

TFG informe inicial

Raquel

November 8, 2017

Contents

1	Data	1
1.1	Initial Data	1
1.2	Generated Data	1
2	Objetivos	2
3	Análisis Base	3
3.1	Intra sinset	3
3.2	Distribución total de las features	6
3.3	Distribución de las features por layer	6
3.4	Distribución de las features por synset	11
3.5	Features per image	13
3.6	Images per feature	15
3.7	Images per feature per synset	18
3.8	Comprobación de que las cosas tienen sentido	23
3.9	Estudio de los outliers de imágenes per feature	23

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.

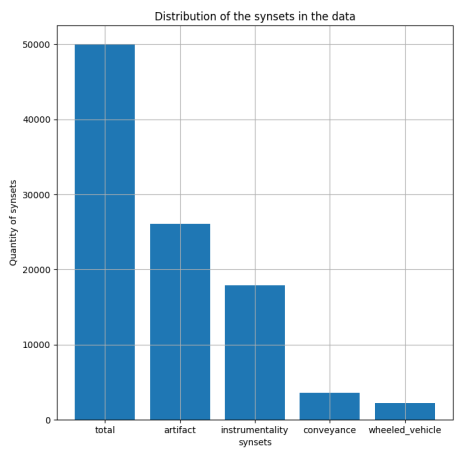
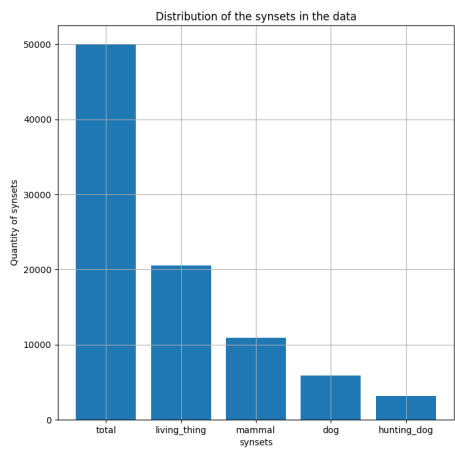
1.2 Generated Data

- **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].pkl** Genera un diccionario tal que:
`dict[features][category][synset] = cantidad e imagenes del synset que tienen esta categoria en cuestión.`
- **intra_synset[synsets].pkldfs**

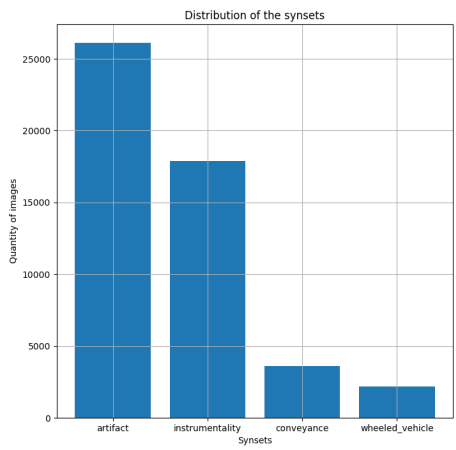
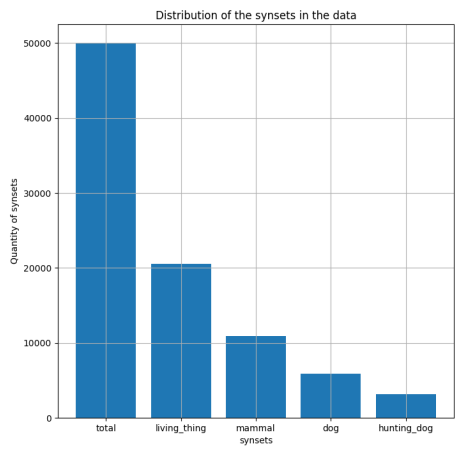
2 Objetivos

