CLAREL: Classification via retrieval loss for zero-shot learning

Boris N. Oreshkin
Element AI
boris.oreshkin@gmail.com

Negar Rostamzadeh
Element AI
negar@elementai.com

Pedro O. Pinheiro
Element AI
pedro@elementai.com

Christopher Pal
Element AI
christopher.pal@elementai.com

Abstract

We address the problem of learning cross-modal representations. We propose an instance-based deep metric learning approach in joint visual and textual space. The key novelty of this paper is that it shows that using per-image semantic supervision leads to substantial improvement in zero-shot performance over using class-only supervision. We also provide a probabilistic justification and empirical validation for a metric rescaling approach to balance the seen/unseen accuracy in the GZSL task. We evaluate our approach on two fine-grained zero-shot datasets: CUB *and* FLOWERS.

1. Introduction

Deep learning-based approaches have demonstrated superior flexibility and generalization capabilities in information processing on a wide variety of tasks, such as vision, speech and language [15]. However, it has been widely realized that the transfer of deep representations to real-world applications is challenging due to the typical reliance on massive hand-labeled datasets. Learning in the low-labeled data regime, especially in the zero-shot [24] and the few-shot [25] setups, have recently received significant attention in the literature. In the problem of zero-shot learning (ZSL), the objective is to recognize categories that have not been seen during the training [14] via modality alignment. This is an especially relevant problem as machine learning is challenged with the long tail of classes, and the idea of learning from pairs of images and sentences, abundant on the web, looks like a natural solution. Therefore, in this paper we specifically target the fine-grained scenario of paired images and their respective text descriptions. The uniqueness of this scenario is in the fact that the co-occurance of image and text provides a rich source of information. The ways of leveraging this source have not been sufficiently explored in the context of ZSL.

In this paper, we specifically target the fine-grained visual description scenario, as defined by Reed et al. [20]. Concretely, given a training set $\mathcal{S} = \{(v_n, t_n, y_n) \mid v_n \in \mathcal{V},\ t_n \in \mathcal{T}, y_n \in \mathcal{Y}, n = 1 \dots N\}$ of image, text and label tuples, we are interested in finding representations $f_\phi : \mathcal{V} \to \mathcal{Z}$ of image, parameterized by ϕ, and $f_\theta : \mathcal{T} \to \mathcal{Z}$ of text, parameterized by θ, in a common embedding space $\mathcal{Z}$. Furthermore, generalized ZSL (GZSL) problem is defined using the sets of seen $\mathcal{Y}^{tr}$ and unseen $\mathcal{Y}^{ts}$ classes, such that $\mathcal{Y} = \mathcal{Y}^{tr} \cup \mathcal{Y}^{ts}$ and $\mathcal{Y}^{tr} \cap \mathcal{Y}^{ts} = \emptyset$. The training set only contains the seen classes, *i.e.* $\mathcal{S}^{tr} = \{(v_n, t_n, y_n) \mid v_n \in \mathcal{V},\ t_n \in \mathcal{T}, y_n \in \mathcal{Y}^{tr}\}$ and the task is to build a classifier function $g : \mathcal{Z} \times \mathcal{Z} \to \mathcal{Y}$. This is different from the ZSL scenario focusing on $g : \mathcal{Z} \times \mathcal{Z} \to \mathcal{Y}^{ts}$. The most acute problem in GZSL setup is the accuracy imbalance between seen and unseen classes. To measure and control the imbalance, three metrics are commonly used to assess the classification performance in the GZSL scenario: the Top-1 accuracy on the seen categories (**s**), the Top-1 accuracy on the unseen categories (**u**) and their harmonic mean, $\mathbf{H} = \mathbf{u} \cdot \mathbf{s}/(\mathbf{u} + \mathbf{s})$. The contributions of this work can be characterized under the following two themes.

Instance-based training loss. Zero-shot learning approaches rely heavily on class-level modality alignment [30]. We propose a new composite loss function that balances instance-based pairwise image/text retrieval loss and the usual classifier loss. The retrieval loss term does not use class labels. We show that most of the GZSL accuracy can be extracted from the instance-based retrieval loss.

Metric rescaling. GZSL approaches suffer from imbalanced performance on seen and unseen classes [16]. Previous work proposed to use a heuristic trick, calibrated stacking [4] or calibration [5], to solve the problem. We provide a sound probabilistic justification for it.

