



# ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

GRADO EN INGENIERÍA EN INGENIERÍA EN  
SISTEMAS AUDIOVISUALES Y MULTIMEDIA

## **TRABAJO FIN DE GRADO**

DEEP LEARNING EN SENSORES RGBD

Autor: David Butragueño Palomar

Tutor:

Curso académico 2017/2018



# Resumen

# Agradecimientos

# Índice general

<b>1. Introducción</b>	<b>1</b>
1.1. Contexto y motivación . . . . .	1
<b>2. Infraestructura</b>	<b>2</b>
2.1. Caffe . . . . .	2
2.1.1. Línea de comandos . . . . .	2
2.1.2. Python . . . . .	3
2.1.3. Capas de una red neuronal . . . . .	3
2.2. JdeRobot . . . . .	9
2.2.1. CameraServer y CameraView . . . . .	10
2.3. Bases de datos . . . . .	10
2.3.1. PascalVOC . . . . .	10
2.3.2. COCO . . . . .	12
<b>3. Detección</b>	<b>16</b>
3.1. Técnica SSD . . . . .	16
3.1.1. Arquitectura . . . . .	17
3.1.2. MultiBox . . . . .	19
3.1.3. Entrenamiento . . . . .	19
3.2. Detector de objetos . . . . .	20
3.2.1. Estructura de la red . . . . .	20
3.2.2. Definición del solucionador . . . . .	21
3.2.3. Entrenamiento de la red . . . . .	22
3.3. SSD en Caffe . . . . .	22
3.4. Comparación <i>bounding boxes</i> real y detectada . . . . .	24
3.4.1. VOC . . . . .	24
3.4.2. COCO . . . . .	24
3.5. Medidor de calidad . . . . .	24
3.5.1. Índice Jaccard . . . . .	24
3.5.2. Cálculo del índice Jaccard . . . . .	26
3.5.3. Resultado final . . . . .	26

3.6. Componente Python . . . . .	26
<b>4. Bibliografía</b>	<b>27</b>

# Índice de figuras

2.1.	Ejemplo convolución con tamaño del núcleo 2x2 . . . . .	5
2.2.	Pooling con función de agrupación del máximo . . . . .	5
2.3.	Función de activación ReLu . . . . .	7
2.4.	Función de activación Sigmoide . . . . .	8
2.5.	Distribución de objetos en PascalVOC . . . . .	11
2.6.	Estructura básica de anotaciones en COCO . . . . .	13
2.7.	Estructura básica de anotaciones en COCO . . . . .	14
2.8.	Distribución de objetos en COCO . . . . .	15
3.1.	Framework SSD . . . . .	17
3.2.	Arquitectura SSD . . . . .	18
3.3.	Arquitectura VGG-16 . . . . .	19
3.4.	Intersección y unión sobre 2 conjuntos A y B . . . . .	25
3.5.	Intersection over Union . . . . .	25
3.6.	Evaluación IoU . . . . .	26

# Capítulo 1

## Introducción

### 1.1. Contexto y motivación



# Capítulo 2

## Infraestructura

### 2.1. Caffe

Caffe es un framework de aprendizaje profundo desarrollado por Berkeley AI Research (BAIR) y por contribuyentes de la comunidad. Fue creado por Yangqing Jia durante su doctorado en UC Berkeley. Caffe está publicado bajo la licencia BSD 2-Clause.

#### 2.1.1. Línea de comandos

Caffe dispone de una interfaz de línea de comandos llamada *cmdcaffe* la cual es la herramienta utilizada por Caffe para el entrenamiento del modelo y el diagnóstico del mismo. Los principales comandos que se pueden ejecutar son:

- **Entrenamiento:** Con el comando *caffe train* es posible aprender modelos desde cero,
- **Test:** El comando *caffe test* puntúa los modelos ejecutándolos en la fase de test e informa de la salida de la red como su puntuación. En primer lugar se informa la puntuación por lotes de datos de entrada y finalmente el promedio general.
- **Comparación:** El comando *caffe time* compara el modelo capa a capa. Esto es útil para comprobar el rendimiento del sistema y medir los tiempos de ejecución relativos a los modelos.

### 2.1.2. Python

La interfaz de Python *pycaffe* contiene el módulo de Caffe y sus propios scripts en la ruta *caffe/python*. Con el comando *import caffe* se importará esta interfaz, pudiendo así cargar diferentes modelos de Caffe, manejar instrucciones de entrada/salida, visualizar redes y numerosas funcionalidades más. Todos los datos y parámetros se encuentran disponibles tanto para lectura como para escritura. Algunas de las tareas que se pueden realizar con esta interfaz son:

- **caffe.Net** es la interfaz central para cargar, configurar y ejecutar modelos.
- **caffe.Classifier** y **caffe.Detector** proporcionan interfaces para tareas de clasificación y detección.
- **caffe.SGDSolver** se trata de la interfaz de resolución.
- **caffe.io** maneja funciones de entrada/salida con preprocesamiento.
- **caffe.draw** visualiza las arquitecturas de red.

### 2.1.3. Capas de una red neuronal

Para crear un modelo de Caffe es necesario definir la arquitectura del mismo utilizando para ello un archivo de definición de buffer de protocolo (prototxt).

#### Capas de datos

Los datos entran en Caffe a través de las capas de datos las cuáles se encuentran en la parte inferior de las redes. Estos datos pueden provenir de bases de datos (LevelDB o LMDB), directamente de la memoria, o, cuando la eficiencia no es crítica, desde archivos en disco en formato HDF5 o formatos de imagen comunes.

