineales, entre otras.

2.1.1.2. Aprendizaje No Supervisado

Es un método de aprendizaje automático en el cual el conjunto de datos de entrenamiento no contiene ninguna etiqueta o información asociada (Abu-Mostafa et al., 2012). Para este tipo de aprendizaje, el conjunto de datos está formado solamente por $I_{no-supervisado} = \{x_1, x_2, \dots x_n\}$, donde cada $x_i \in \mathbb{R}^d$ es un "vector de características". Algunas tareas realizadas con esta técnica son agrupaciones (*clustering*), reducción de las dimensiones, detección de elementos extraños (*outliers*), entre otros.

2.1.1.3. Aprendizaje Semi-Supervisado

Es un método de aprendizaje automático en el cual se tiene un conjunto de datos compuesto de "n" muestras divididas en dos subconjuntos de la forma $I_{etiquetados} = \{(x_1, y_1), (x_2, y_2), \dots (x_p, y_p)\}$ e $I_{no-etiquetados} = \{x_1, x_2, \dots x_q\}$, tal que $I_{semi-supervisado} = I_{etiquetados} \cup I_{no-etiquetados}$; donde cada $x_i \in \mathbb{R}^d$, y_i son etiquetas/clases y $p + q = n$ el total de muestras. El objetivo principal de este tipo de aprendizaje es aprender representaciones relevantes de los datos (Goodfellow et al., 2016). Algunas tareas realizadas por esta técnica pueden ser aumento de datos, generación de datos, pseudo-etiqueta, así como también las tareas de aprendizaje supervisado y no supervisado.

2.1.2. Tareas de Aprendizaje Automático

Explicaremos brevemente las tareas de aprendizaje automático: clasificación de imágenes, detección y segmentación de objetos.

2.1.2.1. Clasificación de imágenes

La tarea de clasificación de imágenes es un problema de aprendizaje automático que define un conjunto de clases (objetos para identificar en imágenes) y entrena un

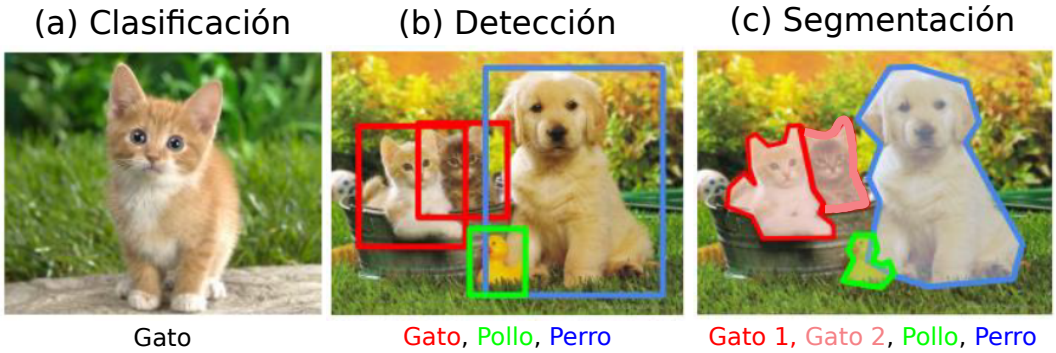


Figura 2.1: Mostramos las diferentes tareas en aprendizaje automático definidas: a) Clasificación: identifica las características de un gato en la imagen. b) Detección de objetos: identifica y localiza diferentes objetos dentro de una misma imagen. c) Segmentación de objetos: Identifica, localiza y recubre todos los píxeles en donde se encuentra cada objeto. Fuente: *Stanford-CS231 Deep Learning course* (CS231n, 2022).

modelo para reconocer los objetos usando etiquetas asociadas a cada imagen (Google-Developers, 2020). Algunas de los modelos más destacables son LeNet (Y. et al., 1990), AlexNet (Krizhevsky et al., 2012), ZFNet (Zeiler y Fergus, 2013a), GoogleNet (InceptionV1) (Szegedy et al., 2014), VGG-Net (Simonyan y Zisserman, 2014), ResNet (He et al., 2015), InceptionV3 (Szegedy et al., 2015), Xception (Chollet, 2017), entre otros.

2.1.2.2. Detección de objetos

La tarea de detección de objetos es una técnica de visión computacional para localizar instancias de objetos dentro de imágenes o vídeos (MatLab-Developers, 2020). Algunos métodos conocidos son R-CNN (Girshick et al., 2014), Fast R-CNN (Girshick, 2015), Faster R-CNN (Ren et al., 2017), *Single Shot MultiBox Detector* (SSD) (Liu et al., 2016) y *You Only Look Once* (YOLO) y derivados (hasta la versión actual 7) (Redmon et al., 2016).

2.1.2.3. Segmentación semántica y de instancias

La tarea de segmentación de imágenes es una técnica de agrupamiento de píxeles que pertenezcan a un mismo objeto dentro de una imagen, También llamado "clasificación a nivel de píxeles". En otras palabras, consiste en dividir una imagen en varias regiones (en grupos de píxeles) denominadas segmentos. (Viso-AI, 2020). Algunos métodos destacables son DeconvNet (Noh et al., 2015), U-Net (Ronneberger et al., 2015), DeepMask (Pinheiro et al., 2015), *Dilated Convolutions* (Yu y Koltun, 2015), *Pyramidal Scene Parsing Network* (PSP-Net) (Zhao et al., 2017), Mask RNN (Hu et al., 2017), Mask R-CNN (He et al., 2017), DeepLab (y derivados) (Chen et al., 2017), entre otros.

2.1.3. Modelos de Aprendizaje Automático

Explicaremos brevemente los modelos lineales y no lineales, así como también los principales componentes de cada uno. De manera adicional abarcamos interpretación de modelos.

2.1.3.1. Modelos Lineales

Sea un conjunto de datos de "n" muestras $I = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$ donde cada $x_i \in \mathbb{R}^d$ es llamada variable independiente y cada y_i asociado son variables dependientes. Un modelo lineal estudia la relación lineal entre variables dependientes y y las variables independientes x y predice valores de nuevas muestras. Estos modelos son llamados regresiones lineales, los cuales describen a la variable dependiente $y_i = f(x_i) = \sum_{j=1}^{d} \phi(x_{ij})\beta_j + \beta_0$, donde $\beta = (\beta_0, \beta_1, \dots, \beta_d) \in \mathbb{R}^{(d+1)}$ son las pendientes de cada variable $\phi(x_{ij})$. Así mismo, se observa la relación lineal entre cada y_i y x_i a través de las variables β y las funciones $\phi()$ que pueden ser funciones lineales o no (Wikipedia, a).

Una forma para evaluar la eficiencia de estos modelos, es utilizando el error obtenido de la función de costo definida de la forma $Err = L(y_i, f(x_i)) + R(\beta_i)$, donde $L(y_i, f(x_i))$ es la función de costo, $R(\beta_i) = \dfrac{(1-\rho)}{2}||\beta_i||_2 + \rho||\beta_i||_1$, donde $\rho \in [0, 1]$ es el parámetro de regularización. Además, se tiene que: (a) Si $\rho = 0$ es regularización **L2** o también llamado **Ridge** (Tikhonov, 1943), (b) Si $\rho = 1$ es regularización **L1** o también llamado **LASSO** (Tibshirani, 1996) y (c) Si $0 < \rho < 1$ es regularización **Elastic Net** (Zou y Hastie, 2005).

Así mismo, la función de costo $L(y, f(x))$ cambia dependiendo del tipo de tarea a realizar. En la tarea de clasificación tenemos las siguientes funciones de costo:

- El método **Logistic Regression** tiene la función de costo definida por $L(y, f(x)) = \sum_{i=1}^{n}[y_i \log(f(x_i)) + (1-y_i)\log(1-f(x_i))]$, esta función también es llamada **Binary crossentropy**.
- El método **Support Vector Classification** tiene la función de costo definida por $L(y, f(x)) = \sum_{i=1}^{n} \text{máx}(0, 1-y_i f(x_i))$, esta función también es llamada **hinge**. Así mismo, también puede ser definida como $L(y, f(x)) = \sum_{i=1}^{n} \text{máx}(0, 1 - y_i f(x_i))^2$ (**huber** o **squared hinge**)

En la tarea de regresión tenemos las siguientes funciones de costo:

- El método **Linear Regression** tiene la función de costo definida por $L(y, f(x)) = \sum_{i=1}^{n} ||y_i - f(x_i)||_2^2$, esta función también es llamada **Least Squares**.

- El método **Support Vector Regression** tiene la función de costo definida por $L(y, f(x)) = \sum_{i=1}^{n} \text{máx}(0, |y_i - f(x_i)|)$, esta función también es llamada **epsilon-sensitive**.

2.1.3.2. Modelos No Lineales

Un modelo no lineal estudia la relación no lineal entre la variables dependientes y_i y las variables independientes x_i. Estos modelos son llamados regresiones no lineales, los cuales describen a la variable dependiente como $y_i = f(x_i, \theta_i) + \epsilon$, donde θ son otros parámetros desconocidos (Wikipedia, 2020). Además, las variables independientes x_i puede ser de cualquier dimensión, por ejemplo en imágenes tendríamos la dimensión $r \times c \times 3$, donde $r \times c$ son el tamaño de la imagen y 3 es la escala de colores.

De manera similar a lo expuesto para modelos lineales, la forma de evaluar modelos no lineales es a través del error calculado de la función de costo definida por $E(\theta_i) = L(y_i, f(x_i, \theta_i))$. Para el caso de donde nuestras x_i son imágenes, la función de aproximación puede ser una red de convolución profunda de la forma $f(x_i, \theta_i) = D(A_1(P_1(C_1(...A_n(P_r(C_s(...)))))))$, donde $C_j()$ son operaciones de convolución (también llamados de filtros), $P_j()$ son funciones de *pooling*, $D_j()$ es una combinación lineal (denominadas capas densas) y $A_j()$ son llamadas funciones de activación. A continuación describimos cada una:

- **Convolución**: $h_{ij}(f, g) = (f * g)(i, j) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} f(m, n) g(i + m, j + n)$, donde $f(m, n)$ corresponde al píxel en la posición (m,n) de la imagen de entrada y $g(i+m, j+n)$ son los píxeles correspondientes a la matriz de convolución. Esta función también es llamada de "filtro".

- ***Pooling***: es una técnica para re-dimensionar una matriz $I \in \mathbb{R}^{W \times H}$ a través de una sub-matriz de *Pooling* $P \in \mathbb{R}^{w_p \times h_p}$. Se define la variable *stride* como distancia entre cada píxel. Sea $I_{i,j}$ la posición del píxel (i, j) de la imagen I y también la matriz de *Pooling*. Para simplificar las notaciones, definimos $k = 0, \ldots, W$ y $l = 0, \ldots, H$ como las posiciones de los píxeles en filas y columnas. Se define las siguientes operaciones de *pooling*:

  - **Max Pooling**: Retorna una matriz de dimensión $(\lfloor \frac{W - w_p}{stride} \rfloor + 1, \lfloor \frac{H - h_p}{stride} \rfloor + 1)$. Donde cada valor de la matiz será de la forma: $\underset{k \le i \le k + w_p, l \le j \le l + h_p}{\text{máx}} I_{i,j}$.
  - **Global Max Pooling**: Retorna un valor asociado a una matriz de la forma: $\underset{i,j}{\text{máx}} I_{i,j}$.
  - **Average Pooling**: Retorna una matriz de dimensión $(\lfloor \frac{W - w_p}{stride} \rfloor + 1, \lfloor \frac{H - h_p}{stride} \rfloor + 1)$. Donde cada valor de la matriz será de la forma: $\frac{1}{w_p * h_p} \sum_{i=k}^{k+w_p} \sum_{j=l}^{l+h_p} I_{i,j}$.
  - **Global Average Pooling**: Retorna un valor asociado a una matriz de la forma: $\frac{1}{W * H} \sum_{i=0}^{W} \sum_{j=0}^{H} I_{i,j}$ (Lin et al., 2013).

(a) (b) (c)

Figura 2.2: Predicción: Perro; explicaciones presentadas por los métodos (a) CAM, (b) GBP y (c) guided-CAM. Fuente: grad-CAM (Selvaraju et al., 2017).

- **Funciones de Activación**: son funciones utilizadas para acotar un conjunto de valores a un dominio específico, las más conocidas son:
  - Tangente hiperbólica (**Tanh**): $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.
  - **Sigmoid**: $f(x) = \frac{1}{1+e^{-x}}$.
  - **ReLU**: $f(x) = \text{máx}(0, x)$.
  - Función linear (**Linear**): $f(x) = x$.

2.1.4. Métodos de Explicación

Los métodos de explicación nos permiten entender el comportamiento y la decisión de un modelo. se pueden dividir en dos clases principales: aquellos del tipo *white-box* que serían los más fáciles de analizar tales como modelos lineales. El segundo tipo son *black-box* los cuales son más difíciles de analizar debido a su complejidad y gran cantidad de parámetros, un ejemplo claro son las redes profundas (Molnar, 2022). Algunos métodos de explicación de este tipo son **Visualización de convoluciones** (Zeiler y Fergus, 2013b), *smooth* (Ancona et al., 2017), *saliency maps* (Simonyan et al., 2013) y *class activation maps* (CAM) (Zhou et al., 2014), entre otros.

Para entender el cálculo, vamos a definir $x \in \mathbb{R}^d$ un vector de características de una imagen, un modelo $S : \mathbb{R}^d \to \mathbb{R}^C$ donde C es el número de clases a evaluar y un método de explicación $E : \mathbb{R}^d \to \mathbb{R}^d$ que define un *explanation map*. Entonces, una explicación basada en gradiente de una variable x es de la forma $E_{grad}(x) = \frac{\partial S}{\partial x}$, la gradiente cuantifica cuánto es el cambio que genera cada dimensión de x en la predicción $S(x)$ en una vecindad pequeña cercana a x (Adebayo et al., 2018). Algunos métodos basados en gradientes son:

- **Gradient Input**: El cáculo es del gradiente es de la forma: $x \odot \frac{\partial S}{\partial x}$, reduce la difusión visual y la saturación de gradiente (Shrikumar et al., 2016).

- **Integrated Gradients**: El cálculo es del gradiente es de la forma: $(x-\bar{x})x\int_0^1 \frac{\partial S(x+\alpha(x-\bar{x}))}{\partial x}$, también reduce la saturación de gradiente mediante la integral de escalas, $\bar{x}$ representa la ausencia de *features* en x (Sundararajan et al., 2017).

- ***Class Activation Map* (CAM)**: El cálculo es del gradiente es de la forma: $M_{cam} = \sum_k w_k^c F^k$, donde M_c es un *class activation map* de la clase c, k representa una capa de convolución y F^k es el resultado de aplicar *Global Average Pooling* (GAP) ($F^k = \frac{1}{Z}\sum_i \sum_j A_{ij}^k$) a cada convolución, donde A^k es el filtro correspondiente a la k-ésima convolución, (i,j) representan los píxeles de la matriz A^k y Z es el producto de las dimensiones de A^k (Zhou et al., 2016a).

- **GradCAM**: Tomando en cuenta el resultado obtenido en CAM, se define $\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial M_c}{\partial A_{ij}^k}$ como el GAP de los gradientes del CAM denominado *partial linearization*, finalmente $M_{gradcam} = ReLU(\sum_k \alpha_k^c A^k)$ (Selvaraju et al., 2017).

- **GBP**: El cálculo es de la forma:
  $B_i^t = (f_i^t > 0) \times (B_i^{t+1} > 0) \times B_i^{t+1}$, donde $B_i^{t+1} = \frac{\partial f^{out}}{\partial f_i^{t+1}}$, $f_i^{t+1} = relu(f_i^t)$ es la activación de la capa actual de la red y B_i^t es denominado GBP de la capa t e i es una muestra (Springenberg et al., 2014).

- **Guided GradCAM**: Es el producto elemento-elemento entre GradCAM y GBP (Selvaraju et al., 2017).

- ***SmoothGrad* (SG)**: Reduce el ruido y difusión visual en los *saliency maps* con un ponderado de las explicaciones de copias ruidosas de x, para una explicación E, se tiene: $E_{sg} = \frac{1}{N}\sum_{i=1}^{N} E(x + g_i)$ donde $g_i \sim \mathcal{N}(0, \sigma^2)$ es el ruido (Smilkov et al., 2017).

En la Figura 2.2 se muestra un ejemplo de algunos métodos tales como CAM, Grad-CAM, y GBP. Los cuales son muy utilizados para interpretar modelos que utilizan imágenes como datos de entrenamiento.

En esta sección hemos presentado conceptos como técnicas de aprendizaje, modelos lineales y no lineales, así como también, las operaciones envueltas en cada una. Estos términos y definiciones las emplearemos más adelante a lo largo de este documento, especialmente a partir de las próximas secciones donde abordamos los trabajos relacionados. A manera de recordatorio, las siguientes secciones están divididas en (a) análisis de la percepción urbana y (b) extracción de características y componentes visuales; y (c) interpretación y visualización de características extraídas.

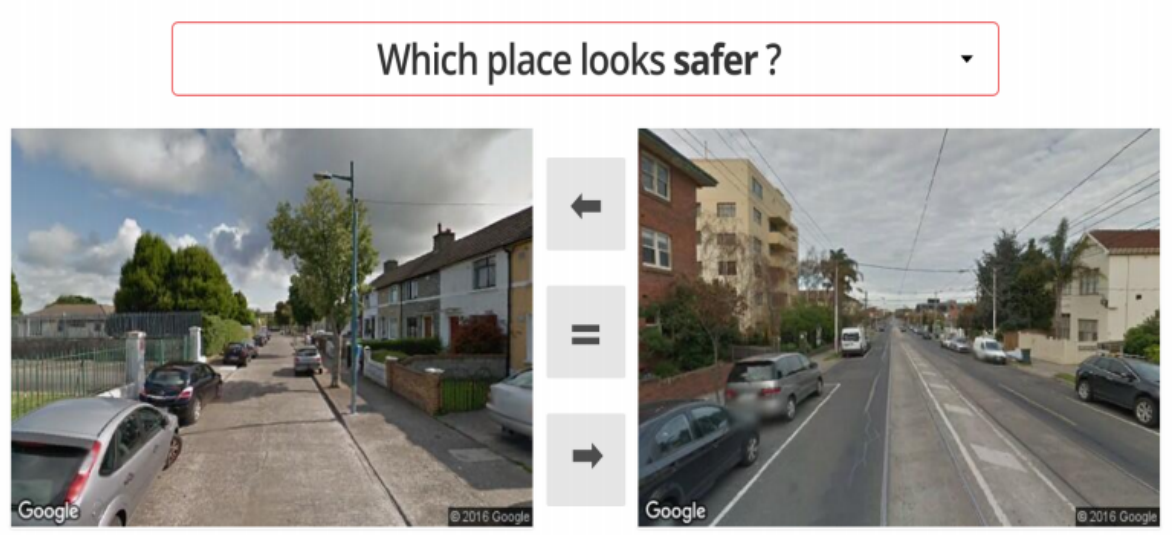


Figura 2.3: Sitio web *Place Pulse*, con la cual se recolecta información acerca de la percepción de las calles a partir de una elección entre dos imágenes de calles. Fuente: *Place Pulse* (Salesses et al., 2013).

2.2. Análisis de la Percepción Urbana

En esta sección se expondrán los trabajos relacionados al análisis de la percepción urbana utilizando diversos métodos para relacionar la apariencia visual de calles y otros datos no visuales, tales como criminalidad. El año 2011, *MIT-Media Lab* inició el proyecto denominado *Place Pulse* (Salesses, 2012), el cual recolectaba informaciones sobre imágenes a través de diversos voluntarios con el objetivo de responder la siguiente pregunta: "*Which Place Looks Safer/Unique/Wealthy*?". Cada voluntario debía seleccionar entre dos imágenes aleatorias descargadas de las ciudades Boston, New York, Linz y Salzburg. El trabajo realizaba un estudio sobre la diferencia de percepciones de cada una de las ciudades evaluadas a partir de sus aspectos visuales creando una medida cuantitativa de los contrastes de una ciudad. Como resultados importantes tenemos la creación del conjunto de datos *Place Pulse* versión 1.0 y que los aspectos visuales entre Boston y New York eran más notorios que entre Linz y Salzburg; los cuales fueron comparados con datos de crímenes dentro de dichas ciudades, mostraban que en lugares con apariencia visual similar se tenía una tasa de criminalidad similar.

En el año 2014, el conjunto de datos *Place Pulse 1.0* motivó el incremento en la cantidad de estudios y análisis sobre la percepción urbana, tales como aprender características específicas para poder predecir el nivel de seguridad en una calle. Uno de estos estudios (Ordonez y Berg, 2014) utilizó como base las comparaciones respecto a las categorías seguro (a partir de ahora *safety*), únicos (a partir de ahora *uniqueness*) y opulentos (a partir de ahora *wealthy*). Adicionalmente a los datos que provee *Place Pulse 1.0*, se recolectó imágenes de New York (8863), Boston (9596) y 2 ciudades adicionales: Chicago (12 502) y Baltimore (11 772), el entrenamiento se realizó con las ciudades de New York y Boston originales del *Place Pulse 1.0*. Este trabajo presentó dos modelos, un clasificador *Support Vector Machine* (SVM) (Boser et al., 1992) y un

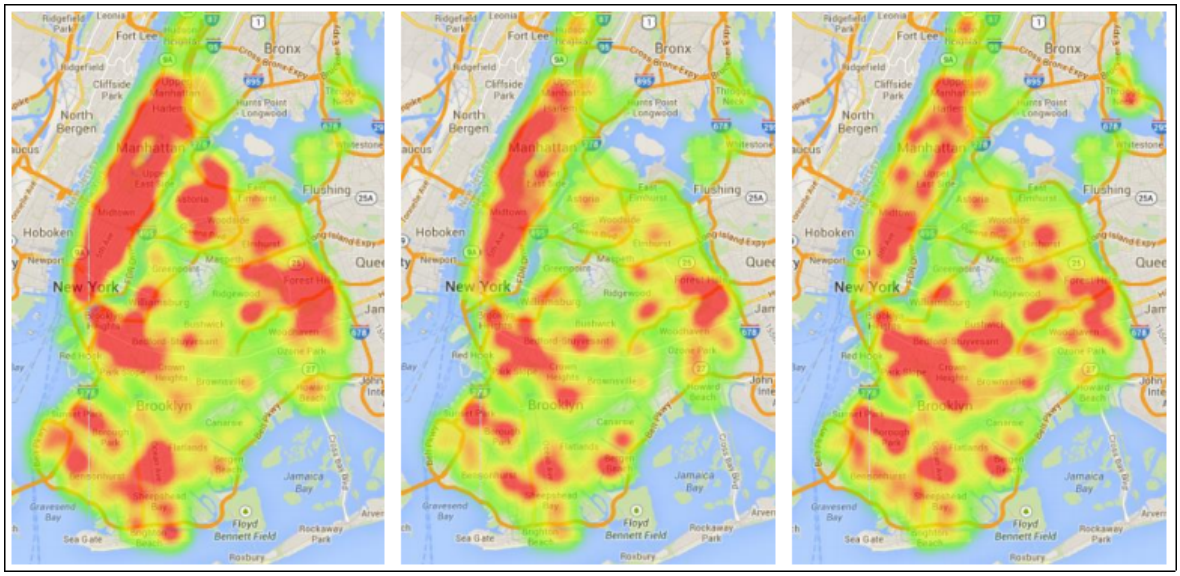


Figura 2.4: Resultados de la evaluación de regresión sobre *Place Pulse 1.0*: (izquierda) puntuaciones de la ciudad de New York; (medio) predicciones de la regresión en New York utilizando un modelo entrenado sobre New York; y (derecha) predicciones de la regresión en New York utilizando un modelo entrenado sobre Boston. Fuente: (Ordonez y Berg, 2014).

regresor *Support Vector Regressor* (SVR) (Smola y Schölkopf, 2004) para la predicción de las puntuaciones (números entre 0 y 10) y percepción respectivamente. Ambos modelos fueron entrenados con regulación l_2 y además, utilizaron como extractores de características los métodos GIST (Oliva y Torralba, 2001), SIFT + Fisher Vectors (Perronnin et al., 2010), *Deep Convolutional Activation Feature* (DeCAF) (Donahue et al., 2014).

Para el entrenamiento, utilizaron 5 validaciones cruzadas entrenadas sobre cada ciudad y realizaron las comparaciones de evaluar en otras. Para realizar el etiquetado, asignaron que una puntuación por debajo de 5.0 sería -1, caso contrario 1. Como resultados principales, se evidenció que la extracción de características con una red profunda como DeCAF tuvo mejor desempeño que otros métodos como GIST o SIFT + Fisher Vectors; otro resultado fue que el regresor tiene mejor precisión para imágenes con puntuaciones entre 4-7. Así mismo, analizaron la predicción de la percepción de manera colectiva a través del método *K-Nearest Neighbors* (KNN) (ver Figura 2.4) mostrando que regiones cercanas poseen una predicción similar. Este trabajo demarca el inicio de la utilización de *Place Pulse 1.0* con *Deep Learning*.

Más adelante, Li et al. (2015a) utilizando los datos de las ciudades Boston y New York de *Place Pulse 1.0* proponen un análisis y exploración acerca de la estética, ambiente y beneficios psicológicos en residencias urbanas, priorizando en cómo las áreas verdes pueden ayudar a incrementar la percepción de seguridad en las calles de lugares como residencias, zonas industriales, lugares públicos, instituciones, etc. Para procesar las imágenes y filtrar áreas verdes, normalizaron los valores de los canales *Red-Green-Blue* (RGB) y calcularon el parámetro *Green Index* definido como $G_I = 2G - R - B$.

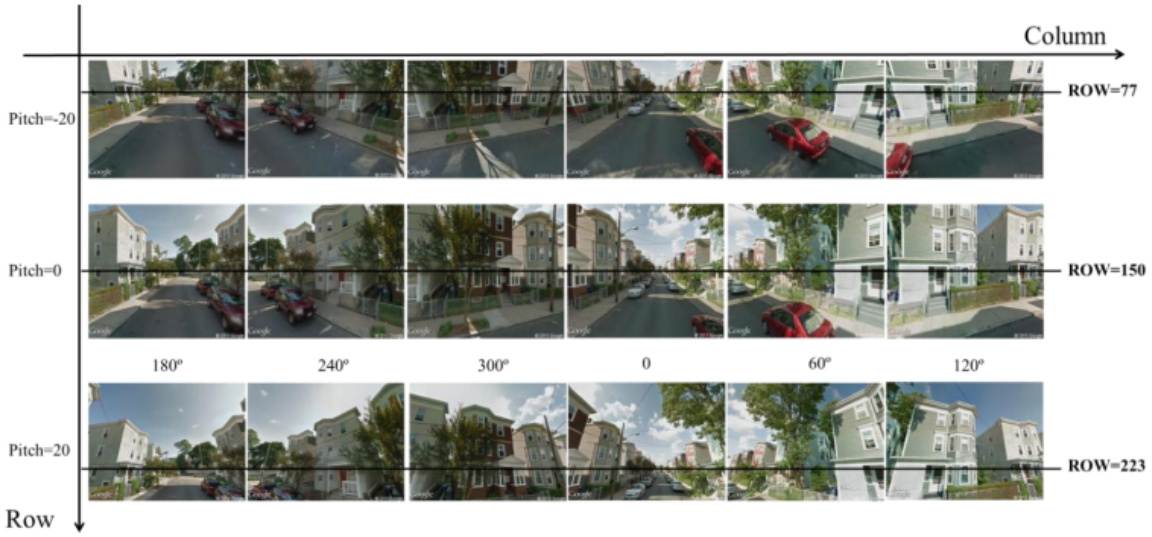


Figura 2.5: Diferentes direcciones y ángulos de una calle de Boston, Las filas representan la variación en altura y las columnas representan el punto de vista. Fuente: (Li et al., 2015a).

