Desafío 11

Detección y clasificación de objetos utilizando YOLOv2 y SSD

Iyán Álvarez

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

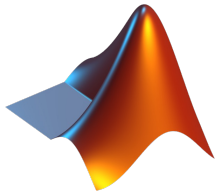
Trabajo realizado

Dos scripts, una función y etiquetado mediante Image Labeler.

El script “YOLOv2_single” entrena un modelo YOLOv2 para la detección de una clase de perro especificado del dataset.

El script “SSD_single” entrena un modelo SSD para la detección de una clase de perro especificado del dataset.

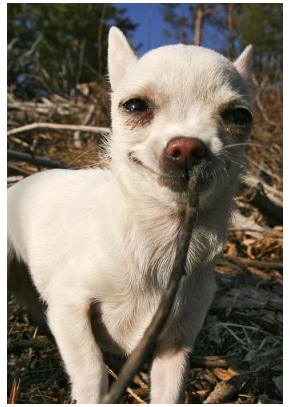
La función “obtener_gTruth_single” lee las anotaciones de una etiqueta almacenadas y las carga en un objeto groundTruth.



Datos

Stanford Dog Dataset es una colección de imágenes de razas de perros ampliamente utilizada en el campo de la visión por computador y el reconocimiento de imágenes. Contiene imágenes de 120 razas de perros diferentes, con un total de alrededor de 20000 imágenes etiquetadas.

La etiqueta "Chihuahua" en el Stanford Dog Dataset se refiere a imágenes que representan esta pequeña y distintiva raza de perro

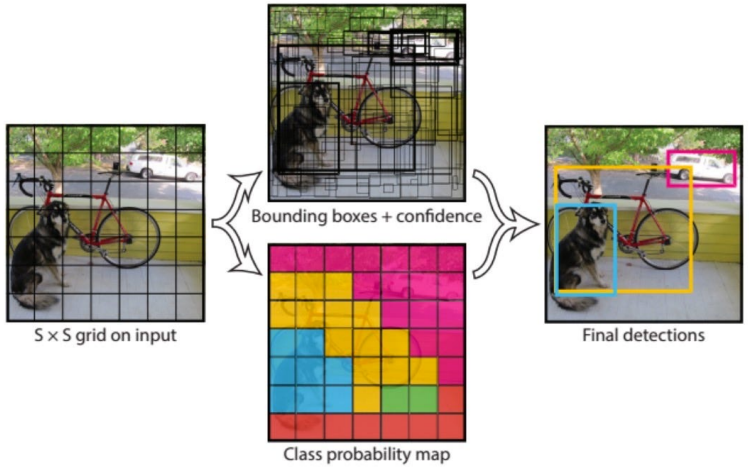


Definición de YOLO

YOLO (You Only Look Once) es un enfoque innovador para la detección de objetos en imágenes.

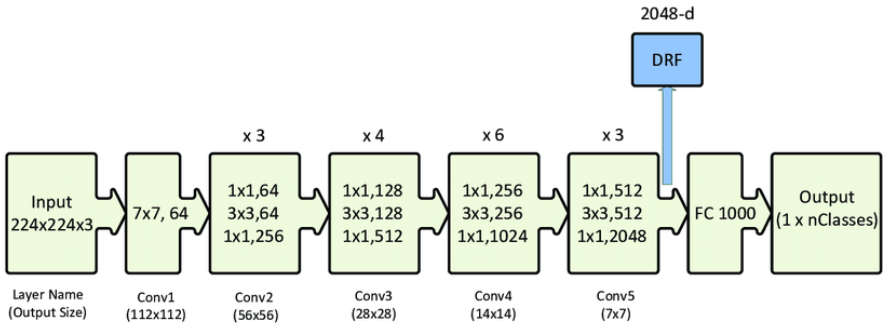
Es una arquitectura CNN para la detección de objetos que divide la imagen en una cuadrícula y predice cajas delimitadoras y probabilidades de clase directamente desde las características de esta cuadrícula.

Utiliza un enfoque de predicción de cajas de anclaje para mejorar la precisión y la estabilidad durante el entrenamiento, ofreciendo un buen equilibrio entre precisión y velocidad.



Adicionalmente, hemos utilizado una red ResNet50 para la feature extraction.

ResNet50 es una arquitectura de red neuronal convolucional profunda desarrollada por Microsoft Research. Es una variante de la familia de redes ResNet (Residual Networks) que se caracteriza por su profundidad y su capacidad para entrenar redes más profundas con un rendimiento mejorado mediante el uso de bloques residuales.