Tareas comunes de preprocesamiento de los datos de entrada, tales como escalado o reflejo, están disponibles especificando *TransformationParameters* por algunas de las capas. Los tipos de capas "bias", "scalez crop" pueden ser útiles para el preprocesamiento de la entrada cuando la opción *TransformationParameters* no está disponible.

- **Image Data:** Lee imágenes sin procesar.
- **Database:** Lee los datos de LevelDB o LMDB.
- **HDF5 Input:** Lee los datos en formato HDF5 permitiendo que estos tengan dimensiones arbitrarias.
- **HDF5 Output:** Escribe datos en formato HDF5
- **Input:** Normalmente utilizada para redes que se están implementando.
- **Window Data:**
- **Memory Data:** Lee archivos directamente desde memoria.
- **Dummy Data:** Utilizado para datos estáticos.

### Capas de visión

Las capas de visión, generalmente toman imágenes como datos de entradas y generan otras imágenes como salida aunque también pueden tomar datos de otros tipos y dimensiones. Una imagen puede tener un canal ( $c = 1$ ) si se trata de una imagen en escala de grises o 3 canales ( $c = 3$ ) si se trata de una imagen RGB. Pero en este contexto, las características distintivas para el tratamiento de las imágenes de entrada serán la altura y la anchura de las mismas. La mayoría de las capas de visión trabajan aplicando una operación particular sobre alguna región de la entrada para producir una región correspondiente a la salida.

- **Convolution Layer:** Convoluciona la imagen de entrada con un conjunto de filtros. Este proceso transforma la matriz de píxeles de la imagen original en otra matriz de características. Esta nueva imagen es esencialmente la original pero cada píxel tiene mucha más información ya que contiene información de la región en la que se encuentra el píxel, no sólo del píxel aislado.

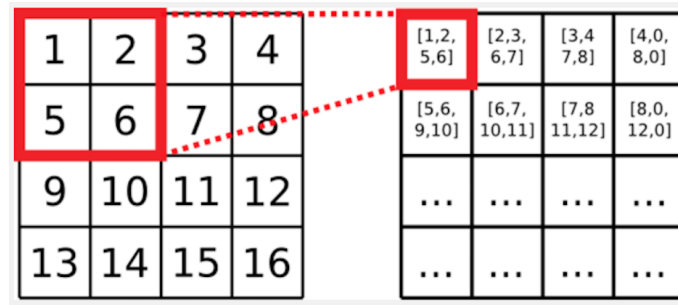


Figura 2.1: Ejemplo convolución con tamaño del núcleo 2x2

- **Pooling Layer:** Realiza *pooling* de los datos de entrada utilizando para ello funciones de máximo, media o estocásticas. La capa de pooling se coloca generalmente después de la capa de convolución. Su utilidad principal radica en la reducción espacial de la imagen de entrada. Para ello, se divide el mapa de características obtenido anteriormente en un conjunto de bloques de  $m \times n$ . A continuación, se aplica una función de agrupación para cada uno de los bloques. Tras este proceso, se obtendrá una matriz de características más pequeño. Dentro de las funciones de agrupación, destacan max pooling, el cual elige el valor más alto dentro del bloque y pooling promedio (average) el cual toma como respuesta de bloque el valor promedio de las respuestas del bloque.

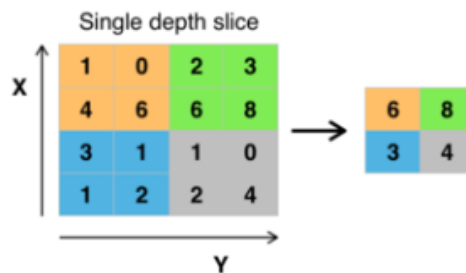


Figura 2.2: Pooling con función de agrupación del máximo

- **Spatial Pyramid Pooling (SPP)**
- **Crop**

- **Deconvolution Layer:** Realiza una convolución transpuesta.
- **Im2Col**

#### Capas recurrentes

- **Recurrent**
- **RRNN**
- **Long-Short Term Memory (LSTM)**

#### Capas comunes

- **Inner Product:** Capa totalmente conectada
- **Dropout**
- **Embed**

#### Capas de pérdida

Estas capas de pérdida conducen al aprendizaje comparando la salida obtenida con el valor de la entrada asignado así un coste para minimizarla.

- **Multinomial Logistic Loss**
- **Infogain Loss**
- **Softmax with Loss**
- **Sum-of-Squares / Euclidean**
- **Hinge / Margin**
- **Sigmoid Cross-Entropy Loss**
- **Accuracy / Top-k layer**
- **Contrastive Loss**

### Capas de normalización

- **Local Response Normalization (LRN):** Normaliza regiones locales de los datos de entrada.
- **Mean Variance Normalization (MVN):** Realiza una normalización de contraste / normalización de instancia.
- **Batch Normalization:** Realiza normalizaciones sobre pequeños lotes de datos de entrada.

### Capas de activación

En general, estas capas son operados que toman un dato de la salida de la capa anterior y generan datos con las mismas dimensiones.

- **ReLU / Rectified-Linear and Leaky-ReLU:** Se trata de una función lineal, rectilínea con pendiente uniforme.