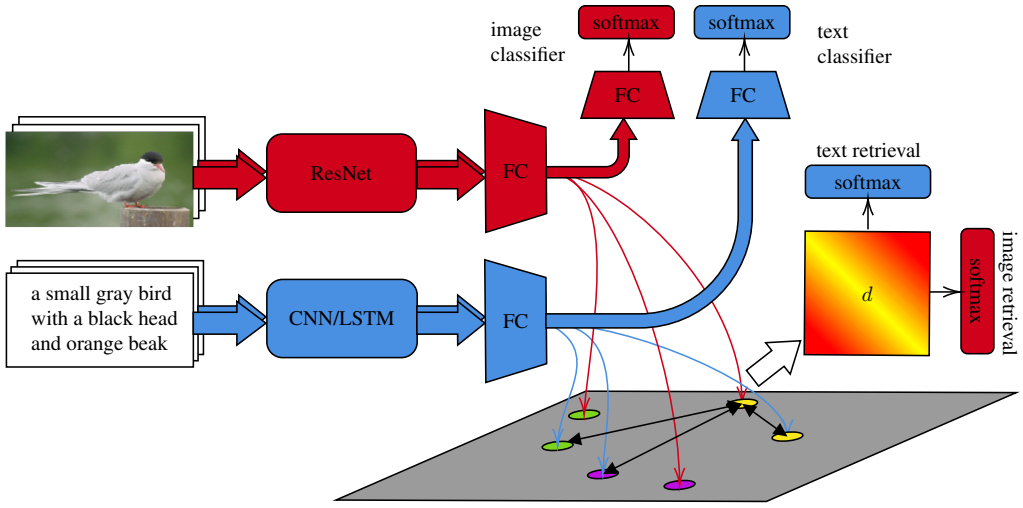

Figure 1: Proposed method. Each batch consists of randomly sampled instances, *i.e.* pairs of images and their corresponding texts. Images are embedded via ResNet and texts are embedded via a CNN/LSTM stack. Image and text features are projected via a fully connected layer into the same dimensional space. The negative distances between text and image features are fed into softmax to train on the image and the text retrieval tasks. In addition, image and text embeddings are trained on auxiliary image and text classification tasks on the class labels corresponding to instances.

2. Proposed Method

To build g, most approaches to joint representation learning rely on class labeling to train a representation. For example, all the methods reviewed by Xian et al. [30] require the access to class labels at train time. We hypothesise that in the fine-grained learning scenario, such as the one described by Reed et al. [20], a lot of information can be extracted simply from pairwise image/text co-occurrences. The class labels really only become necessary when we define class prototypes, *i.e.* at zero-shot test time. Following this intuition, we define a framework based on projecting texts and images into a common space and then learning a representation based on a mixture of four loss functions: a pairwise text retrieval loss, a pairwise image retrieval loss, a text classifier loss and an image classifier loss (see Fig. 1 and Algorithm 1 in Appendix A). The framework enables us, among other things, to experiment with the effects of train-time availability of class labels on the quality of zero-shot representations.

Pairwise cross-modal loss function is based solely on the pairwise relationships between texts and images. Suppose d is a metric $d : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}^+$, v_i is an image and $\tau = \{t_{j'}\}$ is a collection of arbitrary texts sampled uniformly at random, of which text t_j belongs to v_i. We propose the following model for the probability of image v_i and text t_j to belong to the same object instance:

$$p_{\phi,\theta}(i = j|v_i, t_j, \tau) = \frac{\exp(-d(f_\phi(v_i), f_\theta(t_j)))}{\sum_{t_{j'} \in \tau} \exp(-d(f_\phi(v_i), f_\theta(t_{j'})))} .$$

The learning is then based on the cross-entropy log-loss defined on the batch of size B:

$$J_{TR}(\phi, \theta) = -\frac{1}{B} \sum_{i,j=1}^{B} \ell_{i,j} \log p_{\phi,\theta}(i = j|v_i, t_j, \{t_{j'}\}_{j'=1}^{B}),$$

where $\ell_{i,j}$ is a binary indicator of the true match ($\ell_{i,j} = 1$, if $i = j$ and 0 otherwise). Note that the expression above has the interpretation of the text retrieval loss. Exchanging the order of image and text in the probability model $J_{TR}(\phi, \theta)$ leads to the image retrieval loss, $J_{IR}(\phi, \theta)$. The two losses are mixed using parameter $\lambda \in [0, 1]$ as shown in Algorithm 1 in Appendix A. The pairwise retrieval loss functions are responsible for the modality alignment. In addition to those, we propose to include the usual image and text classifier losses responsible for reducing the intraclass variability of representations. The classifier losses are added to the retrieval losses using a mixing parameter $\kappa \in [0, 1]$ as shown in Algorithm 1 in Appendix A.