Entrenamiento de YOLO

Se ha utilizado el 60% de las datos para train, el 10% para validation y el 30% restante para test.

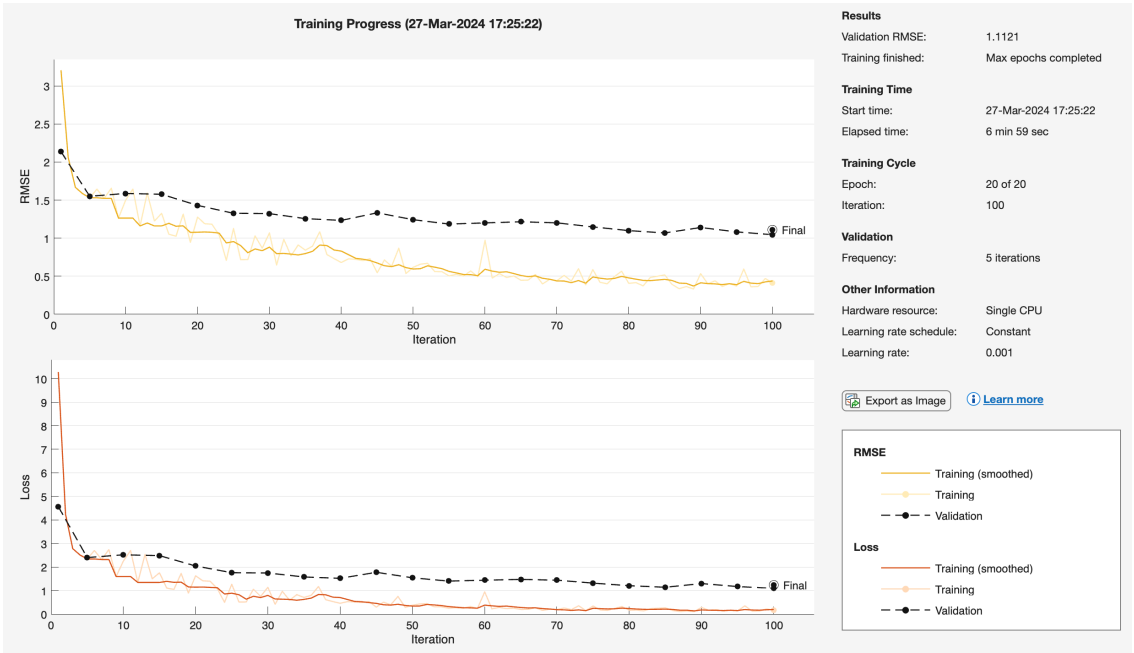
Valores importantes:

- `input_size = [128 128 3];`
- `num_anchors = 3;`
- `feature_extraction_network = resnet50;`
- `feature_layer = 'activation_40_relu';`

```
********************************************************************************
Training a YOLO v2 Object Detector for the following object classes:

* Chihuahua

Training on single CPU.
Initializing input data normalization.
|=========================================================================================================|
|  Epoch  |  Iteration  |  Time Elapsed  |  Mini-batch  |  Validation  |  Mini-batch  |  Validation  |  Base Learning  |
|         |             |   (hh:mm:ss)   |     RMSE     |     RMSE     |     Loss     |     Loss     |      Rate       |
|=========================================================================================================|
|       1 |           1 |       00:00:09 |         3.21 |         2.14 |      10.2815 |       4.5669 |          0.0010 |
|       1 |           5 |       00:00:22 |         1.53 |         1.55 |       2.3426 |       2.4023 |          0.0010 |
|       2 |          10 |       00:00:41 |         1.50 |         1.59 |       2.2509 |       2.5197 |          0.0010 |
|       3 |          15 |       00:01:01 |         1.33 |         1.58 |       1.7562 |       2.4844 |          0.0010 |
|       4 |          20 |       00:01:19 |         1.28 |         1.43 |       1.6286 |       2.0447 |          0.0010 |
|       5 |          25 |       00:01:38 |         1.13 |         1.33 |       1.2689 |       1.7597 |          0.0010 |
|       6 |          30 |       00:01:56 |         1.07 |         1.32 |       1.1414 |       1.7474 |          0.0010 |
|       7 |          35 |       00:02:16 |         0.84 |         1.26 |       0.7083 |       1.5783 |          0.0010 |
|       8 |          40 |       00:02:35 |         0.68 |         1.23 |       0.4614 |       1.5204 |          0.0010 |
|       9 |          45 |       00:02:53 |         0.55 |         1.33 |       0.2977 |       1.7768 |          0.0010 |
|      10 |          50 |       00:03:12 |         0.62 |         1.24 |       0.3802 |       1.5436 |          0.0010 |
|      11 |          55 |       00:03:30 |         0.50 |         1.18 |       0.2532 |       1.4034 |          0.0010 |
|      12 |          60 |       00:03:49 |         0.97 |         1.20 |       0.9437 |       1.4405 |          0.0010 |
|      13 |          65 |       00:04:06 |         0.45 |         1.21 |       0.1990 |       1.4760 |          0.0010 |
|      14 |          70 |       00:04:24 |         0.44 |         1.20 |       0.1946 |       1.4412 |          0.0010 |
|      15 |          75 |       00:04:43 |         0.59 |         1.15 |       0.3439 |       1.3174 |          0.0010 |
|      16 |          80 |       00:05:00 |         0.41 |         1.10 |       0.1644 |       1.2071 |          0.0010 |
|      17 |          85 |       00:05:18 |         0.52 |         1.07 |       0.2673 |       1.1411 |          0.0010 |
|      18 |          90 |       00:05:35 |         0.53 |         1.14 |       0.2839 |       1.2985 |          0.0010 |
|      19 |          95 |       00:05:53 |         0.37 |         1.08 |       0.1370 |       1.1710 |          0.0010 |
|      20 |         100 |       00:06:10 |         0.41 |         1.04 |       0.1691 |       1.0916 |          0.0010 |
|=========================================================================================================|
Training finished: Max epochs completed.
Detector training complete.
********************************************************************************
```

