

# Supplementary Material for “On zero-shot recognition of generic objects”

Tristan Hascoet, Tetusya Takiguchi, Yasuo Ariki

## Appendix A. Structural flaws

Figure 1 of this supplementary material reproduces Figure 1 of the original paper to help the following discussion. This figure illustrates the configuration of visual classes of the standard test splits within the Wordnet hierarchy. It should be noted that the *2-hops* test split is a super-set of the *1-hop* split: it contains both classes annotated in green and blue. Similarly, the *all* test split is a super-set of the *2-hops* test split: it contains all blue, green and black classes. In the generalized ZSL setting, training classes (red) are also included in the test set.

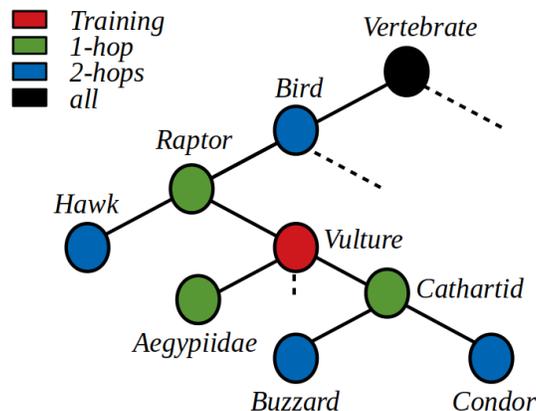


Figure 1: Illustration of the standard test splits configuration.

Figure 2 and 3 illustrate the distribution of ZSL classification outputs on the *2-hops* and *all* test splits respectively. On the *2-hops* standard ZSL test set, 3.6% of test images were correctly classified by the Linear baseline model. This ratio corresponds to the percentage of images of *Raptor* correctly classified as *Raptor*, *Buzzard* images classified as *Buzzard*, etc. We refer to such classification outputs as True Positive (TP). These correspond to the accuracy reported by previous

works on the standard benchmark. 2.3% of test images were classified as one of their parent class: These correspond to images of *Buzzard* or *Hawk* classified as *Raptor* or *Bird* for example. These classification outputs are considered as errors by the current benchmark, while they are semantically correct: a *Hawk* is just a specific kind of *Bird*. 3.7% of test images were classified as one of their children class: images of *Raptor* or *Bird* classified as *Buzzard* or *Hawk*. Such classification outputs are considered as errors by the current benchmark, whereas they may be either semantically correct or incorrect depending on the specific kind of bird in the image. We refer to both of these classification scenarios as False Negative (FN). On the other hand, an image of *Buzzard* classified as *Aegypiidae* is an actual classification error: *Buzzard* and *Aegypiidae* are two distinct, mutually exclusive concepts. We refer to such classification errors as True Negatives (TN).

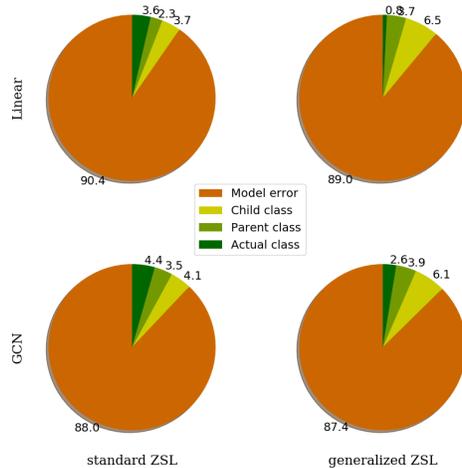


Figure 2: Distribution of classification outputs on the *2-hops* test split.