Luego a través del filtro Otsu (Otsu, 1975) y otras operaciones sobre píxeles para retirar el brillo y contraste (sombras). También, utilizando las respectivas latitud y longitud de cada imagen, utilizando *MassGIS Data-Land Use 2005* (Massachusets-Office-Goverment, 2005) descargaron otros *Field of View* (FOV) aumentando así las imágenes especialmente seleccionadas de lugares como: residenciales, lugares públicos (hospitales, universidades, escuelas, aparcamientos, museos, prisiones, etc.), lugares industriales, cementerios, lugares abiertos y lugares de recreación. Finalmente, con una regresión lineal relacionaron la presencia de áreas verdes y las puntuaciones de percepción demostrando así la importancia e influencia positiva de las áreas verdes en lugares con altas puntuaciones. Como conclusiones se presentó que las áreas verdes impactan positivamente en la percepción de seguridad en su mayoría, sin embargo, pueden ser percibidas como inseguras, debido a que obstruye la visión de un lugar contrastando así las teorías presentadas en Fisher (1992); Nasar et al. (1993).

Otro estudio fuertemente basado en *Place Pulse 1.0*, fue el desarrollo de la plataforma Wmodi (Acosta y Camargo, 2018b). Esta plataforma de manera semejante a *Place Pulse 1.0* recolecta información a manera de encuesta (ver Figura 2.6 (a)), con la única diferencia que Wmodi utiliza imágenes de la ciudad de Bogotá, Colombia. Como pre-selección de imágenes, utilizaron el extractor SIFT (Lowe, 2004) para determinar si existía características mínimas o si eran solo paredes o fondos negros; obteniendo así 5505 imágenes. Así mismo, el sitio web obtuvo alrededor de 17 703 comparaciones, donde cada imagen había sido comparada en promedio 6 veces. Además, se recolectó 5657 empates, 5946 seguras y 6100 inseguras. Para el procesado de las puntuaciones, utilizaron el algoritmo *TrueSkill* (Herbrich et al., 2007) el cual potenciaba la actualización en línea de las puntuaciones; generando un mapa de puntuaciones de percepción de seguridad (ver Figura 2.6 (b)). Utilizaron la red *VGG19* (Simonyan y Zisserman, 2014) y los métodos GIST y *Histogram of Oriented Gradients* (HOG) como extractores

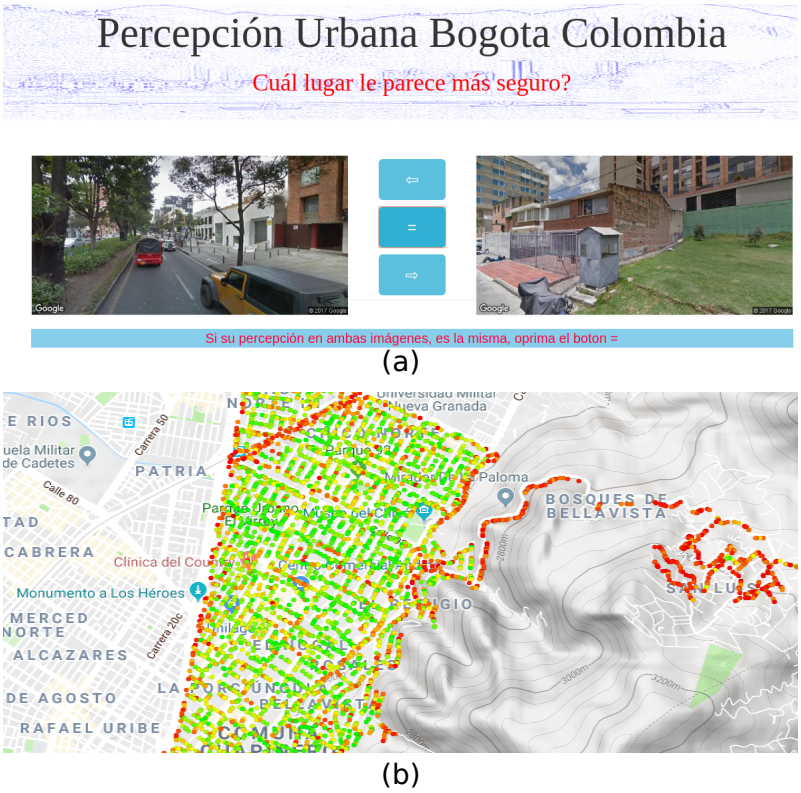


Figura 2.6: (a) Sitio web Wmodi para recolectar información acerca de la percepción de las calles. (b) Mapa de puntuaciones de percepción de seguridad de la ciudad Chapinero (Bogotá-Colombia). Fuente: (Acosta y Camargo, 2018a).

de características para luego ser entrenados en un SVR. Las principales contribuciones son: (i) la percepción de alta inseguridad está relacionada a lugares con poca áreas verdes, lugares con alta densidad de tráfico vehicular, avenidas principales, caminos debajo de puentes, caminos tipo trocha; (ii) la plataforma Wmodi.

Así mismo, tenemos estudios con el enfoque de entender y explorar la correlación entre la percepción urbana y las estadísticas de crímenes tal como *StreetNet* (Fu et al., 2018). Para este estudio, fue construido un conjunto de datos utilizando los índices de robo, robo agravado, robo al paso, robo con arma, entradas sin autorización a domicilios, etc. de las ciudades de New York y Washintong DC. Para jerarquizar la gravedad de los crímenes se utilizó el método *Preference Learning* (Har-Peled et al., 2003) en cada lugar, además de también utiliza referencias geográficas de las calles al alrededor utilizando el *Application Programming Interfaces* (API) *CycloMedia GlobalSpotter* (CycloMedia, 1980), donde el punto central es llamado *sample point*. En la Figura 2.7 (b) se puede observar las referencias geográficas a partir de cada *sample point*. A partir de estos *sample points*, se utiliza el algoritmo DSVR para agrupar los crímenes perpetrados en una determinada zona radial establecida. En la Figura 2.7 (c) se puede observar cómo se agrupan los puntos donde ocurrieron crímenes en un punto referencial entre ellos. Como resultados destacables presenta: (i) creación de conjuntos de datos Washintong DC y New York City (NYC) llamadas DC-1k, DC-2k, NYC-1k y NYC-2k generados utilizando como radio de distancia 1000 y 2000 pies respectivamente; (ii) el

Figura 2.7: Resultados: (a) Predicción de los posibles crímenes a ocurrir a partir de una determinada vista de calle. (b) Puntos con referencias geográficas tomados topológicamente a partir de un crimen cometido. (c) Puntos de referencia geográficas obtenidos luego de aplicar el algoritmo *Direction based Street View Retrieval* (DSVR). Fuente: (Fu et al., 2018).

modelo *StreetNet*, el cual permite predecir qué tipo de posible crimen podría ocurrir basado en los datos de crímenes ocurridos alrededor y en las características del lugar (ver Figura 2.7 (a)).

En esta sección se presentaron trabajos relacionados al estudio y análisis de la percepción urbana, algunos de ellos están basados en el conjunto de datos *Place Pulse 1.0*. Los cuales en su mayoría buscan relacionar alguna característica de la ciudad con las puntuaciones de percepción. Además, la utilización de otros conjuntos de datos relacionados al índice de criminalidad y aspectos visuales presentes en ciertas ciudades. En el presente trabajo, nosotros nos enfocamos en el estudio del conjunto de datos *Place Pulse 2.0*, el cual describiremos más adelante presentando un análisis de la percepción urbana a partir de la metodología propuesta que describiremos en el siguiente capítulo.

2.3. Extracción de Características y Componentes Visuales

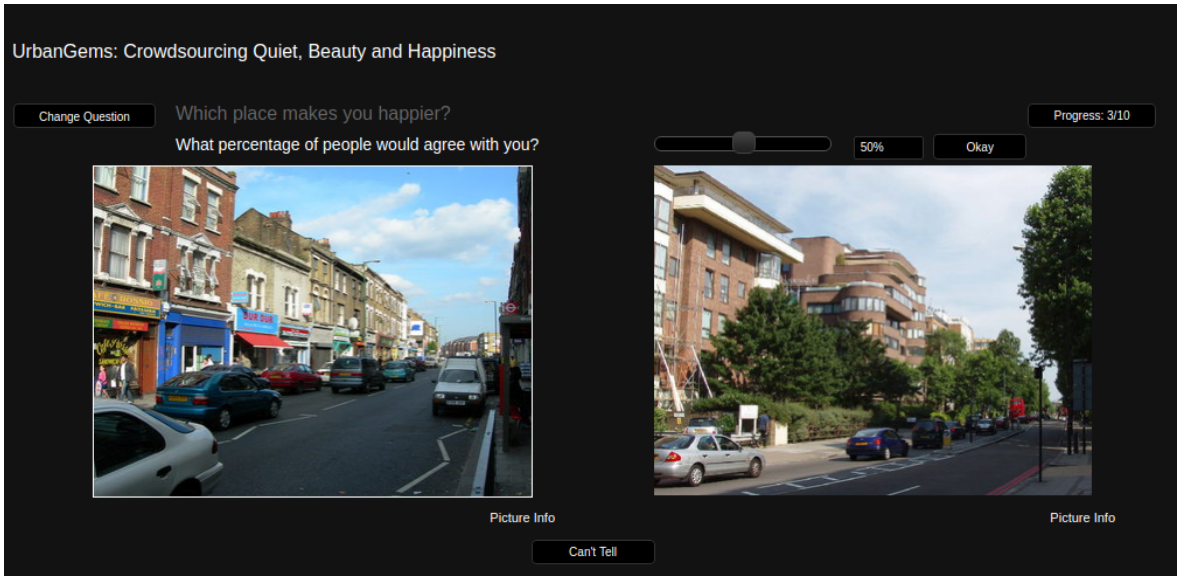
En esta sección se expone trabajos relacionados a la percepción urbana pero que están más enfocados a la extracción de características. También hacemos mención que

Figura 2.8: (a) París-Francia. (b) Praga-República Checa. (c) Londres-Inglaterra. Correspondencia visual entre cada elemento presente en cada ciudad. Fuente: (Doersch et al., 2012).

podemos dividir en dos grandes conjuntos, aquellos de bajo/medio nivel que serian los métodos convencionales y alto nivel, aquellos que usan redes profundas para la extracción. Algunos de los métodos de bajo/medio nivel son GIST (Oliva y Torralba, 2001), SIFT + Fisher Vectors (Perronnin et al., 2010), HOG+*Color descriptor* Dalal y Triggs (2005), *Geometric Probability Map* (Hoiem et al., 2007) y *Color Histograms* (Novak et al., 1992; Chakravarti y Meng, 2009) y métodos de alto nivel son *AlexNet* (Krizhevsky et al., 2012), *VGGNet* (Simonyan y Zisserman, 2014), *ResNet* (He et al., 2015), *PlacesNet* (Zhou et al., 2014).

2.3.1. Extracción de características de bajo/medio nivel

Uno de lo trabajos más resultante es "*What makes Paris look like Paris?*" (Doersch et al., 2012). En este trabajo se pretende entender e identificar cuales son las diferencias que existen entre las diversas edificaciones de ciudades presentes en Europa. A manera de experimento se realizó un cuestionario a 11 personas con 100 imágenes aleatorias descargadas desde GSV en las cuales 50 % eran de París y las demás de diferentes ciudades, el cuestionario consistía en responder si una determinada calle pertenecía o no a París (ignorando textos presentes en la imagen). En promedio acertaron alrededor de 79 %, sin embargo, cuando si tomaban en cuenta textos presentes en las imágenes, el promedio de aciertos incrementó hasta 90 %. Esto demuestra que las personas son sensible ante la información dentro de una imagen (p.ej. letreros, carteles), ayudando a distinguir e identificar de manera más rápida qué ciudad era. Para el experimento final, se recolectó 10 000 imágenes por ciudad, de 12 ciudades: París, Londres, Praga, Barcelona, Milán, New York, Boston, Philadelphia, San Francisco, Sao Paulo, Ciudad de México y Tokio. Para el estudio dividieron la información en dos: (i) imágenes de París; e (ii) imágenes de las otras ciudades. También, asumieron que los patrones vi-



(a)

(b)

Figura 2.9: (a) Sitio web Urbangems; (b) Se observa los resultados de los *visual words* asociados a la categoría belleza que los puntos rojos son los que representan una imagen. Fuente: (a) (UrbanGems, 2014), (b) (Quercia et al., 2014)

suales como árboles, autos, cielo, etc. existirían en diferentes ciudades así que no eran tomadas en cuenta.

Utilizando HOG+*Color descriptor* se extrajeron las características, para luego ser agrupadas con *Locality-Sensitive Hashing* (Gong et al., 2012). Para la creación de los grupos se escogieron de manera aleatoria 25 000 imágenes de ciudades diferente de París aplicándoles un KNN, obteniendo 20 grupos de los cuales mantuvieron solamente aquellos con mayor proporción de vecinos cercanos del conjunto de París. Del total 25 000, se redujo a 1000 imágenes como centros. Para el entrenamiento de las características, se utilizó una SVM con 3 validaciones cruzadas. Como resultados destacables: (i) se evidenció que en muchas ciudades de Europa se presentaban apariencias visuales muy similares. En la Figura 2.8 se observa los atributos visuales correspondientes a Pa-

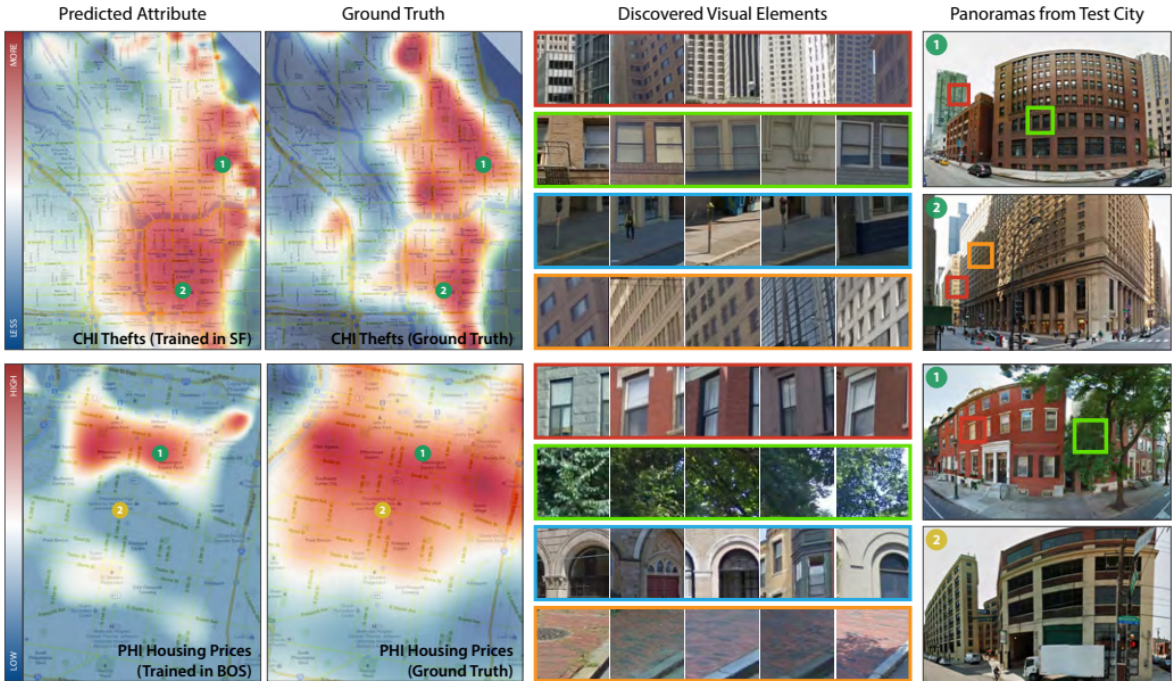


Figura 2.10: Resultados de la predicción de los atributos, la primera fila muestra los resultados del modelo entrenado sobre San Francisco y probado en Philadelphia, la segunda fila muestra un caso de error cuando el modelo de predicción de precios de casas entrenados sobre Boston y probado en Philadelphia, el error de predicción es debido a que algunos elementos visuales presentes en Boston que influyen en la categoría "precios altos" no se encuentran presentes en Philadelphia. Fuente: (Arietta et al., 2014).

rís, Praga y Londres, pudiéndose observar las ligeras pero marcadas diferencias entre los elementos; (ii) un método robusto para diferenciar las características de todas las ciudades estudiadas, a pesar de ser bastante similares.

Otro trabajo sobre la exploración de componentes visuales fue "*What Makes London Look Beautiful, Quiet, and Happy?*" (Quercia et al., 2014); teniendo como objetivo recopilar información acerca de la percepción de las ciudades, y además, analizar un factor de percepción colectiva con la pregunta "*What percentage of people would agree with you?*" y unos colores (p.ej. de las calles) asociados a dicha percepción. En la Figura 2.9 (a) se puede ver el sitio web Urbangems (UrbanGems, 2014) en donde las personas debían escoger entre dos imágenes de un total de 700 000 y dar un porcentaje de las personas que concordaría de la misma forma. A través de las respuestas de 3301 usuarios donde cada usuario realiza una ronda compuesta por 10 comparaciones, obteniendo en promedio una preferencia acerca de belleza (171), silencioso (12) y feliz (16). Las imágenes fueron recolectadas a través de GSV de lugares cercanos a estaciones de metro en un radio de 300 metros.

Una vez obtenida la información equivalente a 17 261 comparaciones, se procedió a analizar qué colores tienen mayor correlación con las imágenes de *beauty* (belleza), *quiet* (silencioso) y *happiness* (felicidad) utilizando los canales RGB de la imagen, así como también las texturas con *Global Edge Histogram* (GEH) (Park et al., 2000) con

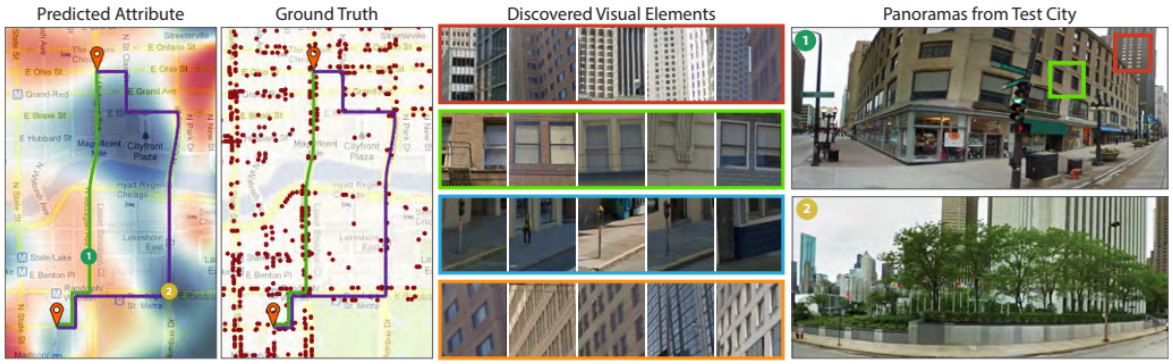


Figura 2.11: Otro resultado es la predicción de rutas seguras (morada) evitando la mayoría de sucesos de robos (círculos rojos) en Chicago diferenciando de la ruta directa (verde), para calcular esta ruta se predijo la tasa de robos en Chicago utilizando el modelo entrenado en San Francisco el cual contiene elementos visuales relacionados al tráfico en los cuales son comunes los robos, en contraste con la ruta predicha que contiene áreas verdes y árboles, las cuales presentan baja tasa de robos. Fuente: (Arietta et al., 2014).

la técnica *region-based MPEG-7 Edge Histogram descriptor* (Manjunath et al., 2001). Tomando en cuenta los puntos en donde los usuarios hicieron *click* dentro de la imagen (denominados puntos de interés), dividieron a las imágenes en 4 contornos: horizontal, vertical, diagonal y no direccional. Para analizar los puntos de interés en las imágenes se utilizó *Speeded Up Robust Features* (SURF) (Bay et al., 2006) agrupándolos con el algoritmo *K-means* en 500 grupos de *visual word* (Jurie y Triggs, 2005). En la Figura 2.9 (b) se puede observar los *visual words* o puntos de interés de una imagen, estos puntos pueden describir una imagen. Como resultados presentaron: (i) los puntos seleccionados asociados a la categoría belleza son casas victorianas, jardines públicos, residenciales y los menos asociados son edificios gubernamentales, puentes y carreteras. (ii) los puntos asociados a la categoría silencioso son árboles, setos, bosques y ventanas residenciales, por el contrario, los menos asociados son sitios de construcción y buses; (iii) los puntos asociados a felicidad son árboles, buses y personas, por el contrario, los menos asociados son puentes, calles y cercas de alambre.

Otro trabajo enfocado en apariencia visual es *City Forensics* (Arietta et al., 2014) el cual se propone un método para predecir, identificar y corroborar una correlación entre la apariencia visual de una ciudad y sus atributos no visuales. Los datos de "atributos no visuales" son aquellos correspondientes a tasas de criminalidad violentas (CrimeMapping, 2012), tasa de robos (CrimeReports, 2013), precios de casas, densidad de la población, presencia de árboles (UrbanForest, 2014), presencia de grafitis (obtenido por reportes) y percepción de peligro. Para obtener más datos, usando GSV se descargó imágenes panorámicas con FOV de 360 grados e ángulo de inclinación de 20 de las ciudades San Francisco, Chicago y Boston. Obteniendo entre 30 000 hasta 170 000 panoramas por ciudad, de los cuales 10 000 son utilizadas para entrenar. A través de HOG+*Color descriptor* se identificaron los atributos visuales y se procedió a anotar etiquetas para cada conjunto no visual (p.ej. si fuera precios de casas, se anotaría como positivo a los valores por encima de la media y negativo por debajo). Se interpoló

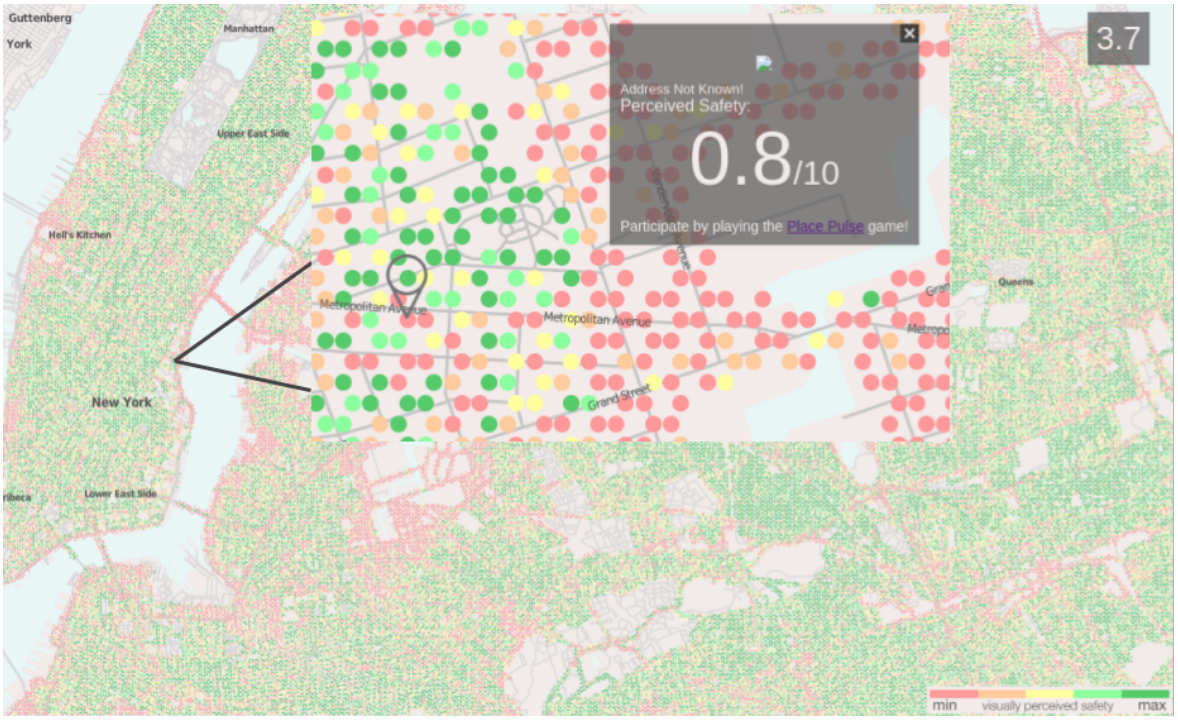
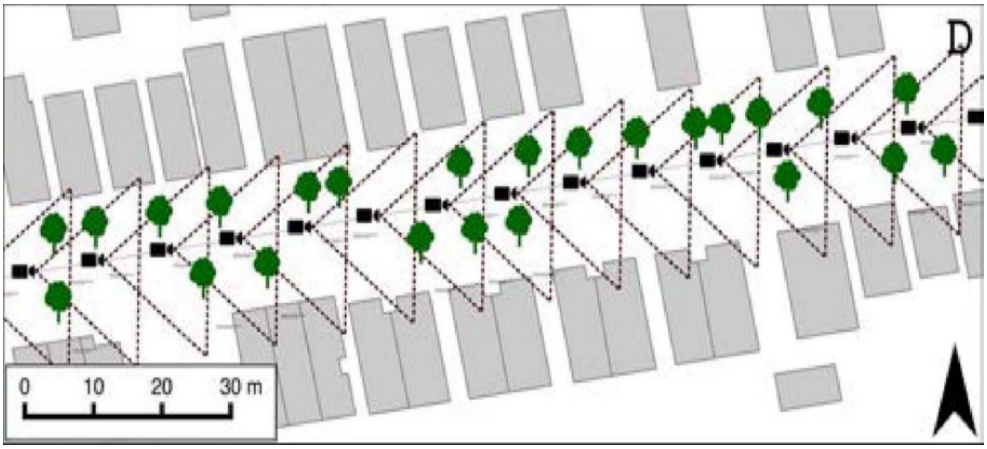


Figura 2.12: Resultados de las predicciones de las puntuaciones de percepción en la ciudad de New York: generación de un *HPM* donde las puntuaciones están entre 0 y 10, siendo un punto rojo la percepción de una calle insegura y verde la percepción de una calle segura. Fuente: (MIT-Media-Lab, 2014).

los atributos visuales usando las latitudes y longitudes de las imágenes a través del método *Radial Basis Function* (RBF) (Broomhead y Lowe, 1988), el cual también fue el modelo para entrenar 10 00 muestras con la proporción de 2000 positivas y 8000 negativas. Finalmente, se entrena una SVM, refinándola con 3 iteraciones de *hard negative mining technique* (Felzenszwalb et al., 2008) para relacionar cada panorama a un atributo no visual. Como resultado destacable tenemos que existe una correlación entre la apariencias visuales de cada ciudad con sus respectivas tasas de criminalidad, tasas de robos, precios de casas, densidad de población, presencia de árboles, presencia de grafitis y la percepción de peligro. Así como también 3 aplicaciones como rutas seguras (ver Figura 2.11 primera columna), limite o divisiones de la ciudad (ver Figura 2.10 tercera columna) y que la presencia de algunos componentes visuales pueden describir una ciudad (p.ej. grafitis, ladrillos, ventanas con estilos y postes de luz describen a Chicago).

En el año 2014, utilizando los datos de las ciudades New York y Boston del conjunto de datos *Place Pulse 1.0*, *StreetScore* (Naik et al., 2014) presentó un estudio comparando cual extractor de características sería el adecuado para las imágenes de ese conjunto. Los extractores comparados fueron GIST, *Geometric Probability Map*, *Texton Histograms* (Martin et al., 2001), *Color Histograms* (Novak et al., 1992; Chakravarti y Meng, 2009), *Geometric Color Histograms* (Rao et al., 1999), HOG (Dalal y Triggs, 2005), Dense SIFT (Lazebnik et al., 2006), LBP (Ojala et al., 2002), *Sparse SIFT histograms* (Sivic y Zisserman, 2004) y SSIM (Matas et al., 2004) los cuales fueron



(a)

(b)

Figura 2.13: (a) Mapa de distancias tomadas entre cada imagen, (b) sitio web Treepedia con los índices de áreas verdes de árboles en la ciudad de Boston. Fuente: (a) (Li et al., 2015b), (b) (MIT-Media-Lab, 2015)

entrenados en una SVR. Para procesar las puntuaciones se utilizó el algoritmo *TrueSkill*, resaltando que en promedio cada imagen fue comparada sólo 6 veces, el cual estaba muy lejos de la convergencia óptima entre 12 y 36 comparaciones. Para las evaluaciones, se descargó aproximadamente 200 imágenes por cada 1.6 km^2, de tal forma, tener la mayor cobertura de una ciudad, logrando obtener aproximadamente 1 millón de imágenes de 27 ciudades diferentes de USA. Como resultados obtuvieron que *Color Histograms*, GIST y *Geometric Color Histograms* presentaban el mejor desempeño. A partir de ese resultado, se define *StreetScore* el cual es una concatenación de las características extraídas por esos 3. En la Figura 2.12 se puede observar el resultado de la predicción de puntuaciones usando *StreetScore* entrenado en Boston y New York, evaluado en New York.

