

ZEST: Zero-shot Learning from Text Descriptions using Textual Similarity and Visual Summarization

Tzuf Paz-Argaman¹, Yuval Atzman², Gal Chechik^{1,2}, and Reut Tsarfaty^{1,3}

¹Bar-Ilan University

²NVIDIA Research

³Allen Institute for Artificial Intelligence

{tzufar, reut.tsarfaty}@gmail.com

Abstract

We study the problem of recognizing visual entities from the textual descriptions of their classes. Specifically, given birds’ images with free-text descriptions of their species, we learn to classify images of previously-unseen species based on specie descriptions. This setup has been studied in the vision community under the name *zero-shot learning from text*, focusing on learning to transfer knowledge about visual aspects of birds from seen classes to previously-unseen ones. Here, we suggest focusing on the textual description and distilling from the description the most relevant information to effectively match visual features to the parts of the text that discuss them. Specifically, (1) we propose to leverage the *similarity* between species, reflected in the similarity between text descriptions of the species. (2) we derive *visual summaries* of the texts, i.e., extractive summaries that focus on the *visual* features that tend to be reflected in images. We propose a simple attention-based model augmented with the similarity and visual summaries components. Our empirical results consistently and significantly outperform the state-of-the-art on the largest benchmarks for *text-based zero-shot learning*, illustrating the critical importance of texts for zero-shot image-recognition.

1 Introduction

In computer vision, *zero shot-learning* (ZSL) for image classification is the problem of classifying images given auxiliary information. An image classification model is trained to classify images from a pre-defined set of classes. At test time, images from new classes are given, and the task is to transfer knowledge learned from seen classes during training to unseen test classes.

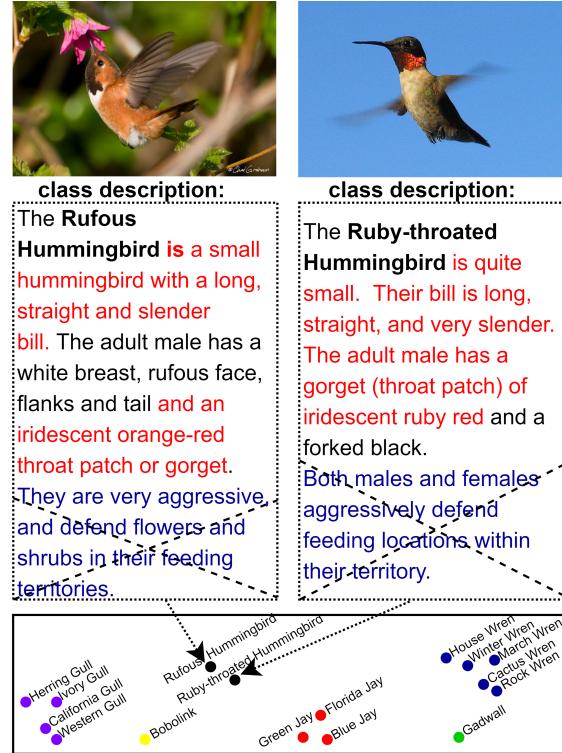


Figure 1: An illustration of textual similarity and visually relevant descriptions in Wikipedia articles: (1) we aim to leverage the similarity within texts (red) via document clustering (bottom box); (2) we aim to extract similar (red) and dissimilar (black) visual descriptions, and remove non-visually relevant (blue) one.

A common setup for ZSL assumes that the auxiliary information is a set of semantically meaningful properties (called attributes) describing the class (e.g., black-beak, long-tail) (Wah et al., 2011; Farhadi et al., 2009). A different ZSL setup uses image captions as auxiliary information (Reed et al., 2016; Felix et al., 2018). Typically, this auxiliary information is manually collected by human raters for each image (test and train alike) and averaged across images. A more realistic approach relies on available online text descriptions of classes (e.g., Wikipedia) (Elhoseiny et al., 2017). It avoids ex-

pensive annotation and exposure to test images.

In this work, we classify bird species according to Wikipedia descriptions. This task raises many challenges: (1) Differences between the birds are very small, which makes it a fine-grained classification task; (2) This is an expert task, and the text contains terminology that is unlikely to be familiar to a layman; and, on top of that (3) The text descriptions of the classes are long, containing few visually relevant sentences.

As opposed to previous work on text-based ZSL employing textual descriptions (Zhu et al., 2018; Elhoseiny et al., 2017) that focused on the *visual* modality, here we focus on the *text* modality, and address a key question in ZSL: *How can we identify text components that are visual in nature?*

To get an intuition about the task setup and our proposed solution, consider the following situation. Imagine you have never seen a zebra but have seen a horse. What if you were given a text describing a zebra: “Zebras have hooves, mane, tail, pointed ears, and white and black stripes”. This description would probably be very close to a description of a horse having “hooves, mane, tail, pointed ears” and you would probably be looking for an image that reminds you of a horse but has “white and black stripes”. So, even without ever seeing a zebra, using text-descriptions of the zebra and knowledge already acquired about horses, one can correctly classify unknown classes like a zebra.