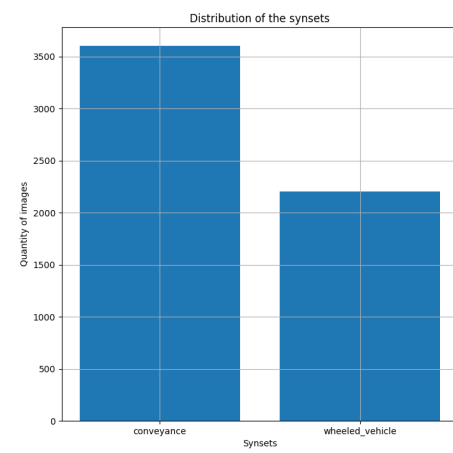
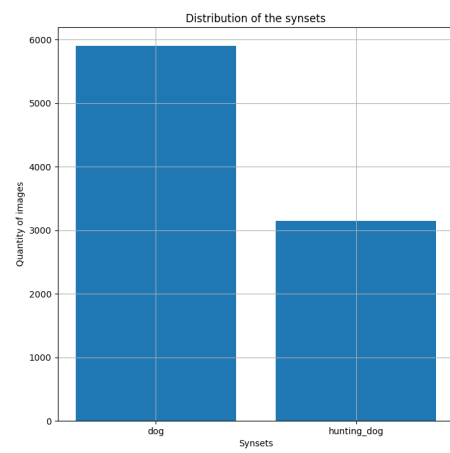
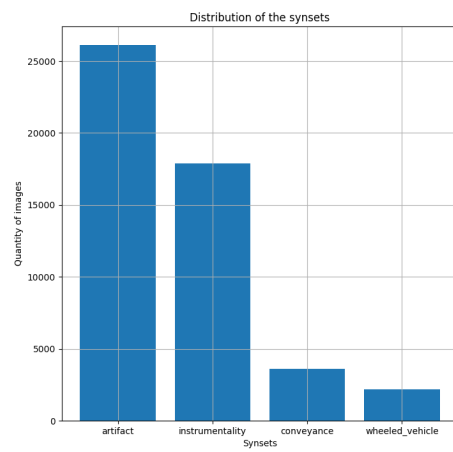
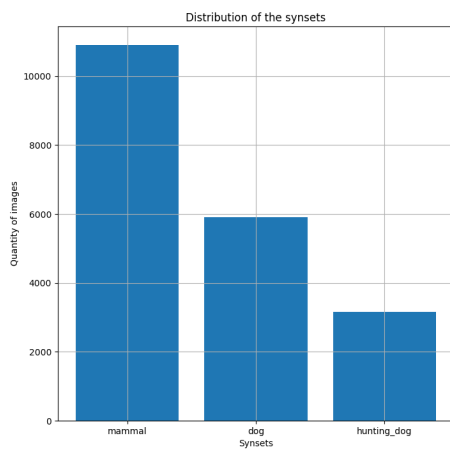
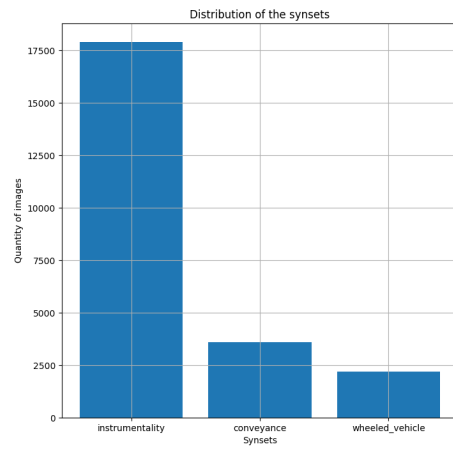
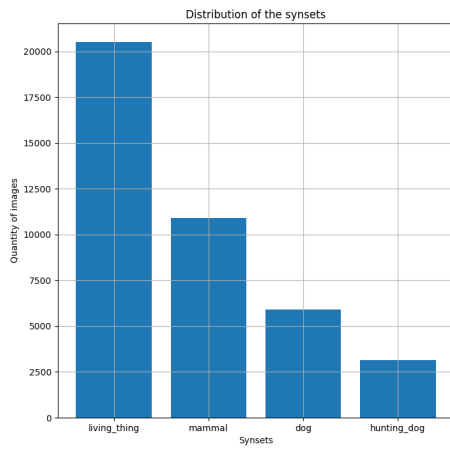
- Estadísticas del dataset por imagen del conjunto de synsets (**hecho**)
- Estadísticas internas para cada synset (que synsets son subconjunto de cual y en que proporción) (**hecho**)
- Estadísticas totales de la distribución de las features respecto a toda la matriz de imágenes. (**solo faltan grafiquillas**)
- 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.

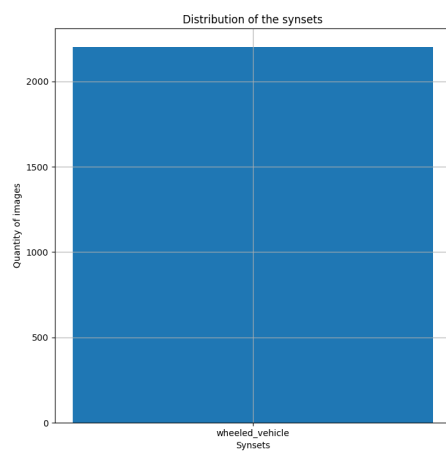
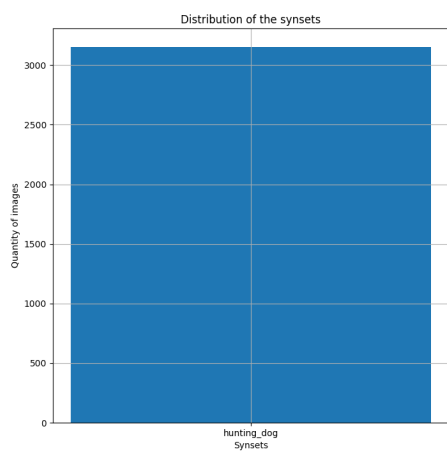
3 Análisis Base



