

TFG informe inicial

Raquel

October 23, 2017

Contents

1	Data	1
1.1	Initial Data	1
1.2	Generated Data	2
2	Objetivos	2
3	Análisis	4
3.1	Intra sinset	5
3.2	Distribución total de las features	7
3.3	Distribución de las features por layer	8
3.4	Distribución de las features por synset	11
3.5	Features per image	13
3.6	Images per feature	15

1 Data

1.1 Initial Data

- Embedding matrix of size (50000, 12416), con 62080000 features.
- **labels** Labels vector of size 50k which every label is in numeric format (0, 999)
- **synsets** = **synset0 synset1 synset2 ...** The set of synsets that we will analyze:
synsets = [*living_things*, *mammal*, *dog*, *hunting_dogs*]
- **categories** = **{-1 0 1 }** The possible values of the features.

1.2 Generated Data

- **synset_index_hyponim.txt** A list with all the hyponims of every synset.
- **synset_index.txt** For each synset a list with the index of the elements of the hyponim list in the embedding.
- Un diccionario con la cantidad de imágenes que tiene cada feature para cada category.
- **features_per_image[synsets].pkl** dfd
- **features_per_layer[synsets].pkldsf**
- **images_per_feature_per_synset[synsets].pklfdsfs**
- **intra_synset[synsets].pklfdsfs**

2 Objetivos

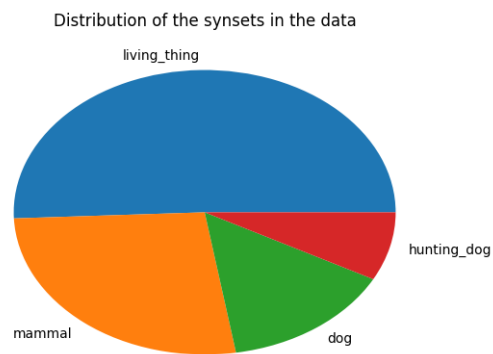
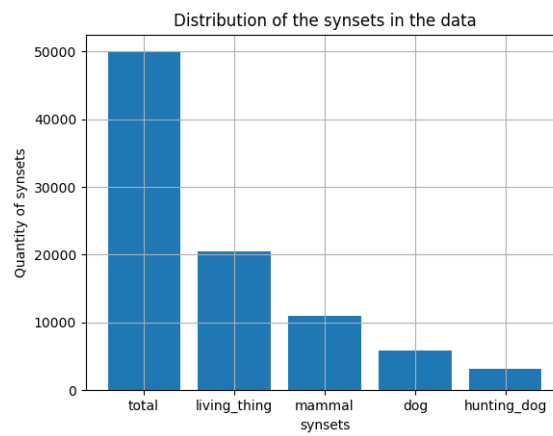
- Estadísticas del dataset por imagen del conjunto de synsets (**hecho**)
- Estadísticas internas para cada synset (que synsets son subconjunto de cual y en que proporción) (**hecho**)
- Estadísticas totales de la distribución de las features respecto a toda la matriz de imágenes. (**solo faltan grafiquillas**)
- Estadísticas totales de la distribución de las features respecto a toda la matriz de imágenes por layer. (**falta escribir + grafiquillas**)
- Estadísticas de las features respecto el subconjunto de imágenes de cada synset. (**faltan grafiquillas**)
- Estadísticas de las features respecto el subconjunto de imágenes de cada synset por layer. (**falta todo**)
- Distribución de imágenes por feature (para 1, -1, 0):
 - respecto a todas las imágenes
 - respecto a cada uno de los subconjuntos de imágenes de cada synset**está calculado, falta escribir**
- Repetir para el resto de embeddings
- Montarlo para que me genere todo automático para cada conjunto de synsets.
- Hacer que guarde todo en maps

- Features por imagen: Para cada imagen cuantas features (de cada category) se activan. Usando todos los valores obtenidos dibujo un histograma para cada category de feature. Con el eje de las x siendo la frecuencia y el de las y la cantidad de imágenes que cumplen eso.
- Imágenes por feature: Para cada features (de cada category), cuantas imágenes toman este valor. Usando todos los valores obtenidos dibujo un histograma para cada category de feature. Con el eje de las x siendo la frecuencia y el de las y la cantidad de imágenes que cumplen eso.
- Comprobar si dentro de un mismo synset hay features que se den con 1 y -1.
- Hacer las distribuciones de la suma.
- Ordenar del código y generar una documentación y tests.