Continuando el estudio presentado en Li et al. (2015a) (descrito en la sección anterior), Li et al. (2015b) extendieron el análisis de áreas verdes imágenes en panoramas; buscando la presencia de áreas verdes en calles, además de su influencia en la percepción de seguridad. Modificaron los FOV cambiando los ángulos de 0, 60, 120, 180, 240 y 300. Utilizando imágenes de las ciudades de East Village, Manhattan District y New York, localizaron 300 ubicaciones generadas aleatoriamente con ArcGIS 10.2 (ArcGis, 1999) con una separación de 300 metros entre cada ubicación; obteniendo aproximadamente 28 448 m^2 en donde cada 100 metros se tendrá una vista global de un lugar (ver Figura 2.13 (a)). El procesamiento de las imágenes es realizado utilizando el método *Green View Index* (GVI) (Yang et al., 2009) y las operaciones a través de los canales RGB de cada imagen a nivel de píxeles de la forma: $G_I = (G - R) * (G - B)$ del cual, si G_I es positivo, ese píxel se considera vegetación. Finalmente, se obtiene un promedio entre el número de píxeles que son considerados vegetación y el número total de píxeles encontrados en todas las imágenes evaluadas. Como resultado se presentó: (i) un *Human Perception Mapping* (HPM) de las ciudades mencionadas destacando la asociación de las áreas verdes y robos, donde efectivamente a mayor concentración de vegetación se tiene menor cantidad de robos; (ii) el sitio web Treepedia (MIT-Media-Lab, 2015) (ver Figura 2.13 (b)) en el cual este análisis se extendió a 30 ciudades.

2.3.2. Extracción de características de alto nivel

Continuando con los métodos de extracción de alto nivel, aquí abordamos los conceptos de redes profundas, ya sean como extractores de características o para entrenar desde cero. Uno de estos trabajos fue el realizado por Porzi et al. (2015), el cual propone identificar elementos visuales y asignar un respectivo "ranking" de percepción (p.ej. de seguridad) a imágenes de calles provistas por el conjunto de datos *Place Pulse 1.0* y otros conjuntos de datos tales como: (i) *ImageNet* (Deng et al., 2009) con 1000 clases entre animales y objetos; (ii) *Places205* (Zhou et al., 2014) con 205 categorías de escenas o entornos (p.ej. restaurante, bosque, cafeterías, etc.); y (iii) SUN (Xiao et al., 2010) con objetos y categorías de escenas. Utilizando *TrueSkill* en las puntuaciones y SSIM, GIST, HOG y *AlexNet* (Krizhevsky et al., 2012) y su variante rCNN (propuesto por los autores) con pesos entrenados previamente en SUN, *Places205* e *ImageNet*. Los autores propusieron una capa de *pooling* definida por $\prod_{\eta_i}(M) = 1 + \lceil \eta_i(wz - 1) \rceil$ donde $0 \leq \eta_i \leq 1$ y $M \in \mathbb{R}^{(w \times z)}$ resultado de las convoluciones. En la Figura 2.14 (a) se puede observar los resultados de aplicar este *pooling* con diferentes valores de η_i. Finalmente, para el entrenamiento utilizaron una SVM para entrenar las características GIST, HOG, SSIM, *AlexNet-ImageNet*, *AlexNet-Places205*, *AlexNet*-SUN, rCNN-*ImageNet*, rCNN-*Places205* y rCNN-SUN. Para después adicionar un *RankingSVM* (Joachims, 2002) regularizado con l_2 (Tikhonov, 1943). Como resultados destacables tenemos: (i) la implementación de una función de *pooling* genérico, permitiendo obtener diferentes regiones de una imagen; (ii) el modelo rCNN que tuvo mejor desempeño con las configuraciones: *AlexNet-Places205* + $rCNN_2[m = 24, \eta{=}(0,\ 0.01,\ 0.05,\ 0.1)]$ que significa características de la segunda capa, 24 filtros lineales y los valores η.

En el año 2016, Dubey et al. (2016) extendieron el conjunto de datos *Place Pulse*

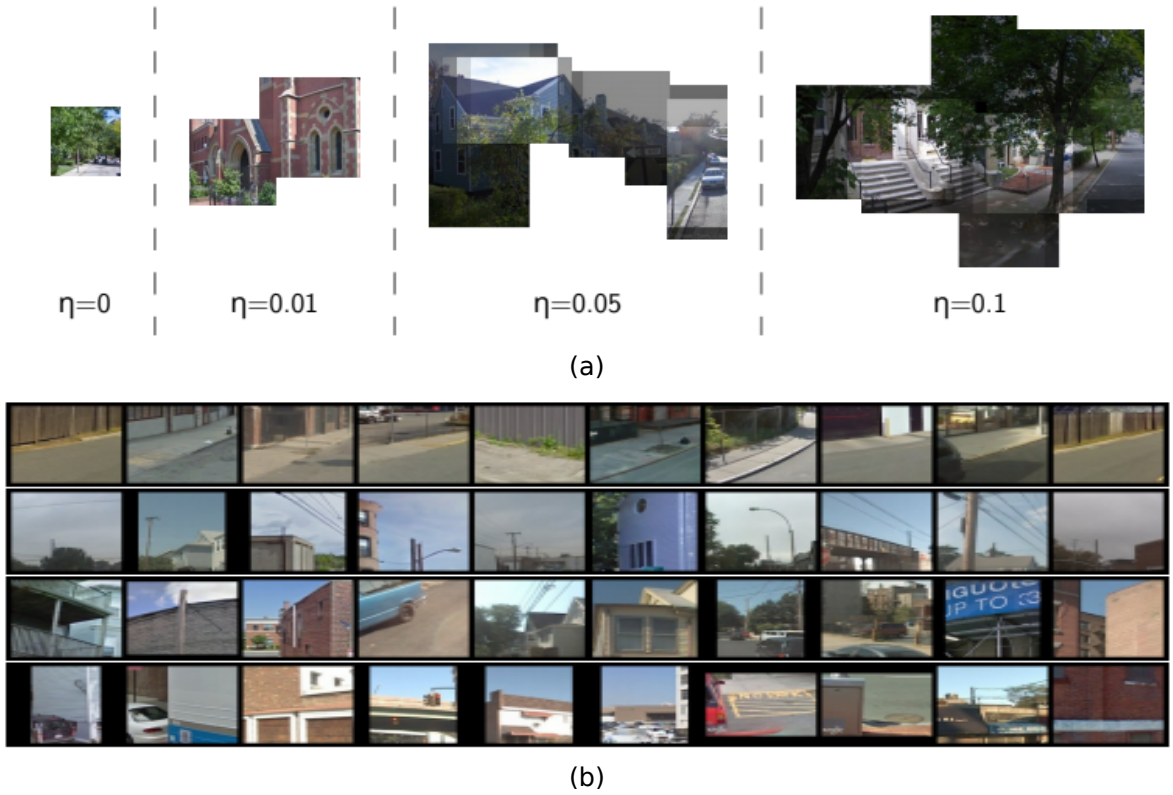


Figura 2.14: (a) Representación visual del método *pooling* mostrando los resultados para diferentes n (factor de *pooling*), se observa que si $\eta = 0$ es el clásico *max pooling* y si $\eta = 1$ es *average pooling*; (b) los patrones más relevantes encontrados con la técnica de *pooling* y altamente relacionados a la percepción de seguridad encontrados por la red *AlexNet-Places205* + $rCNN_2$. Fuente: (Porzi et al., 2015)

1.0 con 4 ciudades, 73 806 comparaciones y 4136 imágenes evaluadas en 4 categorías a *Place Pulse 2.0* con 56 ciudades, 1 223 649 comparaciones y 111 390 imágenes evaluadas en 6 categorías; donde las categorías extendidas son *safe* (seguro), *lively* (bueno para vivir), *boring* (aburrido), *wealthy* (opulento), *depressing* (depresivo) y *beautiful* (bonito). Además, basados en los resultados presentados en *StreetScore* (Naik et al., 2014), los autores indicaron que el análisis realizado en New York y Boston no estaba del todo correcto, debido al número de comparaciones que se contaba en ese momento (6 en promedio) no llegaba al mínimo necesario para tener una convergencia del algoritmo *TrueSkill*. Por lo cual, se propuso dos modelos: (i) *StreetScore-CNN* (SS-CNN) y (ii) *Ranking SS-CNN* (RSS-CNN). La arquitectura de SS-CNN consiste la fusión de las redes pre-entrenadas *AlexNet*, *PlacesNet* (Zhou et al., 2014) y *VGGNet* (Simonyan y Zisserman, 2014); generando la nueva *SiamesesNet* (Koch et al., 2015).La RSS-CNN es utilizada como un un *RankSVM* (Joachims, 2002) como función de salida para obtener al ganador entre la comparación de dos imágenes. En la Figura 2.15 (a) se observa el resultado de HPM obtenido de las predicciones de las redes SS-CNN y RSS-CNN. Como resultados relevantes destacamos: (a) la creación y liberación del conjunto de datos ***Place Pulse 2.0***. (b) No es posible utilizar el algoritmo *TrueSkill* debido a que mínimo se necesita alrededor de 24 o 36 votos por imagen, es decir, entre 1.2 y 1.9

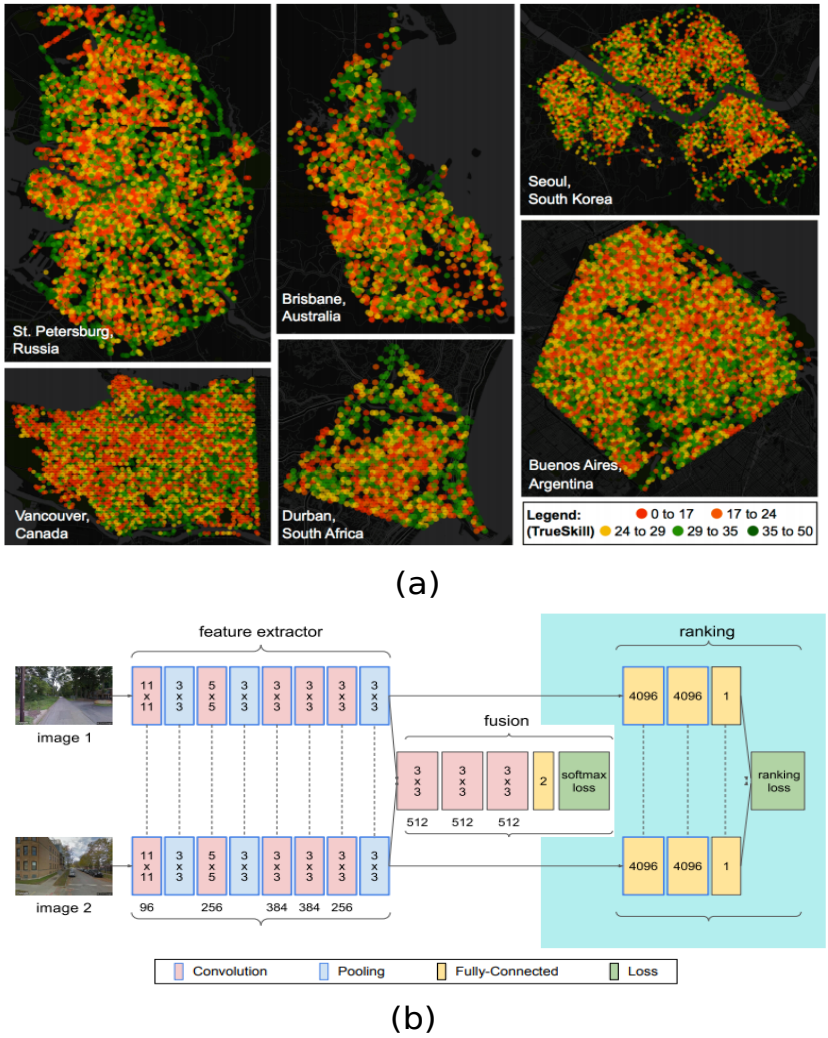


Figura 2.15: (a) Mapa de puntuaciones de seguridad predichos por la red RSS-CNN (*VGGNet*) y procesados con *TrueSkill*, mostrando el nivel de seguridad en una determinada ciudad; (b) arquitectura de las redes propuestas (modelo *AlexNet*). Fuente: (Dubey et al., 2016).

millones comparaciones en total. (c) Las redes SS-CNN y RSS-CNN para la predicción de un ganador entre dos imágenes.

De manera adicional, Liu et al. (2017) proponen un método para identificar regiones dentro de una determinada imagen que tengan una relación con la percepción urbana en ciudades y ambientes urbanos (p.ej. residencias, calles). Los experimentos fueron realizados utilizando imágenes de 5000 ubicaciones diferentes de las ciudades de Chicago, San Francisco, Seattle y New York; donde en cada ubicación se escogieron 8 FOV diferentes (ver Figura 2.16 (a)). Además, cada ubicación fue seleccionada a partir de datos de crímenes recolectados por agencias de gobierno durante un período de 15 años (p.ej. robos, peleas, asaltos, entre otros), así como también las puntuaciones de percepción encontradas por *StreetScore* y *Place Pulse 1.0*; obteniendo un total de 1 434 558 eventos de crímenes de las 4 ciudades generando una puntuación de percep-

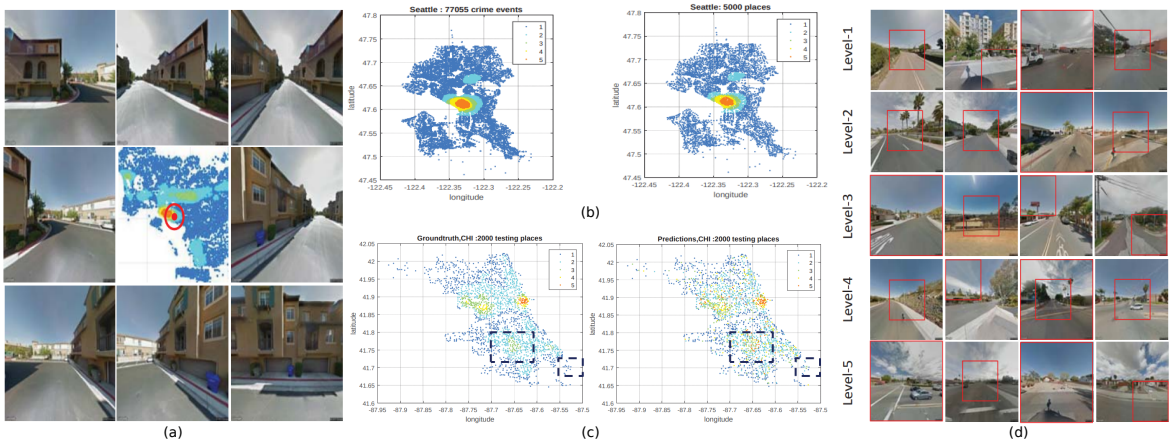


Figura 2.16: (a) Los 8 puntos de vista tomadas a partir de una posición donde ocurrió un crimen. (b) Aplicación del *clustering* sobre la ciudad de Seattle con los eventos de criminalidad (latitud y longitud) agrupándolo en 5000 grupos representados por un punto con latitud y longitud. (c) Comparación de los datos reales y las predicciones de la ciudad de Chicago, se resaltan los lugares donde hubo error de predicción. (d) Ejemplos de regiones asociadas a un nivel de criminalidad y con la mayor puntuación de percepción representadas por los rectángulos rojos en las imágenes. Fuente: (Liu et al., 2017).

ción asociada. Para el procesamiento de los datos, se utilizó el método *Parzen Window* (Parzen, 1962) para estimar la densidad de cada lugar y cuantificar la densidad en 5 niveles, lugares con baja densidad son interpretadas como seguras. Para evitar redundancias en los datos, se aplicó *K-means* en las ubicaciones (latitud y longitud) para la creación de 5000 grupos (ver Figura 2.16 (b)).

Para el procesamiento de datos, se define *bag of street view images* a cada ubicación (conteniendo las 8 FOV); así mismo, cada imagen en un diferente ángulo se compone de un conjunto de regiones (*bag of image regions*), así mismo, una región de imagen se descompone a su vez recursivamente en un conjunto de sub-regiones (*bag of sub-regions*). Este conjunto de datos pasó a denominarse *Place-Centric*, la cual se compone de las 5000 imágenes por cada una de las 4 ciudades donde cada imagen tiene 8 regiones y 40 sub-regiones. Para el entrenamiento de estos datos se utiliza una variación *Multi-Instance regressor* (MiR) (Ray y Page, 2001) denominado *Hierarchical Deep Multi-instance Regression* (HDMiR); donde las entradas son las características extraídas por la red *VGGNet*. En la Figura 2.16 (c) se observan los resultados de la predicción a través de un HPM de los niveles de criminalidad. Como resultados destacables: (i) creación del conjunto de datos *Place-Centric* compuesto de imágenes y registros criminales; (ii) un método denominado HDMiR para la predicción de una puntuación de percepción basado en densidad de ubicaciones por tasa de crímenes.

Otro trabajo enfocado a la apariencia visual es "*What makes an outdoor space beautiful?*" (Seresinhe et al., 2017), el cual presenta un estudio sobre espacios protegidos de Reino Unido (p.ej. áreas naturales, paisajes, campos, entre otros); denominándose escénicos. Este estudio fue realizado a través del juego *Scenic-Or-Not* (UK-gov, 2017),

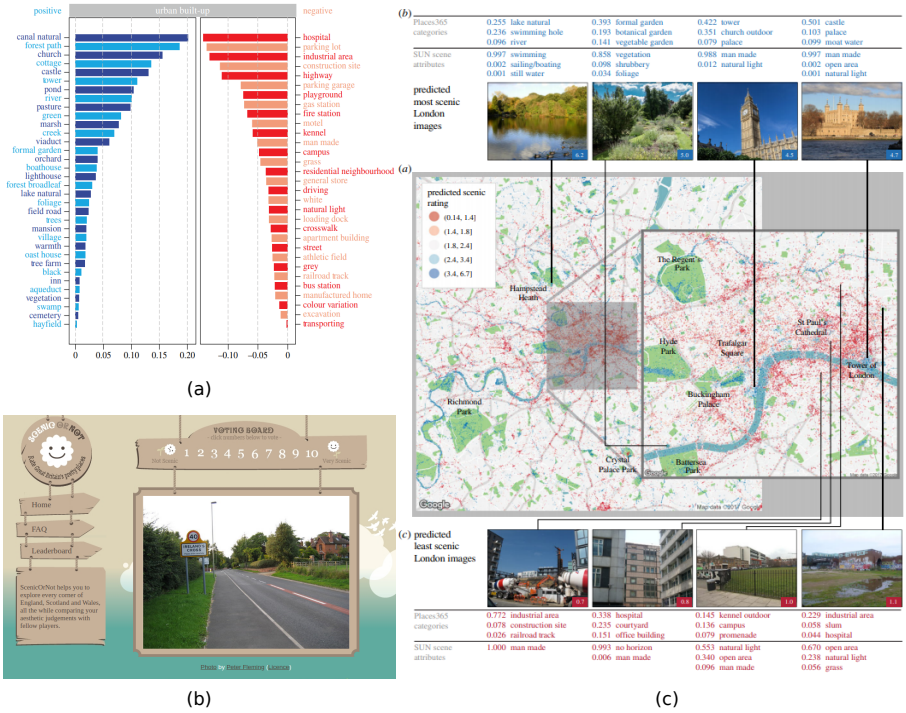


Figura 2.17: (a) Características y atributos que describen una imagen como escénica, mostrando las características consideradas positivas o negativas. (b) Sitio web *Scenic-or-Not*. (c) Mapa general sobre las escenas y sus respectivos atributos. Fuente: (a) (Seresinhe et al., 2017), (b) (UK-gov, 2017; Seresinhe et al., 2017), (c) (Seresinhe et al., 2017)

el cual contiene más de 217 000 imágenes donde una representa 1 km^2 de Gran Bretaña y proviene del sitio web *Geograph* (UK-gov, 2015). Para evaluar e identificar los atributos que contiene cada imagen, se utilizó la red *AlexNet-Places205* con pesos entrenados previamente en el conjunto de datos *Scene UNderstanding* (SUN) (Xiao et al., 2010) para obtener cuáles de los 102 atributos (p.ej. árboles, flores, vegetación, tiendas, entre otros) estaban presentes en cada imagen. Así mismo, contrastaron las características de cada imagen utilizando a *ResNet-152* previamente entrenado sobre *Places365* (Zhou et al., 2017), el cual predice entre 365 categorías de tipo de escenario (p.ej. montañas, lago natural, residencial, estación de tren, entre otros). Denominaron a la composición de los colores negro, azul, marrón, plomo, verde, anaranjado, rosado, morado, rojo, blanco y amarillo presentes en las imágenes como *ElasticNet*. El cual fue utilizado para concatenarlo con las extracciones de *Places205*+SUN y *Places365*, siendo entrenadas con un SVR para predecir las puntuaciones del nivel de escena. Como resultados relevantes tenemos: la identificación de categorías y atributos más importantes de cada imagen (ver Figura 2.17 (a)); mostrando que las características naturales como costas, montañas, ríos naturales y estructuras hechas por el hombre (p.ej. torres, castillos y viaductos) conducen a lugares considerados más escénicos. Por el contrario, escenas con árboles, lugares con áreas verdes como pasto o campos son considerados menos escénicos (ver Figura 2.17 (c)).

En el año 2018, Zhang et al. (2018) utilizaron los datos de *Place Pulse 2.0* y el

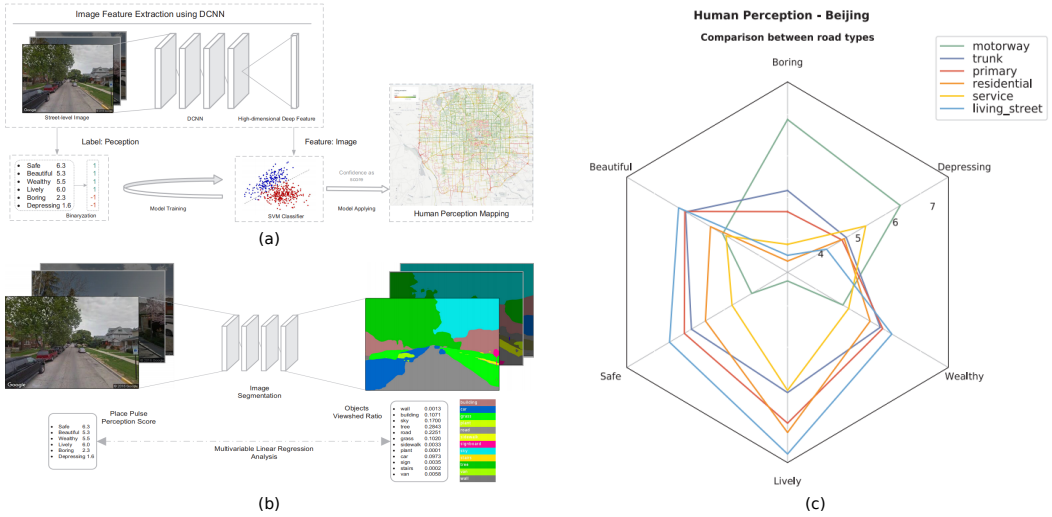


Figura 2.18: (a) Modelo de entrenamiento sobre los datos de *Place Pulse 2.0* y generación de puntuaciones de percepción sobre Beijing. (b) Modelo de entrenamiento sobre las imágenes de Beijing, sus puntuaciones de percepción y los componentes visuales (segmentos) extraídos a partir de las calles. (c) Resultados del tipo de percepción presente en los diferentes tipos de entorno presentes en la ciudad de Beijing-China. Fuente: (Zhang et al., 2018).

segmentador de objetos *PSPNet* (Zhao et al., 2017) para realizar un análisis sobre qué características u objetos influyen en mayor magnitud ante la predicción de la percepción de seguridad en imágenes de las ciudades de Shanghai y Beijing obtenidas a través desde *Tencent Street View Service* (Tencent-Street-View-service, 2016). Cabe destacar que algunas de las imágenes de Beijing fueron obtenidas de *Place Pulse 2.0*. Para obtener un mapa de imágenes, se escogieron imágenes en un intervalo de 50 metros de distancia una de otra, con un tamaño de 400×600 píxeles y ángulos de cámara 0, 90, 180 y 270. Con este proceso se obtuvieron alrededor de 245 388 imágenes de Shanghai y 135 175 imágenes de Beijing. Se extrajeron las puntuaciones de Beijing-*Place Pulse 2.0* en las 6 categorías. Para generar el HPM de Beijing (ver Figura 2.18 (a)) se utilizó un SVR con las características extraídas de *ResNet*. Después, se utilizó la red *PSPNet* para obtener los objetos presentes en las imágenes para entrenarlo usando una *Multi Linear Regressor* (MLR) (Tranmer) (ver Figura 2.18 (b)). Esto permitió entender la percepción por cada tipo de entorno, tales como carreteras, calles, parques, residenciales. En la Figura 2.18 (c) se muestran las 6 categorías mencionadas y el nivel de percepción según cada tipo de ambiente (residenciales, carreteras, etc.). Como resultados destacables obtenemos: (i) el análisis de la presencia de objetos predominantes en diversas partes de la ciudad; (ii) la generación de un HPM y una relación entre presencia de objetos y tipos de lugares.

Por otro lado, muchos estudios psicológicos mencionados anteriormente concluyeron que los grafitis tienen una alta influencia en la presencia de crímenes en una determinada zona, así como también influyen en la percepción de baja seguridad en las calles. Debido a eso, se han realizado estudios relacionados a la presencia de grafiti y su influencia en la ciudad tal como Sao Paulo - Brasil (Tokuda et al., 2019), Belo

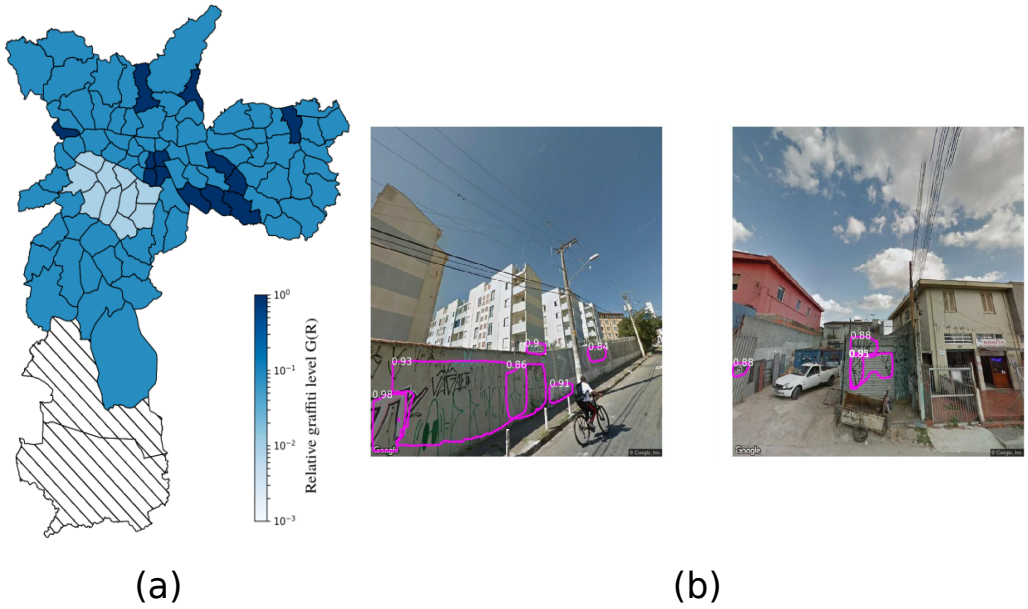


Figura 2.19: (a) Nivel relativo de grafiti en la ciudad de Sao Paulo. (b) Resultados de la detección de grafitis en la ciudad de Sao Paulo, se muestra también la probabilidad de detección. Fuente: (Tokuda et al., 2019)

Horizonte - Brasil Diniz y Stafford (2021) y Medellín - Colombia (Alzate et al., 2021). (I) El primer trabajo estudiado en Sao Paulo, utlizó el detector Mask R-CNN (He et al., 2017) con extractor de características ResNet-101 (He et al., 2015) y pesos entrenados previamente sobre MS-COCO (Lin et al., 2014); el entrenamiento se realizó a unas 10 000 imágenes obtenidas a través de GSV. Una vez entrenado se evalúo el nivel de presencia de grafitis en la ciudad (ver Figura 2.19 (b)) obteniendo una media de presencia utilizando el método GVI, comparándola espacialmente con el *Human Development Index* (HDI) (Índice de desarrollo Humano, el cual mide el índice de crecimiento, tasa de natalidad y mejora de educación) (ver Figura 2.19 (a)); mostrando como resultado que a mayor presencia de grafitis en determinados lugares coinciden con los lugares de bajo HDI. (II) El segundo trabajo estudiado en Belo Horizonte, realiza una comparación geográfica de la presencia de los grafitis y el índice de criminalidad como: agresiones a personas, invasiones de casas, violencia sexual, tráfico de drogas y armas; todos estos datos fueron obtenidos de la policía local en los años 2011, 2015 y 2017. Este estudio se realizó en la ciudad central de Belo Horizonte, utilizando el método *zero-inflated negative binomial regressor* (Garay et al., 2011) para encontrar una correlación entre la presencia de grafitis y crímenes serios, mostrando que en Belo Horizonte existe poca o nula relación entre ambos. (III) El tercer trabajo estudiado en Medellín explora los tipos de grafitis presentes, tales como artísticos y de vandalismo. Utilizaron el modelo *Faster R-CNN* (Ren et al., 2017) y el conjunto de datos STORM (Charalampos et al., 2019). Como resultados (i) mostraron que los grafitis tienen mayor presencia en lugares comerciales correspondientes al centro de la ciudad, zonas industriales y poca presencia en zona residenciales; (ii) también presentaron una versión extendida del conjunto de datos STORM, adicionando solamente 373 imágenes.