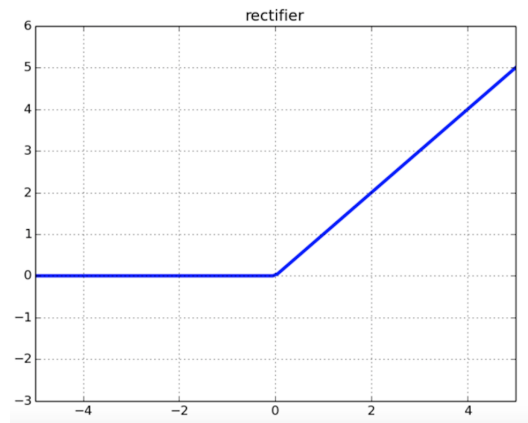


Figura 2.3: Función de activación ReLu

Se puede definir a partir de la siguiente ecuación:

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ x & \text{si } x \geq 0 \end{cases} \quad (2.1)$$

- PReLU
- ELU
- Sigmoid

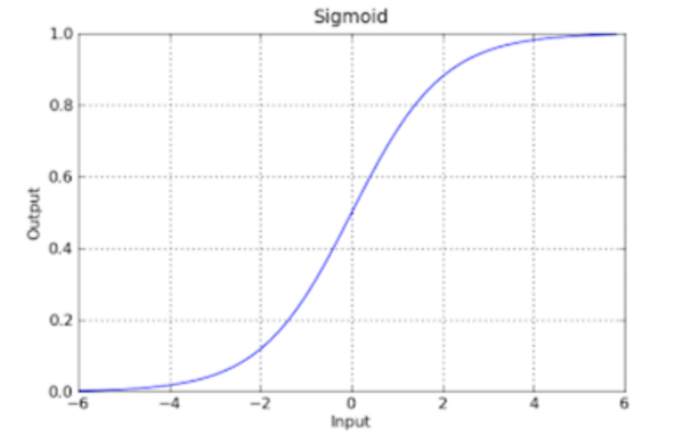


Figura 2.4: Función de activación Sigmoide

Se puede definir a partir de la siguiente ecuación:

$$f(x) = \frac{1}{1 + e^{-x}}$$

- TanH
- Absolute Value
- Power

$$f(x) = (shift + scale * x)^{power} \quad (2.2)$$

- Exp

$$f(x) = base^{(shift + scale * x)} \quad (2.3)$$

- Log

$$f(x) = \log(x) \quad (2.4)$$

- BNLL

$$f(x) = \log(1 + \exp(x)) \quad (2.5)$$

- **Threshold:** Realiza la función de paso en el umbral definido por el usuario.
- **Bias**
- **Scale**

## 2.2. JdeRobot

Se trata de un framework cuyo objetivo es desarrollar aplicaciones en robótica y visión por computadora. También tiene actuación en domótica y en escenarios con sensores, accionadores y software inteligente. Ha sido desarrollado para ayudar en la programación de este software inteligente. Está escrito principalmente utilizando el lenguaje C++ proporcionando un entorno de programación en el que el programa de aplicación está compuesto por una colección de varios componentes asíncronos concurrentes. Estos componentes pueden ejecutarse en diferentes equipos y están conectados mediante el middleware de comunicaciones ICE. Los componentes pueden estar escritos en C++, Python, Java y todos ellos interactúan a través de interfaces ICE explícitas.

JdeRobot simplifica el acceso a dispositivos hardware desde el programa de control. Obtener mediciones de sensores es tan simple como llamar a una función local y ordenar comandos de motor tan fácil como llamar a otra función local. La plataforma adjunta esas llamadas a invocaciones remotas sobre los componentes conectados al sensor o los dispositivos de accionamiento. También, pueden conectarse a sensores y activadores reales o simulados, tanto a nivel local como remoto utilizando para ello la red. Esas funciones construyen la API para la capa de abstracción del hardware. La aplicación robótica obtiene las lecturas del sensor y ordena los comandos del actuador usando esa API para desplegar su comportamiento. Se han desarrollado varios drivers para soportar diferentes sensores, activadores y simuladores. Los robots y sensores actualmente soportados son:



- **Sensores RGBD:** Kinect and Kinect2 de Microsoft, Asus Xtion
- **Robots con ruedas:** Kobuki (TurtleBot) de Yujin Robot y Pioneer de MobileRobotics Inc.
- **ArDrone quadrotor de Parrot**
- **Escáneres laser:** LMS de SICK, URG de Hokuyo y RPLidar
- **Simulador Gazebo**
- **Cámaras Firewire, cámaras USB, archivos de vídeo (mpeg, avi), cámaras IP (como Axis)**

JdeRobot incluye varias herramientas de programación de robots y bibliotecas. En primer lugar, teleespectadores y teleoperadores para varios robots y sus sensores y motores. En segundo lugar, un componente de calibración de cámara y una herramienta de tuning para filtros de color. En tercer lugar, una herramienta llamada VisualHFSM para la programación del comportamiento del robot utilizando la jerarquía Finite State Machines. Además, también proporciona una biblioteca para desarrollar controladores difusos y otra para la geometría proyectiva y el procesamiento de la visión por computadora.

Cada componente puede tener su propia interfaz gráfica de usuario o ninguna en absoluto. Actualmente, las bibliotecas GTK y Qt son compatibles, incluyéndose varios ejemplos de OpenGL para gráficos 3D con ambas bibliotecas.

JdeRobot es un software de código abierto con licencia como GPL y LGPL. También utiliza software de terceros como el simulador Gazebo, ROS, OpenGL, GTK, Qt, Player, Stage, GSL, OpenCV, PCL, Eigen u Ogre.