Table 1 summarizes the ratio of false negative per true positive on each of the standard test split:  $ratio = FN/TP$ . This table shows two interesting trends: First, as noted in the original paper, the ratio is much higher in the Generalized ZSL setting. This is due to the fact that ZSL models tend to classify test images as their parent or children training class. Second, in the standard ZSL setting, the ratio tends to increase with larger test sets: the GCN model ratios are 2.3, 3.8 and 4.1 on the *1-hop*, *2-hops* and *all* test splits respectively. We believe this is due to larger overlaps within the Wordnet hierarchy: In the *1-hop* test set, the only FN classes for *Cathartid* images is *Raptor*. In the *2-hops* test set, *Buzzard*, *Condor*, *Raptor* and *Bird* are all FN classification outputs for *Cathartid* images. This trend, however, does not hold for the Linear model in the Generalized ZSL

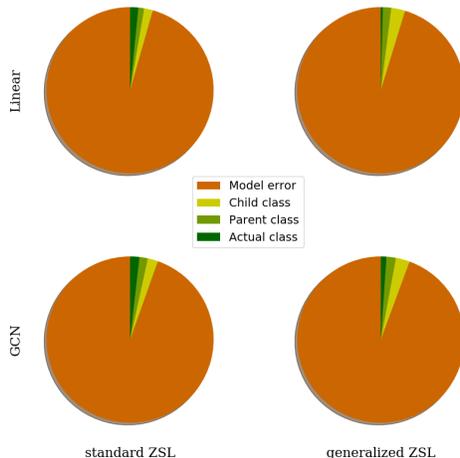


Figure 3: Distribution of classification outputs on the *all* test split.

setting.

Table 1: Ratio of false negatives (FN) per true positives (TP).

Model	Task	1-hop			2-hops			all		
		TP	FN	ratio	TP	FN.	ratio	TP	FN	ratio
Linear	ZSL	14.7	10.2	<b>0.7</b>	3.6	6.0	<b>1.7</b>	1.6	2.8	<b>1.7</b>
	GZSL	1.9	39.2	<b>20.6</b>	0.8	10.23	<b>12.7</b>	0.4	4.27	<b>10.7</b>
GCN	ZSL	21.8	18.6	<b>0.8</b>	4.4	7.6	<b>1.7</b>	1.8	3.6	<b>2.0</b>
	GZSL	10.3	34.2	<b>2.3</b>	2.6	10.0	<b>3.8</b>	1.1	4.5	<b>4.1</b>

## Appendix B. Word embeddings

### Occurrence frequency

We used the full English Wikipedia corpus to estimate the occurrence frequency of words: we scanned the Wikipedia corpus to count the occurrence of each visual class labels (*Hawk*, *Raptor* or *Aegyptiidae*, etc.). We use these occurrence counts as a measure to identify rare and common words. Figure 4 represents the cumulative distribution of visual class label occurrence counts.

As shown in this figure, 24% of Imagenet class labels occur less than 10 times in the full Wikipedia corpus. 45% of Imagenet class labels occur less than 100

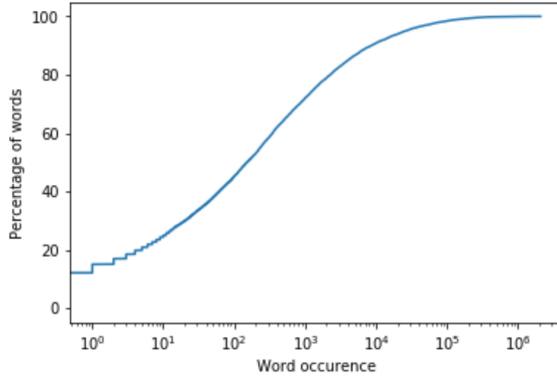


Figure 4: Word occurrence cumulative distribution. The x axis is in logarithmic scale.

times. We found that fine-grain animal species, in particular, exhibit rare word labels (see Figure 1). We expect the word embedding of such classes to provide noisy semantic representations, which has been confirmed by the experiments presented in the original paper.

## Polysemy

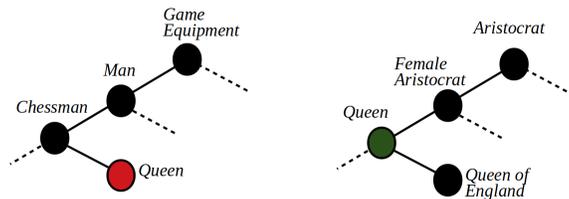


Figure 5: Illustration of two Wordnet concepts sharing the same label Queen.

Figure 9 illustrates several polysemous visual classes of the Imagenet dataset. To deal with polysemy, we want to assign a unique visual class to polysemous words. To do so, we define a similarity score  $s(w, c)$  between words  $w$  and their visual classes  $c$ . Given a polysemous word  $w$ , we assign  $w$  to its visual class  $c$  of highest similarity score:

$$s : W \times C \rightarrow \mathbb{R} \tag{1a}$$

$$c^* = \operatorname{argmax}_{c \in C} s(w, c) \tag{1b}$$

As a similarity score, we use the cosine similarity between word embeddings and the average word embedding of visual class parent and children concepts.

Consider the example of the word *Queen* illustrated in Figure 5. There are 9 visual classes associated with the word *Queen* in the Imagenet dataset. For brevity, we only consider two of the *Queen* visual classes: one as an *Aristocrat*, and one as a *chesspiece*. The similarity score between *Queen* and its *Aristocrat* visual class is given by:

$$s(c, w) = \cos(w_{Queen}, \times(w_{Aristocrat} + w_{Female} + w_{England})/3) \quad (2a)$$

$$s(c, w) = 0.23 \quad (2b)$$

The similarity score between *Queen* and its *Chess* visual class is given by:

$$s(c, w) = \cos(w_{Queen}, w_{Chessman}) \quad (3a)$$

$$s(c, w) = -0.04 \quad (3b)$$

So we assign the word *Queen* to the visual class of highest similarity score: The one corresponding to the *Aristocrat* meaning.

## Appendix C. Visual samples

### Class-wise selection

Xian *et al.* [1] have proposed different test splits based on visual class sample populations. They conjecture that small population classes correspond to fine-grained visual concepts, while large population classes correspond to coarse-grained concepts. Manually inspecting each of these visual classes, we found many fine-grain concepts to have large image sample populations while many coarse grain concepts have small sample populations. As a measure of the "granularity" of visual classes, we propose to use their distance to the root node within the Wordnet hierarchy. Fine-grain classes are lower in the Wordnet hierarchy, hence further away from the root node than coarse-grain classes.

Figure 6 shows the average sample population of visual classes with respect to their distance to the root node in the Wordnet hierarchy. Visual classes within 6 hops of the root node have an average sample population of 490 images. Visual classes within 10 hops of the root node have an average sample population of 700 images. This figure illustrates no clear correlation between visual class granularity and their sample population. In contrast, we found that many low sample population classes instead correspond to visually ambiguous concepts, as illustrated in Figure 10. Hence, we remove low sample population classes from our proposed benchmark to avoid visually ambiguous concepts.

### Sample-wise selection process

We define high-quality image samples as images that can be correctly classified by a supervised model on a non-ZSL classification task. We propose a simple

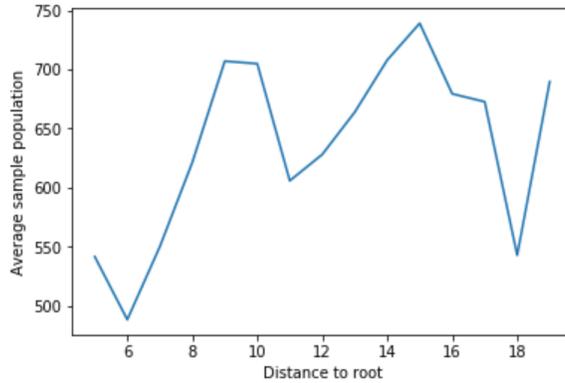


Figure 6: Average sample population per visual class with respect to their "granularity".

procedure to select such image samples. Given a set of labeled samples  $X = \{(x, c)\}$ , our procedure returns a subset  $X' \subset X$  of high-quality images. This selection process is formalized in Algorithm 1, and proceeds as follows:

First, we randomly sample subsets of 1000 visual classes  $C' \subset C$  from the full Imagenet dataset. Classes are sampled so as to contain no overlap in the Wordnet hierarchy: random splits  $C'$  do not contain both parent and their children classes.

Second, we randomly sample 250 images per class as training samples, and use the remaining images as test samples. We fine-tune the last layer of a pre-trained Resnet-50 on the set of training samples, and evaluate the classification output of the model on the test samples.

We consider correctly classified image samples as high-quality test samples for our benchmark and discard the incorrectly classified images. We repeat this operation until all samples  $x \in X$  have been evaluated. The output  $X'$  of this procedure is a subset of high-quality image samples that were correctly classified by the model.

**Input:**

Imagenet Dataset:  $X = \{(x, c) \in \mathbb{R}^{3 \times h \times w} \times C\}$

ILSVRC-pretrained ResNet:  $BaseModel : \mathbb{R}^{3 \times h \times w} \rightarrow C$

**Output:**

High-quality Imagenet subset:  $X' \subset X$

**Init:**

Initialize an empty error set  $Err = \emptyset$  and accurate set:  $Acc = \emptyset$

```

while  $Err \cup Acc \neq X$  do
   $C' = SampleClass(C, 1000)$ 
   $X_{C'} = \{(x, c) | c \in C'\}$ 
   $X_{train}, X_{test} = SampleSplit(X_{C'}, 250)$ 
   $Model = FineTune(BaseModel, X_{train})$ 
  for  $((x, c) \in X_{test})$  {
    if  $Model(x) == c$  then
       $Acc = Acc \cup \{(x, c)\}$ 
    else
       $Err = Err \cup \{(x, c)\}$ 
    end
  }
end
 $X' = Acc$ 

```

**end**

**Algorithm 1:** Sample-wise selection procedure.  $SampleSplit(C, n)$  is a sampling procedure that returns a subset  $C'$  of  $n$  non-overlapping classes (i.e.; no children classes and their parents are contained in  $C'$ ) from the class set  $C$ .  $SampleSplit(X, n)$  is a sampling procedure that returns a training set  $X_{train}$  of  $n$  training samples for each class in  $X$ , and the remaining samples as a test set  $X_{test}$ .  $FineTune(M, X)$  is a procedure that fine-tunes a model  $M$  on the input training set  $X$ .

## Appendix D. Standard benchmark summary

Figure 6 of the main paper summarizes the impact of visual, semantic and structural flaws on the  $top-1$  accuracy of the  $1-hop$  test split.

In these plots, the accuracy score (in green) corresponds to the model accuracy as reported by the standard benchmark. The model error (in orange), represents the classification errors after removing ambiguous images, semantic samples, and structural flaws. For example, the error rate of the GCN model on the generalized setting drops from 90% to 47%. In order to estimate the impact of all three individual factors individually, we ran a set of  $2^3 = 8$  experiments with all possible configurations: with or without considering visual sample quality, semantic sample quality, and structural flaws. The estimated impact reported for each factor corresponds to the mean improvement in classification accuracy brought by this specific factor within all the other factors configuration. Figure 7 and 8 of this supplementary material report similar analysis on the  $top-1$  accuracy of the  $2-hops$  and  $all$  test splits respectively.

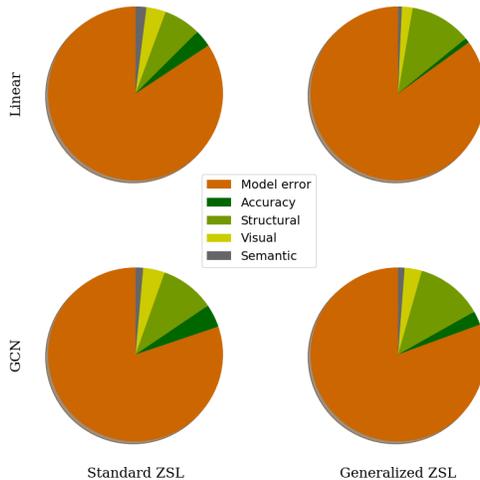


Figure 7: Estimation of the impact of different factors on the reported error of existing models on the  $2$ -hops test split.