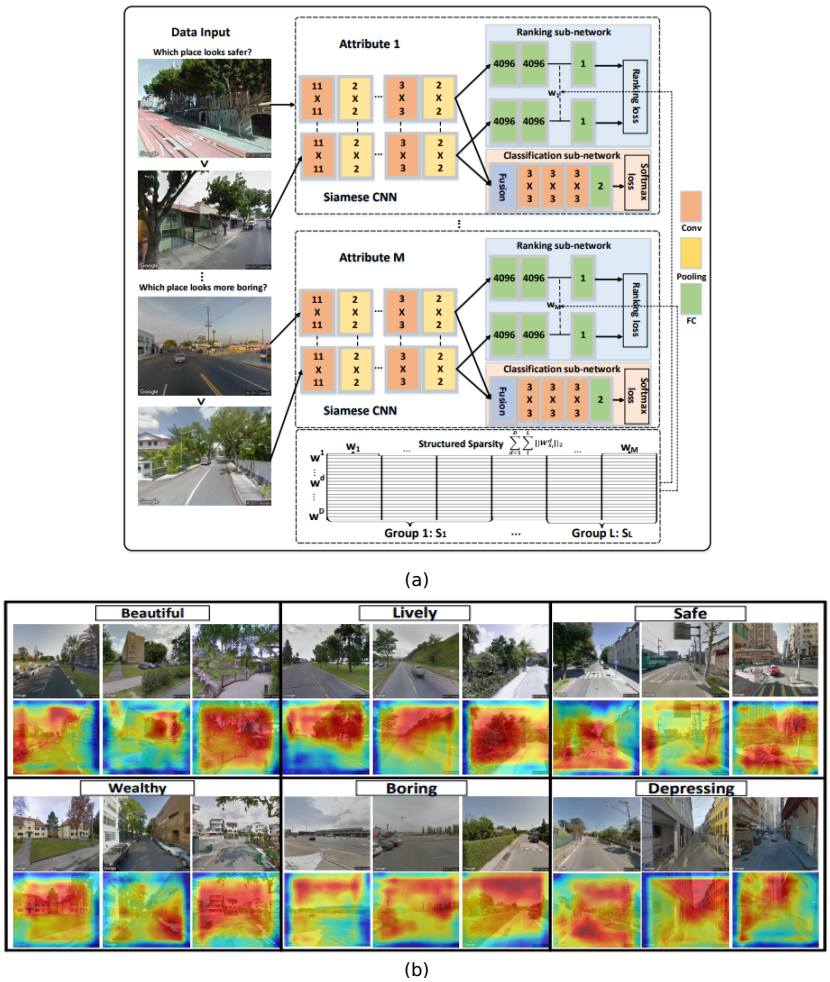
En esta sección se presentaron diferentes trabajos orientados a la predicción de la percepción de calles (p.ej. seguridad) utilizando diferentes métodos de extracción de características basados en redes convolucionales profundas, para después entrenar las características extraídas utilizando diversos modelos como SVM o derivados como *RankSVM* y DCNN. Además, mostraron estudios sobre el impacto de la apariencia visual de las calles, como la presencia de árboles, edificios, áreas verdes o grafitis. Dichos estudios se realizaron con datos reales de históricos de crímenes u otros datos como los precios de las casas, tasas de robos, entre otros; con los cuales realizaron la evaluación de las predicciones de sus modelos.

2.4. Interpretación y Visualización de Características Extraídas

En esta sección se exponen los trabajos sobre la explicación y visualización acerca de cuáles son las características de una determinada correlación entre una imagen y la percepción urbana, tomando como base de estudio la influencia de elementos visuales presentes en las imágenes de calles analizadas. Los métodos de visualización para el entendimiento de redes profundas es a través de métodos que resaltan las características aprendidas en la red, tales como GBP (Springenberg et al., 2014), CAM (Zhou et al., 2016a) y *Grad-CAM* (Selvaraju et al., 2017).

En el año 2019, continuando la idea presentada en Doersch et al. (2012), donde se estudió *Place Pulse 2.0* y las predicciones de posibles ganadores a través de una *SiameseNet*, Min et al. (2019) abordaron la idea de encontrar las características visuales más relevantes de la predicción de una comparación al momento de realizarla. Es decir, si comparamos dos imágenes en la categoría segura, qué característica es la que indica "esta imagen es más segura?". Para ello, propusieron el método denominado MTDRALN el cual aprende todos los atributos de percepción simultáneamente. Este método está compuesto por dos *multi-task siamese networks* (von Platen et al., 2020) con dos tipos de subredes, una de clasificación con pesos entrenados previamente sobre la red *Places205* y otra de ranking usando un *RankSVM*. En estas redes, cada *SiameseNet* aprenderá un atributo relativo de cada par de imágenes a comparar; a través de una matriz dispersa de atributos, permite el fácil y rápido intercambio entre características de un conjunto de atributos $A = \{a_m\}_{m=1}^{M}$ para cada atributo "m" (p.ej. seguro) relacionados y no relacionados.

El objetivo del método es conseguir la matriz dispersa mencionada anteriormente definida como los valores de atributos representada por $W = [w_1, w_2, ..., w_M] \in \mathbb{R}^{D \times M}$ la cual es dividida en grupos de atributos relativos, por ejemplo con *Place Pulse 2.0* se denomina grupo positivo: *safe*, *lively*, *beautiful*, *wealthy* y grupo negativo: *depressing*, *boring*. Para el entrenamiento del modelo, redujeron el conjunto de datos filtrando las 161 882 empates dentro del total de 1 208 808 comparaciones, así mismo, solo analizaron las ciudades de New York, Berlín, Tokio y Moscú. Finalmente, para determinar qué objetos son los más influyentes de cada categoría utilizan la red *PSPNet* para



(a)

(b)

Figura 2.20: (a) Arquitectura de la *MultiTask Deep Relative Attribute Learning Network* (MTDRALN). (b) Resultados de aplicar *Grad-CAM* a las imágenes de cada categoría pudiéndose observar que los atributos relativos engloban casi toda la imagen, mostrando que las imágenes clasificadas en un misma categoría tienen aspecto visual similar. Fuente: (Min et al., 2019).

obtener la segmentación de objetos con el objetivo de calcular una intersección entre las puntuaciones de percepción obtenidas por la *RankSVM* y las áreas generadas por el método de interpretación *Grad-CAM* tal como se puede observar en la Figura 2.20 (b). Como resultado destacable resaltamos: (i) implementación de una doble *SiamesNet* junto a la interpretación usando *Grad-CAM* para entender qué características de las dos imágenes de entrada determinan la categoría predominante en las dos.

En ese mismo año, Xu et al. (2019) propusieron otro método basado también en el análisis de las comparaciones de imágenes y su atributos característicos. El objetivo es identificar qué factores dentro de una imagen influyen directamente en la percepción de las calles resaltando la importancia de la información semántica que proveen los objetos a través de los mapas de características de cada atributo (p.ej. seguro); estos mapas de características son obtenidos por la técnica *Grad-CAM*. Por lo cual, se proponen 2 modelos para la comparación de imágenes, el primero predice al ganador generando

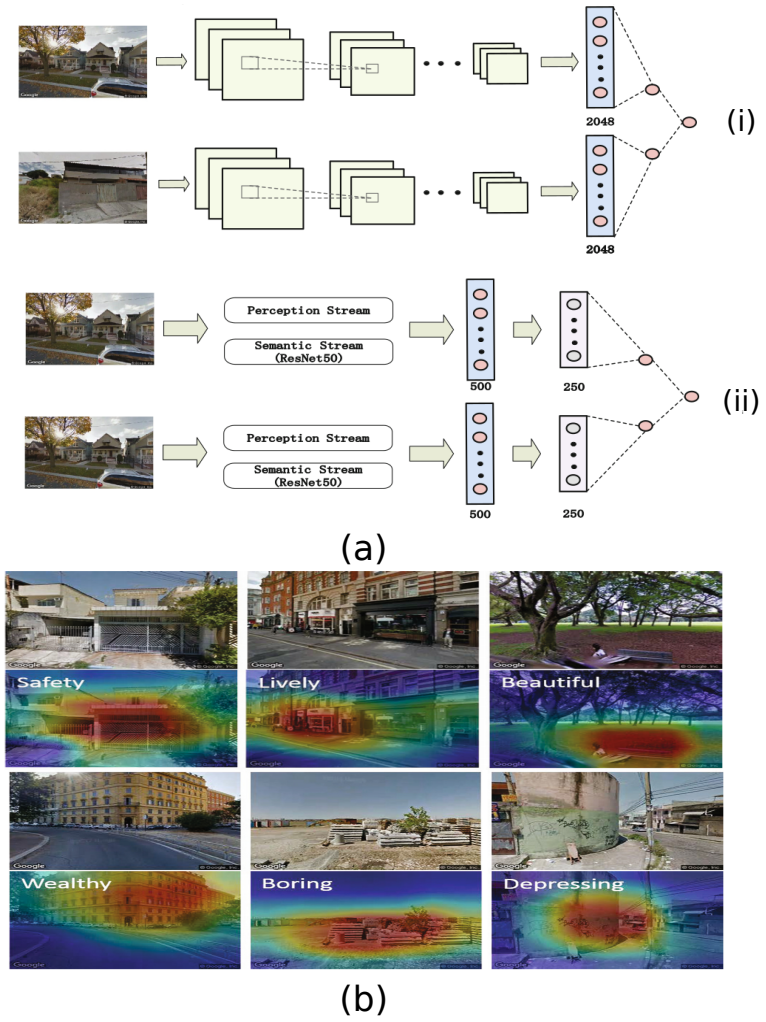


Figura 2.21: (a)-(i) Arquitectura de *Perception Rank Network* (PRN) y (a)-(ii) *Semantic-Aware Perception Network* (SAPN). (b) Resultados de aplicar *Grad-CAM* a algunas imágenes del conjunto dedatos *Place Pulse 2.0*, escogiendo una por cada una de las categorías: *safe, wealth, depression, boring, lively, beautiful*. Fuente: (Xu et al., 2019).

una puntuación de percepción a partir de esa comparación; y el segundo predice una puntuación de percepción para ambas imágenes. Dichos modelos están conformados por 2 sub-redes denominadas *Perception Stream* y *Semantic Stream*.

Ambas sub-redes son un *fine-tunned* de *ResNet-50* previamente entrenada sobre *ImageNet* modificándolas a partir de los bloques 4 y 5 de la red, cambiando cada *max pooling* por un GAP cuyas arquitecturas de ambas redes se muestran en la Figura 2.21 (a). La primera sub-red denominada PRN evalúa las características entre 2 imágenes comparadas entre sí, obteniendo como salida el valor de la regresión de ambas imágenes. La segunda sub-red denominada SAPN evalúa la puntuación de una imagen; y también tiene dos sub-componentes que utilizan el modelo *ResNet-50*. El primer componente *Semantic Stream* (S-Stream) utiliza la salida de tamaño 1000 de *ImageNet*; esta red tiene como salida la puntiación entre las dos. En cambio, el segundo componente se

utiliza el GAP extraído de la última convolución ($1 \times 1 \times 2048$) denominada *Perception Stream* (P-Stream). Como resultados destacables tenemos: la SAPN tiene mejor resultado que la PRN en la predicción de las puntuaciones de percepción. Además, muestra que los modelos entrenados sobre *safety, lively, beautiful, wealthy* tienen mejor resultado evaluados entre ellos mismos que al evaluar uno de estos con las categorías *boring, depressing* y viceversa.

En esta sección vimos trabajos relacionados a la interpretación de modelos como *Siamese Network* y PRN y la predicción de la percepción, ambos modelos buscaban características diferentes entre cada par de imágenes obtenidas de los datos para posteriormente usar un método de interpretación (generalmente el *Grad-CAM* observado en ambos métodos. En la Figura 2.20 (b) y la Figura 2.21 (b)) se muestra los resultados de ambos trabajos, resaltando las características que los modelos consideran relevantes dentro de cada imagen evaluada en una determinada categoría de percepción.

2.5. Consideraciones Finales

En este capítulo hemos realizado una presentación exhaustiva de los trabajos relacionados a nuestro trabajo, el cual tiene dos principales temáticas: Análisis y predicción de la percepción de seguridad y la utilización del conjunto de datos denominado *Place Pulse 2.0*. En su mayoría, encontramos que gran parte de los trabajos estudiados tiene como enfoque principal cómo encontrar un método para predecir la percepción de seguridad urbana utilizando el conjunto de datos *Place Pulse* u otros. La finalidad principal es encontrar qué aspectos de la apariencia visual de las calles pueden influir en esta percepción. Al pasar de los años, cada trabajo ha propuesto un modelo cada vez más complejo para extraer características y resaltarlas, o por el otro lado, buscan complementar información utilizando otros conjunto de datos e intentar describir un panorama más general. En nuestro trabajo, nuestro principal objetivo es analizar, explorar y entender la composición de los datos en *Place Pulse 2.0*; para que una vez comprendido cómo se comportan los datos. Este enfoque difiere de los mencionados anteriormente, puesto que ninguno realizó un análisis exploratorio de los datos antes de proponer algún tipo de solución. Este paso nos permitirá entender qué tipo de modelo o técnica sería el más adecuado frente a los datos estudiados.

Se ha realizado una revisión acerca de los trabajos relacionados a: (a) análisis de la percepción urbana y (b) extracción de características y componentes visuales de imágenes y (c) interpretación y visualización de características extraídas. También hemos descrito algunos estudios realizados sobre la percepción urbana y cómo se intenta cuantificar el nivel de percepción explicando a través de las relaciones encontradas en las características aprendidas por modelos entrenados sobre datos de imágenes de calles con las puntuaciones de percepción. Sin embargo, cabe mencionar que en ninguno de los trabajos descritos y relacionados al conjunto de datos *Place Pulse 2.0* se realizó un análisis de los datos. En el siguiente capítulo describiremos en detalle el análisis realizado al conjunto de datos *Place Pulse 2.0*.

Capítulo 3

Análisis Exploratorio del Conjunto de Datos *Place Pulse 2.0*

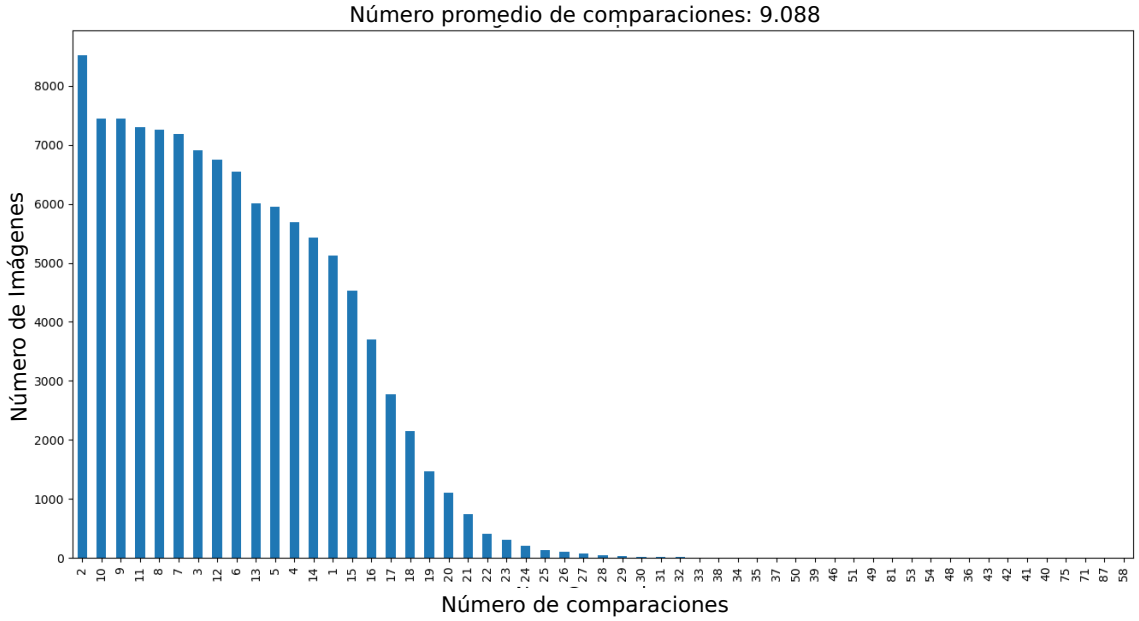


Figura 3.1: Mostramos el número de comparaciones en la categoría *safety* (segura), en la cual se observa que el número de comparaciones no supera en promedio 10 por imagen, además, la mayoría de imágenes fueron comparadas 2 veces solamente. Fuente: El autor.

Tal como escribimos en el Capítulo 1, la motivación del presente trabajo está en estudiar la percepción de seguridad urbana a través del estudio y análisis del conjunto de datos *Place Pulse*; con el cual se propone un estudio y análisis exploratorio para entender el comportamiento de los datos. En rasgos generales, se sabe que *Place Pulse 2.0* es un conjunto de comparaciones entre dos imágenes de ciudades iguales o diferentes, evaluadas en 6 diferentes categorías: seguro, depresivo, aburrido, opulento, bonito y bueno para vivir respectivamente (a partir de ahora, nos referiremos como *safety, depressing, boring, wealthy, beauty* y *lively*) y no necesariamente el mismo número de

veces. También sabemos que el número promedio de comparaciones no supera 10 comparaciones por imagen (ver Figura 3.1), por lo cual algunos algoritmos como *TrueSkill* no funciona del todo bien (Dubey et al., 2016; Naik et al., 2014).

Es de importancia mencionar que en el presente trabajo, estaremos enfocándonos exclusivamente en la categoría de percepción de *safety* (seguridad), debido a que esta categoría posee la mayor cantidad de comparaciones de imágenes. El análisis realizado fue dividido en pequeñas secciones que describiremos a continuación: (i) descripción de los datos; (ii) cálculo de las puntuaciones de percepción; (iii) análisis de los posibles "niveles de generalización geográficas" de los datos; y (iv) análisis de la disparidad de los datos.

3.1. Descripción de los Datos

El conjunto de datos que utilizaremos en nuestro trabajo es obtenido del sitio web *Place Pulse* (MIT-Media-Lab, 2013) teniendo 2 versiones, la primera es *Place Pulse 1.0* del 2013 y la segunda versión es *Place Pulse 2.0* del 2016, en la cual está centrado el presente trabajo. En ambas versiones de *Place Pulse* se compone de 8 campos: para cada comparación se cuenta con las posiciones de las imágenes (latitud y longitud), los identificadores de imágenes (derecha e izquierda), el resultado de la comparación y la respectiva categoría evaluada. En la Figura 3.2 se muestra los datos crudos, es decir, como están los datos sin procesar. Se observa que son comparaciones entre dos imágenes enfatizando al ganador.

Id img izquierda	Id img derecha	ganador	latitud img izquierda	longitud img izquierda	latitud img derecha	longitud img derecha	categoría
513d7e23fdc9f	513d7ac3fdc9f	igual	40.744156	-73.93557	-33.52638	-70.591309	depresivo
513f320cfdc9f	513cc3acfdc9f	izquierda	52.551685	13.416548	29.76381	-95.394621	seguro
513e5dc3fdc9f	5140d960fdc9f	derecha	48.878382	2.403116	53.32932	-6.231007	Bueno para vivir

Figura 3.2: Mostramos la composición del conjunto de datos *Place Pulse*, se observa la comparación entre las dos imágenes y el ganador en cada categoría. Fuente: El autor.

Place Pulse 1.0: A finales del 2013, *Place Pulse 1.0* contiene 73 806 comparaciones, 4136 imágenes y 4 ciudades provenientes de dos países (US y Austria): New York City, Boston, Linz y Salzburg y tres tipos de comparaciones: *safe*, *wealth*, y *unique*.

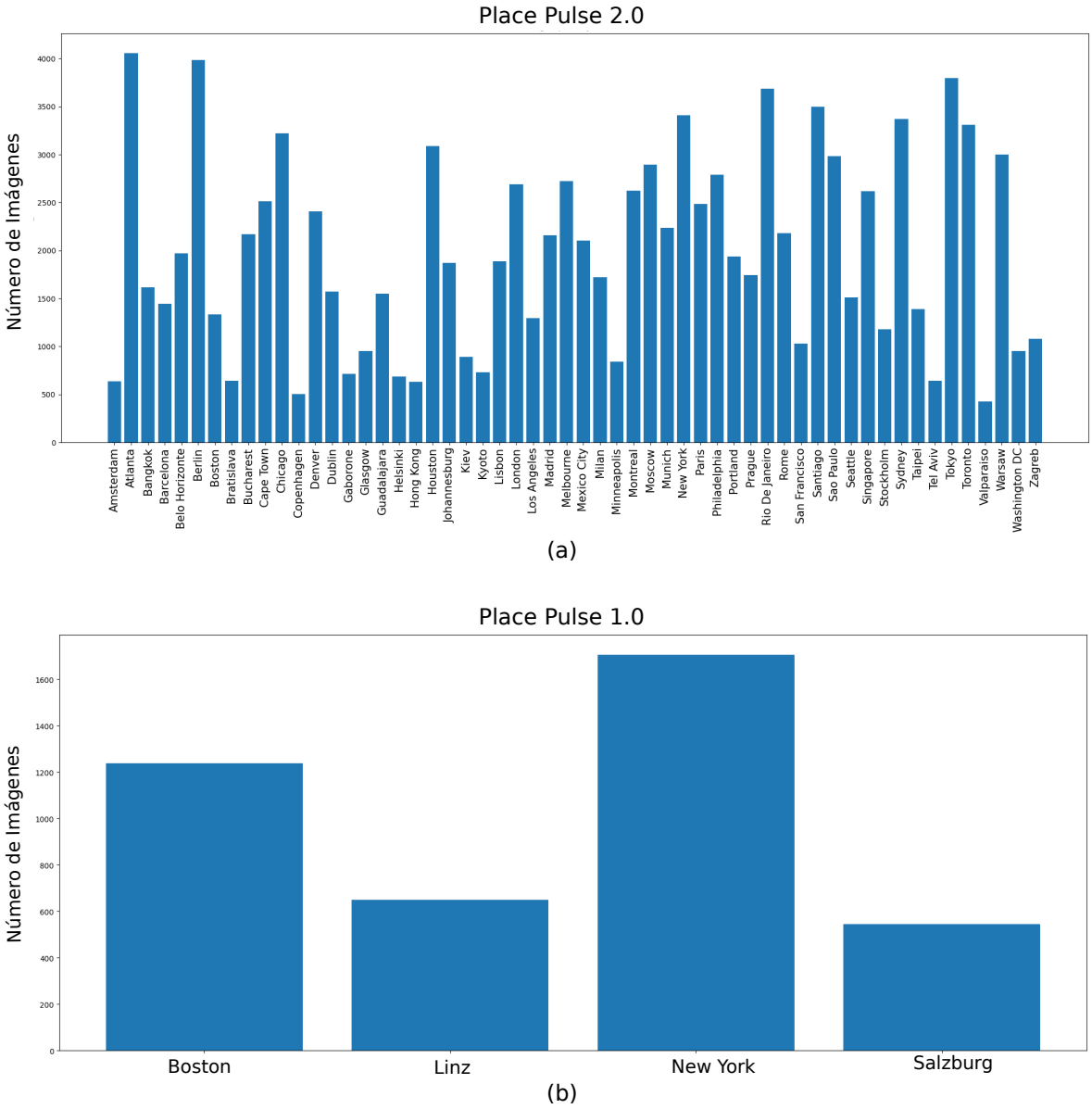


Figura 3.3: (a) Relación entre las ciudades y el número de imágenes dentro del conjunto de datos *Place Pulse 2.0*; (b) Relación ciudades-número de imágenes dentro del conjunto de datos *Place Pulse 1.0*. Fuente: El autor.

Place Pulse 2.0: En el año 2016, *Place Pulse 2.0*, la cual ya contenía alrededor de 1.22 millones de comparaciones de 111 390 imágenes de 56 ciudades provenientes de 32 países entre los 5 continentes, tal como se observa en la Figura 3.4, del cual podemos notar que hay más imágenes de ciudades de Europa y América del Norte que en otros lugares; Así mismo, se tiene seis tipos de categorías ya mencionadas anteriormente.

Figura 3.4: Mapa de las 56 ciudades con imágenes de calles contenidas en el conjunto de datos *Place Pulse 2.0*, se observa que en Europa y América del Norte se cuentan con mayor número de ciudades evaluadas en el sitio web *Place Pulse* (MIT-Media-Lab, 2013) que en otros lugares. Cabe mencionar que puntos con un mismo color, pertenecen al mismo país. Fuente: El autor.

3.2. Cálculo de las Puntuaciones de Percepción

En esta sección describiremos las ecuaciones utilizadas para calcular las puntuaciones ponderadas en cada categoría, cabe mencionar que estas están fuertemente relacionadas al número de veces que una imagen ganó o perdió; según sus comparaciones por imagen. Como ejemplo, en la Figura 3.1 mostramos el número de comparaciones en la categoría segura. A continuación, describiremos y presentamos las ecuaciones que utilizamos para calcular dichas puntuaciones de percepción.

Al tener una imagen i comparada con otras imágenes muchas veces en diferentes categorías, el porcentaje de veces que i fue elegida indica la intensidad de la percepción de la imagen, puesto que de todas las imágenes evaluadas, se obtendrá que tanto % fue considerada con mayor percepción (p.ej. de seguridad) comparado al resto de imágenes. Además, sea $i^{'}$ una imagen comparada con i, la intensidad de $i^{'}$ también afecta a la intensidad de la imagen i, por lo cual, se define las tasa positiva $W_i = \frac{w_i}{w_i+d_i+l_i}$ y la tasa negativa $L_i = \frac{l_i}{w_i+d_i+l_i}$ de una imagen i de una determinada categoría. En donde w_i indica cuántas veces ganó, l_i cuántas veces perdió y d_i empató; Entonces a partir de esto, se calcula el *Q-score* denominado $q_{i,k}$ para cada imagen i de una determinada categoría k:

$$q_{i,k} = \frac{10}{3}(W_i + \frac{1}{w_i}(\sum_{k_1=1}^{w_i} V_w(k_1)) - \frac{1}{l_i}(\sum_{k_2=1}^{l_i} V_l(k_2)) + 1) \tag{3.1}$$

La Ecuación 3.1 se interpreta como la tasa positiva de la imagen i como una aproximación ponderada sobre todas las imágenes k_1 a las que ganó y una penalización de todas las imágenes k_2 con las cuales perdió; donde V_w es el vector de tasas positivas de las imágenes a las que ganó y V_l es el vector de tasas negativas de las imágenes con las que perdió, para finalmente obtener una puntuación entre 0 y 10 obteniendo una escala utilizada en estudios anteriores (Nasar et al., 1993; Nasar, 1998).

