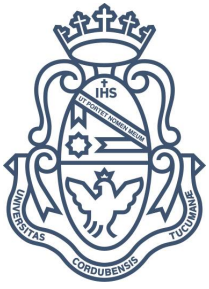
Facultad de Matemática, Astronomía,
Física y Computación

Universidad Nacional de Córdoba



Zero-Shot Object Detection

Tesis

que para obtener el título de

Licenciado en Ciencias de la Computación

presenta

Agustin Horacio Urquiza Toledo

Director: Jorge Sanchez

Córdoba, Argentina 2019

DEDICATORIA

Agradecimientos

Resumen

Pasado el año 2000 se dan dos hechos que harán que las imágenes en Internet de un gran salto. Por un lado se empiezan a popularizar las cámaras digitales y por otro lado las conexiones de Internet subieron su velocidad. Esto genero la necesidad de crear métodos veloces y eficaces que faciliten la extracción de información en este tipo de datos. Luego, a partir del año 2010 con la "Revolución" del Aprendizaje profundo, surgieron una gran cantidad de métodos para realizar esta tarea, entre ellos los Detectores. Pero esto genero la necesidad de tener una gran cantidad de imágenes anotadas, que en algunos casos no resulta viable. **Zero-shot Object Detection** intenta atacar este problema. En esta tesis estudiaremos, analizaremos y probaremos este método.

Índice general

Capítulo 1

Introducción y Motivación

1.1. Historia

La detección de objetos es una de las áreas de la visión por computadora que está creciendo muy rápidamente. Gracias al aprendizaje profundo, cada año, los nuevos algoritmos/modelos siguen superando a los anteriores. Aunque la visión por computadora recientemente tomo gran importancia (el momento decisivo ocurrió en 2012 cuando AlexNet ganó ImageNet), ciertamente no es un nuevo campo científico. Uno de los artículos más influyentes en Visión Informática fue publicado por dos neurofisiólogos, David Hubel y Torsten Wiesel, en 1959. Su publicación, titulada "Receptive fields of single neurons in the cat's striate cortex" en español "Campos receptivos de neuronas individuales en la corteza estriada del gato", describieron las propiedades de respuesta central de las neuronas corticales visuales y cómo la experiencia visual de un gato moldea su arquitectura cortical. Los investigadores establecieron, a través de su experimentación, que existen neuronas simples y complejas en la corteza visual primaria y que el procesamiento visual siempre comienza con estructuras simples como los bordes orientados. En la actualidad, es esencialmente el principio básico detrás del aprendizaje profundo.

Otro echo importante en la historia de la visión por computadora fue,

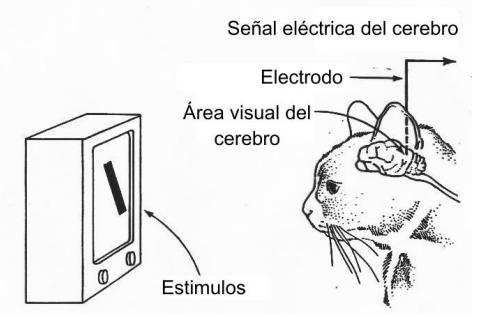


Figura 1.1: En esta imagen se puede ver en que consistía el experimento realizado por David Hubel y Torsten Wiesel

en 1959, Russell Kirsch y sus colegas desarrollaron un aparato que permitía transformar imágenes en cuadrículas de números: las máquinas de lenguaje binario podían entender. En la década de 1960 fue cuando la IA se convirtió en una disciplina académica y algunos de los investigadores, eran extremadamente optimistas sobre el futuro del campo. En este periodo, Seymour Papert, profesor del laboratorio de IA del MIT, decidió lanzar el Proyecto de Verano y resolver, en pocos meses, el problema de la visión artificial. Los estudiantes debían diseñar una plataforma que pudiera realizar, automáticamente, segmentación de fondo y extraer objetos no superpuestos de imágenes del mundo real. Claro esta que el proyecto no fue un éxito. Cincuenta años después, todavía no estamos cerca de resolver la visión por computadora. Sin embargo, ese proyecto fue, el nacimiento oficial de CV como campo científico. A este acontecimiento le siguieron una gran cantidad de investigaciones que hicieron grandes aportes al campo de la vision por computadoras. Como la tesis de doctorado de Roberts, Lawrence [10] en 1963, el paper de David Marr [5] en 1982, entre los mas reconocidos.