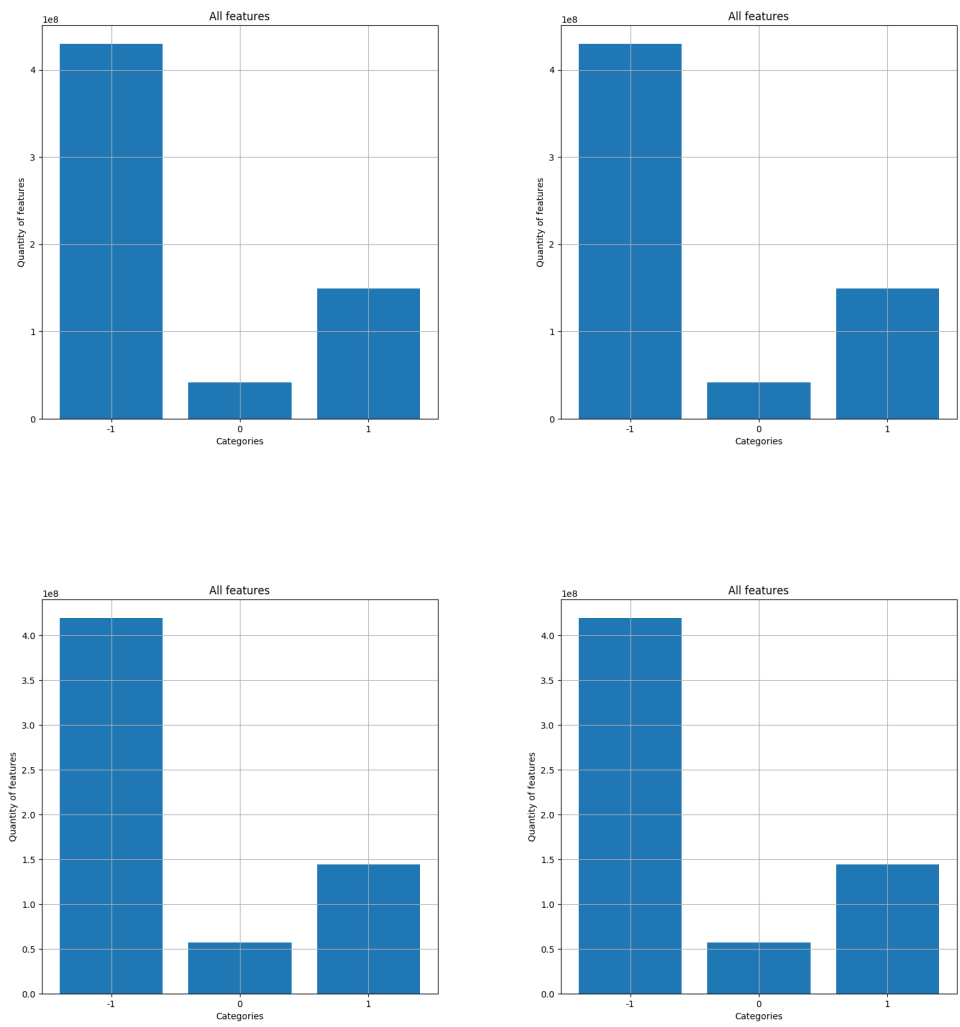
3.1 Intra sinset



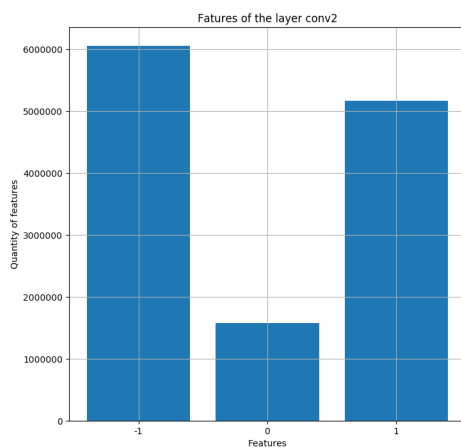
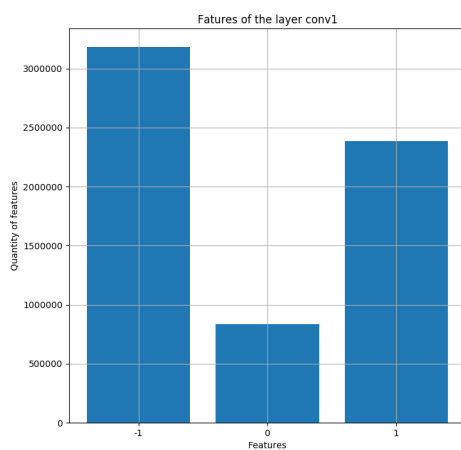
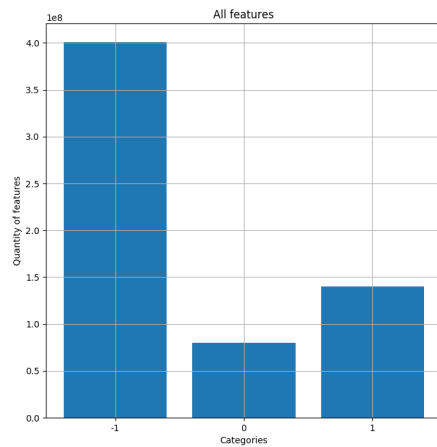