Our proposed solution has two-phases. First, based on the intuition that similar objects (or images thereof) tend to have similar texts, we encode a similarity feature that enhances text descriptions’ separability. In addition, we leverage the intuition that the differences between text descriptions of species would be their most salient visual features, and extract visually relevant descriptions from the text.

Our experiments empirically demonstrate both the *efficacy* and *generalization* capacity of our proposed solution. On two large ZSL datasets, in both the *easy* and *hard* scenarios, the similarity method obtains a ratio improvement of up to 18.3%. With the addition of extracting visually relevant descriptions, we obtain a ratio improvement of up to 48.16% over the state-of-the-art. We further show that our visual-summarization method generalizes from the CUB dataset (Wah et al., 2011) to the NAB dataset (Van Horn et al., 2015), and we demonstrate its contribution to additional models

by a ratio improvement of up to 59.62%.

The contributions of this paper are threefold. First, to the best of our knowledge, we are the first to showcase the critical importance of the text representation in zero-shot image-recognition scenarios, and we present two concrete text-based processing methods that vastly improve the results. Second, we demonstrate the efficacy and generalizability of our proposed methods by applying them to both the *zero-shot* and *generalized zero-shot* tasks, outperforming all previously reported results on the CUB and NAB Benchmarks. Finally, we show that visual aspects learned from one dataset can be transferred effectively to another dataset without the need to obtain dataset-specific captions. The efficacy of our proposed solution on these benchmarks illustrates that purposefully exposing the *visual features in texts* is indispensable for tasks that learn to align the vision-and-language modalities.

2 Background and Related Work

Zero-shot learning (ZSL) aims at overcoming the need to label massive datasets for new categories, by learning the connections between images and prior auxiliary knowledge about their classes. At test-time, this auxiliary information compensates for the lack of previously-attained visual information about the new categories.

Text-based ZSL is a specific multimodal instantiation of this learning task that uses natural language descriptions as the auxiliary information. Models for text-based ZSL are typically composed of three parts: (1) the text representation; (2) the image representation; (3) a compatibility function between the two. While most previous work focused mainly on the latter two components, here we focus on the text.

Most ZSL studies for object recognition are aimed at processing the image modality. For example, Xu et al. (2018); Lei Ba et al. (2015); Qiao et al. (2016); Akata et al. (2016) rely on visual features extracted using Convolutional Neural Network (CNN). More recent studies use object detection to detect the *semantic parts* of the object and extract visual features at the part-level (Elhoseiny et al., 2017; Zhu et al., 2018; Zhang et al., 2016). This approach makes the image more compatible with the text, as it enables text-terms such as “crest” to be linked to the visual representation of parts like “head”.

The auxiliary information provided to ZSL tasks

may be of various kinds, ranging from pre-defined semantic attributes (Lampert et al., 2009; Changpinyo et al., 2020; Atzmon and Chechik, 2018), to captions (Xian et al., 2018; Sarıyıldız and Cinbis, 2019) to Wikipedia article describing the species (Elhoseiny et al., 2017). Here we assume the latter scenario. ZSL studies that rely on Wikipedia articles as auxiliary information improve the visual representation and the compatibility function, and use text representations such as Bag-of-Words and TF-IDF, without further text processing. (Lei Ba et al., 2015; Elhoseiny et al., 2013, 2016, 2017; Zhu et al., 2018). Qiao et al. (2016) used a simple BOW and a $L_{1,2}$ -norm objective to suppress the noisy signal in the text. However, this basic treatment of the text is problematic, as it misses crucial information for detecting the correct class.

Recent studies (Lu et al., 2019; Tan and Bansal, 2019) have shown improved performance on multiple vision-and-language tasks using pre-trained BERT-based models that jointly learn a representation for vision and language. However, they are tuned on relatively short texts and are not optimal for classifying long textual descriptions.

In this work, we proceed in a different, yet complementary, direction to previous work, aiming to purposefully model the contribution of the textual modality to ZSL. We aim to establish the importance of adequately processing the text into a sound representation of visually salient features, in order to increase the vision-and-language compatibility, which can then be effectively learned in an end-to-end manner.

3 Strong Baseline Model

The basic architecture, which term ZEST_{vanilla}, is a simple multiplicative attention mechanism (Luong et al., 2015) inspired by Romera-Paredes and Torr (2015). We model the problem using an attention-based model, where the image is queried against a set of candidate documents.

Formally, let x_1^S, \dots, x_M^S be image feature vectors from a training-set, where $x_i^S \in \mathbb{R}^m$. The set of M training images corresponds to a set of L seen classes. Each class has a *single* “class description” which is a document written by experts in free language (e.g. Wikipedia). We denote d_1^U, \dots, d_L^U as a set of L document feature vectors, where $d_i^U \in \mathbb{R}^{\hat{m}}$. Likewise, let x_1^U, \dots, x_N^U be the image feature vectors from a test set, where $x_i^U \in \mathbb{R}^m$. The set of test images corresponds to a set of K unseen classes.