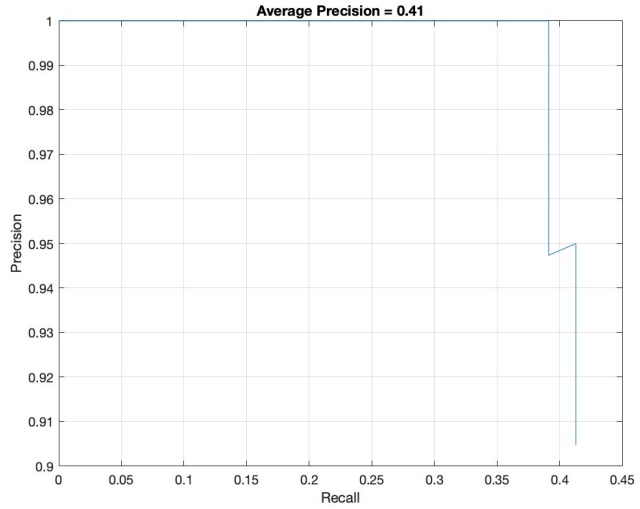


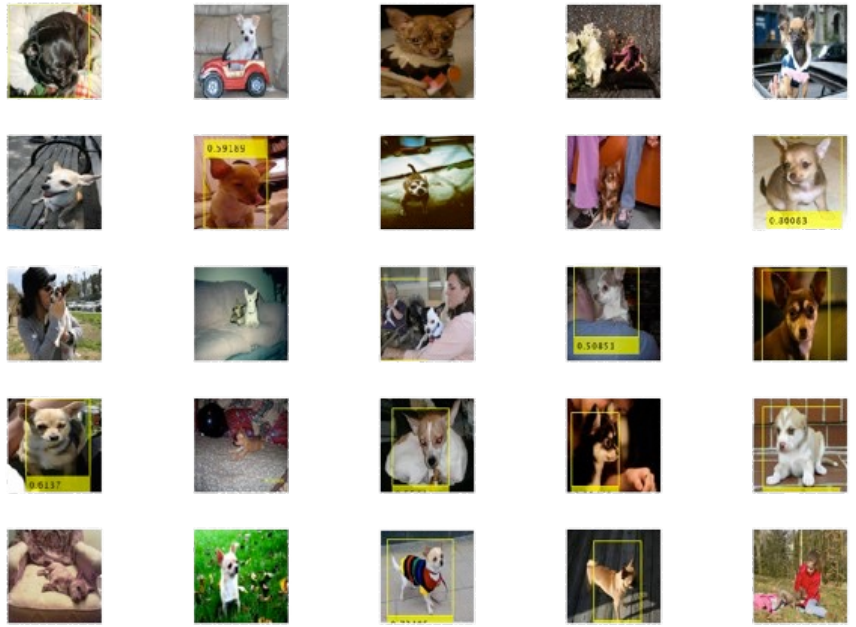
Evaluación de YOLO

El modelo tiene una precisión media de 0.41 en el conjunto de test.