3 Análisis

Tenemos 50000 imagenes, de las cuales:

- el 41.0% son living_thing.
- el 21.8% son mammal.
- el 11.799999999999999% son dog.
- el 6.3% son hunting_dog.



3.1 Intra sinset

Dentro de living_things las imágenes se distribuyen como:

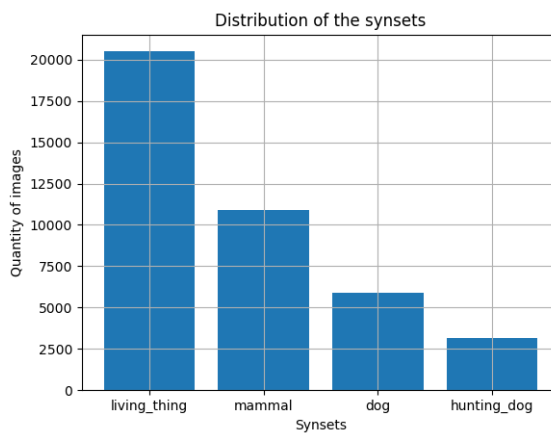
- el 53.1707317073% son mammal.
- el 28.7804878049% son dog.
- el 15.3658536585% son hunting_dog.

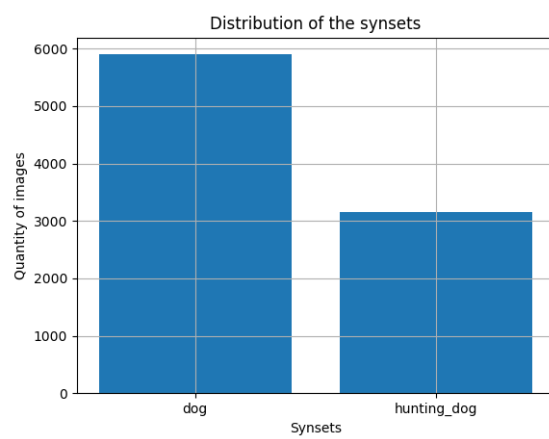
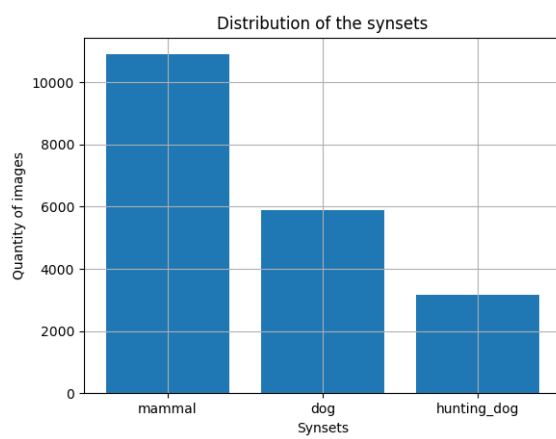
Dentro de mammal las imágenes se distribuyen como:

- el 54.128440367% son dog.
- el 28.8990825688% son hunting_dog.

Dentro de dog las imágenes se distribuyen como:

- el 53.3898305085% son hunting_dog.