Likewise, each class has a *single* “class description”. We denote d_1^U, \dots, d_K^U as a sets of document feature vectors, where $x_i^U \in \mathbb{R}^{\hat{m}}$. Finally, $W \in \mathbb{R}^{m \times \hat{m}}$ is our learned matrix. At inference, the label assignment of an image x_i^U is defined as:

$$\hat{y} = \arg \max_k (x_i^U)^T W d_k^U, k \in \{1 \dots K\} \quad (1)$$

For an image representation x_i^S and a text representation d_j^S , an indicator function $I(x_i^S, d_j^S)$ outputs 1 if image x_i^S corresponds to the class described by d_j^S and 0 otherwise. The matrix W is then learned by minimizing the categorical cross-entropy loss:

$$\sum_{j=1}^L I(x_i^S, d_j^S) \times \log(\text{softmax}(x_i^S W d_j^S)) \quad (2)$$

Image Encoding The image encoder’s goal is to transform the image into a vector representation of the most salient visual features for the classification. We adopt the image encoder for text-based ZSL of Zhang et al. (2016); Zhu et al. (2018); Elhoseiny et al. (2017). It is based on a Fast R-CNN with (Girshick, 2015) a VGG16 backbone for object detection to detect seven semantic parts in the CUB dataset: “head”, “back”, “belly”, “breast”, “leg”, “wing”, “tail”. Each visual part’s encoded features are then concatenated into a feature vector that functions as the image representation for the text-based ZSL.

Text Processing Our basic encoder processes the text into a feature vector. Similar to previous studies, we employ a TF-IDF representation (Salton and Buckley, 1988). We preprocess the text to tokenize words, remove stop words, and stem the remaining words. Then, we extract a feature vector using TF-IDF. This processing procedure is similar to the text processing presented by Zhu et al. (2018). The dimensionalities of TF-IDF features for CUB and NAB are 7,551 and 13,217, respectively.

4 The Proposed Approach

Our solution’s key idea is to replace the general class’s text representation with a text representation focusing on the most salient features for the visual recognition task. To do so, we employ two different (complementary) methods: (i) induce a similarity measure used for clustering; and (ii) extract visually relevant text descriptions. Both methods are incorporated in our proposed end-to-end

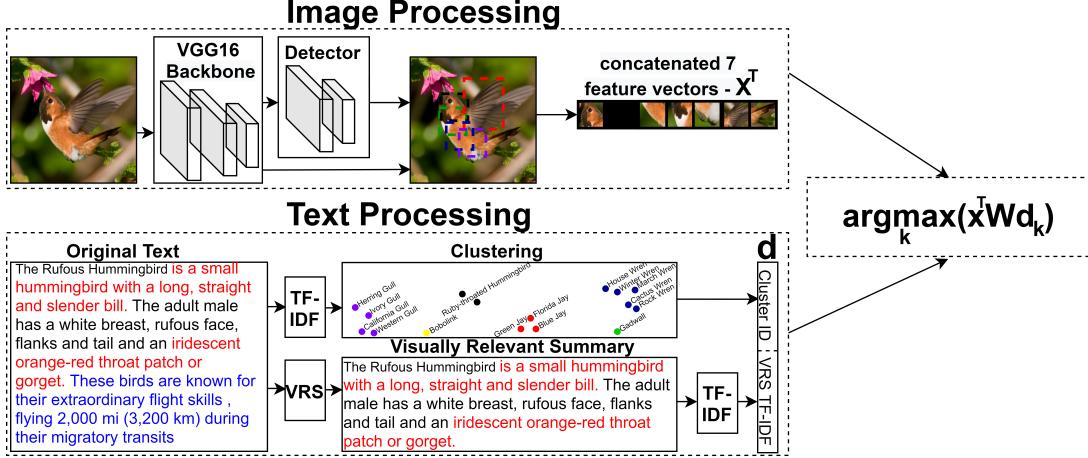


Figure 2: Our $ZEST_{similarity} + VRS$ model with the similarity component and Visually Relevant Summaries (VRS).

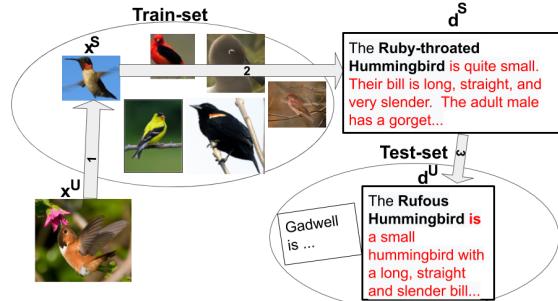


Figure 3: The Nearest Neighbor Similarity (NNS) model links images and texts through in-modality similarities

vision-and-language classification architecture, presented in Figure 2. In what follows we describe the similarity component 4.1, and the extraction of visually relevant summaries 4.2.