Una vez realizado este paso, podemos extraer información del conjunto de datos descrito, En los Cuadros 3.1 y 3.2 se las estadísticas respectivas para cada versión. Por ejemplo, en *Place Pulse 1.0* se obtiene el número de imágenes por ciudad y la puntuación de percepción promedio por cada categoría evaluada. Se observa que *Place Pulse 2.0* cuenta con más información, especialmente a nivel de continente, país y ciudades. Así como también el número de imágenes asociadas a cada categoría y su respectivo promedio. En la Figura 3.3 se muestra la relación entre el número de imágenes por ciudad, en la cual se observa el incremento del número de ciudades y cantidad de imágenes por ciudad. También se observa que el número de imágenes por ciudad son muy disparejos, especialmente en *Place Pulse 2.0*.

Place Pulse 1.0: En el Cuadro 3.1 se puede observar la cantidad de imágenes por ciudad y las puntuaciones promedio por cada categoría obtenidas a partir de una limpieza de información. Se observa que en la categoría *safety* se tiene la mayor puntuación promedio, así como también la mayor cantidad de imágenes evaluadas. Lo cual es verificado observando que las imágenes y sus posiciones existan y sean correspondientes a una determinada calle.

Place Pulse 2.0: En el Cuadro 3.2 se puede observar la cantidad de imágenes por ciudad, país y continente es más amplia que la versión anterior. Además, las puntuaciones promedio por cada categoría son mayores. También destacamos que la categoría *safety* posee la mayor cantidad de comparaciones y la puntuación promedio. Además, vemos que el país USA posee la mayor cantidad de imágenes y ciudades evaluadas en total. Es por eso, que dividimos el continente de América en Sur y Norte; además de que las imágenes son muy diferentes en la apariencia visual,

Place Pulse 1.0				
Ciudades	# de imágenes	promedio *safe*	promedio *wealth*	promedio *unique*
New York	1705	4.47	4.31	4.46
Boston	1237	4.93	4.97	4.76
Linz	650	4.85	5.01	4.83
Salzburg	544	4.75	4.89	5.04
Total	4136			

Cuadro 3.1: Datos estadísticos acerca de las ciudades y puntuaciones promedio de cada categoría de percepción dentro del conjunto de datos *Place Pulse 1.0* obtenidas a partir del archivo JSON descargado del sitio web MIT-Media-Lab (2013).

Place Pulse 2.0			
Continente	#países	# ciudades	# imágenes
Europa	19	22	38 747
América del Norte	3	17	37 504
América del Sur	2	5	12 524
Asia	5	7	11 417
Oceanía	1	2	6097
África	2	3	5101
Total	32	56	111 390

(a)

Place Pulse 2.0		
Categoría	# comparaciones	puntuación promedio
Safety	368 926	5.18
Lively	267 292	5.08
Beautiful	175 361	4.92
Wealthy	152 241	4.89
Depressing	132 467	4.82
Boring	127 362	4.81
Total	1 223 649	

(b)

Cuadro 3.2: Estadísticas de las puntuaciones de percepción: (a) Continentes e imágenes; notar que dividimos América del Norte y América del Sur, (b) Número de comparaciones; notar que *safety* fue la categoría más comparada y con la puntuación medio mayor.

3.3. Análisis de los Niveles de Generalización Geográfica

Hasta este punto, ya hemos calculado las respectivas puntuaciones de percepción para cada ciudad, sin embargo, a partir de que sabemos a qué ciudad pertenece una imagen podríamos extender esta información para saber a qué país pertenece y a qué continente. Como mencionamos en el capítulo anterior, definimos unos "niveles de generalización geográficos" como aquellas regiones que podemos utilizar para segmentar los datos a través de ciudad, país, continente o a nivel global. Siguiendo esta idea, se procedió a calcular las puntuaciones de percepción en cada uno de estos niveles; es decir, a través de los datos de latitud y longitud, filtramos las comparaciones entre dos imágenes cuyas ubicaciones están en el mismo "nivel de generalización geográfica". Para el cálculo de las puntuaciones se prefirió dividir el continente América en dos: América del Norte y América del Sur. Una vez calculadas las puntuaciones de las comparaciones filtrando imágenes comparadas en la misma ciudad, mismo país, mismo continente y a nivel global (por no decir "mismo mundo") se procedió a observar las distribuciones de las puntuaciones encontradas.

En el Cuadro 3.3 se puede observar el impacto en las puntuaciones luego de ser

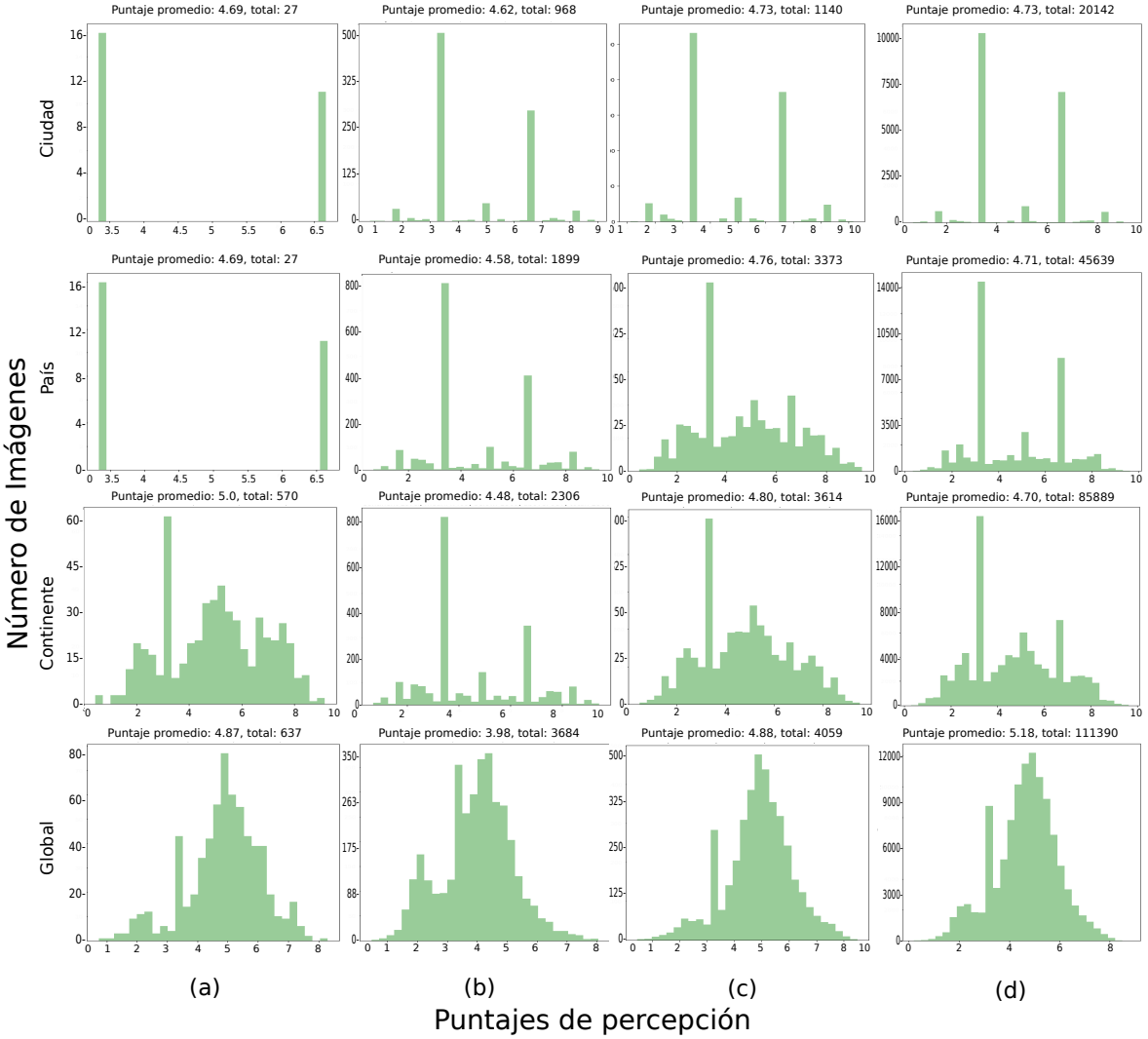


Figura 3.5: Distribución de las puntuaciones de percepción en los diferentes "niveles de generalización geográficos" en 3 ciudades diferentes: (a) Ámsterdam-Países Bajos, única ciudad analizada; (b) Río de Janeiro-Brasil con 3 ciudades; (c) Atlanta-USA, el cual tiene 17 ciudades; (d) todas las ciudades (global). Fuente: El autor.

calculados a través de esos niveles, así mismo, notamos una reducción en el número de imágenes por cada categoría evaluada siendo que la categoría de seguridad es la que mantiene la mayor cantidad de imágenes en todos los casos y el mayor puntaje de percepción promedio (ver Cuadro 3.2 (b)). Se observa que el número de comparaciones de una imagen con otra de la misma ciudad, país o continente es mucho menor que con otras a nivel global. Esto solo corrobora la idea que las imágenes evaluadas de par en par eran seleccionadas de manera aleatoria y no filtradas por una misma localidad. Así mismo, se observa una drástica reducción de las imágenes evaluadas en la misma ciudad y a nivel global (cerca del 82 % del total). Es importante mencionar que en general, todos los países poseen como máximo 3 ciudades, algunos poseen solo una y el único caso con más de 3 ciudades es USA. Sabiendo esto, se procedió a observar la distribución de las puntuaciones obtenidas a partir de cada nivel, notamos que en los donde sólo poseemos una ciudad (la mayoría de países de Europa) o 2 o tres ciudades

Place Pulse 2.0						
Geographic level	*safety*	*lively*	*Beautiful*	*Wealthy*	*Depressing*	*Boring*
City	20 143	14 803	9410	7642	6556	6148
Country	45 640	38 216	28 811	24 326	21 171	20 931
Continent	85 890	79 788	66 792	57 780	52 504	52 031
Global	111 390	111 349	110 767	107 796	105 496	106 364

Cuadro 3.3: Cantidad de imágenes obtenidas por categoría luego de realizar el cálculo en cada "nivel de generalización geográfico". Vemos que la categoría *safety* presenta puntuaciones de percepción para todas las imágenes de *Place Pulse 2.0* a nivel Global (Ver Cuadro 3.2).

(tal como Brasil, Chile, México y Japón) el cálculo de las puntuaciones en los niveles de ciudad y país no tuvieron impacto. Por otro lado, el caso de una ciudad de USA (p.ej. Atlanta) sí mostró un cambio considerable entre ciudad y país, esto es debido a que al tener 17 ciudades, cuenta con mayor número de comparaciones a nivel de país. A nivel de continentes, tenemos que África y América del Sur presentan una pésima distribución, esto es debido no sólo al número reducido de ciudades (3 África y 5 en América del Sur), sino también al número de comparaciones obtenidas. A nivel global, se observa una distribución más equitativa respecto a las puntuaciones calculadas.

Como se ve en la Figura 3.5, comparamos la distribución de cada caso encontrado: países con uno/dos ciudades (Países Bajos-Amsterdam), países con 3 ciudades (Brasil-Rio de Janeiro) y USA con 15 ciudades (único país con más de 3 ciudades). Debido a esto, realizar una especificación entre los diferentes niveles propuestos no es posible porque el número de comparaciones disminuye afectando directamente a las puntuaciones y número de imágenes. Además, se observa que aún después de calcular a nivel global se observa un gran número de imágenes que tienen una puntuación de 3.33; esto se debe a que la mayoría de imágenes fueron comparadas como máximo 2 veces (ver Figura 3.1) de las cuales, no ganó ni una vez. Por ejemplo, a nivel ciudad notamos que el número de comparaciones entre 2 imágenes de Rio de Janeiro, del total de 3684 imágenes, solo obtenemos 968 con puntuaciones 3.33 y 6.66 correspondiente a la gran mayoría de imágenes. A partir de esto, descartamos el posible enfoque de analizar localmente a las ciudades, cuyas imágenes fueron comparadas en mayor cantidad con otras imágenes de diferentes ciudades.

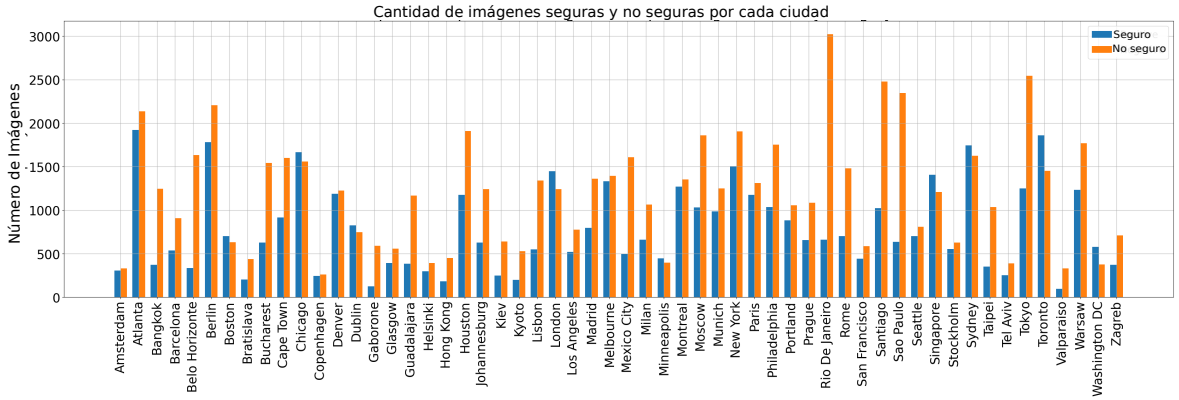


Figura 3.6: Utilizando un umbral de 5.0 para designar si se segura o no segura, se muestra la disparidad en la cantidad de imágenes entre la percepción segura y no segura por cada ciudad. Se observa que en la mayoría de ciudades, la percepción no segura es mucho mayor (p.ej. Río de Janeiro y Sao Paulo). Fuente: El autor.

3.4. Análisis de la Disparidad de los Datos

Como se mostró en la sección anterior, al observar las distribuciones en los niveles de ciudad, país y continente de cada ciudad, percibimos que al tener mayor cantidad de imágenes con una puntuación de 3.33, generaba una disparidad en los datos para los niveles ciudad, país y continente. Sin embargo, a nivel global observamos una mitigación en la cantidad de imágenes dispares pero aún conservando un alto número de imágenes comparadas con puntuaciones de 3.33; Además, se observa que a nivel global, la distribución de las puntuaciones tiene una mejor variedad que en otros niveles. Esto es debido a que en general el promedio global es de 5.188 (vea Cuadro 3.2 (b)). Por lo cual, se decidió utilizar como umbral el valor 5.0 para la división de las clases segura y no segura, esto también es porque en el "nivel de generalización geográfica" global, la mayor cantidad de imágenes se encuentra con valor de 5.0 ± 0.1 (ver Figura 3.5).

En la Figura 3.6 mostramos la disparidad de las puntuaciones utilizando como umbral el número 5.0 para asignar las etiquetas a cada imagen (seguras y no seguras). Se puede observar que la gran mayoría de ciudades presentan una disparidad alta, especialmente en las ciudades como Rio de Janeiro, Belo Horizont y Sao Paulo las cuales curiosamente pertenecen al mismo país. Mientras que en ciudades como Washington DC, Toronto, Sydney, Singapure, Londres, Boston y Chicago presentan (en menor medida) una disparidad favoreciendo a la percepción segura. Adicionalmente, tenemos ciudades como Atlanta, Amsterdam, Denver, Dublín, Montreal, Melbourne y Minneapolis con una proporción muy cercana entre seguras y no seguras. Como comentario final, debido a los resultados obtenidos anteriormente de los "niveles de generalización geográfico" y la disparidad encontrada, nosotros decidimos continuar nuestro análisis y experimentos enfocados en las puntuaciones de percepción calculados a nivel global, no por ciudad, no por país, ni por continente.

3.5. Consideraciones Finales

En este Capítulo se presentó el análisis exploratorio realizado sobre el conjunto de datos *Place Pulse 2.0*, se procedió a calcular las respectivas puntuaciones de percepción en la categoría de seguridad. Estos valores fueron calculados a través de diferentes "niveles de generalización geográficas" tales como ciudad, país, continente y global. Al analizar la distribución de puntuaciones resultante, se observa que la mejor posible se obtiene cuando utilizamos todos los datos. Así mismo, al tomar como umbral el valor de 5.0 se obtiene una disparidad de aproximadamente 11 mil imágenes en la categoría "no segura". Este análisis nos permitió entender las limitaciones del conjunto de datos, tales como: (i) el conjunto de datos está sesgado por la percepción individual de cada voluntario que participó en la creación de este conjunto de datos; (ii) es necesario analizar todos los datos en conjunto, no es posible realizar análisis regionales; (iii) los datos presentan una gran disparidad a partir del umbral escogido (siguiendo las distribuciones de las puntuaciones calculadas). En el siguiente Capítulo presentamos los modelos y métricas que utilizaremos en los experimentos.

Capítulo 4

Predicción de la Percepción de Seguridad Urbana

En este capítulo se presentan los modelos, técnicas y métricas que utilizaremos para la clasificación de la percepción de seguridad. Para ello, definimos un lineamiento de experimentos a realizar, dividiremos en 3 grupos según el tipo de técnica utilizada y tipo de aprendizaje. A nivel general, utilizaremos dos tipos de aprendizaje: (a) Aprendizaje Supervisado y (b) Aprendizaje Semi-Supervisado. En el grupo de Aprendizaje Supervisado, utilizaremos dos técnicas denominadas *transfer-learning* y *fine-tuning*. En el grupo de aprendizaje semi-supervisado utilizaremos un modelo GAN, la cual se compone de dos sub-modelos llamados discriminador y generador. Tal como mencionamos antes, la tarea principal será la clasificación de imágenes entre las clases segura y no segura. Dicho entrenamiento será realizado en todas las 56 ciudades y a nivel global, además, las métricas que utilizaremos para estos experimentos son *Accuracy*, *F1 score* y *Area Under Curve* (AUC) calculado a partir de los valores obtenidos del *Precision-Recall*. Sin embargo, para comparar el desempeño de los modelos, nos basamos principalmente en los valores reportados por AUC y *F1 score*. Éstas métricas son calculadas como:

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{4.1}$$

$$Precision = \frac{T_P}{T_P + F_P} \tag{4.2}$$

$$Recall = \frac{T_P}{T_P + F_N} \tag{4.3}$$

$$F1_{score} = 2\frac{Precision * Recall}{Precision + Recall} \tag{4.4}$$

En donde T_P significa *True Positive*, T_N significa *True Negative*, F_P significa *False Positive* y F_N significa *True Negative*.

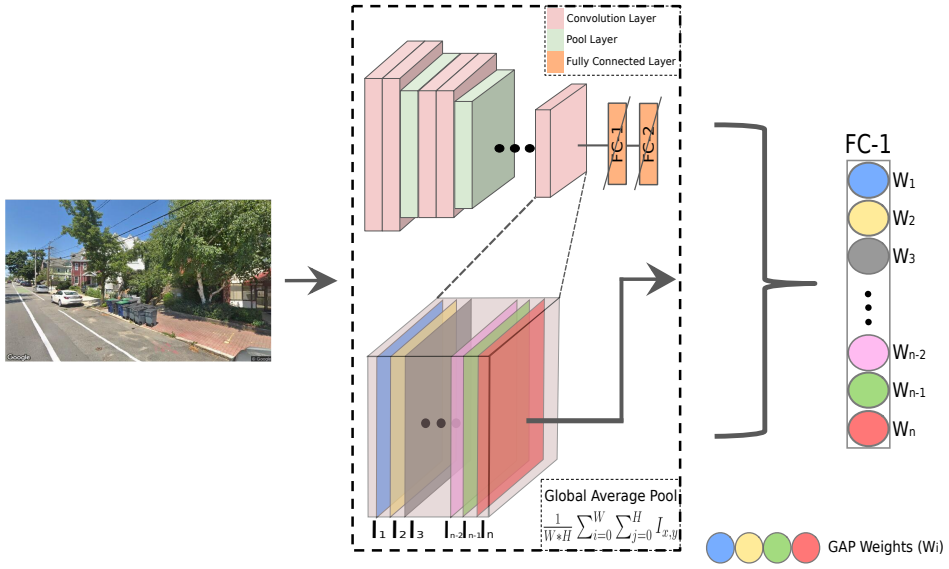


Figura 4.1: Se presenta la modificación del modelo VGG16, denominado "VGG-GAP" a partir de ahora. Para ambos casos, *baseline* y *fine-tuning* utilizaremos esta arquitectura, en el primer caso como un extractor de características; para el segundo, será un reemplazo de las capas *max pooling* y *flatten* del modelo original. Fuente: El autor.

4.1. Modelos del Grupo *Transfer-Learning (Baseline)*

Definiremos a nuestro grupo de modelos *baseline* al conjunto de modelos basados en *transfer learning*. Para los experimentos, usamos principalmente a las red VGG16, ResNet y Excepcition con pesos entrenados previamente en *ImageNet* (Russakovsky et al., 2015) y *Places365* (Zhou et al., 2016b)), los cuales utilizaremos como extractores de características. Decidimos escoger los pesos previamente entrenados en dos bases de datos diferentes, estos son: (i) *ImageNet* presenta un excelente desempeño en el entrenamiento y predicción de imágenes tales como animales u objetos; (ii) *Places365* presenta una relación a la predicción de lugares/escenas, tales como lugares residenciales, calles, parques, etc. Lo cual está directamente relacionado a nuestro conjunto de datos estudiado (imágenes de calles). Decidimos dar un estudio más profundo a la red *VGG16* sobre otras, esto es debido a que la red *VGG16* presentó un mejor rendimiento, desempeño y precisión en conjuntos de datos como *Places365*, SUN y *Scene15* explicados y mostrados detalladamente en el estudio de Ali y Zafar (2018); los cuales son conjuntos de datos compuestos por imágenes de calles o ambientes (resultados reportados en Zhou et al. (2017, 2016c)).

Así mismo, realizamos un pequeño cambio en nuestro modelo *VGGNet baseline* en el removeremos las últimas dos capas densas y añadiremos un GAP, sustentando en que la utilización de esta técnica, respecto a un *max pool* tiene mejor rendimiento al extraer características (Lin et al., 2013). En la Figura 4.1 se puede observar que a partir del último bloque de convolución, se realiza el cálculo del GAP, los cuales utilizamos como características; a este modelo le denominaremos "VGG_GAP" para diferenciarlo

del original "VGG". Para tener una facilidad de reconocer nuestros modelos del grupo *baseline* los denominaremos: "TL_VGG16", "TL_VGG16_GAP", "TL_VGG16_Places" y "TL_VGG16_GAP_Places"; donde aquellos que tienen "_Places" son aquellos que usan los pesos entrenados previamente en *Places365*. Por lo que las arquitecturas serian las siguientes: (i) "TL_VGG16" y "TL_VGG16_Places" son la arquitectura original de VGG16, por lo cual las características extraídas seria un vector de tamaño 4096 de la última capa densa; (ii) "TL_VGG16_GAP" y "TL_VGG16_GAP_Places" al tener el método GAP se extraerá un vector de tamaño 512, el cual es un ponderado general de cada *feature map* obtenidos en la última convolución. Finalmente, una vez extraídas las características con estos 4 modelos, procederemos a realizar nuestro entrenamiento de clasificación de la percepción de seguridad utilizando los modelos lineales y no-lineales: (i) *Logistic Regression*: $L(y, f(x)) = \sum_{i=1}^{n} log(e^{(-y_i f(x_i))} + 1)$; (ii) *Ridge Classifier*: $L(y, f(x)) = sgn(||y - f(x)||_2^2 + ||w||_2^2)$; (iii) *Linear SVC*: $L(y, f(x)) = \sum_{i=1}^{n} max(0, 1 - y_i f(x_i))$ y (iv) *RBF SVC*.

4.2. Modelos del Grupo *Fine-Tuning*

Tal como describimos en la sub-Sección 4.1, empleamos la red *VGG16* debido al buen desempeño reportado en el conjunto de datos *Places365*. Utilizaremos las mismas arquitecturas descritas en la sección anterior *baseline* (ver Figura 4.1), con la diferencia que a este grupo de modelos a los cuales denominaremos como "FT_VGG", "FT_VGG_GAP", "FT_VGG-Places" y "FT_VGG_GAP-Places" serán entrenados congelando algunas capas de convolución (limitado por nuestra memoria y poder computacional). De manera similar, asignaremos prefijo "Places" o nada respectivamente a los pesos previamente entrenados de *Places365* e *ImageNet*. Para los experimentos, congelamos todas las capas de cada arquitectura hasta el bloque de convolución 4, por lo que solamente se entrenará el último bloque y las densas (en el caso de "FT_VGG16" y "FT_VGG16_Places"). Finalmente, en todos los casos añadimos una última capa densa con solos dos salidas (correspondientes a las clases segura y no segura) con función de activación *Softmax*: $f(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$ y función de pérdida *Categorical Cross-Entropy*.

4.3. Modelo GAN Semi-Supervisada

Como mencionamos anteriormente en el Capítulo 3 sobre las posibles limitaciones de *Place Pulse 2.0*, propusimos un método basado en aprendizaje semi-supervisado; el cual podría mitigar y tener un buen desempeño frente a las características de *Place Pulse 2.0* tales como: pocas imágenes, datos con disparidad y poca generalización de datos. Sin embargo, ¿por qué usar un modelo Semi Supervisado? Este conjunto de técnicas que utilizan datos etiquetados y no etiquetados, presentan un mejor desempeño y una mejora considerable durante el aprendizaje en casos donde se tiene disparidad de datos. Debido a que tenemos pocos datos, se propone el uso de una GAN Semi-Supervisada (Salimans et al., 2016).

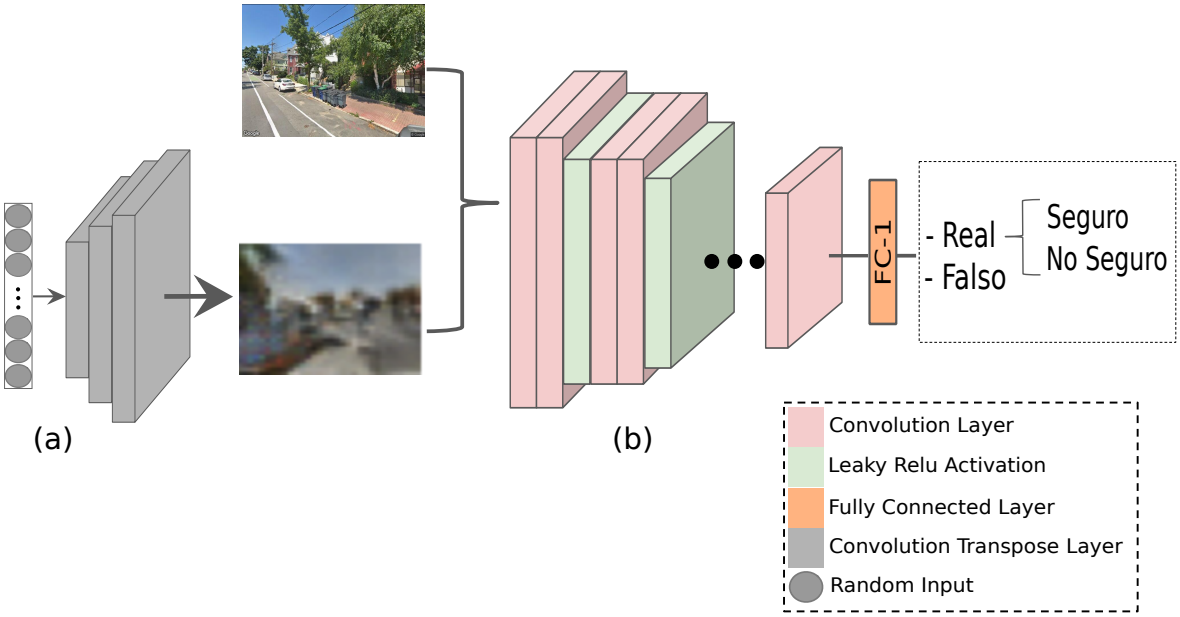


Figura 4.2: Se presenta a manera general el modelo "SSL-GAN" implementado compuesto de dos componentes principales: (a) modelo Generador, este modelo se encarga de generar imágenes basado en las características aprendidas; (b) modelo Discriminador, que se encarga de dos sub-tareas, la primera es la clasificación de una imagen entre segura o no segura (aprendizaje supervisado) y la segunda es la clasificación de una imagen entre real o falsa (aprendizaje no supervisada). Fuente: El autor.