2.1. Balancing Accuracy for Seen and Unseen

Let us define class prototype $\mathbf{p}(y)$ based on the set of texts $\mathcal{T}_y$ belonging to class y, $\mathbf{p}(y) = \frac{1}{|\mathcal{T}_y|} \sum_{t_i \in \mathcal{T}_y} f_\theta(t_i)$. In GZSL, the nearest neighbor decision rule for a given image v and its features $\mathbf{z}_v = f_\phi(v)$ has the following form:

$$\widehat{y} = \arg\min_{y \in \mathcal{Y}} d(\mathbf{z}_v, \mathbf{p}(y)) . \quad (1)$$

To formalize the problem, we first introduce y_v, the true class label of image v. Mathematically, the main GZSL pain point is that $\mathbb{P}\{\widehat{y} \in \mathcal{Y}^{tr}|y_v \in \mathcal{Y}^{ts}\}$ is significantly greater

Table 1: Generalized zero-shot Top-1 classification accuracy.

	CUB			FLOWERS		
	u	**s**	**H**	**u**	**s**	**H**
CADA-VAE [21]	n/a	n/a	53.4	n/a	n/a	n/a
Xian et al. [31]	50.3	58.3	54.0	59.0	73.8	65.6
Xian et al. [32]	48.4	60.1	53.6	56.8	74.9	64.6
Felix et al. [6]	47.9	59.3	53.0	61.6	69.2	65.2
Atzmon et al.[3]	n/a	n/a	n/a	59.6	81.4	68.8
Ours	59.3	52.6	**55.8**	73.0	73.6	**73.3**

Table 2: Zero-shot Top-1 classification accuracy.

	CUB	FLOWERS
f-CLSWGAN [31]	57.3	67.2
f-VAEGAN-D2 [32]	61.0	67.7
cycle-(U)WGAN [6]	58.6	70.3
Ours	**66.7**	**76.8**

than $\mathbb{P}\{\widehat{y} \in \mathcal{Y}^{ts} | y_v \in \mathcal{Y}^{tr}\}$. In other words, the problem is that a given image is more likely to be confused with one of the seen classes if it belongs to an unseen class than vice versa. We propose the following probabilistic representation of the event space for the decision rule in Equation (1):

$$\mathbb{P}\{\widehat{y} \in \mathcal{Y}^{tr} | y_v \in \mathcal{Y}^{ts}\} = \mathbb{P}\left\{ \min_{y \in \mathcal{Y}^{tr}} d(\mathbf{z}_v, \mathbf{p}(y)) < \min_{y \in \mathcal{Y}^{ts}} d(\mathbf{z}_v, \mathbf{p}(y)) \mid y_v \in \mathcal{Y}^{ts} \right\} . \quad (2)$$

To balance $\mathbb{P}\{\widehat{y} \in \mathcal{Y}^{tr} | y_v \in \mathcal{Y}^{ts}\}$ and $\mathbb{P}\{\widehat{y} \in \mathcal{Y}^{ts} | y_v \in \mathcal{Y}^{tr}\}$, we introduce a positive scalar $\alpha \in \mathbb{R}^+$ and scale all the distances corresponding to the seen prototypes by $1 + \alpha$, giving rise to the scaled distance d_α:

$$d_\alpha(\mathbf{z}_v, \mathbf{p}(y)) = \begin{cases} (1+\alpha) d(\mathbf{z}_v, \mathbf{p}(y)), & \text{if } y \in \mathcal{Y}^{tr} \\ d(\mathbf{z}_v, \mathbf{p}(y)), & \text{otherwise} \end{cases}$$

The error probability classifying unseen classes as seen ones for the classifier based on $d_\alpha(\mathbf{z}_v, \mathbf{p}(y))$, $\mathbb{P}\{\widehat{y}_\alpha \in \mathcal{Y}^{tr} | y_v \in \mathcal{Y}^{ts}\}$, is then a monotone non-increasing function of α and we can reduce it by increasing α (please refer to Appendix B for a proof). Consider now $\mathbb{P}\{\widehat{y}_\alpha \in \mathcal{Y}^{tr} | y_v \in \mathcal{Y}^{tr}\}$, which is a probability that we classify an image v from one of the seen classes as still one of the seen classes. Using exactly the same chain of arguments as in Appendix B we can show that the probability is a non-increasing function of α. Hence the probability $\mathbb{P}\{\widehat{y}_\alpha \in \mathcal{Y}^{ts} | y_v \in \mathcal{Y}^{tr}\} = 1 - \mathbb{P}\{\widehat{y}_\alpha \in \mathcal{Y}^{tr} | y_v \in \mathcal{Y}^{tr}\}$ is a non-decreasing function of α. Therefore, we expect that by varying $\alpha > 0$ we can balance the error rates $\mathbb{P}\{\widehat{y}_\alpha \in \mathcal{Y}^{tr} | y_v \in \mathcal{Y}^{ts}\}$ and $\mathbb{P}\{\widehat{y}_\alpha \in \mathcal{Y}^{ts} | y_v \in \mathcal{Y}^{tr}\}$.