La precisión es alta, generalmente cuando identifica un perro, acierta.

El recall es bajo, no logra identificar muchos de los perros presentes en las imágenes.





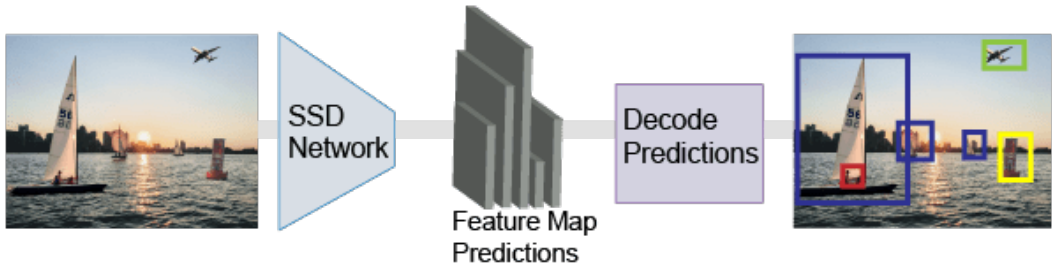
Definición de SSD

SSD, que significa "Single Shot MultiBox Detector" es una arquitectura de red neuronal convolucional (CNN) que realiza la detección de objetos en una sola pasada.

Utiliza múltiples cuadros delimitadores (bounding boxes) y clases de objetos en cada celda de la cuadrícula en varias escalas de características.

SSD puede tener dificultades para detectar objetos pequeños debido a la forma en que maneja las escalas de características.

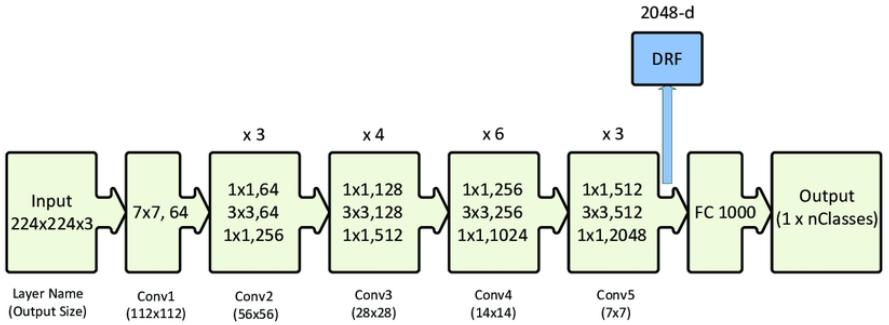
Fue desarrollado por investigadores de Google y se destaca por su capacidad para realizar detecciones de objetos en tiempo real con un alto rendimiento.



Adicionalmente, hemos utilizado una red ResNet50 para la feature extraction.

ResNet50 es una arquitectura de red neuronal convolucional profunda desarrollada por Microsoft Research. Es una variante de la familia de redes ResNet (Residual Networks) que se caracteriza por su profundidad y su capacidad para entrenar redes más profundas con un rendimiento mejorado mediante el uso de bloques residuales.

A diferencia de YOLO, SSD hace uso de varias capas de la red para feature extraction.



Entrenamiento de SSD

Se ha utilizado el 60% de las datos para train, el 10% para validation y el 30% restante para test.