Escogimos utilizar este método debido a las limitaciones mencionadas (pocos datos y datos no balanceados) las cuales una GAN puede fácilmente tratar. La primera limitación: **conjunto de datos no balanceados**, se han visto resultados de la aplicación de GANs en datos no balanceados (Sampath et al., 2021; Zhou et al., 2018), demostrando que este tipo de arquitecturas permiten aprender características de las imágenes y clasificar a qué clase pertenece una determinada imagen. La segunda limitación: **conjunto de datos con pocas muestras**, la utilización de una GAN con datos limitados como *Place Pulse 2.0*, teniendo solo alrededor de 110 mil imágenes sin ningún tipo de *Data Augmentation* es ideal (Karras et al., 2020; Cenggoro et al., 2018). Así mismo, una GAN semi-supervisada nos permite no solamente generar datos, sino que también nos permite clasificar los datos. A diferencia de una GAN *vanilla* la cual está enfocada en generar y diferenciar entre una distribución de datos generada a la distribución de datos de entrada. Una GAN semi-supervisada (a partir de ahora la llamaremos "SSL_GAN") además de realizar dicha tarea de generación y discriminación, también realiza la tarea de clasificación de los datos.

Discriminador					
Capa	Entrada	Canales	*Kernel size*	*Stride*	Activación
Conv	$32 \times 32 \times 3$	32	3×3	1	LeakyReLU
Conv	$32 \times 32 \times 32$	32	3×3	2	LeakyReLU
DropOut (0.2)	$16 \times 16 \times 32$	-	-	-	-
Conv	$16 \times 16 \times 32$	64	3×3	1	LeakyReLU
Conv	$16 \times 16 \times 64$	64	3×3	2	LeakyReLU
DropOut (0.2)	$8 \times 8 \times 64$	-	-	-	-
Conv	$8 \times 8 \times 64$	128	3×3	1	LeakyReLU
Conv	$8 \times 8 \times 128$	128	3×3	2	LeakyReLU
DropOut (0.2)	$4 \times 4 \times 128$	-	-	-	-
Conv	$4 \times 4 \times 128$	256	3×3	1	LeakyReLU
Flatten	$4 \times 4 \times 256$	-	-	-	-
Dense	128	-	-	-	-
DropOut (0.4)	128	-	-	-	-
Dense	3	-	-	-	Softmax
Total Parámetros	1 107 882				

Generador					
Capa	Entrada	Canales	*Kernel size*	*Stride*	Activación
Espacio Latente	100	-	-	-	-
Dense	4096	-	-	-	LeakyReLU
Re-dimensionar	$4 \times 4 \times 256$	-	-	-	-
Deconv	$4 \times 4 \times 256$	256	4×4	2	LeakyReLU
Deconv	$8 \times 8 \times 256$	128	4×4	2	LeakyReLU
Deconv	$16 \times 16 \times 128$	64	4×4	2	LeakyReLU
Conv	$32 \times 32 \times 64$	3	3×3	1	Tanh
Total Parámetros	2 119 811				

Cuadro 4.1: Configuración de los modelos discriminador y generador de nuestra GAN semi-supervisada denominada"SSL_GAN" que serán utilizados para entrenar el conjunto de datos *Place Pulse 2.0*. También cabe destacar la cantidad de parámetros a entrenar por cada modelo, dando el detalle de cada capa utilizada en la construcción de cada modelo. Así mismo, mencionamos que el valor del parámetro α de la función $LeakyReLU$ es 0,2.

En resumen, una GAN semi-supervisada combina un modelo No Supervisado (clasificación entre datos generados con etiqueta real o falso) y otro Supervisado (clasificación de los datos discriminados como reales entre las clases propuestas). En la Figura 4.2 se observa la estructura de nuestra GAN; en la que a partir de un vector de valores aleatorios entre 0 y 1 (también llamado como "ruído") el modelo puede aprender a generar imágenes artificiales casi tan buenas como las del conjunto de datos original. La configuración de nuestra GAN se muestra en el Cuadro 4.1, en el cual podemos observar las arquitecturas para el discriminador y generador en detalle y las operaciones asociadas a cada uno.

4.4. Consideraciones Finales

En este Capítulo se ha presentado los enfoques, métodos y modelos a utilizar en nuestros experimentos de entrenamiento sobre el conjunto de datos *Place Pulse 2.0* previamente descrito, en este estudio abarcamos 3 tipos de técnicas: (i) *transfer-learning*; (ii) *fine-tuning*; y (iii) una GAN semi-supervisada. Así mismo, hemos descrito las métricas que utilizaremos para evaluar el desempeño de los modelos. Estas evaluaciones tienen como finalidad verificar qué tipo de técnica podría ser la más adecuada frente a las limitaciones de *Place Pulse 2.0* presentadas en el Capítulo 3. En el siguiente capítulo veremos en detalle los métodos, hiper-parámetros y experimentos a realizados, los cuales hacen parte de nuestra metodología propuesta. Se verá en detalle el comportamiento de cada un de los modelos de los grupos *transfer-learning*, *fine-tuning* con las respectivas modificaciones usando el GAP y la GAN semi-supervisada.

Capítulo 5

Resultados

En este Capítulo se presentan las evaluaciones y reportes de resultados correspondientes a los modelos presentados en el Capítulo 4. Para realizar los experimentos se utilizó un entorno compuesto de un GPU NVIDIA GeForce GTX 1070, driver 460.91.03, versión de CUDA 11.2 y 8.11 Gb de VRAM; CPU de 12 núcleos Intel(R) Core(TM) i7-8700 CPU @ 3.20GHz cada uno y un total de 31.1 Gb de RAM. En todos los experimentos, salvo en la "SSL_GAN" realizamos un entrenamiento por ciudad y un entrenamiento a nivel global (utilizando todas las ciudades).

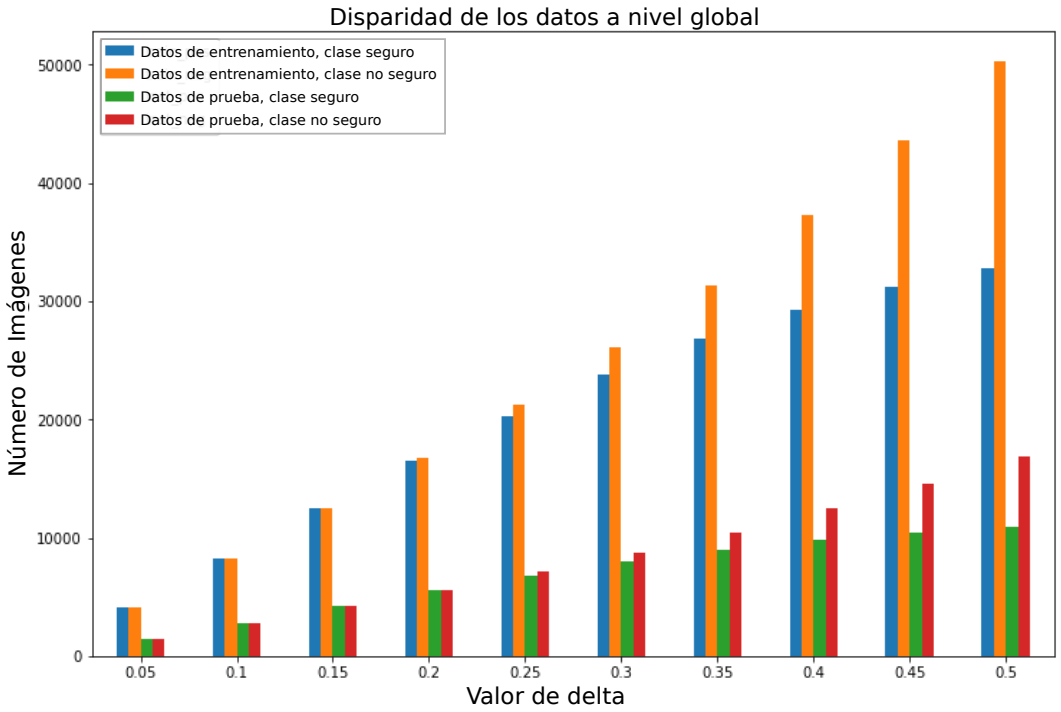


Figura 5.1: Distribución de la disparidad en la cantidad de imágenes a nivel global (juntando todas las ciudades) correspondientes a cada clase. Se evidencia que a mayor valor del parámetro δ, incrementa la disparidad de los datos. Fuente: El autor.

Tal como se estudiaron en trabajos anteriores previamente mencionados, se definió

Resumen de los parámetros de cada modelo							
Nombre	**Hiper-parámetros de los modelos**						**Datos usados**
Método	Entrada	*Batch*	*Opt*	*LR*	*Ep/It*	*CV*	Región
TL_VGG	4096	-	lbfgs	-	1000	5	Global/Ciudad
TL_VGG_GAP	512	-	lbfgs	-	1000	5	Global/Ciudad
FT_VGG	$224 \times 224 \times 3$	128	Adam	$1e^{-3}$	100	5	Global/Ciudad
FT_VGG_GAP	$224 \times 224 \times 3$	128	Adam	$1e^{-3}$	100	5	Global/Ciudad
SSL_GAN_Dis	$32 \times 32 \times 3$	128	Adam	$1e^{-3}$	100	5	Global
SSL_GAN_Gen	100	128	Adam	$1e^{-3}$	100	5	Global

Cuadro 5.1: Lista de hiper-parámetros utilizados en cada modelos durante el entrenamiento: (i) *Batch*: tamaño de datos a entrenar; (ii) LR: tasa de aprendizaje; (iii) *Opt*: Optimizador; (iv) *Ep/It* significa *epochs* o *iteration* (solo en los modelos TL se uiliza iteraciones); y (v) *CV* validación cruzada. Las librerías utilizadas para la implementación de los experimentos fueron *Sklearn* (Pedregosa et al., 2011) y *TensorFlow-Keras* v2.3 (Abadi et al., 2015) para los grupos de modelos "TL" y "FT"/"SSLGAN" respectivamente.

un parámetro adicional denominado δ (delta). Este parámetro δ nos permite escoger un subconjunto del conjunto total de imágenes a partir de las puntuaciones de percepción de la siguiente forma: El rango de valores de δ varía desde 0.05 hasta 0.45 los cuales significan el porcentaje de datos que escogeremos de cada clase, cuando $\delta = 0{,}5$ se entiende que es todos los datos. Por ejemplo, para un valor de $\delta = 0{,}45$ escogeremos las imágenes asociadas al 45 % de las puntuaciones de percepción mayores y las imágenes asociadas al 45 % de puntuaciones de percepción menores.

En la Figura 5.1 se muestra la variación del valor δ y su impacto sobre nuestro conjunto de datos (entrenamiento y prueba). La idea principal de este valor δ es observar el comportamiento del modelo al seleccionar un conjunto de datos con cantidad similar de ambas clases. Así mismo observamos que conforme va incrementando el valor de δ con paso de 0.05, también incrementa la disparidad de clases a nivel de ciudad y a nivel global. Esto nos demuestra que para valores bajos de δ es posible tener una paridad de datos, se observa que a partir de δ=0.2 ya existe una pequeña disparidad en el conjunto de entrenamiento. Para δ =0.5, tenemos que la disparidad es alrededor de 11 mil imágenes en favor de la categoría no segura.

5.1. Experimentos realizados

En primer lugar, tenemos que describir de manera general cómo realizamos nuestros experimentos. Para cada tipo de modelo utilizamos técnicas diferentes para poder encontrar los mejores parámetros que presentan mejores resultados. Para realizar los experimentos utilizamos las funciones *GridSearchCV* y *KerasClassifier* para crear nuestra malla de búsqueda de los mejores parámetros e hiper-parámetros para nuestros

modelos, así mismo, también utilizamos una **validación cruzada de 5 pasos** dividiendo al conjunto de datos en **80 % para entrenamiento y 20 % de prueba**. Para los modelos del grupo *Transfer-Learning*, utilizamos los métodos *LinearSVC*, *RBF SVC*, *Logistic Regression* y *Ridge Classifier*, en todos modificamos la regularización l_2 usando valores de α desde 10^{-4} hasta 10^{2} y el parámetro "solver" variando desde *liblinear* y *lbfgs*.

Para los modelos del grupo *Fine-Tuning* añadimos otros parámetros, estos son el *batch size* variando desde tamaño 8 hasta 128 (en potencias de 2), también variamos el optimizador del modelo entre el conjunto *SGD*, *RMSprop*, *Adagrad*, *Adadelta*, *Adam* y *Adamax* con *learning rate* variando entre 10^{-6} hasta 10^{-1}. El número de *epochs* se mantuvo como constante 100, debido a que utilizábamos el método *earlyStopping*, el cual controlaba el comportamiento del modelo. Así mismo, en el Cuadro 5.1 resumimos todos hiper-parámetros que nos generan la mejor configuración y mejores resultados para cada caso. Finalmente, es de importancia mencionar que para cada validación cruzada se utilizó el método *stratiffied Kfold* para garantizar que en cada validación se tenga una cantidad de muestras con una proporción similar de cada clase segura y no segura.

5.2. Modelos del Grupo *Transfer-Learning* (*Baseline*)

En la Figura 5.2 mostramos los resultados del modelo *baseline* "TL_VGG_GAP" en el cual observamos que que conforme el valor de δ va incrementando desde 0.05 hasta 0.5, el *accuracy* reportado va decreciendo. Sin embargo, en la Figura 5.2 columna (b), el *accuracy* reportado para la ciudad de Río de Janeiro incrementa conforme δ aumenta; vemos este comportamiento similar en otras ciudades con gran cantidad de datos disparejos (ver Figura 3.6). Pero no podemos decir que esos resultados sean correctos, pues al analizar la métrica *AUC*, observamos que existe una caída considerable. Esto nos indica que, partiendo de la idea de la disparidad de datos, el modelo puede clasificar correctamente a las imágenes correspondientes a la clase con mayor cantidad de datos. Por el otro lado, al tener valores pequeños de δ y valores altos en *accuracy* y *AUC* nos indica que el modelo está aprendiendo a clasificar correctamente; esto puede ser debido a que a menor valor de δ, hay mayor proporcionalidad en la cantidad de imágenes de cada clase (ver Figura 5.1).

Además, debemos mencionar que a diferencia de otras ciudades, Rio de Janeiro presenta mucha más variedad en sus calles, desde áreas verdes (zona sur) hasta callejones (zona norte); siendo el caso contrario en la mayoría de ciudades en donde se tiene solamente presencia de áreas verdes (p.ej. Atlanta, Berlin, Amsterdam, entre otros) o rascacielos (p.ej. Boston, New York, Vancouver, entre otros). Como mencionamos, ciudades monótonas a nivel de imágenes como Atlanta se observa un comportamiento similar al que se tiene a nivel global, a mayor valor de δ hay una tendencia a disminuir el valor de las métricas AUC y *accuracy* (ver Figura 5.2 (c)).

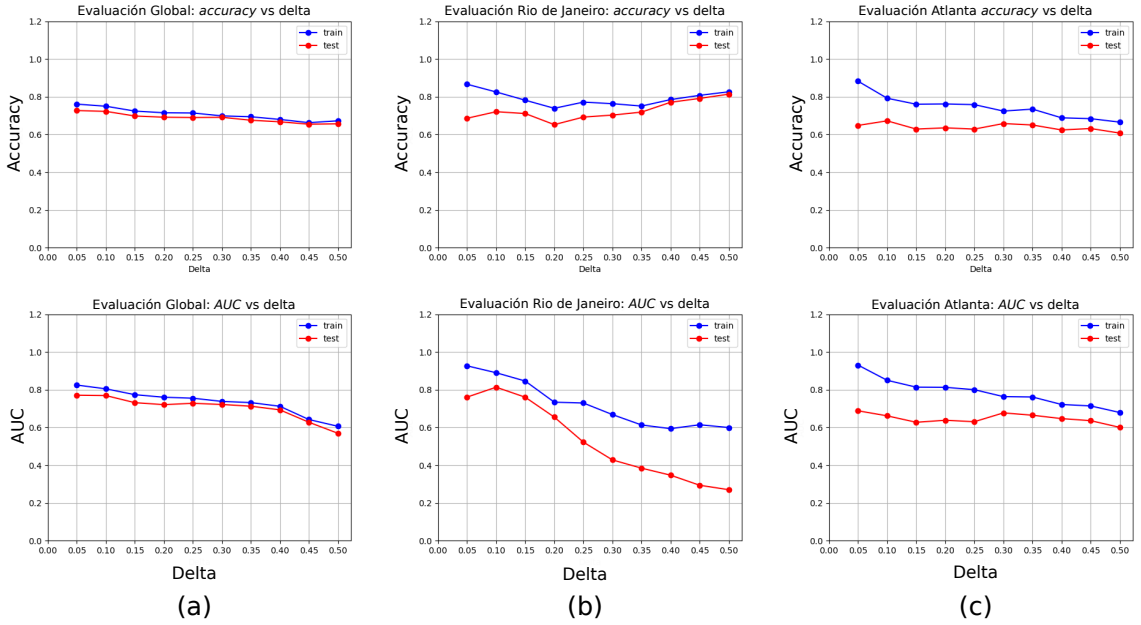


Figura 5.2: Resultados de "TL_VGG_GAP" (mejor modelo). Observamos 3 casos puntuales, en la columna (a) se muestra resultados a nivel global donde se presenta el decrecimiento de la precisión conforme δ incrementa; columna (b) correspondiente a Río de Janeiro, se presenta una mayor precisión en mayores valores de δ; y la columna (c) Atlanta, se presenta un comportamiento similar a la mayoría y al global.

En el Cuadro 5.2 mostramos los respectivos promedios de las 5 validaciones cruzadas reportando las métricas obtenidas luego de evaluarlos con los datos de entrenamiento y prueba en cada modelo a nivel global. Debemos mencionar que incluimos modelos como *ResNet50* y *Xception* con pesos previamente entrenados en *ImageNet* para hacer el análisis comparativo respecto a las arquitecturas que propusimos. Sin embargo, no mostraron ningún resultado destacable por lo que fueron descartados para los modelos *fine-tuning*. Se observa que los modelos basados en *VGGNet* (ya sea entrenado previamente en *ImageNet* o *Places*) tienen mejor desempeño que *Xception* y ligeramente mejor que *ResNet*. Además, el modelo "TL_VGG16_GAP" entrenado usando un *Linear SVC* obtuvo los mejores resultados a nivel de *accuracy* y AUC (a pesar de un valor relativamente bajo). Así mismo, se observa que en todos los casos el modelo *rbf SVC* tuvo un pésimo desempeño, no consiguió aprender ni a diferenciar las imágenes entre ambas clases.

		auc		*accuracy*		*f1 score*	
Modelos	Método	entrena	prueba	entrena	prueba	entrena	prueba
VGG	*LinearSVC*	63.62	56.50	68.85	65.22	54.78	**49.41**
	Logistic	60.63	**57.52**	67.25	**65.72**	51.42	49.07
	Ridge Classifier	64.72	54.75	69.44	64.38	56.50	49.34
	RBF SVC	45.14	42.42	52.13	52.37	46.93	46.59
VGG_GAP	*LinearSVC*	59.01	**57.93**	66.51	**66.09**	49.52	49.06
	Logistic	58.07	57.57	65.95	65.59	46.06	45.61
	Ridge Classifier	59.20	57.93	66.59	65.89	50.27	**49.76**
	RBF SVC	42.93	41.70	50.25	50.35	47.16	46.75
VGG Places	*LinearSVC*	64.44	57.14	69.48	65.79	56.39	51.20
	Logistic	61.74	**58.35**	68.16	**66.44**	53.77	**51.28**
	Ridge Classifier	65.20	55.76	69.84	64.86	57.56	50.67
	RBF SVC	47.32	45.25	56.56	55.69	44.78	44.21
VGG_GAP Places	*LinearSVC*	60.26	**59.76**	67.38	**66.96**	51.65	51.04
	Logistic	59.40	58.97	66.81	66.62	49.16	48.90
	Ridge Classifier	60.45	59.15	67.45	66.94	52.23	**51.53**
	RBF SVC	44.40	42.47	52.59	52.54	43.39	45.05
ResNet50	*Linear SVC*	61.62	59.10	68.10	**66.42**	53.63	50.80
	Logistic	60.04	**59.15**	67.25	66.37	51.47	49.70
	Ridge Classifier	62.11	58.38	68.36	66.08	54.59	**51.00**
	RBF SVC	45.36	44.07	53.46	53.57	44.99	44.98
Xception	*LinearSVC*	55.29	**53.25**	64.43	**63.33**	41.66	39.69
	Logistic	53.48	52.75	63.56	63.14	36.72	35.87
	Ridge Classifier	57.23	52.22	65.22	63.04	45.63	42.11
	RBF SVC	45.57	44.99	49.12	49.12	55.01	**55.05**

Cuadro 5.2: Cada columna entrena y prueba reporta el valor promedio de evaluar en cada conjunto de datos los modelos obtenidos de las 5 validaciones cruzadas. Los modelos ResNet y Xception fueron pre-entrenados sobre *ImageNet*.

5.3. Modelos del Grupo *Fine-Tuning*

Tal como describimos en el Capítulo 3 y describimos en el Cuadro 5.1 entrenamos los modelos *fine-tuning* en todas las ciudades y a nivel global. Por lo cual en las Figuras 5.4 y 5.3 se observa un mapa de color sobre el *accuracy* obtenido por cada modelo entrenado en cada ciudad y a nivel global, evaluado en la misma ciudad, las demás ciudades y en todas (nivel global). Vemos la necesidad de descartar *Xception* debido a su pésimo rendimiento y métricas reportadas. También se observa que existen ciudades que mantienen un *accuracy* elevado (comparado al promedio p.ej. Rio de Janeiro y Belo Horizonte); así mismo, podemos observar que entre ambos tipos de modelos "FT_VGG" y "FT_VGG_GAP" las ciudades Taipei, Singapore, Philadelphia, Londres, Kiev, Dublin y Cape Town mantienen un *accuracy* baja en todos las evaluaciones.

En el Cuadro 5.3 se reportan los valores promedio de las 5 validaciones cruzadas realizadas en todos los modelos, de los cuales excluimos *Xception* debido a los bajos valores reportados en el Cuadro 5.2, solo *ResNet50* mostró una similitud con los demás modelos descritos. Sin embargo, el mejor modelo obtenido hasta este punto es "FT_VGG_GAP_Places" el cual superó por poco a "FT_VGG_Places". Observamos también que a pesar de mantener un *accuracy* similar a los modelos "TL", el *auc* y *F1 score* incrementaron sus valores, demostrando así que los modelos *fine-tuning* mostraron un mejor desempeño frente a los modelos *transfer-Learning*, los cuales se observan en las métricas reportadas, dicho de otra forma, tienen un valor de *accuracy* cercano pero estos modelos son más prometedores, puesto que consiguen predecir la disparidad correctamente.

	auc		*accuracy*		*f1 score*	
Modelos "FT"	entrena	prueba	entrena	prueba	entrena	prueba
VGG	77.83	77.42	74.01	64.71	74.01	64.69
VGG_GAP	76.14	75.59	69.40	66.88	69.41	66.87
VGG_Places	74.95	74.75	68.71	67.26	68.71	67.27
VGG_GAP_Places	77.98	**77.5**	70.52	**67.28**	70.52	**67.28**
ResNet50	76.36	72.71	70.36	65.64	67.35	64.98

Cuadro 5.3: Valores promedio de cada métrica obtenidos luego de evaluar cada modelo en los datos de entrenamiento y de prueba. A pesar de los resultados tan cercanos entre todos los modelos, observamos un drástico aumento en las métricas, así como también entendemos que el modelo consigue distinguir con mayor eficacia ambas clases.