4.1 The Importance of being Similar

Our proposed method leverages the similarities between images and texts. That is, when the images look similar, the texts describing their classes are also similar, and vice versa. Here, we propose to reconstruct this similarity link.

To this end, we propose two models: (1) a strong baseline based on two nearest neighbors, which create a link between images and texts; (2) adding a similarity component to our model $ZEST_{vanilla}$. For both models we use the Image Encoder (section 3) to process the images x , and the Text Processing (section 3) to process documents D.

4.1.1 Nearest Neighbor Similarity (NNS)

Figure 3 presents our Nearest Neighbor Similarity (NNS) method, which aims to reconstruct the parallel similarity links between the vision and text latent spaces.

The algorithm is as follows. Given an image x^U from an unseen class in the zero-shot phase, we first look for the nearest neighbor image in the set of training images, using cosine similarity. The closest image from the training set x_k^S corresponds to a document from training d_k^S . We then look for the nearest neighbor text in from test set d_y^U and predict the corresponding class label y.

4.1.2 $ZEST_{similarity}$

A different way to incorporate textual similarity into the classification is to embed it into our model $ZEST_{vanilla}$, to benefit from it in the learning procedure. To this aim, we want to add on top of our text feature vector a representation of the text’s similarity to its neighbors.

The Basic Encoder captures similarities and differences at the word-level. However, to find similarities at the document level we add to this vector our *similarity* component, which applies unsupervised clustering to all class descriptions in the training and test texts. We use two different clustering methods that capture different aspects of text similarities. The cluster indexes are then embedded as a BOW (hence cluster embedding).

We hypothesize that the similarity component will work well on the “easy” scenario - where closely related birds are seen during training, and their text can cluster together to indicate these similarities.

4.2 The Importance of Being Seen

Here we extract visually relevant features *from the text*, making the texts that enter the classification more compatible with the salient visual information typically reflected in images.

While the similarity method takes advantage of the similarity between objects seen in training and objects seen at test time, here we want to address the harder scenario, where similar objects are observed together during test time only (e.g. zebras and mules), and they may be very different from those observed during training.

To differentiate between classes in the test set we need to emphasize the parts that are different, both in the image and the text — and these are typically their most salient visual features.

4.2.1 Visually Relevant Summaries (VRS)

Our method for enhancing the textual description is based on visually relevant extractive summaries. Extractive summarization is the task of extracting a small number of sentences that summarize a given document. In this work, we define *visually relevant extractive summarization* (VRS) as the task of extracting only sentences that represent visually relevant language. The term *visually relevant language* (VRL) was coined by Winn et al. (2016) to indicate sentences which are visually descriptive with respect to the object (i.e., bird species).

A naïve approach for VRS would be to extract sentences with parts that we know are visually salient in our domain (e.g., the 7 parts employed by the vision recognition representation). However, this naïve approach has several drawbacks. First, bird parts can be described using many different terms and paraphrases; additionally, a bird can be described by its property values (e.g., black), without any mention of the attribute (e.g., beak). Instead, we propose to use the similarity of sentences in the documents and compare them to naturally occurring sentences ('in the wild') containing VRL.

Note that we cannot rely on descriptions of particular species due to the zero-shot setup. We must do with descriptions of objects in the *general domain* of objects we are interested in classifying.

4.2.2 ZEST_{similarity}+VRS

One way to obtain naturally-occurring descriptions of birds is from captions that describe bird images. Critically, these captions need *not* be from our dataset, they can describe any bird image.

We propose to use a set of L bird captions to create an unsupervised classifier. The classifier will receive a set of sentences (assembled as a document), and for each sentence, the classifier will predict whether the sentence is relevant, that is, whether it contains descriptions that can be *seen* in a bird image.

For each document, we propose to calculate the pairwise similarity between captions and sentences in the Wikipedia description, and based on this similarity, assign a VRS-score to each sentence.

We calculate the VRS-score of a sentence s_j to a caption by computing the cosine similarity of the embeddings of both the captions ($c_{0:L}$) and sentences ($s_{0:M}$) in the document. For a fixed-size sentence embeddings, we use a pre-trained siamese-and-triplet network (Reimers and Gurevych, 2019; Schroff et al., 2015) on top of a pre-trained BERT network (Devlin et al., 2019).

The VRS-score of sentence s_j with respect to all available captions $c_{1:L}$ is thus defined to be:

$$score(s_j) = \frac{1}{L} \sum_{i=1}^L \frac{c_i \cdot s_j}{\|s_j\| \|c_i\|} \quad (3)$$

We then take the highest k scoring sentences from $s_{0:K}$ to be the visually relevant extractive summary of the document. We can then concatenate the similarity embedding to the VRS summary of the text, and perform the multiplicative attention on this revised encoding of the documents and the same image encoding as before.