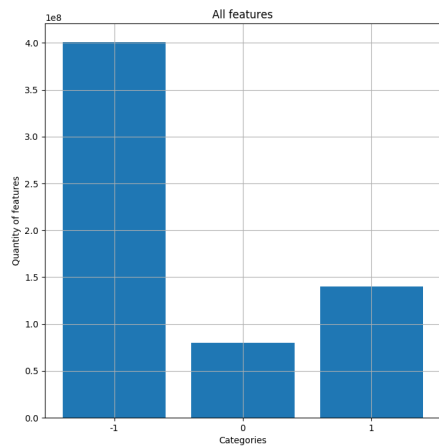


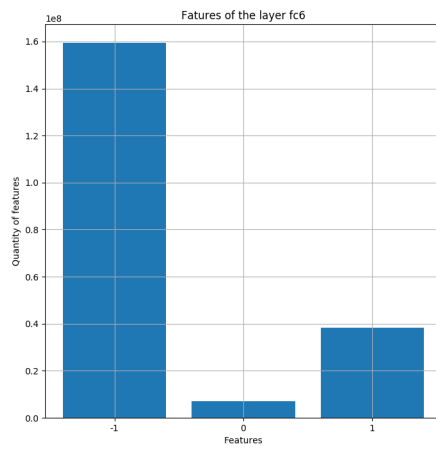
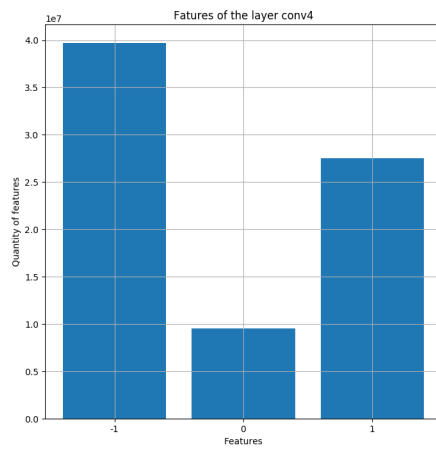
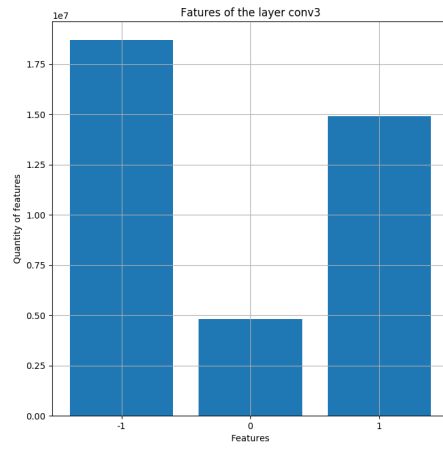