### 2.2.1. CameraServer y CameraView

## 2.3. Bases de datos

### 2.3.1. PascalVOC

Esta base de datos contiene **11.530** imágenes de entrenamiento y validación que representan **27.450** objetos diferentes distribuidos en **20 clases**. Los datos de entrenamiento proporcionados consisten en un conjunto de imágenes; cada imagen tiene un archivo de anotación que proporciona un cuadro

delimitador o *bounding box* y una etiqueta de las 20 clases para cada objeto presente en la imagen. Por lo tanto, la tarea de detección consiste en predecir el cuadro delimitador y la etiqueta de cada objeto en la imagen de prueba.

Es importante saber cuántas imágenes aparecen en cada uno de los 20 objetos para saber si la base de datos está bien escalada, o si, por el contrario, algunos objetos aparecen con más frecuencia que otros. En la base de datos VOC2012, las distribuciones de imágenes y objetos por clase son aproximadamente iguales en todos los conjuntos de entrenamiento / validación y prueba. Específicamente, la distribución de objetos para la tarea de detección se muestra en la siguiente imagen.

	train		val		trainval		test	
	Images	Objects	Images	Objects	Images	Objects	Images	Objects
<b>Aeroplane</b>	327	432	343	433	670	865	-	-
<b>Bicycle</b>	268	353	284	358	552	711	-	-
<b>Bird</b>	395	560	370	559	765	1119	-	-
<b>Boat</b>	260	426	248	424	508	850	-	-
<b>Bottle</b>	365	629	341	630	706	1259	-	-
<b>Bus</b>	213	292	208	301	421	593	-	-
<b>Car</b>	590	1013	571	1004	1161	2017	-	-
<b>Cat</b>	539	605	541	612	1080	1217	-	-
<b>Chair</b>	566	1178	553	1176	1119	2354	-	-
<b>Cow</b>	151	290	152	298	303	588	-	-
<b>Diningtable</b>	269	304	269	305	538	609	-	-
<b>Dog</b>	632	756	654	759	1286	1515	-	-
<b>Horse</b>	237	350	245	360	482	710	-	-
<b>Motorbike</b>	265	357	261	356	526	713	-	-
<b>Person</b>	1994	4194	2093	4372	4087	8566	-	-
<b>Pottedplant</b>	269	484	258	489	527	973	-	-
<b>Sheep</b>	171	400	154	413	325	813	-	-
<b>Sofa</b>	257	281	250	285	507	566	-	-
<b>Train</b>	273	313	271	315	544	628	-	-
<b>Tvmonitor</b>	290	392	285	392	575	784	-	-
<b>Total</b>	5717	13609	5823	13841	11540	27450	-	-

Figura 2.5: Distribución de objetos en PascalVOC

En cuanto a los archivos de anotaciones, estos tienen un formato XML. En primer lugar, aparecen algunas etiquetas informativas acerca de la imagen, como la etiqueta *filename* que indica el nombre de la imagen, o la etiqueta *database* que indica el nombre de la base de datos.

```
<folder>VOC2012</folder>
<filename>2007_009084.jpg</filename>
<source>
  <database>The VOC2007 Database</database>
  <annotation>PASCAL VOC2007</annotation>
  <image>flickr</image>
</source>
```

Después, aparece el grupo de etiquetas *size*, donde se indican el ancho de la imagen, *width*, el alto, *height* y la profundidad, *depth*.

```
<size>
  <width>500</width>
  <height>375</height>
  <depth>3</depth>
</size>
```

Finalmente, aparecen las etiquetas referentes a cada uno de los objetos existentes en la imagen. Destacan la etiqueta *name*, la cual indica el nombre de la etiqueta del objeto, y el grupo *bndbox*, que representa la *bounding box* a partir de las coordenadas *x\_min*, *y\_min*, *x\_max* y *y\_max*.

```
<object>
  <name>motorbike</name>
  <pose>Right</pose>
  <truncated>0</truncated>
  <difficult>0</difficult>
  <bndbox>
    <xmin>43</xmin>
    <ymin>77</ymin>
    <xmax>481</xmax>
    <ymax>375</ymax>
  </bndbox>
</object>
```

### 2.3.2. COCO

COCO es un conjunto de datos creado para la detección y segmentación de objetos y para generación de subtítulos a gran escala. Algunas de las características de esta base de datos son:

- Más de 300.000 imágenes
- 1.5 millones de instancias de objetos
- 80 categorías de objetos

Actualmente, COCO utiliza 3 tipos de anotaciones: instancias de objetos, puntos claves de objetos y leyendas de imágenes. Las anotaciones se almacenan usando el formato de archivo JSON. Todas las anotaciones comparten la estructura de datos básica definida a continuación:

```
{
  "info"          : info,
  "images"        : [image],
  "annotations"   : [annotation],
  "licenses"      : [license],
}

info{
  "year"          : int,
  "version"       : str,
  "description"   : str,
  "contributor"   : str,
  "url"           : str,
  "date_created"  : datetime,
}

image{
  "id"            : int,
  "width"         : int,
  "height"        : int,
  "file_name"     : str,
  "license"       : int,
  "flickr_url"    : str,
  "coco_url"      : str,
  "date_captured" : datetime,
}

license{
  "id"            : int,
  "name"          : str,
  "url"           : str,
}
```

Figura 2.6: Estructura básica de anotaciones en COCO

Para la tarea de detección, es de especial interés la anotación usando instancias de objetos. Cada anotación de instancia contiene una serie de campos, incluida la identificación de categoría y la máscara de segmentación del objeto. El formato de segmentación depende de si la instancia representa un solo objeto ( $\text{iscrowd} = 0$  en cuyo caso se usan polígonos) o una colección de objetos ( $\text{iscrowd} = 1$  en cuyo caso se usa RLE). Hay que tener en cuenta que un solo objeto ( $\text{iscrowd} = 0$ ) puede requerir múltiples polígonos. Las anotaciones de multitudes ( $\text{iscrowd} = 1$ ) se utilizan para etiquetar grandes grupos de objetos (por ejemplo, una multitud de personas). Además, se proporciona un cuadro delimitador para cada objeto (las coordenadas del cuadro se miden desde la esquina superior izquierda de la imagen y están indexadas en 0).