A bird's eye overview of our overall architecture is presented in Figure 2. The text that enters the *similarity* (clustering) component is the original Wikipedia document, *not* the document's VRS summary. Documents contain many non-visual descriptions that are unobserved in the images. However, these non-visual descriptions might still be essential to capture the similarity between documents. For example, similar-looking birds are likely to be in the same habitat. Thus, the VRL sentence extraction and the similarity enhancement operate in parallel on the original document.

5 Experiments

5.1 Experiment setting

Datasets: We evaluate our method ¹ on the Caltech UCSD Birds-2011 dataset (CUB) Wah et al. (2011) and the North America's birds dataset

¹Our code can be found at <https://github.com/tzuf/ZEST>.

(NAB) (Van Horn et al., 2015), using class descriptions obtained from Wikipedia and the AllaboutBirds website², collected by Elhoseiny et al. (2017). Both are fine-grained datasets of birds but from different species. The CUB dataset contains 11,788 images of 200 bird species, and the NAB is a larger dataset of birds with 48,562 images of 404 classes³. The texts of both CUB and NAB are long, containing non-visual information. CUB has an average of 869 tokens and 42 sentences in class documents. NAB has an average of 1277 tokens and 58 sentences in class documents.

Two split Settings We use the two splits presented by Elhoseiny et al. (2017): (1) Super Category-Shared (SCS), also referred to as the ‘easy’ split; and (2) Super-Category-Exclusive (SCE), also referred to as the ‘hard’ split. In the SCS, for each class in the test set, at least one class in the training set belongs to the same category (categories are organized taxonomically). For example, in Figure 1, the Rufous Hummingbird and the Ruby-throated Hummingbird are both from the Hummingbird category. In the SCE, all classes in a category are in the same set. Namely, if a class is in the test set, then other classes from the same category are also in the test set, and will never be seen during training. Intuitively, classes from the same category have high similarity in both images and texts, so while in SCS similar images have been seen during training, in the SCE a class from an entirely new category is seen for the first time.

Training Details: The parameters of our model include cluster parameters. We use two clustering methods: (1) Density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996); (2) Hierarchical DBSCAN (McInnes et al., 2017). The DBSCAN algorithm takes two parameters: (1) “minimal cluster” - the number of samples in a neighborhood for a point to be considered as a core point; (2) “max distance” the maximum distance between two samples for one to be considered as in the neighborhood of the other. The “minimal cluster” is chosen to be two as two birds are the minimal similarity we want (similar to the NNS model). The “max distance” parameter we optimize on validation sets (10% of data) according to the two splits. In addition, the similarity model includes

²<https://dl.allaboutbirds.org>

³Elhoseiny et al. (2017) merged the original 1,011 classes according to the subtle division of classes.

a threshold for performing the similarity component, also optimized over the validation set. The VRS algorithm includes a sentence score threshold for the number of sentences to be extracted. This threshold was chosen on the validation set.

The weights W were initialized with normalized initialization (Glorot and Bengio, 2010). The cross-entropy loss function was optimized with Adam optimizer (Kingma and Ba, 2015).

Human Summarization: To evaluate our proposed VRS extraction method, we designed an oracle experiment using ground-truth visually relevant summarization. To this end, two independent human experts manually annotated the CUB dataset by reading each sentence in the document and marking the sentence as yes\|no VRL. We set guidelines to resolve disagreements (e.g. hatchlings descriptions were marked as not VRL). On average, only 11.9% of the sentences were found to include VRL.

Image Captions: To create visual summaries we use image captions of birds from the CUB train set, provided by Reed et al. (2016). Each image in the CUB dataset has been annotated with ten fine-grained captions. These captions describe only the birds’ visual appearance while avoiding mentioning the names of the bird species. E.g., “This bird has a long beak, a creamy breast, and body, with brown wings”. In this work, we use the first five captions of each image.

To showcase this approach’s generality, we use these captions in both in-domain (CUB) and out-of-domain (NAB) scenarios. In all cases, we avoid using captions of unseen (test) bird classes. In NAB, we effectively use captions from CUB to extract VRS for entirely-different species presented in NAB. Note that *only* models that include the VRS component (+VRS) employ these image captions. We report the accuracy achieved per the number of captions used in the VRS, to indicate the number of captions that are realistically needed.