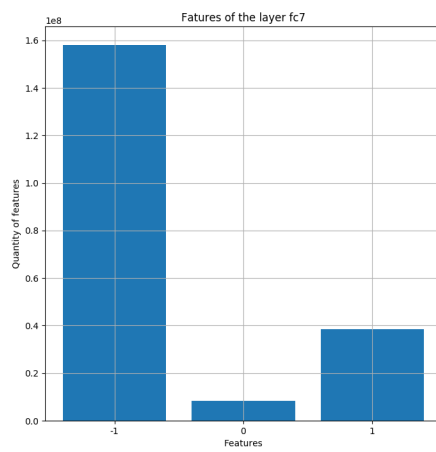
3.2 Distribución total de las features



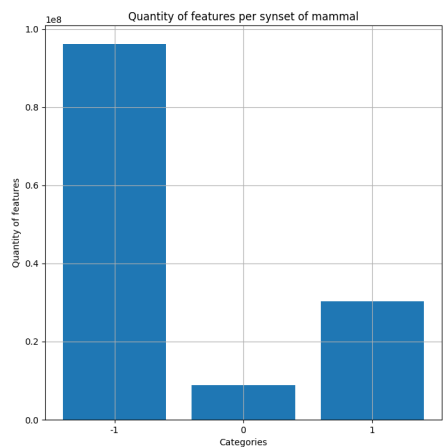
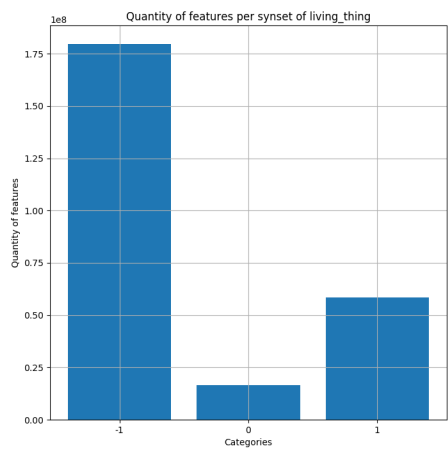
3.3 Distribución de las features por layer

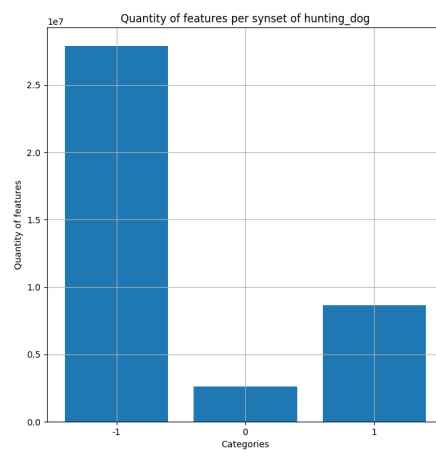
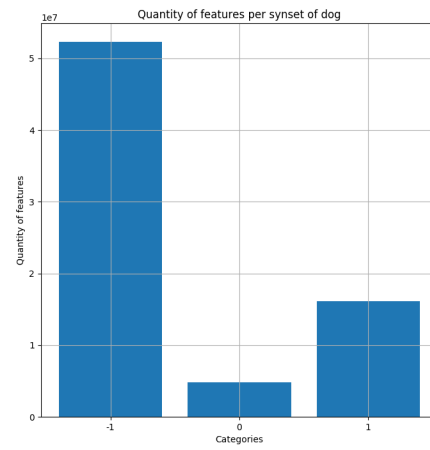






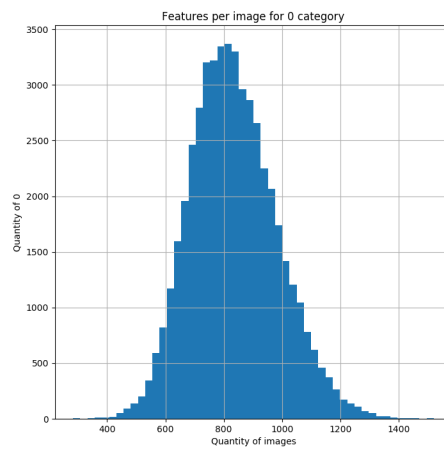
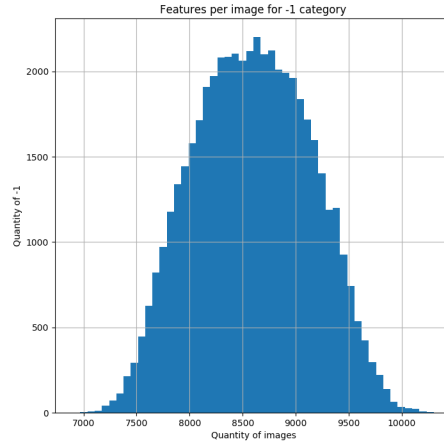
3.4 Distribución de las features por synset