Finalmente, el campo de categorías de la estructura de anotación almacena la asignación de los nombres de categoría y supercategoría.

```
annotation{
  "id"           : int,
  "image_id"     : int,
  "category_id"  : int,
  "segmentation" : RLE or [polygon],
  "area"         : float,
  "bbox"         : [x,y,width,height],
  "iscrowd"      : 0 or 1,
}

categories[
  {
    "id"       : int,
    "name"     : str,
    "supercategory" : str,
  }
]
```

Figura 2.7: Estructura básica de anotaciones en COCO

En referencia a los datos, el conjunto de datos MS COCO se divide en dos partes aproximadamente iguales. La primera mitad del conjunto de datos se lanzó en 2014, mientras que la segunda mitad se lanzó en 2015. La versión 2014 contiene **82.783** imágenes para el entrenamiento, **40.504** para validaciones y **40.775** imágenes de prueba (aproximadamente 1/2 de imágenes de entrenamiento, 1/4 de validación y 1/4 de prueba). Hay casi 270 mil personas segmentadas y un total de 886 mil instancias de objetos segmentados en los datos de entrenamiento y validación. La versión acumulada de 2015 contendrá un total de **165.482** imágenes de entrenamiento, **81.208** de validación y **81.434** de prueba.

La distribución de los objetos en esta base de datos se puede obtener desde su sitio web. En la sección *Explorar*, es posible elegir y combinar cada uno de los objetos y observar cuántas imágenes aparecen. La distribución de cada uno de los objetos en el conjunto de entrenamiento / validación se muestra en la siguiente imagen:

	images		images		images		images
person	66808	cat	4298	wine glass	2643	dinning table	12338
backpack	5756	dog	4562	cup	9579	toilet	3502
umbrella	4142	horse	3069	fork	3710	tv	4768
handbag	7133	sheep	1594	knife	4507	laptop	3707
tie	3955	cow	2055	spoon	3682	mouse	1964
suitcase	2507	elephant	2232	bowl	7425	remote	3221
bicycle	3401	bear	1009	banana	2346	keyboard	2221
car	12786	zebra	2001	apple	1662	cell phone	5017
motorcycle	3661	giraffe	2647	sandwich	2463	microwave	1601
airplane	3083	frisbee	2268	orange	1784	oven	2992
bus	4141	skis	3202	broccoli	2010	toaster	225
train	3745	snowboard	1703	carrot	1764	sink	4865
truck	6377	sports ball	4431	hot dog	1273	refrigerator	2461
boat	3146	kite	2352	pizza	3319	book	5562
traffic light	4330	baseball bat	2603	donut	1585	clock	4863
fire hydrant	1797	baseball glove	2729	cake	3049	vase	3730
stop sign	1803	skateboard	3603	chair	13354	scissors	975
parking meter	742	surfboard	3635	couch	4618	teddy bear	2234
bench	5805	tennis racket	3561	potted plant	4624	hair drier	198
bird	3362	bottle	8880	bed	3831	toothbrush	1041

Figura 2.8: Distribución de objetos en COCO

# Capítulo 3

## Detección

### 3.1. Técnica SSD

Single Shot MultiBox Detector (SSD) es un framework diseñado para detectar objetos en imágenes utilizando para ello únicamente una red neuronal profunda. Fue lanzado a finales de noviembre de 2016 y obtuvo resultados muy positivos en términos de rendimiento y precisión en tareas de detección de objetos obteniendo más del 74 % en *mAP* (mean Average Precision) a 59 *frames* por segundo en conjuntos de datos estándar como PascalVOC y COCO. Para una mejor comprensión de esta técnica, se puede comenzar con definir el significado de su nombre:

- **Single Shot:** Hace referencia a que la tareas de localización y clasificación de objetos se realizan en un único paso hacia adelante de la red.
- **MultiBox:** Es el nombre de una técnica para regresión de *bounding box* desarrollada por Szegedy.
- **Detector:** La red neuronal se trata de un detector de objetos que también clasifica esos objetos detectados.

En el momento de la predicción, la red neuronal genera las puntuaciones para cada categoría de objetos en cada cuadro predeterminado y produce ajustes en este cuadro para ajustarse mejor a la forma del objeto. Además, la red combina predicciones de múltiples mapas de características con diferentes resoluciones para manejar de forma natural objetos de varios tamaños.