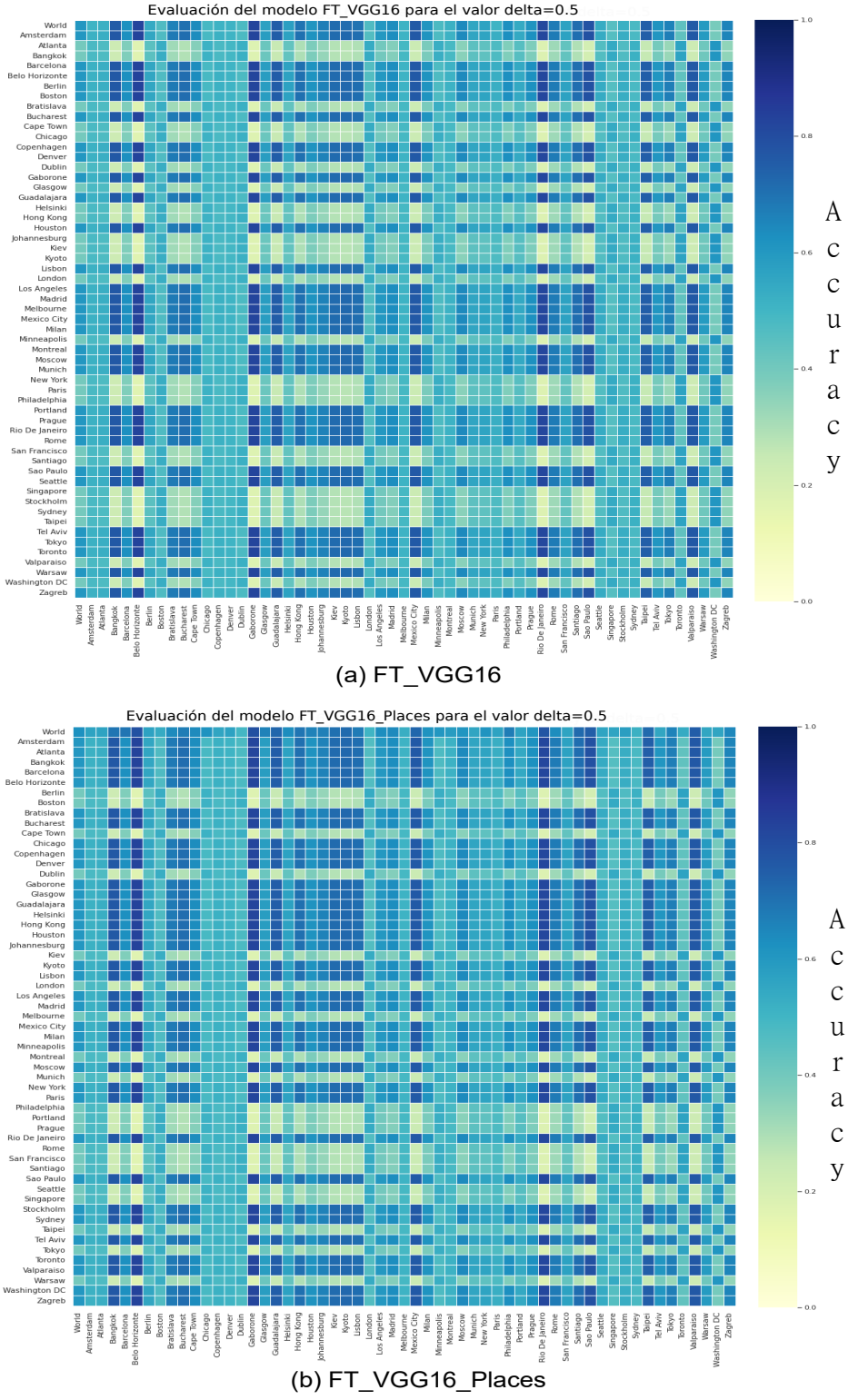
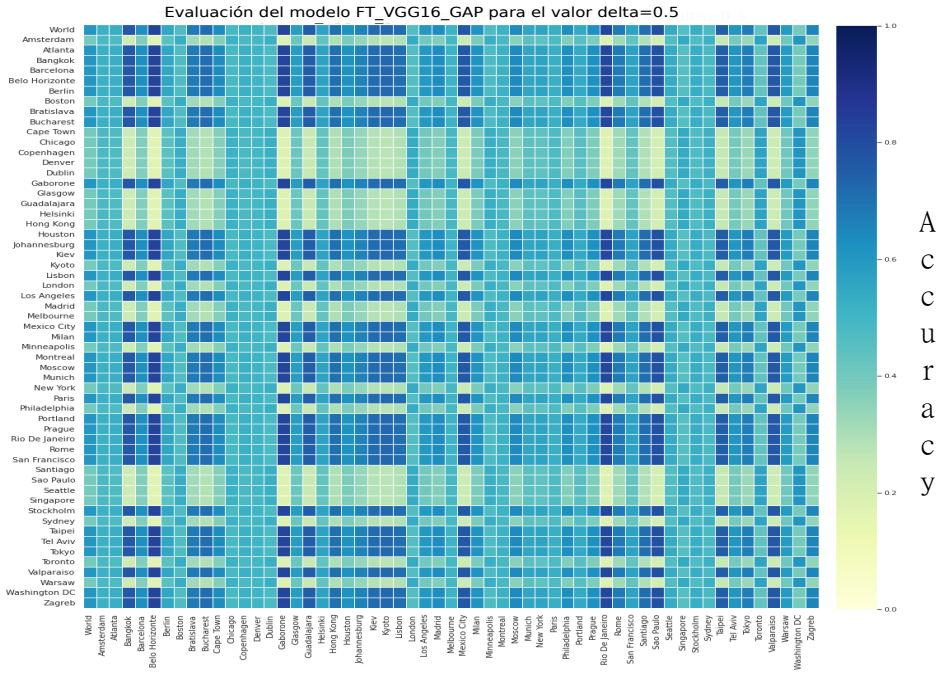
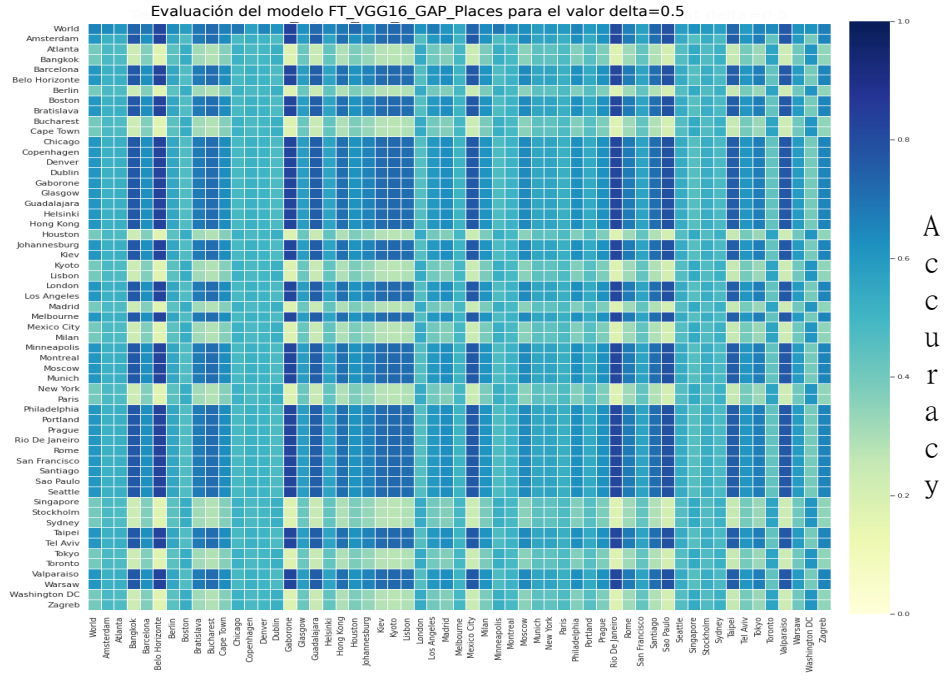


Figura 5.3: Observamos en el mapa de color compuesto de los valores del *accuracy* de cada modelo entrenado (fila) evaluado en cada ciudad (columnas) y también a nivel global "*World*". (a) Resultados de las evaluaciones del modelo "FT_VGG" en cada ciudad. (b) Resultados de las evaluaciones del modelo "FT_VGG_Places" en cada ciudad. Fuente: El autor.



(a) FT_VGG16_GAP



(b) FT_VGG16_GAP_Places

Figura 5.4: Observamos en el mapa de color compuesto de los valores del *accuracy* de cada modelo entrenado (fila) evaluado en cada ciudad (columnas) y también a nivel global "*World*". (a) Resultados de las evaluaciones del modelo "FT_VGG_GAP" en cada ciudad. (b) Resultados de las evaluaciones del modelo "FT_VGG_GAP_Places" en cada ciudad. Fuente: El autor.

5.4. Modelo GAN Semi-Supervisada

Debido al tiempo que conlleva entrenar este modelo (aproximadamente entre 2∼4 días por cada validación cruzada) se decidió entrenar utilizando todos los datos, a fin de comparar los resultados generales con los otros modelos ya mostrados. Tal como mostramos anteriormente en los Cuadros 4.1 y 5.1, describimos la configuración de nuestros modelos discriminador y generador, así como también los hiper-parámetros utilizados. Una vez entrenada la GAN, las métricas reportadas por el modelo discriminador final están presentes en el Cuadro 5.4, donde se observa que en la última época el modelo realiza un sobre-entrenamiento (*overfitting*) de los datos, por lo cual nuestros resultados son bajos comparados a los modelos anteriores, sin embargo, la métrica AUC es mayor que los resultados anteriormente reportados.

		auc		*accuracy*		*f1 score*	
Modelo	CV	entrena	prueba	entrena	prueba	entrena	prueba
SSL_GAN 32x32x3	0	80.95	80.97	90.26	59.06	90.26	59.04
	1	81.43	81.45	89.42	61.50	89.42	61.48
	2	81.43	**81.45**	89.56	**62.58**	89.56	**62.57**
	3	80.59	80.66	90.01	61.52	90.01	61.54
	4	80.61	80.63	89.38	61.14	89.38	61.13

Cuadro 5.4: Métricas obtenidas de las 5 validaciones cruzadas evaluadas en el conjunto de datos de entrenamiento y de prueba, se observa que el AUC reportado es mucho mayor que los anteriores con un valor por encima de 80 %.

Sabiendo que los resultados del Cuadro 5.4 son las evaluaciones en la última época, procedimos a buscar el momento en que comenzó el *overfitting* notando que fue alcanzado en diversas iteraciones en cada validación. En la Figura 5.5 resaltamos las iteraciones en donde fue alcanzado el mejor resultado para los históricos de *accuracy* y *loss*. En promedio, la iteración con las mejores métricas reportadas está alrededor de la iteración 25 mil. Dichas iteraciones con las mejores métricas se muestran en el Cuadro 5.5.

			auc		*accuracy*		*f1 score*	
Modelo	CV	iteración	entrena	prueba	entrena	prueba	entrena	prueba
SSL_GAN 32x32x3	0	23 788	73.89	73.89	78.90	78.12	78.90	78.12
	1	58 550	80.21	**80.22**	92.18	**81.25**	92.18	**81.25**
	2	21 951	73.60	73.60	81.25	79.68	81.25	79.68
	3	23 180	73.53	73.53	76.56	78.90	76.56	78.90
	4	8602	69.84	69.84	74.21	78.90	74.21	78.90

Cuadro 5.5: Métricas reportadas luego de evaluar cada modelo con mayor *accuracy* reportado durante el entrenamiento. Se observa cierta estabilidad en los valores reportados.

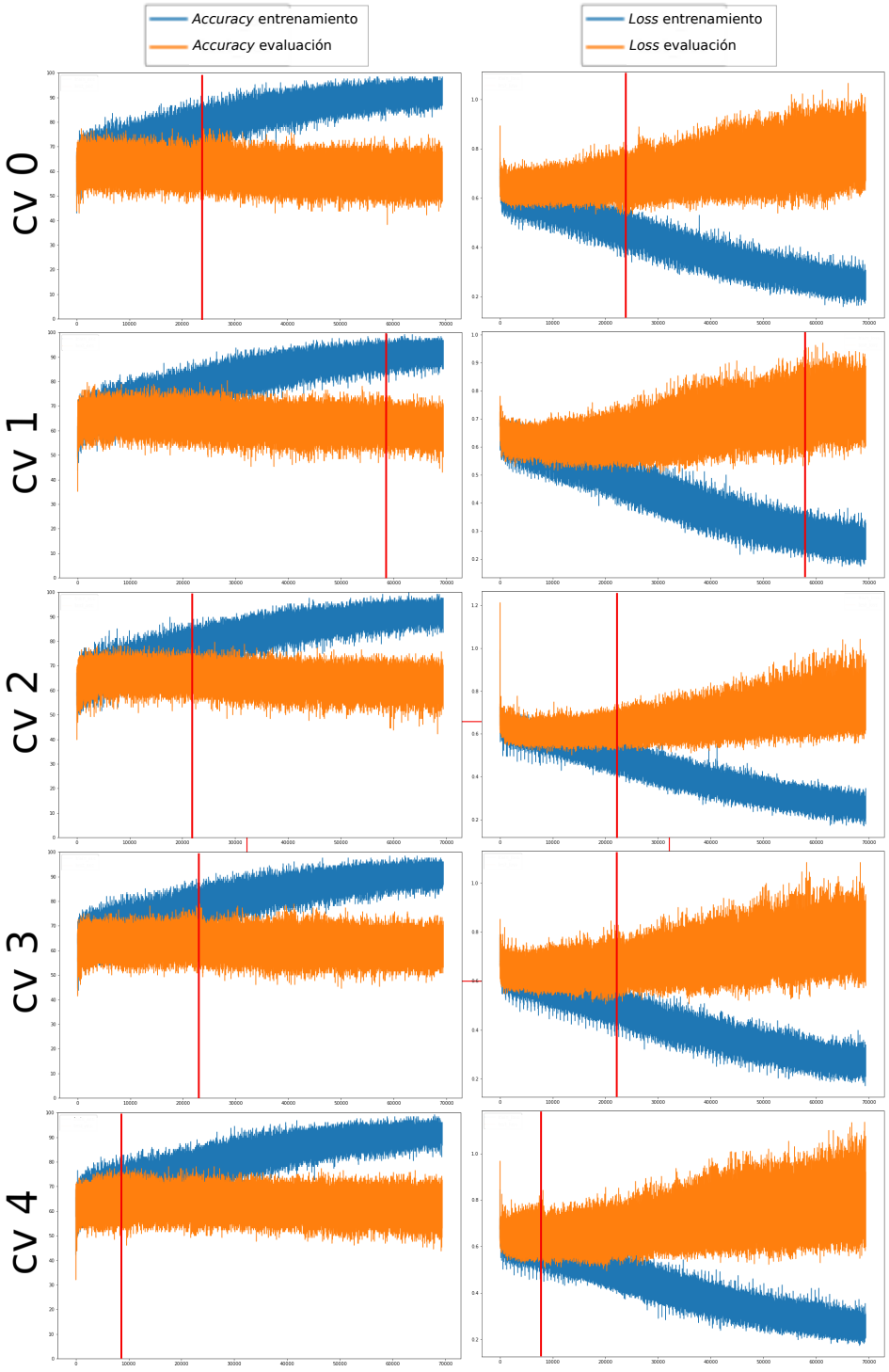


Figura 5.5: Histórico de *accuracy* y *loss* en cada validación cruzada (CV), resaltamos la iteración o paso de entrenamiento con mayor valor (columna izquierda) a través de una línea roja. En la columna derecha, resaltamos la misma iteración o paso para el *loss*. Conforme lo reportado en el Cuadro 5.5. Fuente: El autor.

Figura 5.6: (a) Imágenes reales del conjunto de datos; (b) Imágenes generadas en el último paso de entrenamiento. Se observa la alta calidad de las imágenes generadas de tamaño $32 \times 32 \times 3$, haciendo complicado la posibilidad de distinguirlas visualmente. Fuente: El autor.

Finalmente, en la Figura 5.5 se observa a el histórico del *accuracy* y *loss* del entrenamiento, enmarcamos en cada figura una línea roja correspondiente a la iteración donde se alcanzó el mayor *accuracy*. Una observación importante es el hecho de que *auc* es mayor que *F1 score*. Esto es debido a la disparidad de datos previamente analizada, siendo el conjunto de datos con mayor cantidad de ejemplos la categoría no segura; entonces podemos afirmar que nuestro modelo es más robusto identificando ejemplos seguros y no seguros. También observamos que las métricas *auc* y *F1 score* presentan un valor cercano entre sí, dándonos a entender que estos modelo alcanzan una estabilidad al momento de evaluar los datos. Esta estabilidad esta fuertemente relacionada a: (i) qué tan bien identifica las clases, (ii) cuántas muestras consigue clasificar correctamente, y (iii) la relación entre aciertos y errores en la predicción. De manera adicional, en la Figura 5.6 mostramos un conjunto de imágenes generadas por nuestro modelo generador 5.1 (b). Se observa la buena la calidad de las imágenes generadas de tamaño $32 \times 32 \times 3$, las cuales pueden compararse y confundir por imágenes reales.

5.5. Sistema Web

En el presente trabajo, se vio la necesidad de tener un sistema web con el propósito de conciliar una interacción y visualización de los resultados de los entrenamientos de los datos de manera rápida y simple. Así mismo, es posible observar los resultados y las métricas reportadas para cada valor de la variable δ. También es posible visualizar los resultados de cada validación cruzada reportando las métricas definidas previamente, las gráficas de AUC calculadas a partir del *Precision-Recall* y el promedio

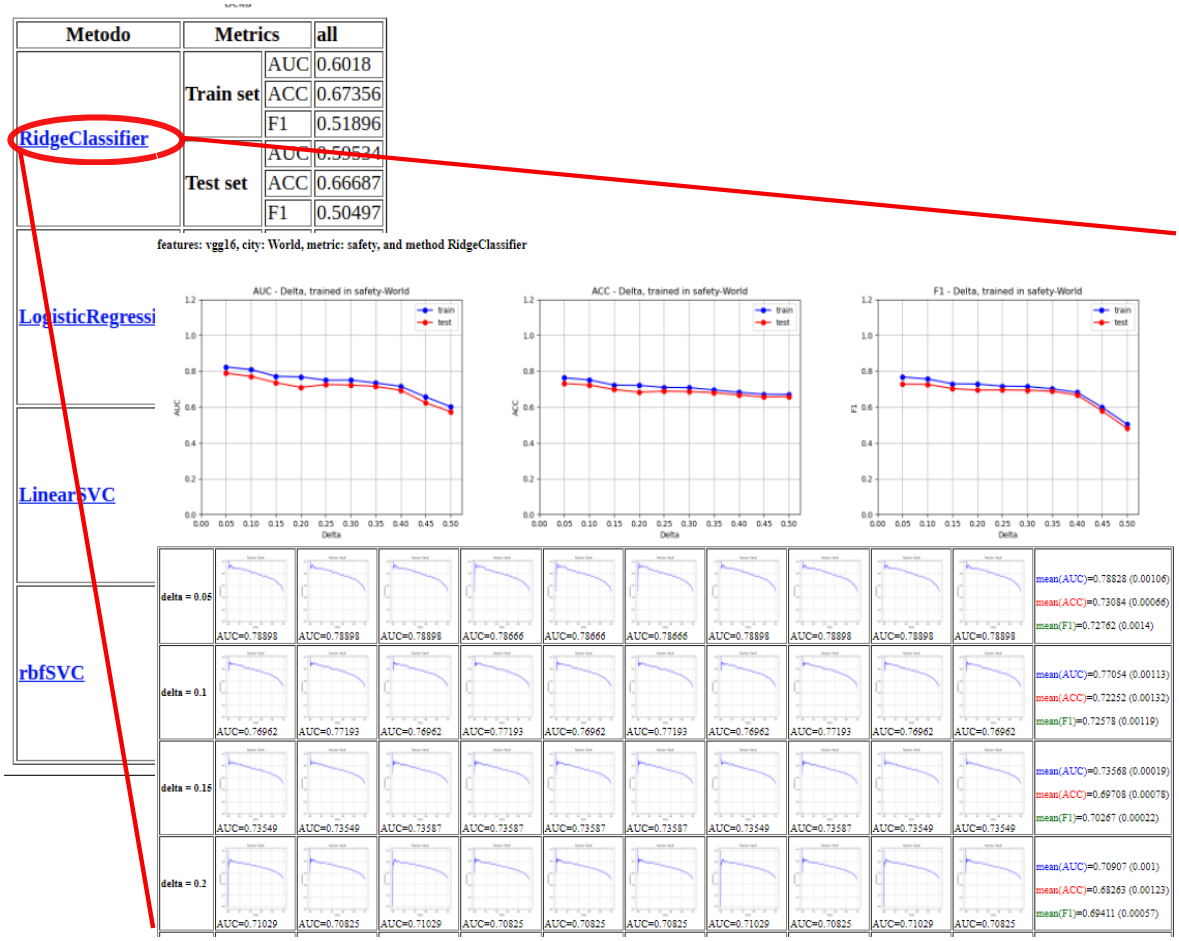


Figura 5.7: Esta pestaña corresponde a los resultados de las validaciones cruzadas y un resumen de las métricas reportadas, así como también las gráficas asociadas a el valor de cada métrica para cada valor de δ. Nota: esta imagen corresponde a un modelo del grupo *transfer-learning*. Fuente: El autor.

de dichas métricas. Además, un resumen de los resultados de cada método utilizado. El sistema web tiene un diseño simple, enfocado principalmente en presentar resultados de los entrenamientos, así como una comparación directa entre cada método utilizado. El sistema web está compuesto de 3 principales paneles: (i) Resultados de *Baseline*; (ii) Resultados de los modelos *Fine-Tuning*; y (iii) Resultados de la "SSL_GAN". En la Figura 5.7 mostramos la apariencia del sitio web. Comenzando con una tabla donde se reportan las métricas obtenidas de cada método en los conjuntos de datos de entrenamiento y de prueba. Se aprecia que para ver en detalle los resultados del entrenamiento, basta apretar sobre el nombre del método y será redirigido a una pestaña con detalle tal como las validaciones cruzadas, las gráficas de cada *Precision-Recall* y la gráfica general del *AUC*, *accuracy* y *F1 score* reportado para cada valor de δ.

5.6. Consideraciones Finales

Se han presentado los resultados de las evaluaciones de los modelos previamente descritos en el Capítulo 4, así como también, las métricas reportadas en cada modelo utilizando una validación cruzada de 5 conjuntos. Para los experimentos se dividió los datos en 80 % para entrenamiento y 20 % de prueba. A partir de los resultados obtenidos, observamos que el modelo semi-supervisado presenta un comportamiento más estable con respecto a los demás (al observar los valores de las métricas reportadas). Observamos que no solamente el *AUC*, sino también el *accuracy* y *F1 score* resultaron con un valor alto y muy cercano, lo cual era el esperado. A continuación presentamos las discusiones y limitaciones.

Capítulo 6

Discusiones y Limitaciones

6.1. Discusiones

En el presente trabajo se ha descrito una metodología que permite estudiar y analizar el conjunto de datos *Place Pulse 2.0* con el objetivo de encontrar y resaltar las posibles limitaciones que pueda presentar; esta motivación es debido a que en la gran mayoría de trabajos revisados, siempre se enfocan más en la búsqueda de un modelo (cada vez más complejo) que tenga el mejor desempeño con *Place Pulse 2.0*. Sin embargo, ninguno realizó ningún análisis previo a los datos.

6.1.1. Análisis exploratorio del conjunto de datos *Place Pulse 2.0*

El análisis de *Place Pulse 2.0* empieza calculando los 111 390 puntuaciones de percepción en la categoría de seguridad en todas las calles a través de las Ecuación 3.1 descritas en el Capítulo 3, cuyos valores finales están en un rango de 0 (poco seguro) a 10 (muy seguro). Durante el proceso de cálculo, se planteó la idea de analizar los datos definiendo regiones geográficos denominados "niveles de generalización geográfica" que abarcan las comparaciones realizadas en 2 imágenes en los siguientes niveles: (i) misma ciudad; (ii) mismo país (incluyendo todas las ciudades de ese país); (iii) mismo continente; y (iv) global (todos los datos). Una vez realizado estos cálculos, pudimos observar 2 problemáticas: (a) la pérdida de información: tal como se muestra en el Cuadro 3.3, vemos que conforme utilizamos una región más pequeña, la cantidad de imágenes disminuye considerablemente; (b) la distribución de las puntuaciones de percepción es poco confiable, tal como se muestra en la Figura 3.5. Vemos que en ciudades con pocas imágenes comparadas (p.ej. Amsterdam) presentan puntuaciones con mayor concentración en 3.33 y 6.66; esto es debido a que el cálculo de puntuaciones de percepción también depende del número de comparaciones realizadas (ver Figura 3.5 (a)).

Este comportamiento se evidencia en todas las ciudades. A nivel de país la distribución va a cambiar dependiendo de cuántas ciudades estén en dicho país. Por ejemplo,

Brasil tiene 3 ciudades y para los datos de Rio de Janeiro se observa un cambio ligero pero insuficiente (ver Figura 3.5 (b)). Así mismo, el caso de USA que contiene 17 ciudades se observa que no es suficiente el número de comparaciones (ver Figura 3.5 (c)). A nivel de continente ya podemos observar un cambio significativo en la distribución. No obstante, en continentes con pocas ciudades como América del Sur, África y Asia aún se observa esta insuficiencia de comparaciones. Caso contrario de América del Norte (17 ciudades) y Europa (22 ciudades). A partir de estos resultados, concluimos que no es posible utilizar *Place Pulse 2.0* sin considerar comparaciones a nivel global (todas las ciudades) debido a las pocas comparaciones realizadas entre imágenes de la misma ciudad.

Otro resultado de observar las distribuciones es la cantidad de imágenes concentradas en el intervalo de 4.5 y 5.5, tal como describimos en la Sección 3.2 se asemeja a una distribución de Gauss con centro cercano al valor de 5.0, siendo este valor próximo al promedio de todas las imágenes (5.18). Por este motivo, para etiquetar a las imágenes en las categorías seguro y no seguro se estableció como umbral 5.0. A partir de este umbral, se observa que tenemos una disparidad de datos, es decir, el número de imágenes de clase segura era diferente al de la clase no segura. Tal como muestra la Figura 3.6 ocurre en casi todas las ciudades. Este umbral no es posible de cambiar puesto que al decrecer el umbral, estaríamos dando prioridad a las imágenes que fueron comparadas en mayor cantidad, y por el contrario, al incrementar el umbral estaríamos incrementando la disparidad de imágenes.

6.1.2. Predicción de la percepción de seguridad urbana

Seguido al análisis de los datos, nos centramos en evaluar diferentes tipos de enfoques basados en modelo de redes convolucionales, el cual pueda presentar un buen rendimiento frente a la naturaleza de *Place Pulse 2.0*, así mismo, un buen desempeño en la tarea de clasificación de la percepción de seguridad urbana (o simplemente percepción de seguridad). Para eso, se planteó un *pipeline* de experimentos basados en dos tipos de aprendizaje ya mencionados: (i) Aprendizaje Supervisado y (ii) Aprendizaje Semi-Supervisado. Las métricas utilizadas para evaluar nuestros modelos son AUC, *F1 Score* generadas por *Presicion-Recall* y *Accuracy*. Se decidió utilizar dichas métricas debido a que la tarea principal es clasificación de dos categorías (segura y no segura). Para el aprendizaje supervisado se optó por utilizar dos técnicas: **transfer-learning** y **fine-tuning**, las cuales utilizaron redes como *VGGNet*, *ResNet50* y *Xception*. Las tres redes entrenadas previamente en *ImageNet*, además, para el modelo *VGGNet* también se usaron los pesos entrenados previamente de *Places365* debido a la naturaleza de los datos; los cuales son imágenes de lugares exteriores e interiores tales como zonas residenciales, calles o restaurantes.

El primer resultado corresponde a los modelos **transfer-learning**, denominados "TL", los cuales consisten en utilizar modelos basados en DCNN como extractores de características (salidas de la última capa). Tal como se describió en la Sección 5.2, entrenamos dichos extractores utilizando 4 métodos lineales y no lineales, cuyos resul-

tados están en el Cuadro 5.2. Observamos que a pesar de tener un *accuracy* de 65 % las métricas *F1 score* y *AUC* no muestran un buen desempeño y esto es debido a que el modelo está prediciendo con mejor exactitud la clase con mayor número de ejemplos. De manera adicional, se decidió experimentar con otros modelos tales como *ResNet50* y *Xception* entrenados previamente sobre *ImageNet* tomando en cuenta dos puntos: el número de parámetros (menos de 23 millones cada uno) y el mayor rendimiento que *VGGNet* sobre dicha base de datos. Sin embargo, solo *ResNet50* mostró resultados similares a los 4 modelos principales, siendo *Xception* descartado. Los principales resultados de los modelos "TL" nos ayudó a entender el comportamiento de los datos al ser entrenados. Utilizando la variación del parámetro δ definido en el Capítulo 5 y mostrado en la Figura 5.1, observamos los valores de las métricas reportadas desde δ=0.05 hasta δ=0.5, mostrándonos que a pesar de tener un alto *accuracy* en δ=0.5, no significa que sea un modelo robusto con buen desempeño (ver Figura 5.2). Así mismo, verificamos que para el "nivel geográfico" tal como ciudad, países y continente no era sostenible utilizar variaciones en el valor de δ, puesto que en algunas ciudades se tenían baja cantidad de imágenes que imposibilitaba el entrenamiento.

El segundo resultado corresponde a los modelos **fine-tuning**, denominados "FT", utilizaron las mismas arquitecturas definidas para los *transfer-learning* a excepción de *Xception*. Para los experimentos, congelamos y entrenamos a partir del quinto bloque de convolución para los modelos basados en *VGGNet* y en el caso de *ResNet50* a partir del bloque 14 residual. Los resultados mostrados en el Cuadro 5.3 muestran un *accuracy* cercano a los reportados en los modelos del grupo "TL". No obstante, hubo una mejora notable en *AUC* y *F1 score* mostrando así que el aprendizaje y predicción del modelo respecto a ambas clases mejoró, debido a que estas métricas reflejan que tan bueno es un modelo para diferenciar las clases evaluadas. En las Figuras 5.4 y 5.3 se muestra un *colormap* de los *accuracy* de cada modelo entrenado evaluado en cada ciudad. La motivación principal era observar el desempeño de modelos entrenados y evaluados en diferentes ciudades. Se observa que el modelo Global para todos los 4 modelos entrenados, mantiene un buen desempeño, y pues, es debido a que Global incluye los datos de todas las ciudades. En las demás ciudades se observa que existen ciudades en las cuales tienen alta *accuracy* y en otras no, y que dependiendo del modelo, eso varía. Además, se observa que todos los modelos mantienen el mismo desempeño evaluado en ciudades como Chicago, Copenhagen, Denver, Dublin, Minneapolis, Montreal, Seattle, New York y Portland.