Pero los aportes mas influyentes a este campo empezaron a surguir en los dos mil. En 2001 Paul Viola y Michael Jones presentaron el primer detector de rostros que funcionó en tiempo real. Aunque no se

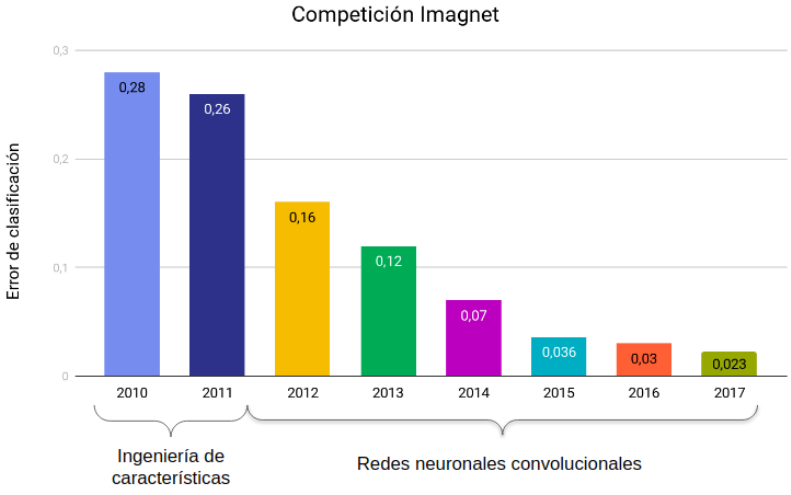
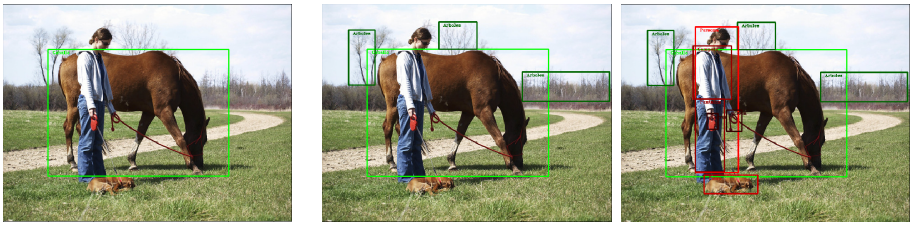


Figura 1.2: La imagen muestra la evolución de los modelos propuestos en la competencia ILSVRC

basa en el aprendizaje profundo, el algoritmo tenia una relación con el mismo, ya que, al procesar imágenes, aprendió qué características podría ayudar a localizar caras, inspirados en el experimento de David Hubel y Torsten Wiesel. En 2006 comenzó la competencia de Pascal VOC, que permitió evaluar el desempeño de diferentes métodos para el reconocimiento de la clase de objeto. En 2010 siguiendo los pasos de Pascal VOC, inicio el concurso de reconocimiento visual a gran escala ImageNet (ILSVRC). En 2010 y 2011, la tasa de error del ILSVRC en la clasificación de imágenes rondaba el 26 % Pero en 2012, un equipo de la Universidad de Toronto ingresó a la competencia un modelo de red neuronal convolucional (AlexNet) y eso cambió todo. El modelo, similar en su arquitectura al LeNet-5 de Yann LeCun, logró una tasa de error del 16,4 %. En los años siguientes, las tasas de error en la clasificación de imágenes en ILSVRC cayeron a un pequeño porcentaje y los ganadores, desde 2012, siempre han sido redes neuronales convolucionales.