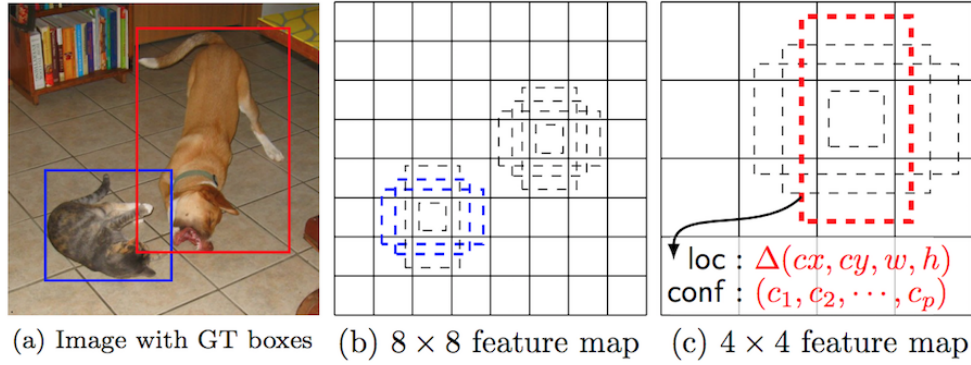


Figura 3.1: Framework SSD

(a) SSD solo necesita una imagen de entrada y los cuadros delimitadores reales para cada objeto durante el entrenamiento. De una manera convolucional, se evalúa un pequeño conjunto de cuadros predeterminados (p.ej. 4) de diferentes relaciones de aspecto para cada ubicación de éstos en varios mapas de características a diferentes escalas (p. ej  $8 \times 8$  y  $4 \times 4$  en (b) y (c)). Para cada cuadro predeterminado se predicen tanto los desplazamiento de forma como la confianza para todas las categorías de objetos  $((c_1, c_2, \dots, c_p))$ . En el momento del entrenamiento, se comparan estos cuadros predeterminados con los cuadros reales.

### 3.1.1. Arquitectura

El modelo SSD se basa en una red convolucional que produce un conjunto de recuadros delimitadores o *bounding boxes* de tamaño fijo y puntuaciones para la presencia de instancias de clases de objetos dentro de esos cuadros, seguido por un paso de supresión no máxima para producir las detecciones finales. Las primeras capas de la red se basan en una arquitectura VGG-16 pero descartando las capas totalmente conectadas de ésta. La razón por la que se usa VGG-16 es su gran desempeño en la clasificación de imágenes de alta calidad. En lugar de las capas totalmente conectadas de VGG, se agregaron un conjunto de capas convolucionales auxiliares, lo que permite extraer características en múltiples escalas y disminuir progresivamente el tamaño de la imagen de entrada. Posteriormente, se agrega una estructura auxiliar a la red para producir detecciones. Esta estructura tiene las siguientes características claves:

- **Mapas de características con múltiples escalas para la detección:** Se agregan varias capas de convolucionales al final de la estructu-



ra estándar comentada anteriormente. Estas capas disminuyen de forma progresiva el tamaño de la imagen de entrada y permiten predecir detecciones a múltiples escalas.

- **Predictores convolucionales para la detección:** Cada capa de características añadida puede producir un conjunto fijo de predicciones de detección utilizando para ello un conjunto de filtros convolucionales. Estos filtros están indicados en la parte superior de la arquitectura de la red SSD en la figura 3.2. Para una capa de características de tamaño  $m \times n$  con  $p$  canales, el elemento básico para predecir los parámetros de una posible detección es un núcleo pequeño de  $3 \times 3 \times p$  el cuál produce una puntuación para una categoría determinada. En cada una de las posiciones por donde se aplica el filtro, éste produce un valor de salida.

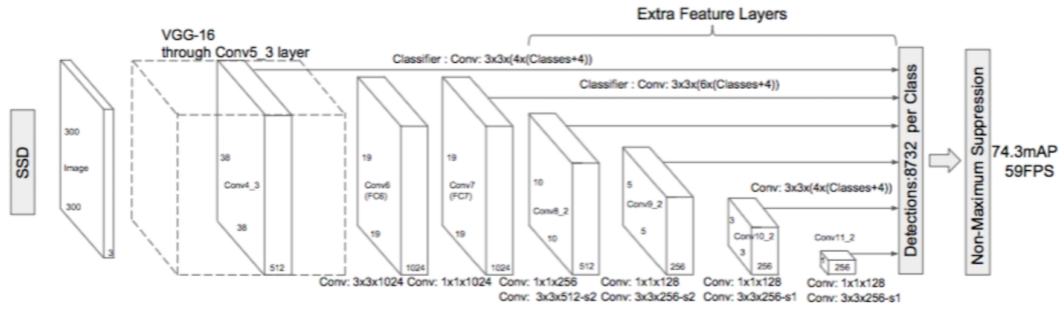


Figura 3.2: Arquitectura SSD

- **Cajas predeterminadas y relaciones de aspecto:** Se asocian un conjunto de cuadros delimitadores por defecto con cada celda del mapa de características. Los cuadros predeterminados marcan el mapa de características de un manera convolucional, por lo que la posición de cada cuadro con respecto a su celda es fija. En cada celda del mapa de características, se predicen los desplazamientos relativos a las formas de los cuadros predeterminados en cada una de las celdas, así como la puntuación por clase que indica la presencia de una instancia de clase en cada uno de los cuadros delimitadores. Específicamente, para cada cuadro  $k$  en una posición determinada se calcula la puntuación de clase  $c$  y los 4 desplazamientos relativos a la forma del cuadro por defecto original. Como resultado se obtienen  $(c + 4)k$  filtros que se aplican alrededor de cada ubicación en el mapa de características, produciendo

$(c + 4)kmn$  de salidas para un mapa de  $m \times n$ . Se permiten diferentes formas para los cuadros predeterminados en los mapas de características con el objetivo de discretizar el espacio para las posibles formas que tengan los cuadros delimitadores de salida.