El tercer resultado corresponde al modelo **GAN semi-supervisada**, la cual consiste en utilizar un conjunto de los datos con las respectivas etiquetas entre 1 para seguro y 0 para no seguro que se estrenarán en el modelo supervisado, también utilizaremos otro conjunto de datos reales sin ninguna etiqueta asociada, las cuales servirán para el entrenamiento del generador y del modelo no supervisado. Así como un GAN *vanilla*, nuestra "SSL_GAN" funciona de manera similar, salvo la diferencia del discriminador. El discriminador se encarga no sólo de discernir entre real o falsa, también identificará a qué clase pertenece una imagen evaluada, ya sea una generada o una del conjunto de datos. Tal como explicamos en la Sección 4.3, la elección de un modelo semi-supervisado fue debido a las limitaciones previamente encontradas en el conjunto de datos. Debido a la diferencia de tiempo de entrenamiento entre los diferentes modelos, la "SSL_GAN"

solo se entrenó utilizando todos los datos. En el Cuadro 6.1 se da una idea aproximada del tiempo de entrenamiento utilizado por cada modelo en cada conjunto de datos (global o ciudad), se observa que el tiempo de la "SSL_GAN" es mucho mayor comparados a los demás modelos. Mencionamos que el tiempo reportado correspondiente a "56 ciudad" es un tiempo promedio de entrenamiento de todas las ciudades por separado.

	Tiempo de entrenamiento para cada modelo	
Método	Datos utilizados	Tiempo promedio
SSL_GAN (32×32)	Global	∼1 semana y media
FT_VGG	Global	∼8 horas
FT_VGG	56 Ciudades	∼6 horas
FT_VGG_GAP	Global	∼7 horas
FT_VGG_GAP	56 Ciudades	∼5 horas
TL_VGG	Global	∼15 minutos
TL_VGG	56 Ciudades	∼10 minutos
TL_VGG_GAP	Global	∼9 minutos
TL_VGG_GAP	56 Ciudades	∼6 minutos

Cuadro 6.1: Cuadro de tiempos promedios de entrenamiento realizado para cada modelo realizando las 5 validaciones cruzadas. Para el caso de "TL" estamos reportando el promedio total de entrenar los 4 modelos para cada caso.

6.2. Limitaciones

El estudio de la percepción urbana es un campo muy complejo puesto que no es posible describir una percepción general (Wilson y Kelling, 1982) y la percepción para cada persona varía dependiendo de el entorno en donde vive una persona (Keizer et al., 2008), es decir, que la percepción es muy relativa y diferenciada para cada persona. Siendo así, presentamos las limitaciones encontradas en el presente estudio, las cuales están fuertemente ligadas al conjunto de datos. Como mencionamos anteriormente, el conjunto de datos *Place Pulse 2.0* analizado nos permitió entender y establecer una metodología enfocada directamente en los datos, en vez del método convencional de pensar en algún modelo complejo que se adecúe.

6.2.1. Percepción individual de los participantes

La construcción del conjunto de datos *Place Pulse* fue a partir de comparaciones entre dos imágenes a través de un sitio web. Para esto, diversos voluntarios realizaron la votación en un conjunto de imágenes totalmente aleatorias. Dicho esto, es posible que algunas imágenes hayan sido comparadas y votadas por algún o algunos voluntarios específicos. Esto genera una dificultad debido a que la percepción de seguridad de una persona es influenciada por el entorno o ambiente en donde vive, generando un criterio

individual sesgado. Esto es un problema al intentar realizar un estudio especializado por ciudad, país, continente debido a que las imágenes eran comparadas y votadas por diversos usuarios de diversos lugares.

Esta limitación no fue posible de solucionar, puesto que es un problema inherente al conjunto de datos, comenzando desde su idealización y cómo fue construido (no tomaron en cuenta percepciones individuales, individuos de una región similar, etc.).

6.2.2. Poca cantidad de datos/imágenes

A nivel general, *Place Pulse 2.0* se compone de 1.22 millones de comparaciones en total, en el Cuadro 3.2 (b) mostramos las estadísticas respectivas de cada categoría, viendo que la categoría correspondiente a seguridad presenta 368 926 comparaciones, siendo esto alrededor del 30.14 % del total de comparaciones. Así mismo, a pesar de que en total fueron comparadas 111 390 imágenes en dicha categoría, el conjunto de datos solo posee 110 988 imágenes. Comparando con otros conjuntos de datos de imágenes con millones de datos, 111 mil imágenes no se compara, además, conjuntos de datos tales como CIFAR10 o MNIST con 60 000 y 50 000 respectivamente, tienen un número de proporción similar de imágenes por cada clase. En nuestro caso, el número de imágenes no es homogéneo por ciudad, es decir, tenemos el caso de la ciudad de Atlanta con 4 034 y casos como Amsterdam con 637 imágenes (ver Figura 3.3 (a)). Lo cual, al no tener un conjunto homogéneo para cada caso, no permite poder tener un análisis especializado en cada caso.

Esta limitación fue posible mitigar a través de la utilización del modelo semi-supervisado "SSL_GAN", la poca cantidad de imágenes y la desproporción de imágenes por ciudad fue amortiguada a través de las imágenes sintéticas generadas en cada iteración, siendo estas utilizadas en el entrenamiento.

6.2.3. Generalización a través de características de las ciudades

Debido al número de ciudades y la gran variedad del número de imágenes de cada ciudad presentes en *Place Pulse 2.0*, no es posible encontrar un modelo que consiga generalizar una predicción con alta precisión. Esto se presenta en casos como Atlanta o Berlín, cuyas imágenes están compuestas (en mayoría) por una carretera y a los lados árboles o pasto; Por el otro lado, tenemos ciudades totalmente urbanizadas tipo Boston. Así mismo, características únicas (o diferenciadas) se encuentran en Tokyo y Kyoto, donde las calles presentan una clara diferencia a las demás ciudades, ya sea de Europa o América del Norte donde se cuentan con rascacielos y edificios muy altos.

Además, tal como describimos en el Cuadro 3.3 conforme estudiamos los posibles "niveles de generalidad geográficos" entendimos que teníamos una pérdida de información mayor en cuanto más específico sea el nivel. Por ejemplo, a nivel global tenemos en el "nivel geográfico" global tenemos 111 390 imágenes y en el "nivel geográfico" ciudad

tenemos 20 143 imágenes, significando una reducción de casi 82 % en el número de imágenes de la categoría "seguridad".

Esta limitación no fue posible de mitigar debido a que está fuertemente ligada a la construcción del conjunto de datos, evaluar dos imágenes de manera aleatoria sin tomar en cuenta de qué imágenes son comparadas ni cuantas veces son comparadas, hace que de manera obligatoria el conjunto de datos se tenga que analizar en total. Dejando así una posibilidad nula de realizar estudios específicos a lo largo de diferentes "nivel de generalización geográfica".

6.2.4. Disparidad del conjunto de datos

De las limitaciones expuesta anteriormente, a partir del cálculo de las puntuaciones de percepción realizadas, observamos a partir de la Figura 3.5 que para el nivel global, las ciudades tienen una distribución similar a una distribución de Gauss con centro cercano en 5.0, ya sea en países con 2 o más ciudades o de una ciudad. La mediana de las puntuaciones es el valor 4.72 el cual es muy próximo a 5.0. Al utilizar nuestro umbral de 5.0, tenemos una disparidad de 11 233 imágenes con respecto a la mediana. Utilizar la mediana como umbral, es decir, dividir 50 % como seguro y 50 % como no seguros no marca ninguna diferencia relevante al momento de entrenar los datos, además, observando manualmente a un subconjunto de las 11 233 imágenes correspondientes a dichos puntuaciones no tienen una apariencia visual que merecería ser etiquetada como segura.

Esta limitación fue posible mitigar a través de la utilización del modelo semi-supervisado "SSL_GAN", la disparidad de datos presentes fue mitigada a través de la generación de imágenes sintéticas por parte del generador durante, que se iban adicionando al entrenamiento en cada iteración.

6.3. Consideraciones Finales

En este Capítulo hemos descrito en detalle las discusiones de los experimentos realizados y además, las limitaciones encontradas en el conjunto de datos *Place Pulse 2.0* que fueron expuestos en el Capítulo 3. También discutimos de qué manera fue posible solucionar dichas limitaciones encontradas a partir de nuestro análisis. Infelizmente no se consiguió solucionar todas, puesto que la naturaleza de los datos y su construcción impiden realizar un análisis más profundo en base a la percepción individual.

Capítulo 7

Conclusiones

En este trabajo se presentó un análisis exploratorio del conjunto de datos *Place Pulse 2.0*; a través de los diferentes estudios y enfoques que explicamos en detalle en el Capítulo 3. Como resultado del análisis exploratorio se encontraron limitaciones en el conjunto de datos estudiado. De las cuatro limitaciones encontradas y explicadas, sólo pudimos resolver computacionalmente tres: Poca muestras de imágenes en general, desproporción en la cantidad de imágenes por ciudad y disparidad de las clases. La única limitación no resuelta está fuertemente relacionada a cómo el conjunto de datos fue construido, basado en la percepción de cada persona y su elección sobre qué imagen es la más segura. Es decir, el conjunto de datos está limitado a la percepción individual de cada persona. Se observó que durante la construcción de los datos se compararon imágenes aleatorias siendo evaluadas por un usuario aleatorio, así mismo, se encontró que el número de imágenes utilizadas por cada ciudad no es proporcional, teniendo ciudades con aproximadamente 4 mil imágenes (Sao Paulo) y otras con menos de 700 (Amsterdam), lo cual impide realizar un análisis individual para cada ciudad. Esta desproporción no permite una generalización lo cual está forzando una dependencia entre sí (a través del cálculo de las puntuaciones de percepción). Sin embargo, fue posible combatir esta desproporción de datos a través del aprendizaje semi-supervisado, un modelo generativo como GAN nos permite extender de manera artificial el conjunto de datos a través de la generación de nuevos datos.

En consecuencia, se analizó y presentó los resultados de las evaluaciones de diferentes modelos clasificadores utilizando técnicas como *Transfer-Learning*, *Fine-Tuning* y GANs. Las evaluaciones fueron realizadas reportando 3 métricas principales: *F1 score* la cual es una media entre *Precision-Recall*; *Accuracy* la cual reporta cuántas imágenes fueron predichas correctamente en cada categoría; y AUC para determinar la proporción en que ambas categorías fueron correctamente clasificadas. La métrica principal utilizada fue el AUC debido a que al tener una disparidad de datos, así mismo, se observa que conforme se entrena un modelo especializado se obtiene un mejor valor del AUC. Notamos que se obtiene un incremento desde un $\sim$59 % (*Transfer-Learning*) hasta un $\sim$81 % (GAN), también vemos ese comportamiento en las otras dos métricas: *accuracy* se incrementa desde $\sim$66 % hasta $\sim$81 % y el *F1 score* incrementa desde $\sim$51 % hasta

∼81 %. Esto demuestra nuestra hipótesis inicial, indicando que era necesario de un modelo que pueda resolver las limitaciones de los datos. Para ese caso, un modelo GAN frente a datos con disparidad mencionada y discutida en detalle en el capítulo anterior tiene un óptimo desempeño. Finalmente, destacamos que nuestra GAN es un modelo estable frente al conjunto de datos de *Place Pulse 2.0* cuya naturaleza son imágenes de calles de diferentes ciudades. Como herramienta final, se presentó un sistema web con el cual es posible realizar una interacción fácil y visualización concisa de los resultados obtenidos por cada modelo, tales como los resultados de cada evaluación realizada en las validaciones cruzadas (reportando las 3 métricas utilizadas). A manera de trabajo a futuro, pensamos extender el trabajo añadiendo más datos y aumentando la resolución de la GAN. esto servirá para identificar características más específicas de cada lugar.

Bibliografía

Abadi, M., Agarwal, A., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abu-Mostafa, Y. S., Magdon-Ismail, M., et al. (2012). *Learning From Data*. AMLBook.

Acosta, S. y Camargo, J. (2018a). wmodi. http://wmodi.com/. [Último acceso: 11-Agosto-2022].

Acosta, S. F. y Camargo, J. E. (2018b). Predicting city safety perception based on visual image content. In *Iberoamerican Congress on Pattern Recognition*, pages 177–185. Springer.

Adebayo, J., Gilmer, J., et al. (2018). Sanity checks for saliency maps.

Ali, N. y Zafar, B. (2018). 15-scene image dataset. figshare. dataset. https://doi.org/10.6084/m9.figshare.7007177.v1.

Alzate, J. R., Tabares, M. S., et al. (2021). Graffiti and government in smart cities: a deep learning approach applied to medellín city, colombia. In *International Conference on Data Science, E-learning and Information Systems 2021*, pages 160–165.

Ancona, M., Ceolini, E., et al. (2017). A unified view of gradient-based attribution methods for deep neural networks. ETH Zurich.

Andersson, V. O., Birck, M. A., et al. (2017). Investigating crime rate prediction using street-level images and siamese convolutional neural networks. In *Latin American Workshop on Computational Neuroscience*, pages 81–93. Springer.

ArcGis (1999). Arcgis. https://www.arcgis.com/index.html. [Último acceso: 11-Agosto-2022].

Arietta, S. M., Efros, A. A., et al. (2014). City forensics: Using visual elements to predict non-visual city attributes. *IEEE transactions on visualization and computer graphics*, 20(12):2624–2633.

Bay, H., Tuytelaars, T., et al. (2006). Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer.

Boser, B. E., Guyon, I. M., et al. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM.

Broomhead, D. y Lowe, D. (1988). Multivariable functional interpolation and adaptive networks, complex systems, vol. 2.

Cenggoro, T. W. et al. (2018). Deep learning for imbalance data classification using class expert generative adversarial network. *Procedia Computer Science*, 135:60–67.

Chakravarti, R. y Meng, X. (2009). A study of color histogram based image retrieval. pages 1323 – 1328.

Charalampos, P., Panagiotis, K., et al. (2019). Storm graffiti/tagging detection dataset. https://doi.org/10.5281/zenodo.3238357.

Chen, L.-C., Papandreou, G., et al. (2017). Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258.

CrimeMapping (2012). Crime mapping website. https://www.crimemapping.com/map. [Último acceso: 11-Agosto-2022].

CrimeReports (2013). Crime report website. https://www.crimereports.com/. [Último acceso: 11-Agosto-2022].

CS231n (2022). Stanford cs231n. http://cs231n.stanford.edu/schedule.html. [Último acceso: 11-Agosto-2022].

CycloMedia (1980). Street smart api. https://www.cyclomedia.com/en/urban-road-safety-index. [Último acceso: 11-Agosto-2022].

Dalal, N. y Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.

Deng, J., Dong, W., et al. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Diniz, A. M. A. y Stafford, M. C. (2021). Graffiti and crime in belo horizonte, brazil: The broken promises of broken windows theory. *Applied Geography*, 131:102459.

Doersch, C., Singh, S., et al. (2012). What makes paris look like paris?

Donahue, J., Jia, Y., et al. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655.

Dubey, A., Naik, N., et al. (2016). Deep learning the city : Quantifying urban perception at A global scale. *CoRR*, abs/1608.01769.

EuroStat (2016). Crime statistics. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Crime_statistics. [Último acceso: 11-Agosto-2022].

Felipe Moreno-Vera, Bahram Lavi, J. P. (2021a). Quantifying urban safety perception on street view images. In *International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*.

Felipe Moreno-Vera, Bahram Lavi, J. P. (2021b). Urban perception: Can we understand why a street is safe? In *Mexican International Conference on Artificial Intelligence (MICAI)*.

Felzenszwalb, P., McAllester, D., et al. (2008). A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

Fisher, B.S.; Nasar, J. (1992). *Fear of crime in relation to three exterior site features prospect, refuge, and escape.*, volume 24, pages 35–65.

Fu, K., Chen, Z., et al. (2018). StreetNet: preference learning with convolutional neural network on urban crime perception. In *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 269–278. ACM.

Garay, A. M., Hashimoto, E. M., et al. (2011). On estimation and influence diagnostics for zero-inflated negative binomial regression models. *Computational Statistics & Data Analysis*, 55(3):1304–1318.

Girshick, R. (2015). Fast R-CNN. *2015 IEEE International Conference on Computer Vision (ICCV)*.

Girshick, R., Donahue, J., et al. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*.

Glaeser, E. L., Kominers, S. D., et al. (2018). Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry*, 56(1):114–137.

Gong, Y., Lazebnik, S., et al. (2012). Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2916–2929.

Goodfellow, I. J., Bengio, Y., et al. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA. http://www.deeplearningbook.org.

Google-Developers (2020). Ml practicum. https://developers.google.com/machine-learning/practica/image-classification?hl=es_419. [Último acceso: 11-Agosto-2022].

Google-Motorolla (2019). Crime map. https://www.crimereports.com/home/. [Último acceso: 11-Agosto-2022].

Har-Peled, S., Roth, D., et al. (2003). Constraint classification for multiclass classification and ranking. In *Advances in neural information processing systems*, pages 809–816.

He, K., Gkioxari, G., et al. (2017). Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*.

He, K., Zhang, X., et al. (2015). Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Herbrich, R., Minka, T., et al. (2007). Trueskill(tm): A bayesian skill rating system. In *Advances in Neural Information Processing Systems 20*, pages 569–576. MIT Press.

Hoiem, D., Efros, A. A., et al. (2007). Recovering surface layout from an image. *International Journal of Computer Vision*, 75(1):151–172.

Hu, Y.-T., Huang, J.-B., et al. (2017). Maskrnn: Instance level video object segmentation. In *Advances in Neural Information Processing Systems*, pages 325–334.

Joachims, T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM.

Jurie, F. y Triggs, B. (2005). Creating efficient codebooks for visual recognition. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 604–610. IEEE.

Karras, T., Aittala, M., et al. (2020). Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*.

Keizer, K., Lindenberg, S., et al. (2008). The spreading of disorder. *Science (New York, N.Y.)*, 322:1681–5.

Koch, G., Zemel, R., et al. (2015). Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2.

Krizhevsky, A., Sutskever, I., et al. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., et al., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Lazebnik, S., Schmid, C., et al. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE.

Li, X., Zhang, C., et al. (2015a). Does the visibility of greenery increase perceived safety in urban areas? evidence from the place pulse 1.0 dataset. *ISPRS International Journal of Geo-Information*, 4(3):1166–1183.

Li, X., Zhang, C., et al. (2015b). Assessing street-level urban greenery using google street view and a modified green view index. *Urban Forestry & Urban Greening*, 14(3):675–685.

Lin, M., Chen, Q., et al. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Lin, T.-Y., Maire, M., et al. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Lindal, P. J. y Hartig, T. (2013). Architectural variation, building height, and the restorative quality of urban residential streetscapes. *Journal of Environmental Psychology*, 33:26–36.

Liu, W., Anguelov, D., et al. (2016). SSD: single shot multibox detector. *European Conference on Computer Vision (ECCV)*, abs/1512.02325.

Liu, X., Chen, Q., et al. (2017). Place-centric visual urban perception with deep multi-instance regression. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 19–27. ACM.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Lynch, K. (1984). Reconsidering the image of the city. In *Cities of the Mind*, pages 151–161. Springer.

Manjunath, B. S., Ohm, J.-R., et al. (2001). Color and texture descriptors. *IEEE Transactions on circuits and systems for video technology*, 11(6):703–715.

Martin, D., Fowlkes, C., et al. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE.

Massachusets-Office-Goverment (2005). Massgis data-land use 2005. https://www.mass.gov/. [Último acceso: 11-Agosto-2022].

Matas, J., Chum, O., et al. (2004). Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767.

MatLab-Developers (2020). Mathworks. https://www.mathworks.com/discovery/object-detection.html. [Último acceso: 11-Agosto-2022].

Mawby, R. (2014). *Crime and Disorder, Security and the Tourism Industry*, pages 383–403. Springer.

Min, W., Mei, S., et al. (2019). Multi-task deep relative attribute learning for visual urban perception. *IEEE Transactions on Image Processing*, 29:657–669.

MIT-Media-Lab (2013). Place pulse. http://pulse.media.mit.edu/data/. [Último acceso: 11-Agosto-2022].

MIT-Media-Lab (2014). Streetscore. http://streetscore.media.mit.edu/. [Último acceso: 11-Agosto-2022].

MIT-Media-Lab (2015). Treepedia. http://senseable.mit.edu/treepedia. [Último acceso: 11-Agosto-2022].

Mohammed, A.-M. y Sookram, S. (2015). The impact of crime on tourist arrivals—a comparative analysis of jamaica and trinidad and tobago. *Social and Economic Studies*, 64(2):153–176.

Molnar, C. (2022). *Interpretable Machine Learning*. 2 edition.

Moreno-Vera, F. (2021). Understanding safety based on urban perception. In *International Conference on Intelligent Computing*, pages 54–64. Springer.

Naik, N., Philipoom, J., et al. (2014). StreetScore: predicting the perceived safety of one million streetscapes. *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Nasar, J., Fisher, B., et al. (1993). *Proximate physical cues to fear of crime.*, volume 26, pages 161–178.

Nasar, J. L. (1998). The evaluative image of the city.

Noh, H., Hong, S., et al. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528.

Novak, C. L., Shafer, S. A., et al. (1992). Anatomy of a color histogram. In *CVPR*, volume 92, pages 599–605.

Numbeo (2019). World database of crime index. https://www.numbeo.com/crime/rankings_by_country.jsp?title=2019. [Último acceso: 11-Agosto-2022].

Ojala, T., Pietikainen, M., et al. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987.

Oliva, A. y Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175.

Ordonez, V. y Berg, T. L. (2014). Learning high-level judgments of urban perception. *European Conference on Computer Vision (ECCV)*.

Otsu, N. (1975). *A threshold selection method from gray-level histograms.*, volume 11, pages 23–27.

Park, D. K., Jeon, Y. S., et al. (2000). Efficient use of local edge histogram descriptor. In *Proceedings of the 2000 ACM workshops on Multimedia*, pages 51–54.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076.

Pedregosa, F., Varoquaux, G., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Perronnin, F., Sánchez, J., et al. (2010). Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer.

Pinheiro, P. O., Collobert, R., et al. (2015). Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998.

Porzi, L., Rota Bulò, S., et al. (2015). Predicting and understanding urban perception with convolutional neural networks.

Quercia, D., O'Hare, N. K., et al. (2014). Aesthetic capital: what makes london look beautiful, quiet, and happy? In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 945–955. ACM.

Rao, A., Srihari, R. K., et al. (1999). Geometric histogram: A distribution of geometric configurations of color subsets. In *Internet Imaging*, volume 3964, pages 91–101. International Society for Optics and Photonics.

Ray, S. y Page, D. (2001). Multiple instance regression. In *ICML*, volume 1, pages 425–432.

Redmon, J., Divvala, S., et al. (2016). You only look once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ren, S., He, K., et al. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149.

Ronneberger, O., Fischer, P., et al. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Russakovsky, O., Deng, J., et al. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Salesses, M. P. (2012). *Place Pulse: Measuring the collaborative image of the city*. PhD thesis, Massachusetts Institute of Technology.

Salesses, P., Schechtner, K., et al. (2013). The collaborative image of the city: Mapping the inequality of urban perception. *PLOS ONE*.

Salimans, T., Goodfellow, I., et al. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242.

Sampath, V., Maurtua, I., et al. (2021). A survey on generative adversarial networks for imbalance problems in computer vision tasks. *Journal of big Data*, 8(1):1–59.

Sampson, R. J., Morenoff, J. D., et al. (2002). Assessing "neighborhood effects": Social processes and new directions in research. *Annual review of sociology*, 28(1):443–478.

Schroeder, H. W. y Anderson, L. M. (1984). Perception of personal safety in urban recreation sites. *Journal of leisure research*, 16(2):178–194.

Selvaraju, R. R., Cogswell, M., et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

Seresinhe, C. I., Preis, T., et al. (2017). Using deep learning to quantify the beauty of outdoor places. *Royal Society open science*, 4(7):170170.

Shrikumar, A., Greenside, P., et al. (2016). Not just a black box: Learning important features through propagating activation differences.

Simonyan, K., Vedaldi, A., et al. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Simonyan, K. y Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*.

Sivic, J. y Zisserman, A. (2004). Video data mining using configurations of viewpoint invariant regions. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE.

Skogan, W. G. (1992). *Disorder and decline: Crime and the spiral of decay in American neighborhoods*. Univ of California Press.

Smilkov, D., Thorat, N., et al. (2017). Smoothgrad: removing noise by adding noise.

Smola, A. y Schölkopf, B. (2004). A tutorial on support vector regression, statist. *Comput*, 14:199–222.

Springenberg, J. T., Dosovitskiy, A., et al. (2014). Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*.

Stalidis, P., Semertzidis, T., et al. (2018). Examining deep learning architectures for crime classification and prediction. *arXiv*.

Sundararajan, M., Taly, A., et al. (2017). Axiomatic attribution for deep networks.

Szegedy, C., Liu, W., et al. (2014). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Szegedy, C., Vanhoucke, V., et al. (2015). Rethinking the inception architecture for computer vision. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tencent-Street-View-service (2016). Map qq. https://map.qq.com/. [Último acceso: 11-Agosto-2022].

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tikhonov, A. N. (1943). On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198.

Tokuda, E. K., Silva, C. T., et al. (2019). Quantifying the presence of graffiti in urban environments. *CoRR*, abs/1904.04336.

Tranmer, M. Multiple linear regression.

UK-gov (2015). Geograph. http://www.geograph.org.uk/. [Último acceso: 11-Agosto-2022].

UK-gov (2017). Scenic-or-not. http://scenicornot.datasciencelab.co.uk/. [Último acceso: 11-Agosto-2022].

Ulrich, R. S. (1979). Visual landscapes and psychological well-being. *Landscape research*, 4(1):17–23.

UrbanForest (2014). Urban forests map website. http://urbanforestmap.org/. [Último acceso: 25-Agosto-2019].

UrbanGems (2014). Urbangems. http://urbangems.org/. [Último acceso: 14-Octubre-2019].

USA, D. o. J. (2012). Mapping crime: Principle and practice. https://www.ncjrs.gov/pdffiles1/nij/178919.pdf. [Último acceso: 11-Agosto-2022].

Viso-AI (2020). Object segmentation. https://viso.ai/deep-learning/image-segmentation-using-deep-learning. [Último acceso: 11-Agosto-2022].

von Platen, P., Tao, F., et al. (2020). Multi-task siamese neural network for improving replay attack detection. *arXiv preprint arXiv:2002.07629*.

Wikipedia. Linear model. https://es.wikipedia.org/wiki/Modelo_lineal. [Último acceso: 11-Agosto-2022].

Wikipedia. Supervised learning. https://en.wikipedia.org/wiki/Supervised_learning. [Último acceso: 11-Agosto-2022].

Wikipedia (2020). Non-linear models. https://es.wikipedia.org/wiki/Regresi%C3%B3n_no_lineal. [Último acceso: 11-Agosto-2022].

Wilson, J. Q. y Kelling, G. L. (1982). Broken windows. *Atlantic monthly*, 249(3):29–38.

Xiao, J., Hays, J., et al. (2010). Sun database: Large-scale scene recognition from abbey to zoo. pages 3485–3492.

Xu, Y., Yang, Q., et al. (2019). Visual urban perception with deep semantic-aware network. In *International Conference on Multimedia Modeling*, pages 28–40. Springer.

Y., L., Boser, B., et al. (1990). Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, pages 396–404. Morgan Kaufmann.

Yang, J., Zhao, L., et al. (2009). Can you see green? assessing the visibility of urban forests in cities. *Landscape and Urban Planning*, 91(2):97–104.

Yu, F. y Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Zeiler, M. D. y Fergus, R. (2013a). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Zeiler, M. D. y Fergus, R. (2013b). Visualizing and understanding convolutional networks.

Zhang, F., Zhou, B., et al. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, 180:148–160.

Zhao, H., Shi, J., et al. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890.

Zhou, B., Khosla, A., et al. (2016a). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

Zhou, B., Khosla, A., et al. (2016b). Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*.

Zhou, B., Lapedriza, A., et al. (2016c). Places365 results. https://github.com/CSAILVision/places365/. [Último acceso: 11-Agosto-2022].

Zhou, B., Lapedriza, A., et al. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464.

Zhou, B., Lapedriza, A., et al. (2014). Learning deep features for scene recognition using places database. In Ghahramani, Z., Welling, M., et al., editors, *Advances in Neural Information Processing Systems 27*, pages 487–495. Curran Associates, Inc.

Zhou, T., Liu, W., et al. (2018). Gan-based semi-supervised for imbalanced data classification. In *2018 4th International Conference on Information Management (ICIM)*, pages 17–21. IEEE.

Zou, H. y Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.