Baselines: Our approach is compared against ten leading algorithms (see Table 1): MCZSL (Akata et al., 2016), WAC-Linear (Elhoseiny et al., 2013), Wac-Kernel (Elhoseiny et al., 2016), ES-ZSL (Romera-Paredes and Torr, 2015), SJE (Akata et al., 2015), Sync_{fast} (Changpinyo et al., 2016), Sync_{OVO} (Changpinyo et al., 2016), ZSLNS (Qiao et al., 2016), and GAZSL (Zhu et al., 2018).

methods	CUB		NAB	
	SCS	SCE	SCS	SCE
MCZSL Akata et al. (2016)	34.7	-	-	-
WAC-Linear Elhoseiny et al. (2013)	27.0	5.0	-	-
WAC-Kernel Elhoseiny et al. (2016)	33.5	7.7	11.4	6.0
ESZSL Romera-Paredes and Torr (2015)	28.5	7.4	24.3	6.3
SJE Akata et al. (2015)	29.9	-	-	-
ZSLNS Qiao et al. (2016)	29.1	7.3	24.5	6.8
SynC _{fast} Changpinyo et al. (2016)	28.0	8.6	18.4	3.8
SynC _{OVO} Changpinyo et al. (2016)	12.5	5.9	-	-
ZSLPP Elhoseiny et al. (2017)	37.2	9.7	30.3	8.1
GAZSL Zhu et al. (2018)	43.7	10.3	35.6	8.6
Nearest Neighbor Similarity (NNS)	40.402	5.551	37.002	5.517
ZEST _{vanilla}	39.16	11.77	27.61	10.18
ZEST _{similarity}	47.48	11.77	38.2	10.18
ZEST _{similarity} +VRS	48.57	15.26	38.51	10.23

Table 1: Top-1 accuracy (%) on CUB and NAB datasets with two split settings. We report the mean over three random initializations. The standard-deviation for ZEST_{similarity} for the CUB is 0.337 and 0.368; for NAB 0.625 and 0.174 (for the SCS and SCE splits accordingly).

methods	CUB		NAB	
	SCS	SCE	SCS	SCE
GAZSL	43.74	10.3	35.6	8.6
GAZSL+parts summarization	19.54	9.557	23.32	7.2
GAZSL+parts summarization+similarity	38.25	9.557	33.05	7.2
GAZSL+our VRS	43.72	16.44	37.28	9.237
GAZSL+HUMAN	35.98	21.81	-	-
GAZSL+HUMAN+similarity	47.32	21.81	-	-
ZEST _{vanilla}	39.16	11.77	27.61	10.18
ZEST _{vanilla} +our VRS	42.58	15.26	32.24	10.23
ZEST _{similarity}	47.48	11.77	38.2	10.18
ZEST _{similarity} +parts summarization	42.27	10.93	37.02	8.055
ZEST _{similarity} +our VRS	48.57	15.26	38.51	10.23
ZEST _{similarity} +HUMAN	48.99	17.2	-	-

Table 2: Visually Relevant Summarization (VRS) with GAZSL, ZEST_{vanilla}, and ZEST_{similarity}.

Generalized Zero-Shot Learning: The conventional zero-shot learning task considers only unseen classes during the zero-shot phase. However, in a realistic scenario, seen objects might also appear (Chao et al., 2016). In Generalized Zero-Shot Learning (GZSL), test data might also come from seen classes, and the labeling space is the union of both types of seen and unseen classes. GZSL is thus considered a more challenging problem setting than ZSL due to the model’s bias towards the seen classes. We follow the metric present by Chao et al. (2016) to evaluate our models on the GZSL task. We evaluate the accuracy of a Seen-Unseen accuracy Curve (SUC) and use Area Under SUC to measure the general capability of ZSL methods.

5.2 Results

Table 1 presents the top-1 accuracy for each of the models. The table is divided into four sections, which are (from top to bottom): (1) previous work; (2) our baselines ;(3) our models with previous setup (for comparison to previous work); (4) Our

methods	CUB		NAB	
	SCS	SCE	SCS	SCE
ZEST _{vanilla}	39.16	-	27.61	10.18
ZEST _{vanilla} +bird category	43.71	-	36.73	-
Zest _{similarity} only 1 cluster	46.55	-	35.94	-
Zest _{similarity} full (2 cluster)	47.48	-	38.2	-

Table 3: Zest model with different similarity methods

methods	CUB		NAB	
	SCS	SCE	SCS	SCE
ESZSL	0.185	0.045	0.092	0.029
ZSLNS	0.147	0.044	0.093	0.023
WAC _{kernal}	0.225	0.054	0.007	0.023
WAC _{linear}	0.239	0.049	0.235	-
SynC _{fast}	0.131	0.040	0.027	0.008
SynC _{OvO}	0.017	0.010	0.001	-
ZSLPP	0.304	0.061	0.126	0.035
GAZSL	0.354	0.087	0.204	0.058
ZEST _{similarity}	0.443	0.1	0.267	0.067
ZEST _{similarity} +VRS	0.437	0.147	0.26	0.084
ZEST _{similarity} +HUMAN	0.445	0.163	-	-

Table 4: Generalized Zero-Shot Learning: AUC of Seen-Unseen Curve.

model with additional data - captions.