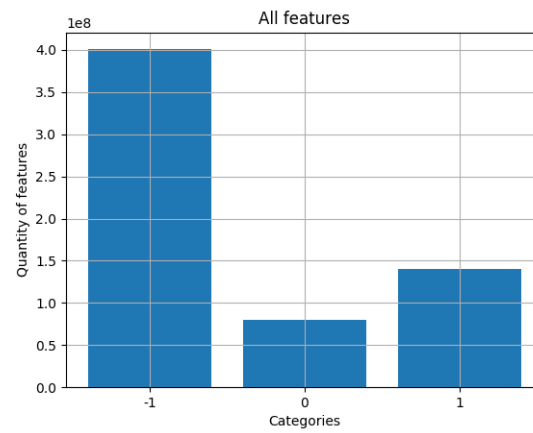




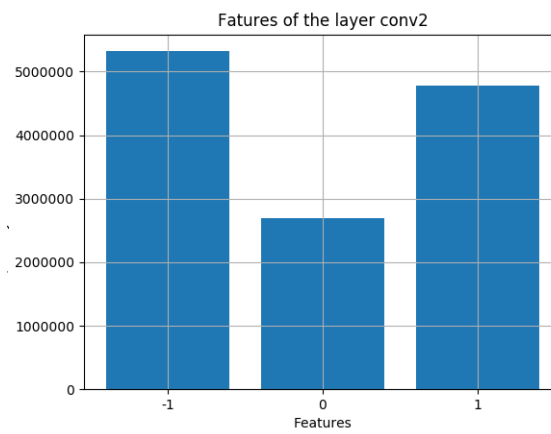
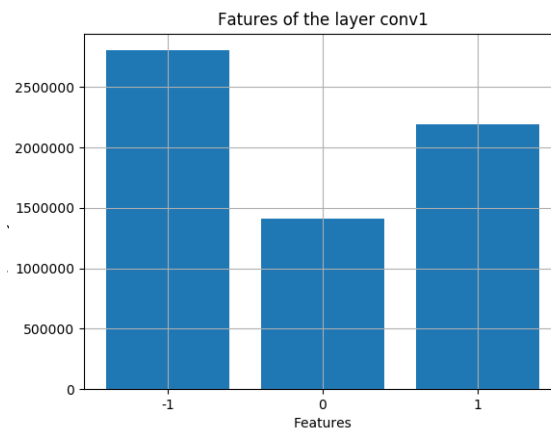
3.2 Distribución total de las features

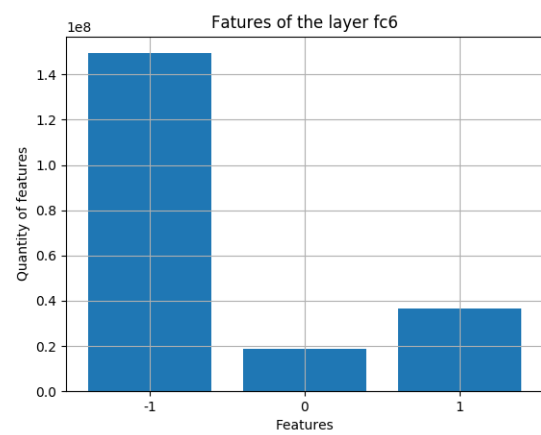
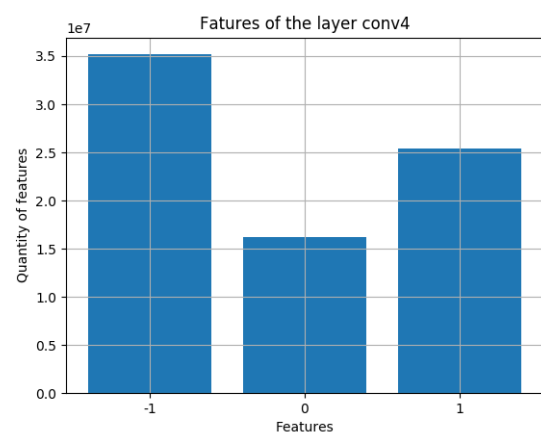
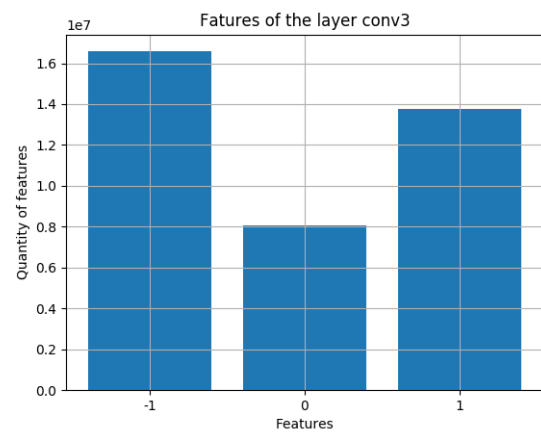
Del total de 620800000 features:

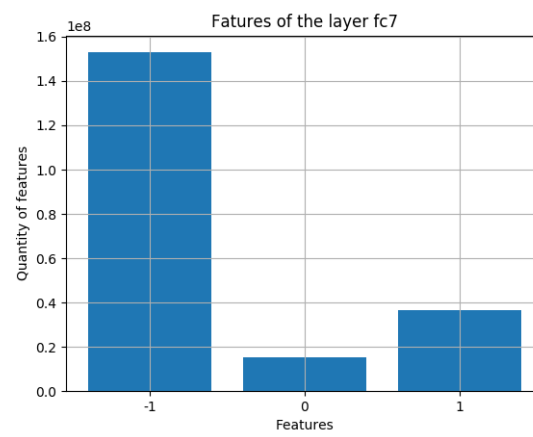
- Features de category -1: 419255437 el 67.534703125 %
- Features de category 0: 56916672 el 9.16827835052 %
- Features de category 1: 144627891 el 23.2970185245 %



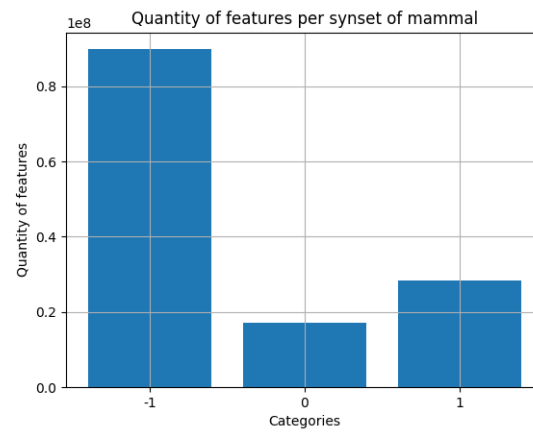
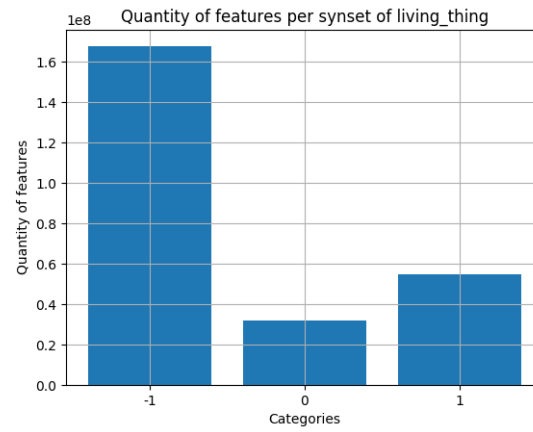
3.3 Distribución de las features por layer