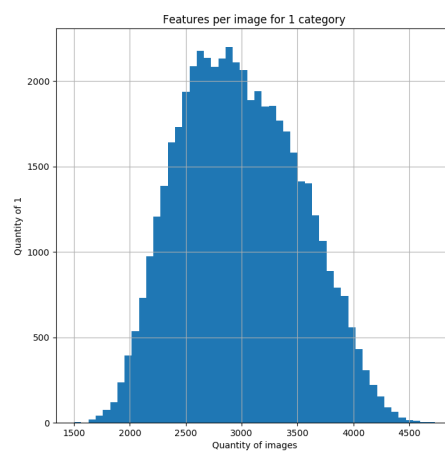




3.5 Features per image

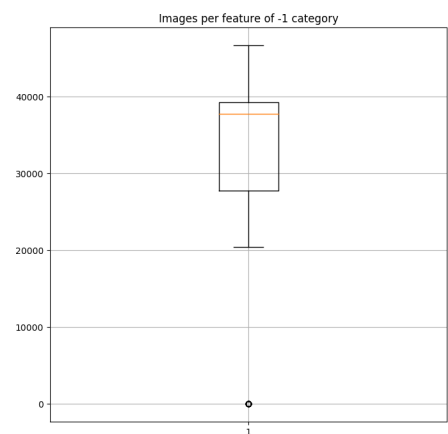
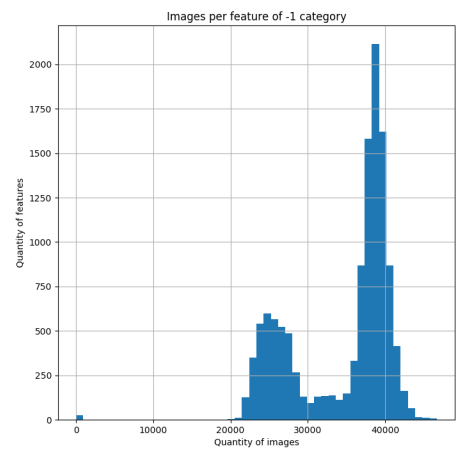
There are for each image the total counting of features active, for the different possible categories $(-1, 0, 1)$:

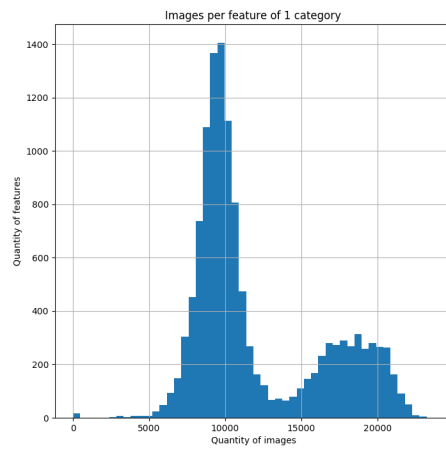
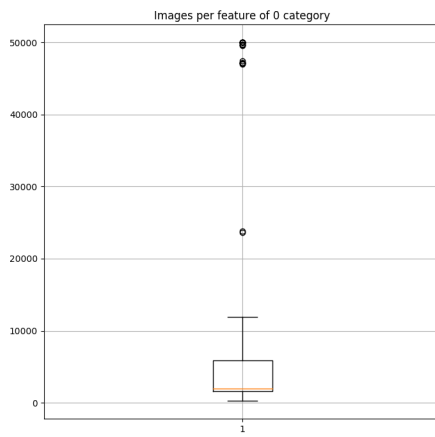
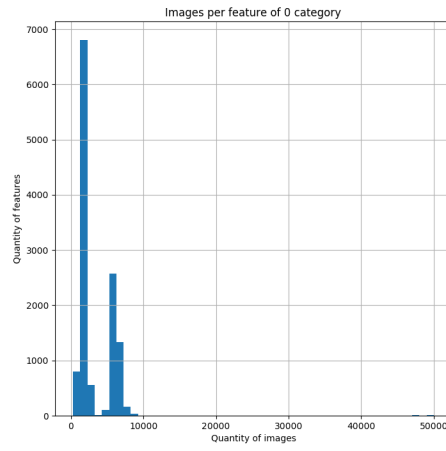


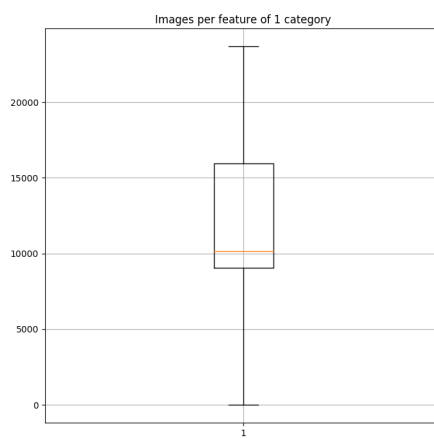


3.6 Images per feature

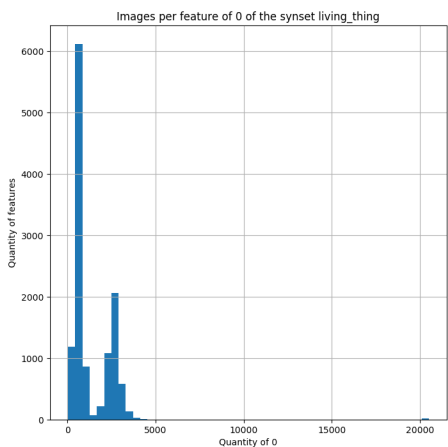
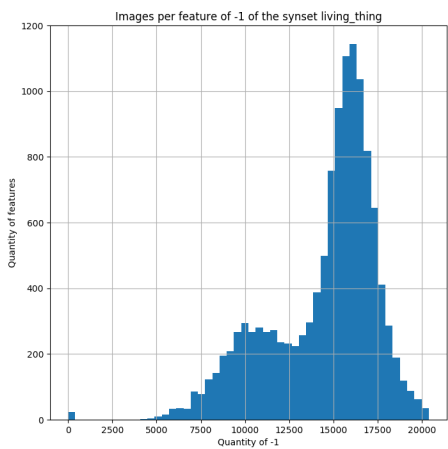
Now I calculate for each feature how many images activate in each specific category:

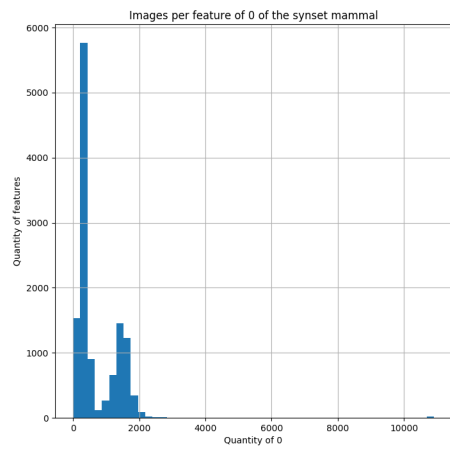
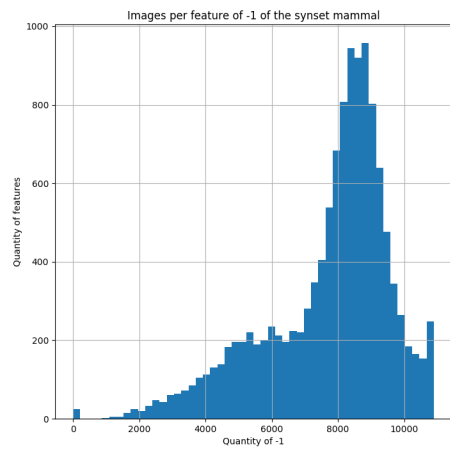
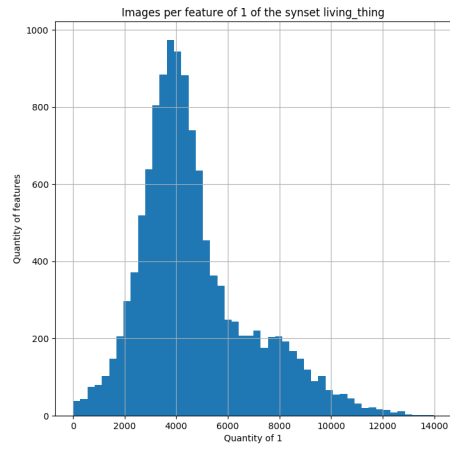


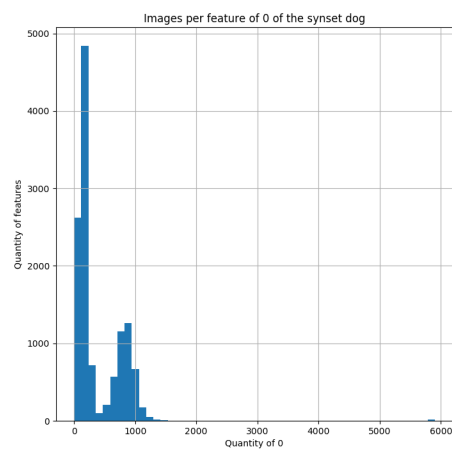
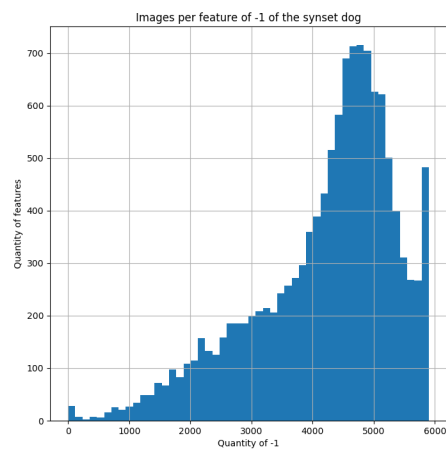
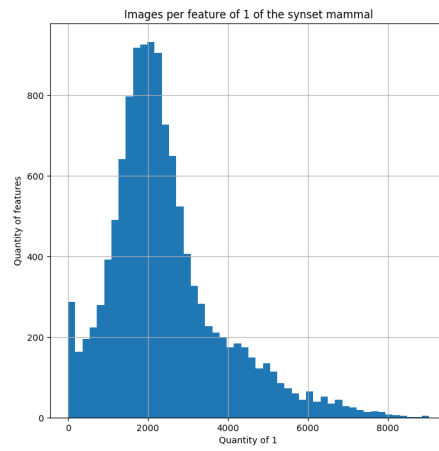