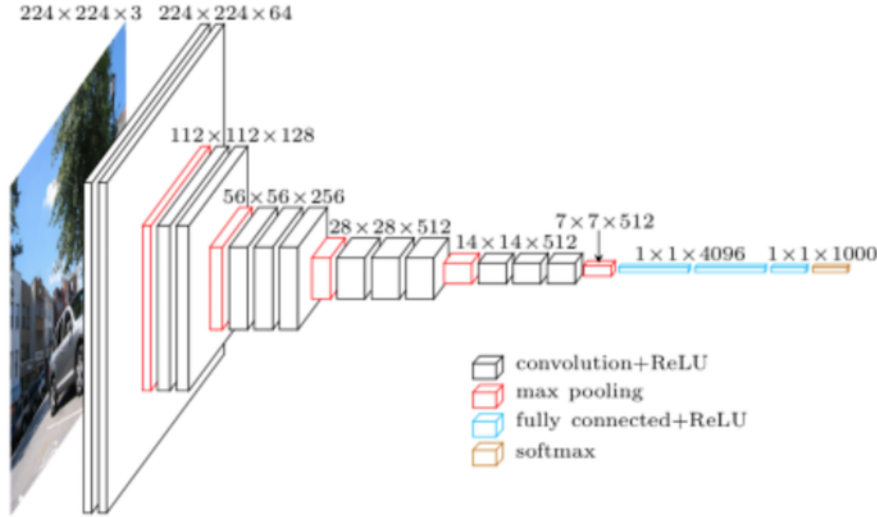


Figura 3.3: Arquitectura VGG-16

### 3.1.2. MultiBox

### 3.1.3. Entrenamiento

La principal diferencia entre el entrenamiento usando la técnica SSD y el entrenamiento de un detector típico, es que la información a cerca de los cuadros delimitadores reales debe asignarse a salidas específicas dentro del conjunto fijo de salidas del detector. Tras realizar esta asignación, la función de pérdida y la propagación hacia atrás son aplicadas de extremo a extremo. El proceso de entrenamiento también implica elegir el conjunto de cuadros y escalas predeterminados para la detección.

- **Har Negative Mining:**
- **Aumento de datos:** El aumento de datos, tanto en la técnica SSD como en muchas otras aplicaciones de aprendizaje profundo, ha sido crucial para que la red neuronal sea mucho más robusta frente a diferentes tamaños y formas de objetos a la entrada de ésta. Con este fin,

se generaron ejemplos de entrenamiento con parches de la imagen original con diferentes relaciones de IoU (p.ej. 0.1, 0.3, 0.5, etc). Además, cada imagen se gira horizontalmente de forma aleatoria con una probabilidad de 0.5, asegurando así que los objetos aparezcan con la misma probabilidad a la izquierda que a la derecha.

## 3.2. Detector de objetos

### 3.2.1. Estructura de la red

```
name: "VGG_VOC0712_SSD_300x300_deploy"
```

```
layer {
  name: "conv1_1"
  type: "Convolution"
  bottom: "data"
  top: "conv1_1"
  param {
    lr_mult: 1.0
    decay_mult: 1.0
  }
  param {
    lr_mult: 2.0
    decay_mult: 0.0
  }
  convolution_param {
    num_output: 64
    pad: 1
    kernel_size: 3
    weight_filler {
      type: "xavier"
    }
    bias_filler {
      type: "constant"
      value: 0.0
    }
  }
}
```

```
layer {
  name: "relu1_1"
  type: "ReLU"
```

```
    bottom: "conv1_1"
    top: "conv1_1"
}
```

```
layer {
  name: "pool1"
  type: "Pooling"
  bottom: "conv1_2"
  top: "pool1"
  pooling_param {
    pool: MAX
    kernel_size: 2
    stride: 2
  }
}
```

### 3.2.2. Definición del solucionador

```
train_net: "models/VGGNet/VOC0712/SSD_300x300/train.prototxt"
test_net: "models/VGGNet/VOC0712/SSD_300x300/test.prototxt"
test_iter: 619
test_interval: 10000
base_lr: 0.00100000000475
display: 10
max_iter: 120000
lr_policy: "multistep"
gamma: 0.10000000149
momentum: 0.899999976158
weight_decay: 0.0005000000023749
snapshot: 80000
snapshot_prefix: "models/VGGNet/VOC0712/SSD_300x300
                  /VGG_VOC0712_SSD_300x300"
solver_mode: CPU
device_id: 0
debug_info: false
snapshot_after_train: true
test_initialization: false
average_loss: 10
stepvalue: 80000
stepvalue: 100000
stepvalue: 120000
iter_size: 1
```

```
type: "SGD"
eval_type: "detection"
ap_version: "11point"
```

### 3.2.3. Entrenamiento de la red

## 3.3. SSD en Caffe

En primer lugar, se definen 3 ficheros que tendrán gran importancia a lo largo del proceso de detección:

- **labelmap.prototxt:** Definirá las etiquetas que pueden ser asignadas en la detección.
- **model\_def:** Definirá la estructura básica de la red neuronal.
- **model\_weights:** Se trata de un fichero binario que contiene el estado actual de los pesos para cada capa de la red neuronal.

Posteriormente, se define la función *get\_labelname*, que es la encargada de devolver las etiquetas que se encuentran definidas en el fichero labelmap.prototxt.

```
def get_labelname(self, labelmap, labels):
    num_labels = len(labelmap.item)
    labelnames = []
    if type(labels) is not list:
        labels = [labels]
    for label in labels:
        found = False
        for i in xrange(0, num_labels):
            if label == labelmap.item[i].label:
                found = True
                labelnames.append(labelmap.item[i].display_name)
                break
        assert found == True
    return labelnames
```

Finalmente, se define la función *detection\_test*, la cual toma una imagen como entrada y retorna esta misma imagen con las bounding boxes de los objetos detectados superpuestas.

- Primero, esta función realiza una serie de transformaciones para adaptar la imagen de entrada a la red neuronal entrenada. Para ello, se utiliza la función *Transformer* proporcionada por caffe.
- Después, la imagen transformada se introduce en la red usando el siguiente código.

```
self.net.blobs['data'].data[...] = transformed_image
```

- Después de esto, comienza la parte de la detección. Para ello, se utiliza la siguiente función de Caffe, la cual devuelve las detecciones obtenidas.

```
detections = self.net.forward()['detection_out']
```

- Esta salida es parseada en varios vectores que contienen las etiquetas de los objetos detectados, su confianza y las coordenadas de su *bounding box*.

```
det_label = detections[0,0,:,1]
det_conf = detections[0,0,:,2]
det_xmin = detections[0,0,:,3]
det_ymin = detections[0,0,:,4]
det_xmax = detections[0,0,:,5]
det_ymax = detections[0,0,:,6]
```

- Es importante no seleccionar todas las detecciones obtenidas, ya que algunas de ellas con un valor de confianza bajo pueden no ser tan precisas como se desee. Para ello, se usa un filtro que elige las detecciones con un nivel de confianza mayor a 0.6.

```
# Get detections with confidence higher than 0.6.
for i in range(0, len(det_conf)):
    if (det_conf[i] >= 0.6):
        top_indices.append(i)

top_conf = det_conf[top_indices]
top_label_indices = det_label[top_indices].tolist()
top_labels = self.get_labelname(self.labelmap,
                                top_label_indices)
```

```

top_xmin = det_xmin[top_indices]
top_ymin = det_ymin[top_indices]
top_xmax = det_xmax[top_indices]
top_ymax = det_ymax[top_indices]

```

### 3.4. Comparación *bounding boxes* real y detectada

En el apartado 3.3 se ha explicado como se obtiene la *bounding box* de los objetos detectados. Para obtener la *bounding box* real, hay que recuperar las coordenadas del objeto, las cuales se encuentran en las anotaciones de la imagen. Como se ha explicado en el apartado 2.3, las anotaciones para VOC y COCO tienen formatos de archivos diferentes, para la primera base de datos se trata de ficheros XML y para la segunda ficheros JSON, por lo que se utilizarán 2 scripts diferentes para recorrer el fichero determinado y obtener las coordenadas originales. Tras ello, se podrá observar fácil y gráficamente la diferencia entre la *bounding box* real y la detectada, consiguiendo así una aproximación visual de la calidad del detector.

#### 3.4.1. VOC

#### 3.4.2. COCO

### 3.5. Medidor de calidad

#### 3.5.1. Índice Jaccard

El índice Jaccard, también conocido como **Intersection over Union**, es una estadística utilizada para comparar la similitud y la diversidad de conjuntos de muestras. Este índice mide la similitud entre conjuntos de muestras finitas y se define como el tamaño de la intersección dividido por el tamaño de la unión de los conjuntos de muestras.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \quad (3.1)$$

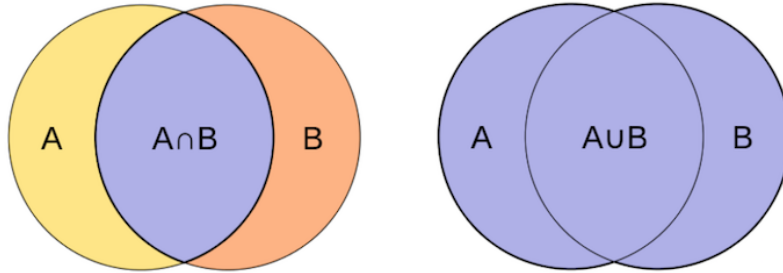


Figura 3.4: Intersección y unión sobre 2 conjuntos A y B

Si no hay coincidencias en ambos conjuntos, el índice Jaccard se definirá como  $J(A,B) = 1$ .

$$0 \leq J(A,B) \leq 1 \quad (3.2)$$

Por lo tanto, el índice Jaccard o **Intersection over Union** (IoU) se puede definir gráficamente de la siguiente manera:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

Figura 3.5: Intersection over Union

Para determinar la calidad de un detector, se utilizarán la *bounding box* real, obtenida de las anotaciones, y la *bounding box* detectada, calculando de esta manera el índice Jaccard. Por lo tanto, podemos evaluar la calidad de un detector a partir del valor obtenido:



- Si el índice Jaccard tiene un valor alrededor de **0.4**, se considerará una calidad de detección escasa.
- Si el índice Jaccard tiene un valor alrededor de **0.7**, se considerará una calidad de detección buena.
- Si el índice Jaccard tiene un valor alrededor de **0.9**, se considerará una calidad de detección excelente.



Figura 3.6: Evaluación IoU

### 3.5.2. Cálculo del índice Jaccard

### 3.5.3. Resultado final

## 3.6. Componente Python

## Capítulo 4

## Bibliografía

Edgar Nelson Sánchez Camperos y Alma Yolanda Alanís García: Redes Neuronales: Conceptos fundamentales y aplicaciones a control automático”