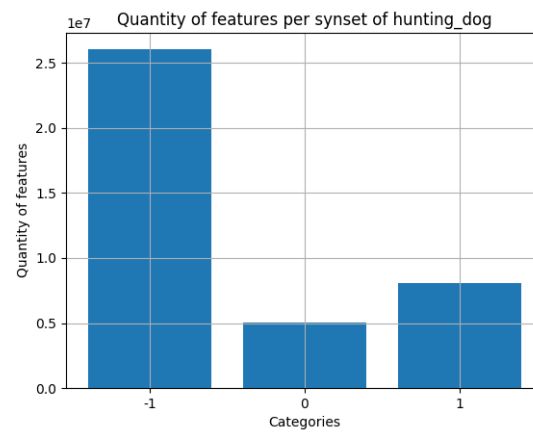
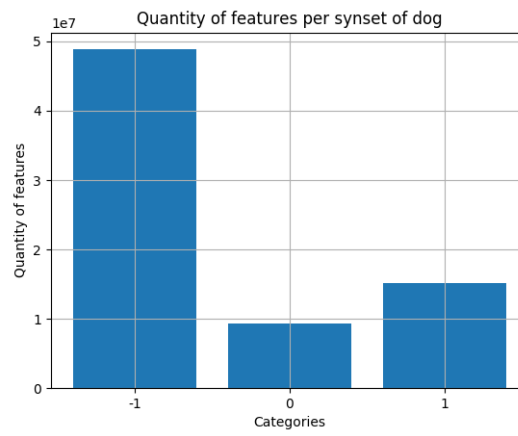






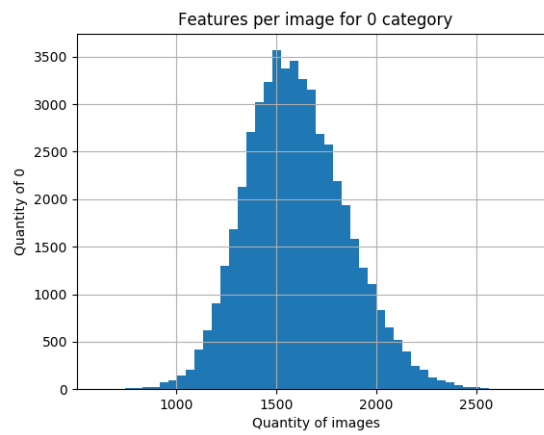
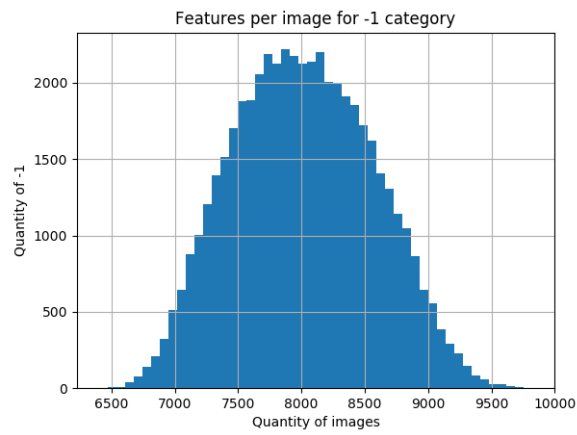
3.4 Distribución de las features por synset

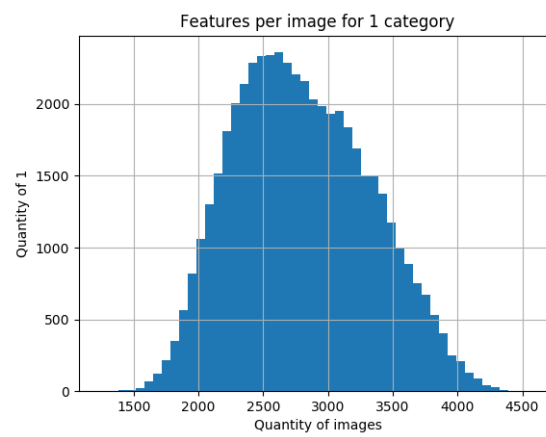




3.5 Features per image

There are for each image the total counting of features active, for the different possible categories $(-1, 0, 1)$:





3.6 Images per feature

Now I calculate for each feature how many images activate in each specific category:

