On zero-shot recognition of generic objects

Tristan Hascoet tristan.hascoet@gmail.com
Yasuo Ariki ariki@kobe-u.ac.jp
Tetsuya Takiguchi takigu@kobe-u.ac.jp
Graduate School of System Informatics, Kobe University, Japan

Abstract

Many recent advances in computer vision are the result of a healthy competition among researchers on high quality, task-specific, benchmarks. After a decade of active research, zero-shot learning (ZSL) models accuracy on the Imagenet benchmark remains far too low to be considered for practical object recognition applications. In this paper, we argue that the main reason behind this apparent lack of progress is the poor quality of this benchmark. We highlight major structural flaws of the current benchmark and analyze different factors impacting the accuracy of ZSL models. We show that the actual classification accuracy of existing ZSL models is significantly higher than was previously thought as we account for these flaws. We then introduce the notion of structural bias specific to ZSL datasets. We discuss how the presence of this new form of bias allows for a trivial solution to the standard benchmark and conclude on the need for a new benchmark. We then detail the semi-automated construction of a new benchmark to address these flaws.

1. Introduction

Datasets play a leading role in computer vision research. Perhaps the most striking example of the impact a dataset can have on research has been the introduction of Imagenet [2]. The new scale and granularity of Imagenet's coverage of the visual world has paved the way for the success and wide spread adoption of CNN [8, 11] that have revolutionized generic object recognition.

The current best-practice for the development of a practical object recognition solution consists in collecting and annotating application-specific training data to fine-tune a large Imagenet-pretrained CNN on. This data annotation process can be prohibitively expensive for many applications which hinders the wide-spread usage of these technologies. ZSL models generalize the recognition ability of traditional image classifiers to unknown classes, for which no image sample is available for training. The promise of ZSL for generic object recognition is huge: to scale up the recognition capacity of image classifiers beyond the set of annotated training classes. Hence ZSL has the potential to be of great practical impact as they would considerably ease the deployment of object recognition technologies by eliminating the need for expensive task-specific data collection and fine-tuning processes.

Despite its great promise, and after a decade of active research [10], the accuracy of ZSL models on the standard Imagenet benchmark [3] remain far too low for practical applications. To better understand this lack of progress, we analyzed the errors of several ZSL baselines. Our analysis leads us to identify two main factors impacting the accuracy of ZSL models: structural flaws in the standard evaluation protocol and poor quality of both semantic and visual samples. On the bright side of things, we show that once these flaws are taken into account, the actual accuracy of existing ZSL models is much higher than was previously thought.

On the other hand, we show that a trivial solution outperforms most existing ZSL models by a large margin, which is upsetting. To explain this phenomenon, we introduce the notion of structural bias in ZSL datasets. We argue that ZSL models should aim to develop compositional reasoning abilities, but the presence of structural bias in the Imagenet benchmark favors solutions based on a trivial one to one mapping between training and test classes. We come to the conclusion that a new benchmark is needed to address the different problems identified by our analysis and, in the last section of this paper, we detail the semi-automated construction of a new benchmark we propose.

To structure our discussion, we first briefly review preliminaries on ZSL in Section 3. Section 4 details our analysis of the different factors impacting the accuracy of ZSL models on the standard benchmark. Section 5 introduces the notions of structural bias, and propose a way to measure and minimize its impact in the construction of a new benchmark. Finally, Section 6 summarizes the construction of our proposed benchmark. For space constraint, we only include the main results of our analysis in the body of this paper. We refer interested readers to the supplementary material for additional results and details of our analysis.

2. Related Work

2.1. ZSL datasets

Early research on ZSL has been carried out on relatively small scale or domain specific benchmarks [9, 14, 19], for which human-annotated visual attributes are proposed as semantic representations of the visual classes. On the one hand, these benchmarks have provided a controlled setup for the development of theoretical models and the accurate tracking of ZSL progress. On the other hand, it is unclear whether approaches developed on such dataset would generalize to the more practical setting of zero-shot generic object recognition. For instance, in generic object recognition, manually annotating each and every possible visual class of interest with a set of visual attributes is impractical due to the diversity and complexity of the visual world.

The Imagenet dataset [2] consists of more than 13 million images scattered among 21,845 visual classes. Imagenet relies on Wordnet [12] to structure its classes: each visual class in Imagenet corresponds to a concept in Wordnet. Frome *et al*. [3] proposed a benchmark for ZS generic object recognition based on the Imagenet dataset, which has been widely adopted as the standard evaluation benchmark by recent works [13, 20, 15, 1, 21, 7, 18]. Using word embeddings as semantic representations, they use the 1000 classes of the ILSVRC dataset as training classes and propose different test splits drawn from the remaining 20,845 classes of the Imagenet dataset based on their distance to the training classes within the Wordnet hierarchy: the *2-hops*, *3-hops* and *all* test splits.

Careful inspection of these test splits revealed a confusion in their name: The *2-hops* test split actually consists of the set of 1589 test classes directly connected to the training set classes in Wordnet, i.e; within *1 hop* of the training set. Similarly, the *3-hops* test set actually corresponds to the test classes within *2-hops*. In this paper, we will refer to the standard test splits by the name of their true configuration: *1-hop*, *2-hops* and *all*, as illustrated in Figure 1.

2.2. Dataset bias

Bias in datasets can take many forms, depending on the specific target task. Torralba *et al*. [17] investigates bias in generic object recognition. The notion of structural bias we introduce in Section 5 is closely related to the notion of negative set bias they analyze.

As more complex tasks are being considered, more insidious forms of bias sneak into our datasets. In VQA, the impressive results of early baseline models have later been shown to be largely due to statistical biases in the question/answers pairs [4, 6, 5]. Similar to these works, we will show that a trivial solution leveraging structural bias in the Imagenet ZSL benchmark outperforms early ZSL baselines.

Xian *et al*. [21] identify structural incoherences in small-scale ZSL benchmarks and proposes new test splits to remedy them. Closely related to our work, they also observe a correlation between test class sample population and classification accuracy in the Imagenet ZSL benchmark. However, their analysis mainly focuses on small-scale benchmarks and the comparison of existing ZSL models, while we analyze the ZSL benchmark for generic object recognition in more depth.

3. Preliminaries

ZSL models aim to recognize unseen classes, for which no image sample is available to learn from. To do so, ZSL models use descriptions of the visual classes, i.e., representations of the visual classes in a semantic space shared by both training and test classes. To evaluate the out-of-sample recognition ability of models, ZSL benchmarks split the full set of classes C into disjoint training and test sets. ZSL benchmarks are fully defined by three components: a set of training and test classes (C_{tr}, C_{te}), a set of labeled images X, and a set of semantic representations Y:

$$C_{tr} \cup C_{te} \subset C \tag{1a}$$
$$C_{tr} \cap C_{te} = \emptyset \tag{1b}$$
$$Y = \{y_c \in \mathbb{R}^d \quad \forall c \in C\} \tag{1c}$$
$$X = \{(x, c) \in \mathbb{R}^{3\times h\times w} \times C\} \tag{1d}$$
$$Tr = \{(x, y_c) \mid c \in C_{tr}\} \tag{1e}$$
$$Te = \{(x, y_c) \mid c \in C_{te}\} \tag{1f}$$

ZSL models are typically trained to minimize a loss function $\mathcal{L}$ over a similarity score E between image and semantic features of the training sample set with respect to the model parameters θ.

$$\theta^* = argmin_\theta \mathbb{E}_{(x,y)\in Tr} \mathcal{L}(E_\theta(x, y) + \Omega(\theta)) \tag{2}$$

In the standard ZSL setting, test samples x_{te} are classified among the set of unseen test classes by retrieving the class description y of highest similarity score:

$$c = argmax_{c\in C_{te}} E(x_{te}, y_c) \tag{3}$$

In the generalized ZSL setting, test samples are classified among the full set of training and test classes:

$$c = argmax_{c\in C} E(x_{te}, y_c) \tag{4}$$

Xian *et al*. [20] have shown that many ZSL models can be formulated within a same linear model framework, with different training objectives and regularization terms. More recently, Wang *et al*. [18] have proposed a Graph Convolutional Network (GCN) model that has shown impressive improvements over the previous state of the art. In our study, we will present results obtained with both a baseline linear model [15] and a state of the art GCN model [18, 7].

4. Error analysis

In the previous section, we have mentioned that ZSL benchmarks are fully defined by three components: a set of labeled images X, a set of semantic representations Y, and the set of training and test classes (C_{tr}, C_{te}). In this section, we analyze each of the standard benchmark components individually: We first highlight inconsistencies in the configuration of the different test splits and show that these inconsistencies lead to many false negatives in the reported evaluation of ZSL models outputs. Next, we identify a number of factors impacting the quality of the word embeddings of visual classes and argue that visual classes with poor semantic representations should be excluded from ZSL benchmarks. We then observe that the Imagenet dataset contains many ambiguous image samples. We define what a *good* image sample means in the context of ZSL and propose a method to automatically select such images.

4.1. Structural flaws

Figure 1 illustrates the configuration of test classes of the standard test splits within the Wordnet hierarchy. This configuration leads to an obvious contradiction: test sets include visual classes of both parents and their children concepts. Consider the problem of classifying images of birds within the *hop-1* test split as in Figure 1. The standard test splits give rise to two possibly inconsistent scenarios:

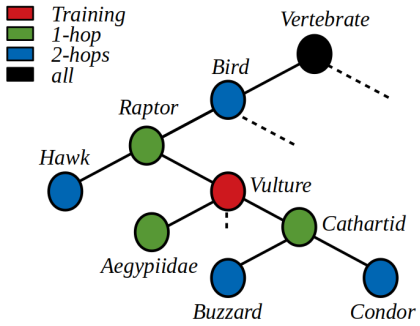

Figure 1. Illustration of the standard test splits configuration

A ZSL model may classify an image of the children class *Cathartid* as its parent class *Raptor*. The standard benchmark considers such cases as classification errors, while the classification is semantically correct.

A ZSL model may classify an image of the parent class *Raptor* as one of its children class: *Cathartid*. Classification may be semantically correct or incorrect, depending on the specific breed of raptor in the image, but we have no way to automatically assess it without additional annotation. The standard benchmark considers such cases as classification errors, while the classification is semantically undefined.

We refer to both of the above cases as false negatives. Figure 2 illustrates the distribution of ZSL classification

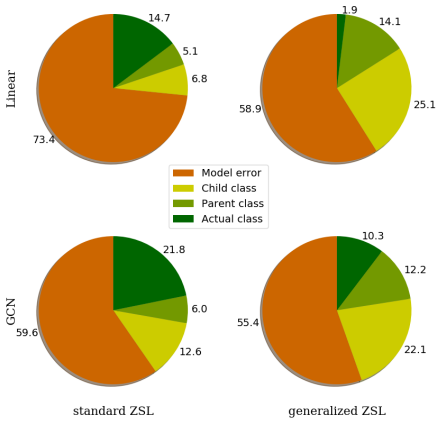

Figure 2. Distribution of the classification outputs of different ZSL models on the *1-hop* test split. An image x can be either be classified into its actual label c, the parent class of c, one of its children class, or an unrelated class. Only the latter case constitutes a definitive error.

outputs among these different scenarios on the 1-hop test split. On the standard ZSL task for instance, the reported accuracy of the GCN model is 21.8% while the actual (semantically correct) accuracy should be somewhere in between 27.8% and 40.4%.

The ratio of false negatives per accuracy increases dramatically in the generalized ZSL setting. The linear baseline reported accuracy is only 1.9%, while the actual (semantically correct) accuracy lies between 16.0% and 41.1%. This is due to the fact that ZSL models tend to classify test images into their parent or children training class: for example, *Cathartid* images tend to be classified as *Vulture*. Appendix A of the supplementary material presents results on the other standard splits on which we show that the ratio of false negative per reported accuracy further increases with with larger test splits.

4.2. Word embeddings

In this section, we identify two factors impacting the quality of word embeddings and analyse their affect on ZSL accuracy: polysemy and occurrence frequency. These problems naturally arise in the definition of large scale object categories so they are inherent problems of ZS recognition of generic objects. However, we argue that ZSL benchmarks should provide a curated environment with high quality, unambiguous, semantic representations and that solutions to tackle the special case of polysemous and rare words should be separately investigated in the future.

4.2.1 Occurrence frequency

Word embeddings are learned in an unsupervised manner from the co-occurrence statistics of words in large text corpora. Common words are learned from plentiful statistics so we expect them to provide more semantically meaningful representations than rare words, which are learned from scarce co-occurrence statistics. We found many Imagenet class labels to be rare words (see Appendix B of the supplementary materials), with as many as 33.7% of label words appearing less than 50 times in Wikipedia. Here, we question whether the few co-occurrence statistics from which such rare word embeddings are learned actually provide any visually discriminative information for ZSL.

To answer this question, we evaluate ZSL models on different test splits of 100 classes: we split the Imagenet classes into different subsets based on the occurrence frequency of their label word. We independently evaluate the accuracy of our model on each of these splits and report the ZSL accuracy with respect to the average occurrence frequency of the visual class labels in Figure 3.

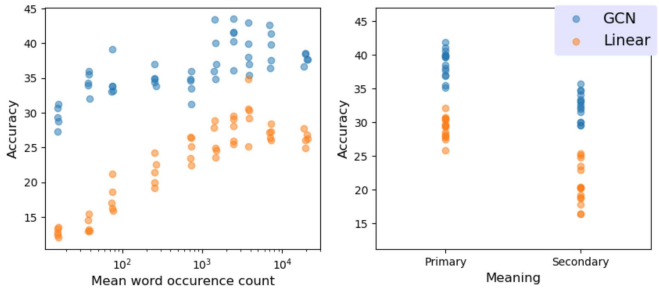

Figure 3. Each dot in these figures represent the top-1 accuracy (y-axis) of a 100 classes test split with respect to the test split characteristics (x-axis): Left: Mean occurrence frequency of the test class labels. Right: test classes of primary meaning, such as cairn (monument), or secondary meaning, such as cairn (dog)

Our results highlight a strong correlation ($r = 0.89$) between word frequency and the Linear baseline accuracy as test splits made of rare words strikingly under-perform test splits made of more common words, although accuracy remains well above chance (1%), even for test sets of very rare words. Results are more nuanced for the GCN model (correlation coefficient $r = 0.74$), which can be explained by the fact that GCN uses the Wordnet hierarchy information in addition to word embeddings.

4.2.2 Polysemy

The English language contains many polysemous words, which makes it difficult to uniquely identify a visual class with a single word. We found that half of the ImageNet word labels are shared with at least one other Wordnet concept, and that 38% of ImageNet classes share at least one word label with other visual classes. Figure 4 illustrates the example of the word "cairn". Two visual classes share the same label "cairn": One relates to the meaning of cairn as a stone memorial, while the other refers to a dog breed. This is problematic as both of these visual classes share the same representation in the label space, so they are essentially defined as the same class although they correspond to two visually very distinct concepts.

To deal with polysemy, we assume that all words have one primary meaning, with possibly several secondary meanings. We consider word embeddings to reflect the semantics of their primary meaning exclusively, and discard visual classes associated with the secondary meanings of their word label. To automatically identify the first meaning of visual class labels. we implement a solution based on both Wordnet and word embeddings statistics detailed in the supplementary material.

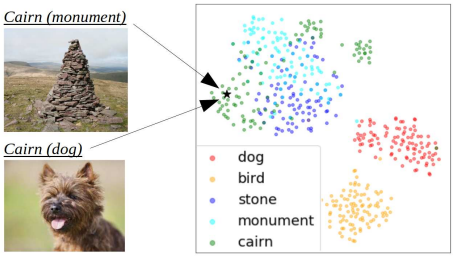

Figure 4. Illustration of polysemous words. Each color represents the 100 nearest neighbors of a given word. "Cairn" and its closest neighbors are clustered around the stone and monument related vocabulary, far away from dog-related vocabulary so we assign the top visual class as primary meaning of the word cairn.

We conduct an experiment to assess both the impact of polysemy on ZSL accuracy and the efficiency of our solution. As in the previous section, we evaluate our ZSL models on different test splits of 100 classes: We separately evaluate test classes identified as the primary meaning of their word label and test classes corresponding to the secondary meaning of their word label. Figure 3 reports the accuracy obtained on these different test splits. We can see a significant boost in the ZSL accuracy of test classes whose word labels are identified as primary meanings. In comparison, test splits made exclusively of secondary meanings performed poorly. This confirms that polysemy does indeed impact ZSL accuracy, and suggests that our solution for primary meaning identification allows addressing this problem.

4.3. Image samples

The ILSVRC dataset consists of a high-quality curated subset of the Imagenet dataset. The current ZSL benchmark uses ILSVRC classes as training classes and classes drawn from the remainder of the Imagenet dataset as test

sets, assuming similar standards of quality from these test classes. Upon closer inspection, we found these test classes to contain many inconsistencies and ambiguities. In this section, we detail a solution to automatically filter out ambiguous samples so as to only select quality samples for our proposed benchmark.

4.3.1 Class-wise selection

Xian *et al*. [20] have first identified a correlation between the sample population of visual classes and their classification accuracy. They conjecture that small population classes are harder to classify because they correspond to fine-grained visual concepts, while large population classes correspond to easier, coarse-grained concepts. Manual inspection of these classes lead us to a different interpretation: First, we found no significant correlation between sample population and concept granularity (Appendix C). For example, fine-grained concepts such as specific species of birds or dogs tend to have high sample populations. On the other hand, we found many visually ambiguous concepts such as "ringer", "covering" or "chair of state" to have low sample populations. Such visually ambiguous concepts are harder for crowd-sourced annotators to reach consensus on labeling, resulting in lower population counts.

In Figure 5, we report the ZSL accuracy of our models on different test splits with respect to their average population counts. This figure shows a clear correlation between the sample population and the accuracy of both models, with low accuracy for low sample population classes. We use the sample population as a rough indicator to quickly filter out ambiguous visual classes and only consider classes with sample population superior to 300 images as valid candidate classes in our proposed dataset.

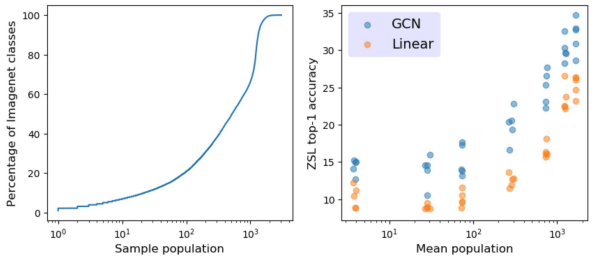

Figure 5. ZSL accuracy with respect to sample population sizes. Left: Distribution of Imagenet class population size. 6.1% of Imagenet classes have less than 10 samples, 21.1% have less than 100 samples. Right: ZSL accuracy of different test splits with respect to their mean sample population size.

4.3.2 Sample-wise selection

Even among the selected classes, we found many inconsistent and ambiguous images to remain (Appendix C), so we would like to further filter quality test images sample-wise. But what makes a good candidate image for a ZSL benchmark? How can we measure the quality of a sample? We argue that ZSL benchmarks should only reflect the *zero-shot ability* of models: ZSL benchmarks should evaluate the accuracy of ZSL models *relatively to the accuracy of standard non-ZSL models*. Hence, we define a good ZSL sample as an image unambiguous enough to be correctly classified by standard image classifiers trained in a supervised manner.

To automatically filter such quality samples, we fine-tune and evaluate a standard CNN in a supervised manner on the set of candidate test classes. We consider consistently misclassified samples to be too ambiguous for ZSL and only select samples that were correctly classified by the CNN Details of this selection process are presented in Appendix C of the supplementary material.

4.4. Dataset Summary

Figure 6 summarizes the impact of the different factors we analyzed on the top-1 classification error of both our baseline models on the "1-hop" test split. The error rate of the Linear model on the standard ZSL setting drops from 86% to 61% after removing ambiguous images, semantic samples, and structural flaws. The error rate of the GCN model on the generalized setting drops from 90% to 47%.

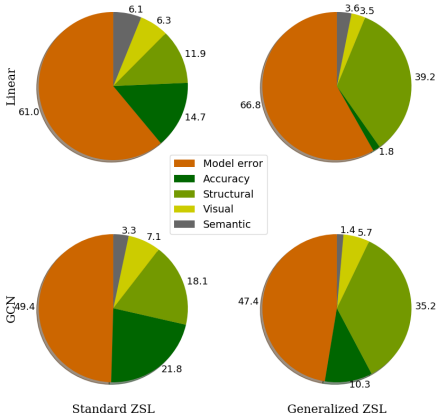

Figure 6. Estimation of the impact of different factors on the reported error of existing models on the 1-hop test split

The GCN model is particularly sensitive to the structural flaws of the standard benchmark, but less sensitive to noisy word embeddings than the linear baseline. This can be easily explained by the fact that GCN models rely on the explicit Wordnet hierarchy information as semantic data in addition to word embeddings. Additional results and details on the methodology of our analysis are given in Appendix D of the supplementary material.

5. Structural bias

ZSL models are inspired by the human ability to recognize unknown objects from a mere description, as it is often illustrated by the following example: Without having ever seen a zebra, a person would be able to recognize one, knowing that zebras look like horses covered in black and white stripes. This example illustrates the human capacity to *compose* visual features of *different* known objects to define and recognize previously unknown object categories.

Standard image classifiers encode class labels as *local representations* (one-hot embeddings), in which each dimension represents a different visual class, as illustrated in Figure 8. As such, no information is shared among classes in the label space: visual class embeddings are equally distant and orthogonal to each other. The main idea behind ZSL models is to instead embed visual classes into *distributed representations*: In label space, visual classes are defined by multiple visual features (horse-ish shape, stripes, colors) shared among classes. Distributed representations allow to define and recognize unknown classes by *composition* of visual features shared with known classes, in a similar manner as the human ability described above.

The embedding of visual classes into distributed feature representations is especially powerful since it allows to define a combinatorial number of test classes by composition of a possibly small set of features learned from a given set of training classes. Hence, we argue that the key challenge behind ZSL is to achieve ZS recognition of unknown classes by composition of known visual features, following their original inspiration of the human ability, and as made possible by distributed feature representations. In this section, we will see that not all ZSL problems require such kind of compositional ability. On the standard benchmark, we show that a trivial solution based on local representations of visual classes outperform existing approaches based on word embeddings. We show that this trivial solution is made possible by the specific configuration of the standard test splits and introduce the notion of structural bias to refer to the existence of such trivial solutions in ZSL datasets.

5.1. Toy example

Figure 7 illustrates a toy ZSL problem in which, given a training set of *Horse* and *TV monitor* images, the goal is to classify images of *Zebra* and *PC laptop*. Let's consider training an image classifier on the training set and directly applying it to images from the test set. We can safely assume that most zebra images will be classified as horses, and most laptop samples as TV monitors. Hence, a trivial solution to this problem consists in defining a one to one mapping between test classes and their closest training class: *Horse=Zebra* and *TV monitor=PC laptop*. This example makes it fairly obvious that not all ZSL problems require the ability to compose visual features to solve.

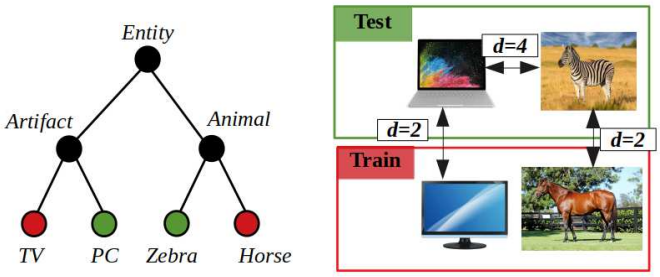

Figure 7. Illustration of the toy example. Left: Wordnet-like class hierarchy. Training classes are shown in red and test class in green. Right: Illustration of image samples. The black captions represent the distance between classes as their shortest path length.

Classification problems define a close-world assumption: As all test samples are known to belong to one of the test classes, classifying an image x into a given test class c means that x is more likely to belong to c than other classes of the test set. In other words, classification is performed relatively to a negative set of classes [17]. What made this trivial ZSL solution possible is the fact that test classes of our toy example are very similar to one of the training class, relatively to their negative set. This allowed us to identify a one-to-one mapping by similarity between training and test classes. We refer to this trivial solution as a similarity-based solution, in opposition to solutions based on the composition of visual features.

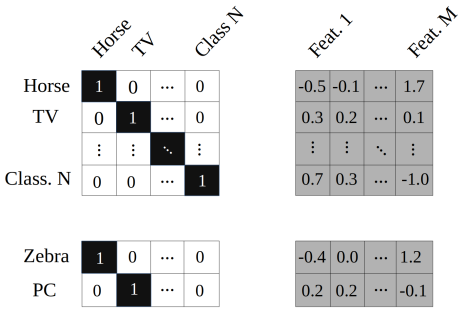

Figure 8. Illustration of local (one-hot, on the left) and distributed (right) representations of visual classes. The similarity-based solution encodes both training and test classes as local representations. Composition-based solutions need distributed representations.

As illustrated in Figure 8, the similarity mapping between test and training classes can be directly embedded in the semantic space using local representations. The trivial solution consists in assigning to test classes the exact same semantic representation as their most similar training class. Consider applying these semantic embeddings within a ZSL framework to our toy problem: classifying a test image x as a Horse relatively to the negative set of TV within the training set becomes strictly equivalent to classifying x as Zebra relatively to its negative set PC within the test set. Hence,

any existing ZSL model using these local embeddings instead of distributed representations like word embeddings Y would converge to the same solution.

5.2. Standard benchmark

Besides our toy example, how well would this trivial solution perform on the standard benchmark? To implement it, we used the Linear baseline model [15] with local representations inferred from the Wordnet hierarchy (see Appendix E), but any model would essentially converge to a similar solution. Table 1 compares the accuracy of this trivial solution to state of the art models as reported in [21, 7]. The trivial similarity-based solution outperforms existing ZSL models by a significant margin. Only GCN-based models [7], which we discuss in the next section, seem to outperform our trivial solution.

Table 1. Top-1 accuracy on the standard test splits (top) as reported for linear baselines in [21], (middle) as reported for GCN-based models in [7] and (down) obtained by our trivial solution

model	1-hop	2-hops	all
SYNC [1]	9.26	2.29	0.96
CONSE [13]	7.63	2.18	0.95
ESZSL [15]	6.35	1.51	0.62
LATEM [20]	5.45	1.32	0.5
DEVISE[3]	5.25	1.29	0.49
CMT [16]	2.88	0.67	0.29
GCNZ [18]	19.8	4.1	1.8
ADGPM [7]	**26.6**	**6.3**	**3.0**
Trivial	20.27	3.59	1.53

5.3. Measuring structural bias

In our toy example, we have hinted at the fact that structural bias emerges for test sets in which test classes are relatively similar to training classes, while being comparably more dissimilar to each other (to their negative set). To confirm this intuition, we define the following structural ratio:

$$r(c) = \frac{min_{c' \in C_{tr}} d(c, c')}{min_{c' \in C_{te}} d(c, c')} \tag{5a}$$

$$R(C_{te}) = \frac{1}{|C_{te}|} \sum_{c \in C_{te}} r(c) \tag{5b}$$

In which c represents a visual class, C_{te} and C_{tr} represent test and training sets respectively, and d is a distance reflecting similarity between two classes. Here, $r(c)$ represents the ratio of the distance between c and its closest training class to the distance between c and its closest test class. In our experiments, we use the the shortest path length between two classes in the Wordnet hierarchy as a measure of distance d, although different metrics would be interesting to investigate as well. We compute the structural ratio of a test set $R(C_{te})$ as the mean structural ratio of its individual classes. Figure 9 shows the top-1 accuracy achieved by baseline models on different test sets with respect to their structural ratio R. As for previous experiments, we report our results on test splits of 100 classes.

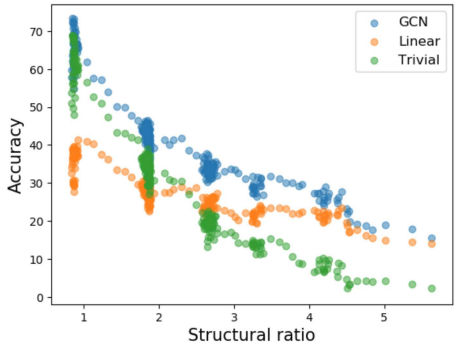

Figure 9. ZSL accuracy on different test sets with respect to their structural ratio $R(C_{te})$.

On test splits of low structural ratio, the trivial solution performs remarkably well, on par with the state of the art GCN model. Such test splits are similar to the toy example in which each test class is closely related to a training class while being far away from other test classes in the Wordnet hierarchy. As an example, the structural ratio of the test split in our toy example is $R(C_{te}) = 1/2 \times (2/4 + 2/4) = 0.5$, which corresponds to the highest accuracies achieved by the trivial solution. We say that such test split is structurally biased towards similarity-based trivial solutions.

However, the accuracy of the similarity-based trivial solution decreases sharply with the structural ratio until it reaches near chance accuracy for the highest ratios. Hence maximizing the structural ratio of test splits seems to be an efficient way to minimize structural bias. Although their accuracy decrease with larger structural ratios, both GCN and Linear models remain well above chance. These results suggest that ZSL models based on word embeddings are indeed capable of compositional reasoning. At the very least, they are able to perform more complex ZSL tasks than the trivial similarity-based solution. Interestingly, as the trivial solution converges towards chance accuracy, the GCN model accuracy seems to converge towards the accuracy of the ZSL baseline. This suggests that the main reason behind the success of GCN models is that they efficiently leverage the Wordnet hierarchy to exploit structural bias.

The *1-hop* and *2-hops* test splits of the standard benchmark consist of the set of test classes closest to the training classes within the Wordnet hierarchy. This leads to test splits of very low structural ratio, similar to our toy example. For instance, the *1-hop* test split has a structural ratio of 0.55. It is an example of structural bias even more extreme than our toy example as test classes are either children or parent classes of a training class. In the next section, we

propose a new benchmark with maximal structural ratio in order to minimize structural bias.

6. New Benchmark

6.1. Proposed Benchmark

In this section, we briefly detail the semi-automated construction of a new benchmark designed to fix the different flaws of the current benchmark highlighted by our analysis. For space constraints, a number of minor considerations could not be properly presented in this paper. We detail these additional considerations in Appendix F of the supplementary material. Appendix F also provides additional details regarding the different parameters and the level of automation of each of the construction process. Appendix G provides details on the code and data we release. Following Frome *et al*. [3], we use the ILSVRC dataset as training set, and propose a new test set. The selection of this new test set proceeds in two steps:

In a first step, we select a subset of candidate test classes $C' \subset C$ from the remaining 20,845 Imagenet classes based on the statistics of image samples and word labels: We first filter out semantic samples $Y' \subset Y$ corresponding to rare or polysemous words of secondary meaning (Section 4.2). We then discard visual classes of low sample population and filter out ambiguous image samples using supervised learning to select $X' \subset X$ (Section 4.3). The set of candidate test classes is the subset of visual classes $C' \subset C$ for which sufficiently high quality image and semantic samples were selected.

In a second step, we define the test split $C_{te} \subset C'$ as a *structurally consistent* set of *minimal structural bias*: The test set was carefully selected so as to contain no overlap among its own classes nor with the training classes in order to provide a structurally consistent test set for the generalized ZSL setting. This test set consists of 500 classes of maximal structural ratio $R(C_{te})$ so as to minimize structural bias.

6.2. Evaluation

Table 2. Evaluation on the proposed benchmark. Accuracy in the generalized ZSL setting are reported as harmonic means over training and test accuracy following [21]

Model	ZSL		G-ZSL	
	@1	@5	@1	@5
Trivial	1.2	3.9	0	0
CONSE [13]	10.65	25.10	0.12	19.34
DEVISE [3]	11.15	29.52	**7.87**	26.10
ESZSL [15]	13.54	32.61	4.59	25.53
GCN-6 [18]	9.58	27.19	4.81	23.35
GCN-2 [7]	14.09	35.12	4.96	**30.35**
ADGPM [7]	**14.10**	**36.03**	4.90	29.96

Table 2 presents the evaluation of a number of baseline models on the newly proposed benchmarks. A few notable results stand out from this table: First, different from the standard benchmark, CONSE [13] performs worse than DEVISE [3]. The relatively high accuracy reported by the CONSE model on the standard benchmark is most likely due to the fact that word embeddings of test classes are statistically close to the word embedding of their parent/children test classes so that CONSE results more closely fit the trivial similarity-based trivial solution. We expect model averaging methods to benefit the most from the structural bias in the standard benchmark.

Second, the impressive improvements reported by GCN-based models over linear baselines are significantly reduced, although GCN models still outperform linear baselines. This result corroborates the observation, in Section 5, that GCN models tend to converge towards the results of linear baseline models for high structural ratio.

7. Conclusion and Discussion

ZSL has the potential to be of great practical impact for object recognition. However, as for any computer vision task, the availability of a high quality benchmark is a prerequisite for progress. In this paper, we have shown major flaws in the standard generic object ZSL benchmark and proposed a new benchmark to address these flaws. More importantly, we introduced the notion of structural bias in ZSL dataset that allows trivial solutions based on simple similarity matching in semantic space. We encourage researchers to evaluate their past and future models on our proposed benchmark. It seems likely that sound ideas may have been discarded for their poor performance relative to baseline models that benefited most from structural bias. Some of these ideas may be worth revisiting today.

Finally, we believe that a deeper discussion on the goals and the definition of ZSL is still very much needed. There is a risk in developing complex models to address poorly characterized problems: Mathematical complexity can act as a smokescreen of complexity that obfuscates the real problems and key challenges behind ZSL. Instead, we believe that practical considerations grounded in common sense are still very much needed at this stage of ZSL research. The identification of structural bias is a first step towards a sound characterization of ZSL problems. One practical way to continue this discussion would be to investigate structural bias in other ZSL benchmarks.

Aknowledgement

This work was supported in part by JSPS KAKENHI (Grant No. JP17K00236 and No. JP17H01995).

References

[1] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5327–5336, 2016.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009.

[3] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[4] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016.

[5] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1988–1997. IEEE, 2017.

[6] K. Kafle and C. Kanan. An analysis of visual question answering algorithms. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1983–1991. IEEE, 2017.

[7] M. Kampffmeyer, Y. Chen, X. Liang, H. Wang, Y. Zhang, and E. P. Xing. Rethinking knowledge graph propagation for zero-shot learning. *arXiv preprint arXiv:1805.11724*, 2018.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[9] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009.

[10] H. Larochelle, D. Erhan, and Y. Bengio. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3, 2008.

[11] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[12] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[13] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[14] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE, 2012.

[15] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

[16] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013.

[17] A. Torralba and A. A. Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011.

[18] X. Wang, Y. Ye, and A. Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6857–6866, 2018.

[19] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-ucsd birds 200. 2010.

[20] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016.

[21] Y. Xian, B. Schiele, and Z. Akata. Zero-shot learning-the good, the bad and the ugly. *arXiv preprint arXiv:1703.04394*, 2017.