1.2. Detectores y ZSD

La detección de objetos es un sub-problema de la visión artificial, que estudia cómo detectar la presencia de objetos en una imagen sobre su apariencia visual. Debido a la complejidad de poder detectar todas las instancias de todos los posibles objectos en una imagen, existen diferentes tareas que tratan de disminuir la dificultad. Par poder explicar los distintos problemas, es necesario distinguir dos conjuntos. Los datos de entrenamiento, consta de las imágenes que se usaran para entrenar el modelo, con sus respectivas etiquetas, es decir, que objetos se encuentran en la imagen, localización de los objetos, descripción de la imagen, o cualquier información extra que requiera la tarea. Las imágenes de prueba, es el conjunto donde se observara o medirá la eficiencia del modelo ya entrenado. Supongamos que las etiquetas, solo cuenta con dos tipos de informacion, que clase de objeto es, es decir si es un perro, auto, persona, etc. y su localización en la imagen. Todas las clases que aparecen en los datos de entrenamiento llamaremos clases visibles o vistas. Toda aquella clase que no sea una clase vista la llamearemos imbisible o no vista.
La **clasificación**, consta en un modelo capas de retornar que objecto hay en una imagen. **Clasificación + localización**, ademas de poder clasificar tiene que ser capas de ubicar el objecto en la imagen. Ambos modelos clasifican en una sola clase vista. **El reconocimiento de imagen**, predice que objetos perteneciente a las clases visibles están presente en la imagen. **La detección de objetos**, tiene que ser capas ademas, de poder localizar dichos objetos. El tradicional **reconocimiento por disparo cero**, tiene que poder reconocer clases no vistas. Por ultimo la **detección de objetos por disparo cero** o ZSD por sus siglas en ingles, debe localizar y clasificar todas las instancias de objetos en la imagen, sin depender si es una clase vista o no. La figura muetra un ejemplo de las distintas tareas antes mencionadas.



(a) **Clasificación** (b) **Detección de objetos** (c) **ZSD**

Figura 1.3: En esta figura se muestra un ejemplo de las tareas mencionadas. En la escala de los verdes son las clases vistas {Caballo, Árbol}, y en rojo las clases invisibles {Perro, Persona, Campera, Pantalón, Correa}

.

Existen otros problemas que no mencionamos acá, como la segmentación. Ya que en este documento trataremos la tarea de ZSD. En el capitulo Capítulo 2 formalizaremos lo aquí explicado.

1.3. Estado del arte

ZSD es un problema que tomo impulso recién en los últimos años, aunque sus estudios se remontan desde mucho antes, como ya mencionamos. Existen muchas técnicas y propuestas para poder resolver este problema, cuando se empezó a leer sobre este tema a fines del 2018 la mas utilizada era usando Multi-modales. Esta es una tecnica que se esta usando mucho como [1] quien utilizó imágenes, texto y sonido para generar representaciones discriminatorias profundas que se comparten en las tres modalidades. Del mismo modo, [16] utilizó imágenes y descripciones de texto para una mejor localización de la entidad visual basada en el lenguaje natural. La idea (para resolver el problema de ZSD) es utilizar un espacio de representación de visión y lenguaje compartido para obtener descripciones de región de imagen y palabra que pueden compartirse en múltiples dominios de visión y lenguaje. Para lograr esto se utiliza por un lado **Las Incrustacio-**

nes de palabras: asignan palabras a una representación vectorial continua codificando similitud semántica entre palabras. Estos vectores de palabras funcionan bien en tareas tales como medir similitudes semánticas y sintácticas entre palabras. Entre los modelos mas famosos se encuentran Glove[8] y Word2vec[6]. Por otro lado se tiene que poder extraer los **vectores con representaciones visuales**, entre los mejores modelos se encuentran VGG [11] ResNet [12], Inception [13]. Todos estos modelos usan redes profundas para extraer dichas características.