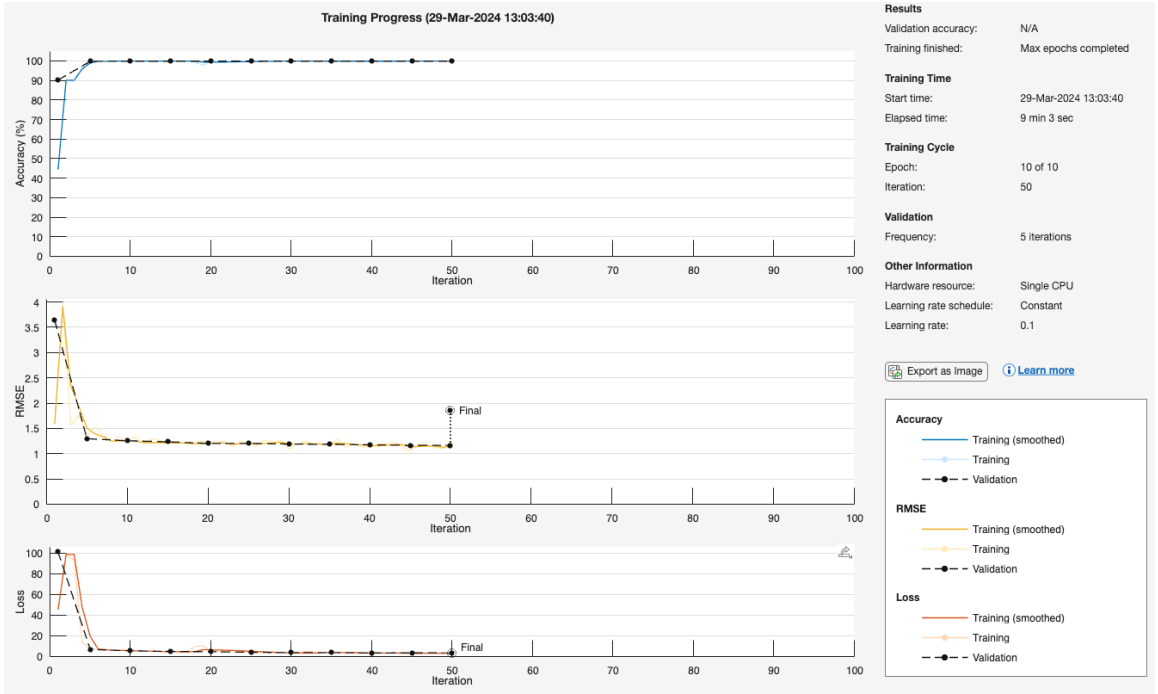
Valores importantes:

- `input_size = [300 300 3];`
- `feature_extraction_network = resnet50;`

```
*************************************************************************
Training an SSD Object Detector for the following object classes:

* Chihuahua

Training on single CPU.
Initializing input data normalization.
|=========================================================================================================================================|
|  Epoch  |  Iteration  |  Time Elapsed  |  Mini-batch  |  Mini-batch  |  Mini-batch  |  Validation  |  Validation  |  Validation  |  Base Learning  |
|         |             |   (hh:mm:ss)   |     Loss     |   Accuracy   |     RMSE     |     Loss     |   Accuracy   |     RMSE     |      Rate       |
|=========================================================================================================================================|
|       1 |           1 |       00:00:14 |      45.3328 |       44.38% |         1.59 |     101.0386 |       90.16% |         3.64 |          0.1000 |
|       1 |           5 |       00:00:48 |       7.1489 |       98.90% |         1.33 |       6.8742 |       99.73% |         1.30 |          0.1000 |
|       2 |          10 |       00:01:35 |       5.5843 |       99.82% |         1.28 |       5.5651 |       99.82% |         1.26 |          0.1000 |
|       3 |          15 |       00:02:20 |       4.5995 |       99.86% |         1.21 |       4.9198 |       99.82% |         1.23 |          0.1000 |
|       4 |          20 |       00:03:08 |       5.0822 |       99.84% |         1.22 |       4.6928 |       99.82% |         1.20 |          0.1000 |
|       5 |          25 |       00:04:03 |       4.1566 |       99.85% |         1.22 |       4.0754 |       99.82% |         1.20 |          0.1000 |
|       6 |          30 |       00:04:59 |       3.3794 |       99.87% |         1.09 |       3.8173 |       99.82% |         1.19 |          0.1000 |
|       7 |          35 |       00:05:47 |       3.5083 |       99.85% |         1.17 |       3.8442 |       99.82% |         1.18 |          0.1000 |
|       8 |          40 |       00:06:38 |       3.4054 |       99.88% |         1.12 |       3.4103 |       99.83% |         1.18 |          0.1000 |
|       9 |          45 |       00:07:26 |       3.0888 |       99.90% |         1.05 |       3.6171 |       99.83% |         1.16 |          0.1000 |
|      10 |          50 |       00:08:11 |       3.7136 |       99.82% |         1.19 |       3.4597 |       99.83% |         1.16 |          0.1000 |
|=========================================================================================================================================|
Training finished: Max epochs completed.
Detector training complete.
*************************************************************************
```



Evaluación de SSD

El modelo tiene una precisión media de 0.32 en el conjunto de test.

La precisión es media, generalmente cuando identifica un perro, acierta, aunque sus cajas delimitadoras pueden no ser ajustadas.

El recall es medio, logra identificar la mayoría de los perros presentes en las imágenes.