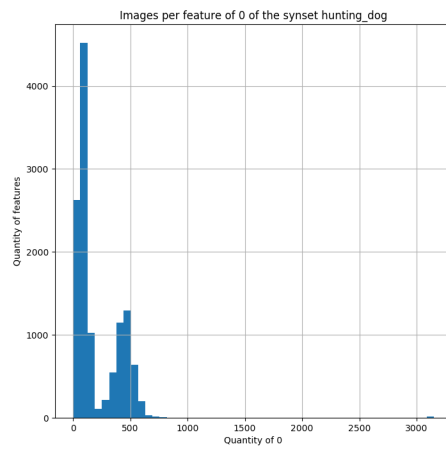
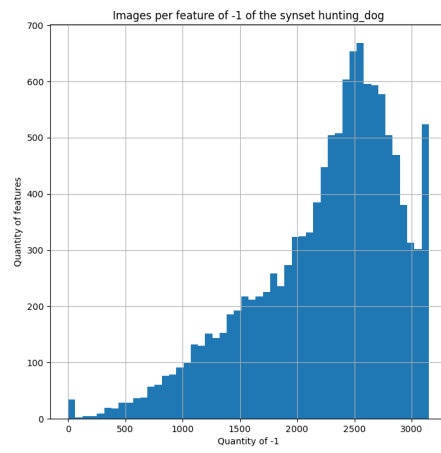
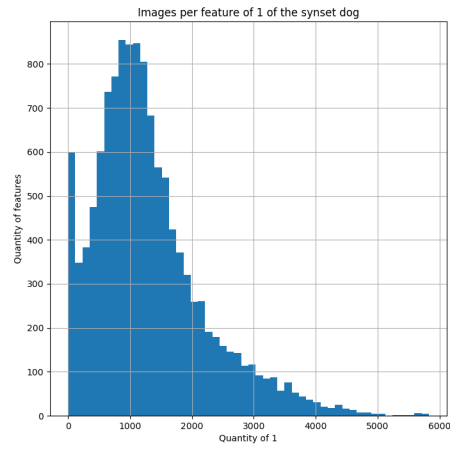


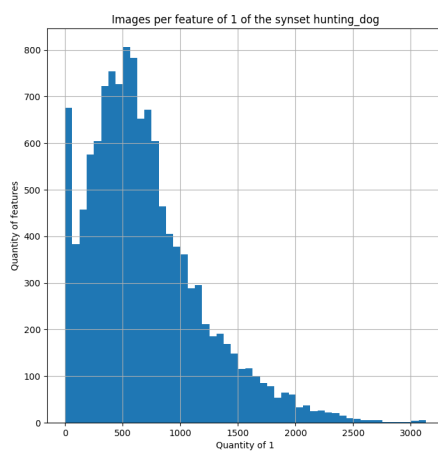
3.7 Images per feature per synset











3.8 Comprobación de que las cosas tienen sentido

- Hay alguna imagen que no tenga ninguna feature con valor cero? No, ninguna
- Hay algun synset que tenga valor 1 y -1 para la misma feature? Si, bastantes.
Estamos tomando:

La feature tiene el valor i para el synset si lo tiene para alguna imagen del synset (Estamos estudiando la submatriz correspondiente al synset)

3.9 Estudio de los outliers de imagenes per feature

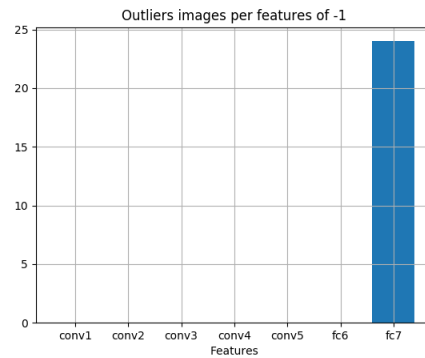


Figure 1: Category: -1

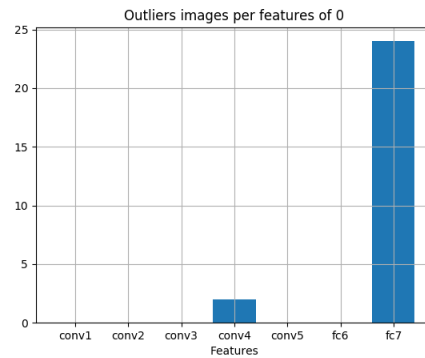


Figure 2: Category: 0

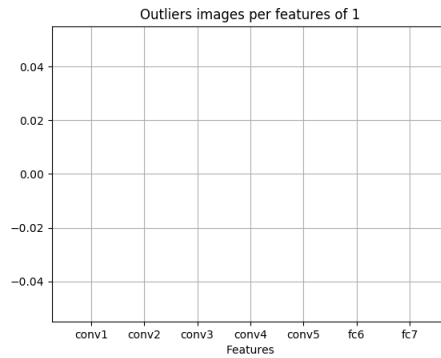


Figure 3: Category: 1