A fines del 2018 se encontro tres trabajos paralelos que apuntaban a resolver el problema de ZSD. [9] [17] [2]. Luego de leer los todos, por una decision personal y con ayuda de mi director eligimos atacar el problema basándonos en el Paper [2]. Asi tambien sacamos muchos conceptos sobre disparo cero generalizado de [15]. En Zero-Shot Object Detection[2], enfrentan el problema de disparo cero de manera similar a la que tenemos los humanos de reconocer un objeto dada una descripción semántica. Es decir asociamos tanto la palabra que representa el objeto con su aspecto visual. Se utiliza dos extractores para simular estas cualidades, uno semántico y otro visual. Luego, en el momento del entrenamiento se proporcionan ejemplos visuales para algunas clases visuales, pero durante la prueba se espera que el modelo reconozca instancias de clases que no se vieron, con la restricción de que las nuevas clases estén semánticamente relacionadas con las clases de entrenamiento.

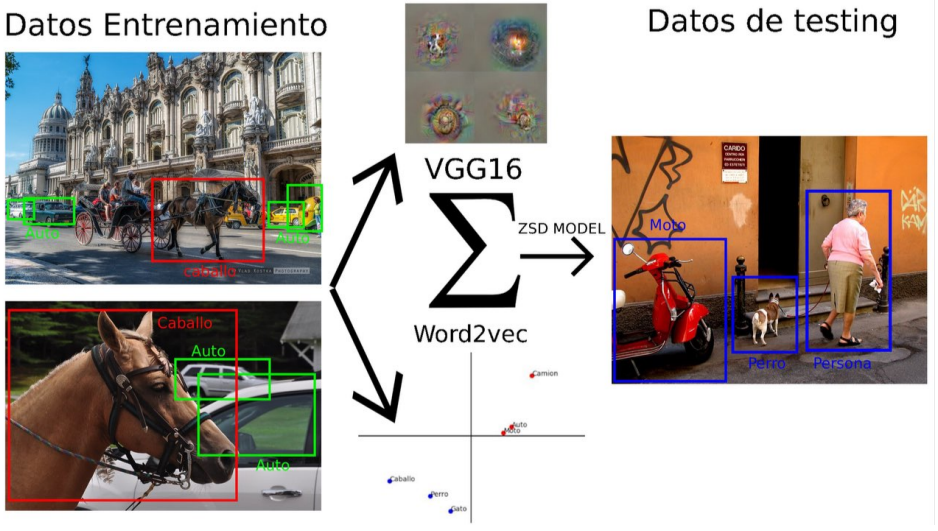


Figura 1.4: Se describe la tarea de detección de objetos por disparo cero, donde los objetos "Auto" y "Caballo" se observan durante el entrenamiento y "Persona", "Perro" y "Moto" son clases invisibles. El enfoque localiza estas clases invisibles aprovechando las relaciones semánticas entre las clases visibles e invisibles y su aspecto visual.

1.4. Motivación

Hoy en día, hay una gran cantidad de modelos, capaces de detectar objetos en una imagen, como son las redes YOLO o Faster R-CNN. Estos, como otros no mencionados, poseen una excelente performance. Pero tienen una gran limitación, necesitan una gran cantidad de imágenes anotadas, para cada clase que se quiere detectar. Por unos minutos dejemos llevarnos nos por la imaginación y supongamos que se quiere crear un programa capas de reconocer todos los objeto en una imagen, pero objetos de cualquier índole, animales, plantas, artículos de limpieza, o cualquier cosa que se te venga a la mente. Seria casi imposible, si es que no lo es, generar un data set que contenga una cantidad considerable de imágenes de todos los objetos posible. Esta

idea puede sonar muy descabellada, o no, pero no se puede negar su potencial y su gran cantidad de usos. Se podría relacionar productos similares a uno que se observa en una imagen, y recomendarnos estos, sin la necesidad de ser lo mismo o lucir parecidos. Si entrenamos un modelo con un conjunto reducido de plantas, luego se puede usar este programa para que extrapole su detección a todas las especies existente en el planeta. Estas son algunas ideas de problemas que se pueden resolver usando ZSD. En nuestro caso en particular estamos, usando multi-modales para resolver este problema, tratando de encontrar una relación entre dos tipos de datos, texto e imágenes. Multi-modales por si solo, tienen innumerables casos de usos.