NNS Model: According to Table 1, the NNS model achieves competitive results on the SCS - 40.402% and 37.002% on CUB and NAB correspondingly. The high scores on the SCS, where similar birds have been seen during training, is expected — as this method relies on similarities within texts and images. In contrast, the NNS suffers from low accuracy on the SCE, where different categories of birds have been seen during training. As the NNS model relies on text and image similarities, it is intuitively appealing that low accuracy on the SCE stems from the fact that birds from different categories are less likely to look alike.

ZEST_{vanilla} In contrast to the very sophisticated approaches of Zhu et al. (2018), the vanilla cross-entropy based approach outperforms all previous methods on the SCE-split on both CUB (+14.27% ratio of improvement) and NAB (+18.37% ratio of improvement). As the SCE-split is a more challenging split, this sheds light on the strength as well as limitations of this simple framework.

ZEST_{similarity} We then combined strengths of ZEST_{vanilla} and NNS models over the two different scenarios: “hard” and “easy”, respectively.

The ZEST_{similarity} model adds the cluster index embedding to the TF-IDF representation, only if a significant percentage of the documents from the

Sentence	HUMAN	VRS Model
1 After nesting, north american birds move in flocks further north along the coasts, returning to warmer waters for winter.	x	x
2 Red foxes and coyotes readily predate colonies that they can access, the later being the only known species to hunt adult pelicans (which are too large for most bird predators to subdue).	x	x
3 when foraging, they dive bill-first like a kingfisher often submerging completely below the surface momentarily as they snap up prey.	x	x
4 It is one of only three pelican species found in the western hemisphere.	x	✓
5 Due to their small size, they are vulnerable to insect-eating birds and animals.	x	✓
6 Hummingbirds show a slight preference for red, tubular flowers as a nectar source.	x	✓
7 The head is white but often gets a yellowish wash in adult birds.	✓	✓

Table 5: Qualitative analysis, showing seven sentences from three randomly selected summaries. The table shows HUMAN and VRS model markings of the sentence as yes(✓)\no(x) VRL.

test-set are clustered with documents from the train set. The threshold picked over the validation set is a 15%. Thus, in the case of the SCE-split, no or few similarities are found, and the ZEST_{similarity} preforms at the same level as the ZEST_{vanilla} model. The threshold parameter was optimized on the validation set.

The two clustering algorithms we applied find real similarities, achieving high accuracy when tested on predicting the correct label according to the ground-truth taxonomical category. The HDBSCAN, and DBSCAN achieved 88% and 84.5% accuracy on the CUB, and 93.07% and 95.05% on the NAB, accordingly.

Interestingly though, different clustering find different sources of similarities, that are essentially *additive*. In Table 3 we can see a comparison between different similarity enhancing methods. The ZEST_{vanilla+bird category} method is a BOW of the bird category added to the original text embedding and then passed as before to a ZEST_{vanilla} model. The use of two clusters that capture different similarities performs better than embedding the bird category in the text representation, by a ratio improvement of up to 8.63%. This suggests that our ZEST_{similarity} method captures similarities that are beyond the bird category.

Finally, in Table 4 we present the results of ZEST_{similarity} in the GZSL setup. On both datasets and splits, the ZEST_{similarity} achieves state-of-the-art results with up to 30.88% ratio improvement.

ZEST_{vanilla+VRS} and ZEST_{similarity+VRS} use the captions from training images in the CUB in order to generate visually relevant extractive summaries of the original Wikipedia documents.

We test the summarized representation on the ZEST_{vanilla} model, the ZEST_{similarity}, and the GAZSL (Zhu et al., 2018) model. In Table 2 we show the experimental results. We compare the models before and after the use of the Visually Relevant Extractive Summarization component. We see an improvement in accuracy in both models on both datasets and on both splits.

In contrast to the ZEST_{similarity}, the GAZSL does not have a component that embeds similarities. The VRS reduces similarity by removing non-VRL that might be similar between documents. The HUMAN summary is an especially lean summary with only 11.9% sentences extracted. Thus, the similarity between texts of similar objects diminishes. The GAZSL+HUMAN in the SCS-split performs poorly due to the diminished similarity. In contrast, The GAZSL+HUMAN+our VRS adds the similarity that was lost and the performance improves.

To assess the quality of the VRS summarization performance, we treat HUMAN summarization as the ground truth. The VRS method succeeds in removing 49.4% of the sentences in the CUB dataset with 96.23% recall and 22.59% precision. For comparison, removing 49.4% of the sentences randomly produces a recall of 50.6% and a precision of 11.9%.