## Appendix E. Trivial solution

To apply the trivial solution of the toy example to the standard benchmark, we need a similarity mapping  $f$  between training and test classes. To define such mapping, we used the shortest path length between nodes of the Wordnet hierarchy as a measure of distance  $d$ . We assign to test classes the semantic embedding of their closest training class, as formalized in equations (4.):

$$f : C_{te} \rightarrow C_{tr} \quad (4a)$$

$$f : c \rightarrow \operatorname{argmin}_{c' \in C_{tr}} d(c, c') \quad (4b)$$

$$y_c = y_{f(c)} + e, \forall c \in C_{te} \quad (4c)$$

However, this procedure leads to many test classes sharing the exact same semantic representations. Consider the example of *Cathartid* and *Aegyptidae* classes in Figure 1. Both classes are closest to the *Vulture* training classes so they share the same semantic vector  $y_{Vulture}$ . This leads to undefined behaviors in the classification process. To differentiate between such classes, we add a small Gaussian noise  $e$  to the semantic embeddings of test classes, following equation (4c).

The trivial solution can be implemented by any existing ZSL model using these semantic embeddings. The results reported in the original paper were

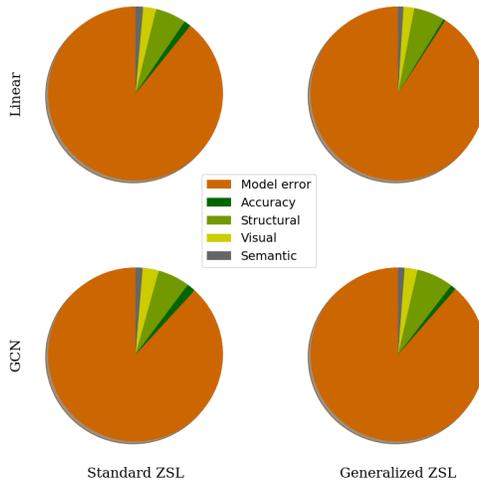


Figure 8: Estimation of the impact of different factors on the reported error of existing models on the *all* test split

computed using the Linear baseline.

## Appendix F. Dataset construction

### Additional considerations

A number of additional factors were taken into consideration in the construction of our proposed benchmark. For space constraints, we could not include these considerations in the original paper, so we briefly present them in this Appendix.

**Sample population:** The number of images per test class in the standard benchmark’s test splits is very uneven. Some test classes have as little as one sample image, while some classes have thousands of images. This leads to highly biased evaluations as test classes of high sample population have a larger impact on the reported classification accuracy. We select 100 quality samples for each test class to ensure an evenly distributed test set.

**Mutual exclusion:** To prevent false negative classification outputs, test classes should be mutually exclusive. The hierarchical structure of Wordnet allows us to automatically create test splits that do not include both parent and test classes, so we can automatically remove such mutually non-exclusive classes from the test sets. However, this is not sufficient to guarantee the mutual exclusivity of test classes. For example, the Imagenet dataset includes classes

such as *Man*, *Woman*, *White Person*, or *Engineer*. We do not want to include such kinds of classes in our benchmark because classifying an image of *Woman* as *White Person* or *Engineer* would result in false negative outputs. These classes, although not directly related to each other in the Wordnet hierarchy, are not mutually exclusive. The Wordnet hierarchy does not provide the logical constructs to automatically detect such instances, so we manually inspect the set of candidate test classes and remove them from the test set.

**Scale considerations:** We favor images of generic objects captured at the scale of human perceptions: we remove classes of images taken at microscopic scale (biological cells, bacteria, etc.), or classes of images at astronomical scales (supernova).

**Shape considerations:** We favor objects that can be recognized by their characteristic shape and remove classes that require reading comprehension to identify. For example, we remove a number of medicines, such as *VitaminD* or branded contents like *Pepsi Cola*. Figure 11 illustrates a few such classes.

## Dataset construction Summary

Table 2 summarizes the different steps of the creation of our benchmark. It details the level of automation, the different parameters involved in each step, as well as the approximate ratio of visual classes selected within each of these steps.

Table 2: Summary of the benchmark construction steps

	Step	Automation	Parameters	Filter ratio
Semantic	Frequency	Auto	$f > 500$	82%
	Polysemy	Auto	-	91%
Visual	Class-wise	Auto	$n > 300$	63%
	Sample-wise	Auto	$n_C = 1000, n_{tr} = 250$	100%
	Shape	Manual	-	95-99%
	Scale	Manual	-	99%
Structural	Hierarchy	Auto	-	82%
	Mutual Exclusivity	Manual	-	95-99%

The majority of the visual classes filtered out from our benchmark were automatically discarded based on their weak semantic features, low sample population or structural constraints to avoid both parents and children classes be included in the test set. Only the semantic and visual sample selection steps are parameterized. We select word labels occurring at least 500 times within the Wikipedia corpus to avoid rare words. We only select visual classes with a sample population superior to 300 images.

## Appendix G. Code & Data

The full Imagenet dataset, as considered in the *all* test split consists of over 13 million images, which is very time-consuming to download and process. In contrast, small-scale benchmarks like AWA, CUB or SUN come with off-the-shelf semantic and visual features. Furthermore, they are orders of magnitude smaller than the Imagenet dataset which makes it much easier for researchers to evaluate their models on. As a result, many recent works on ZSL have only reported the evaluation of their models on small-scale benchmarks, instead of the standard Imagenet benchmark.

To encourage researchers working on ZSL to evaluate their model on our proposed benchmark, we release pretrained semantic and visual features<sup>1</sup>. The dataset is small enough to fit in the memory of most modern computer hardware so it allows for fast prototyping and evaluation. To work on the original raw images, we provide the URL of test images with a Python script for download.

In addition to this data, we also provide code for visual class selection and fast manipulation of the Wordnet hierarchy. This should allow researchers interested in the investigation of different factors impacting ZSL accuracy to quickly build different test splits.

## References

- [1] Xian, Yongqin and Schiele, Bernt and Akata, Zeynep, Zero-shot learning-the good, the bad and the ugly. arXiv:1703.04394, 2017.

---

<sup>1</sup>Download instructions are available at <https://github.com/TristHas/GOZ>

Labels	Definition
queen, queen regnant, female monarch	A female sovereign ruler
	
queen	The only fertile female in a colony of social insects
	
queen	(chess) The most powerful piece
	
queen	One of four face cards in a deck bearing a picture of a queen
	
head, caput	The upper part of the human body or the front part of the body in animals
	
head	The striking part of a tool
	

Figure 9: Examples of polysemous classes

N. Images	Labels	Definition
1	Platform	Combination of a particular computer and operating system
		
8	Continental divide	The watershed of a continent
		
8	Amoralist	Someone who adheres to the doctrine that ordinary moral distinctions are invalid
		
1	Groove, Channel	a long narrow furrow cut either by a natural process or by a tool
		

Figure 10: Examples of low sample population, visually ambiguous classes.

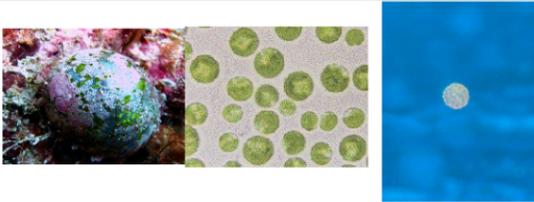
Labels	Definition
supernova	A star that explodes and becomes extremely luminous in the process
	
cell	(biology) The basic structural and functional unit of all organisms
	
vitamin C, C, ascorbic acid	Vitamin found in fresh fruits
	
vitamin D, calciferol, viosterol, ...	A fat-soluble vitamin that prevents rickets
	
Pepsi, Pepsi Cola	Pepsi Cola is a trademarked cola
	

Figure 11: Examples of manually discarded classes. Cell and Supernova correspond to microscopic and astronomic scale images. Vitamin D, Vitamin C, and Pepsi were discarded as they require reading comprehension to identify.