Capítulo 2

Definición del problema

2.1. Redes neuronales convolucionales

Las redes neuronales convolucionales CNN por sus siglas en ingles, es un tipo de modelo de aprendizaje profundo para procesar datos que tiene un formato de cuadrícula, como las imágenes. Está inspirado en la organización de la corteza visual de los animales, diseñada para aprender de forma automática y adaptativa, patrones en jerarquías, de bajo a alto nivel. Por lo genera una red CNN se compone de tres tipos de capas: convolución, agrupación y capas completamente conectadas. Las dos primeras, realizan extracción de características, mientras que la tercera, asigna las características extraídas en la salida final. La capa de convolución desempeña un papel clave en CNN, se compone de una pila de operaciones matemáticas, como la convolución, que es un tipo especializado de operación lineal. En las imágenes en 2D estas redes son muy utilizadas, por su alta eficiencia extrayendo características en cualquier parte de la imagen.
Algunos ejemplos de redes CNN son, VGG16 que posee 13 capas de convolución, 5 de agrupación y una totalmente conectada. AlexNet conocida por ganar la competencia 2012 ImageNet LSVRC-2012 por un amplio margen, contiene 5 capas convolucionales, 3 capas de agrupación y 3 capas completamente conectadas.

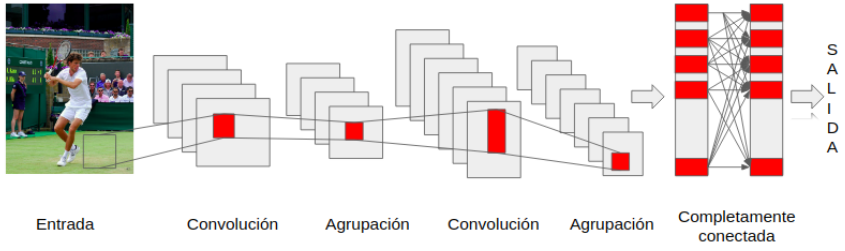


Figura 2.1: Esta imagen muestra una arquitectura simplificada de una rede neuronal convolucional.

Las redes CNN se a utilizado para resolver distintos problemas como, la detección de objetos, Fast R-CNN [4]. La comprensión visual de escenas de calles urbanas [3]. En este trabajo utilizamos la salida de las redes CNN (la capa completamente conectada), como un vector de características visual de la imagen. Debido a que son muy eficaces reconociendo patrones, los vectores de dos imágenes que tienen un aspecto similar, también tienden a tener una semejanza.

2.2. Word embedding

Al igual que con las imágenes, que utilizamos las redes CNN, para obtener un vector que represente a la misma, es necesario un procedimiento para poder representar las palabras, con algún objeto matemático. Hay muchas formas de representar palabras, pero la mas conocida es word embedding, es una técnica de aprendizaje en el campo de procesamiento del lenguaje natural (PLN). Es capas de capturar el contexto de una palabra en un documento, calcular similitud semántica y sintáctica con otras palabras, etc. Word2Vec [7] es una de la implementación mas conocida. Fue desarrollado por Tomas Mikolov en 2013.
El objetivo, es que las palabras con un contexto similar ocupen po-

siciones espaciales cercanas, mientras que palabras que no tienen un contexto similar estén espaciadas. Para lograr esto, se introduce cierta dependencia de una palabra de las otras palabras. Se utilizan texto para entran estos modelos, asi las palabras en el contexto de una palabra especifica, obtendrían una mayor proporción de esta dependencia. En este trabajo, aprovechamos la capacidad de capturar similitudes semántica que tiene word embedding, para relacionar las clases vistas con las clases no vistas. Utilizamos un modelo pre-entrenado generado a partir de word2vec para representar las palabras de las distintas clases.