Table 5 shows a qualitative analysis of our VRS results. In sentences 1-3, the VRS model correctly marked the sentences as non-VRL: sentence 1 is a typical case of non-visually-relevant language describing birds migration; in sentence 2 the VRS model correctly marks the sentence as non-VRL despite the mention of color (red) — since the color does *not* refer to the object to be classified (the bird); in sentence 3 the VRS model correctly marks

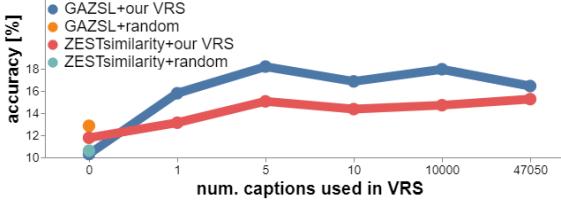


Figure 4: Accuracy per number of captions used to *focus* summarization, measured on the hard SCE split of CUB. Showing that as little as 5 captions in total are sufficient to focus the summarization process.

the sentence as non-VRL despite the mention of a body part (bill) — since that description it is not visually relevant in that particular context. Sentences 4-6 show examples of false-positive predictions of the VRS model. E.g., in sentence 5 the VRS model incorrectly predicted VRL, which we attribute to the mention: “their small size”. In sentence 6 the VRS model incorrectly marks the sentence as VRL, a mistake we attribute to the mention of the flower’s “red” color.

We then compare both $ZEST_{similarity}$ and the GAZSL to the use of HUMAN summarization in the CUB dataset and see additional improvement in both models on the two splits. The gap between the performance on the VRS and the Human summarization indicates that improvement in the summarization of documents will improve the models’ performance, and is, therefore, a promising path for text-based zero-shot learning research.

Finally, we experiment to assess the number of captions that are realistically needed for the VRS method. The results, presented in Table 4, show that only a few (~ 5) sentences (captions) from arbitrary birds are needed to achieve the maximum accuracy with this method. Testing the VRS with five arbitrary captions from CUB dataset on the NAB dataset with SCS-split, we achieved a 39.28% accuracy.

For comparison, Reed et al. (2016) showed that their model needed at least 512 captions per class to achieve the maximum accuracy - i.e., had it used all the captions available.

6 Conclusion

This work aims to establish a better way to represent the language modality in *text-based ZSL for image classification*. Our approach only relies on semantic information about visual features, and not on the visual features themselves. Specifically, our two orthogonal text-processing methods, employ-

ing *textual similarity* and *visually-relevant summaries*, lead to significant improvements across models, splits, and datasets, and illustrate that adequate text-processing is essential in text-based ZSL tasks. We conjecture that text-processing methods will be essential in a range of vision and language-based tasks, and hope this work will assist future research in better representing the language modality in various multi-modal tasks.

Acknowledgments

The research of the first and last author is funded by the European Research Council (ERC grant #677352) and the Israel Science Foundation (ISF grant #1739/26), and the research of the third author is also funded by the Israel Science Foundation (ISF grant #737/2018), for which we are grateful.

References

- Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. 2016. Multi-cue zero-shot learning with strong supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 59–68.
- Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. 2015. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2927–2936.
- Yuval Atzmon and Gal Chechik. 2018. Probabilistic and-or attribute grouping for zero-shot learning. In *Proceedings of the Thirty-Forth Conference on Uncertainty in Artificial Intelligence*.
- Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. 2016. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5327–5336.
- Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. 2020. Classifier and exemplar synthesis for zero-shot learning. *International Journal of Computer Vision*, 128(1):166–201.
- Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European Conference on Computer Vision*, pages 52–68. Springer.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

- Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Mohamed Elhoseiny, Ahmed Elgammal, and Babak Saleh. 2016. Write a classifier: Predicting visual classifiers from unstructured text. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2539–2553.
- Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2584–2591.
- Mohamed Elhoseiny, Yizhe Zhu, Han Zhang, and Ahmed Elgammal. 2017. Link the head to the “beak”: Zero shot learning from noisy text description at part precision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6288–6297. IEEE.
- Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231.
- Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1785. IEEE.
- Rafael Felix, Vijay B. G. Kumar, Ian Reid, and Gustavo Carneiro. 2018. Multi-modal cycle-consistent generalized zero-shot learning. In *The European Conference on Computer Vision (ECCV)*.
- Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.
- Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE.
- Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, et al. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4247–4255.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.
- Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.
- Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.
- Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton Van Den Hengel. 2016. Less is more: zero-shot learning from online textual documents with noise suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2249–2257.
- Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161.
- Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.
- Mert Bulent Sarıyıldız and Ramazan Gokberk Cinbis. 2019. Gradient matching generative networks for zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.
- Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona,

- and Serge Belongie. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 595–604.
- C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.

Olivia Winn, Madhavan Kavanur Kidambi, and Smaranda Muresan. 2016. Detecting visually relevant sentences for fine-grained classification. In *Proceedings of the 5th Workshop on Vision and Language*, pages 86–91.

Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. 2018. Feature generating networks for zero-shot learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Han Zhang, Tao Xu, Mohamed Elhoseiny, Xiaolei Huang, Shaoting Zhang, Ahmed Elgammal, and Dimitris Metaxas. 2016. Spda-cnn: Unifying semantic part detection and abstraction for fine-grained recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1143–1152.

Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. 2018. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1004–1013.