2.3. Propuestas de objetos

En problemas de detección de objetos, generalmente tenemos que encontrar todos los objetos posibles en la imagen como todos los autos todas las bicicletas, etc. La localización de objetos se refiere a identificar la ubicación de uno o varios objetos en la imagen. Un algoritmo de localización de objetos generará las coordenadas de la ubicación de los objetos con respecto a la imagen. En visión artificial, la forma más popular de representar la ubicación de los objetos es con la ayuda de cuadros delimitadores (Bounding Boxs). Existen muchos algoritmos y redes que intenta resolver este problema como por ejemplo, ventana deslizante, Edge-Boxes [18], Selective search [14] ect. En ZSD las propuestas de objetos cumple un papel importante, ya que se necesita extraer todas las instancias de los objetos, pero también tiene que discriminar fondos como cielo, fondo de ciudad, veredas, ect. Es muy difícil encontrar un equilibrio ya que un algoritmo poco “permisivo” ignorara muchas instancias de objetos y por el otro extremo, se incluirá fondos y de esta manera introducir ruido en nuestro modelo. En este proyecto se usa Edge-Boxes, ya que este genera una cantidad significativamente menor a algoritmos del estilo de ventana deslizante. Aun asi, procesar todas estas propuestas es engorroso. Esto da lugar a una técnica que filtra las propuestas, denominada Supresión

no máxima (NMS) 2.2b. Los criterios de selección de NMS se pueden elegir para llegar a resultados particulares. El criterio mas común es Intersección sobre Unión (IoU), en la imagen 2.2a se muestra como se calcula el IoU sobre dos Bounding Boxs.

(a) Iou (b) Supresión no máxima

Figura 2.2: (a)Esta imagen muestra como se calcula el criterio Intersección sobre Unión. (b) Se muestra la salida de la propuesta de objetos y el resultado después de NMS.

2.4. Multimodales

Nuestra experiencia del mundo es multimodal, vemos objetos, escuchamos sonidos, sentimos la textura, olemos los olores y probamos los sabores. La modalidad se refiere a la forma en que algo sucede o se experimenta y un problema de investigación se caracteriza como multimodal cuando incluye múltiples modalidades. Para que la inteligencia artificial avance en la comprensión del mundo que nos rodea, necesita poder interpretar y relacionar estas señales multimodales. Aunque la combinación de diferentes modalidades o tipos de información para mejorar el rendimiento parece una tarea intuitivamente atractiva, en la práctica, es un desafío combinar el nivel variable de ruido y los conflictos entre las modalidades. Además, las modalidades tienen una influencia cuantitativa diferente sobre el resultado de la predicción. La idea general es, partir de dos objetos matemáticos distintos uno de cada modal y poder transformar a ambos para lograr que pertenezcan a un tercer objeto que es la representación multimodal. Las imágenes suelen estar asociadas con etiquetas y explicaciones de texto. En este

trabajo nos aprovechamos esto y tratamos de encontrar un espacio comun entre el vector que representan a la imagen del objeto y el que representa la sintaxis del mismo.

2.5. Detección de objeto por disparo cero (ZSD)

Capítulo 3

Diseño y Arquitectura

Capítulo 4

Experimentos

4.1. Análisis de resultados

Capítulo 5

Conclusiones y trabajo futuro

5.1. Aportes

5.2. Trabajo futuro

Bibliografía

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

[2] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018.

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[4] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[5] David Marr. Vision: A computational investigation into the human representation and processing of visual information. New York, NY: W.H. Freeman and Company, 1982.

[6] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[8] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[9] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *Asian Conference on Computer Vision*, pages 547–563. Springer, 2018.

[10] Lawrence G Roberts. *Machine perception of three-dimensional solids*. PhD thesis, Massachusetts Institute of Technology, 1963.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[13] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[14] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[15] Yongqin Xian, Christoph Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 07 2017.

[16] Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, and Honglak Lee. Discriminative bimodal networks for visual localization and detection with natural language queries. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 557–566, 2017.

[17] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Zero shot detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[18] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.