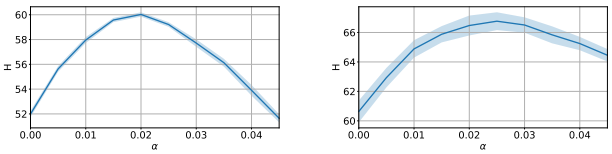

Figure 2: **H** against α on the validation set, the average and 95% confidence intervals of 10 repeats. **H** exhibits a distinct inverted U-shape w.r.t. α. CUB (left) and FLOWERS (right).

3. Related Work

ZSL approaches aim at recognizing objects belonging to classes unseen during training [14, 18]. This has been extended to the GZSL framework in which the decision space consists of both seen and unseen classes [22, 30]. The classical zero-shot approaches build a joint visual-semantic space, relying on a linear cross-modal compatibility function (e.g. dot-product between query embedding and semantic prototypes or a variation of a hinge loss) [7, 2, 1, 20]. Non-linear variants of the compatibility have also been explored [27, 22]. Extending previously proposed cross-modal transfer approaches based on auto-encoders [11] and cross-domain learning [9], more recent line of work [21, 31, 32, 6, 23] relies on combining these approaches and their variations with dataset augmentation tools such as GAN [8] and VAE [12]. It is argued that the use of those tools helps to resolve one of the prominent problems in GZSL scenario: classifying images from unseen classes as one of the seen classes. There exist approaches that try to tackle this same problem via temperature calibration [16] originally proposed by Hinton et al. [10]. Chao et al.[4] and Das et al. [5] proposed approaches to seen/unseen accuracy balancing that are very similar to ours, based on heuristic arguments. We extend this line of work here by providing a probabilistic justification for the balancing effect observed when applying metric rescaling. Atzmon et al. [3] propose a more sophisticated way to deal with seen/unseen imbalance via adaptive confidence smoothing and gating. In this work, we show that the simpler metric rescaling approach can still be used to achieve impressive results on the GZSL task.

4. Experimental Results

Datasets. We focus on learning embeddings for fine-grained visual descriptions and test them in ZSL/GZSL scenario. To test the quality of trained embeddings we focus on datasets that provide paired images and text descriptions, such as Caltech-UCSD-Birds (CUB) [26] and Oxford Flowers (FLOWERS) [17], that were augmented with textual descriptions by Reed et al. [20]. We use the GZSL splits proposed by Xian et al. [30]. The attribute-based datasets, such as SUN [19] and AWA [13] do not contain this information and are out of the scope of the current paper.

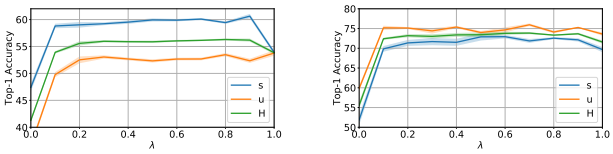

Figure 3: **H** against λ, average of 10 repeats. $\lambda = 0$ corresponds to the case of disabled text retrieval loss. CUB (left) and FLOWERS (right).

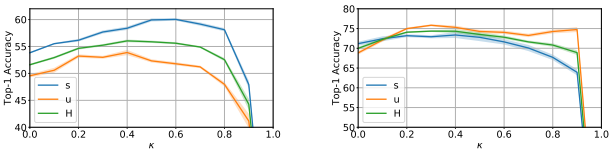

Figure 4: **H** against κ, average of 10 repeats. $\kappa = 0$ corresponds to the case of classification loss having weight 0. CUB (left) and FLOWERS (right).

Table 3: Generalized zero-shot Top-1 classification accuracy.

			CUB			FLOWERS		
α	λ	κ	**u**	**s**	**H**	**u**	**s**	**H**
0.0	0.5	0.5	38.3	65.3	48.3	55.1	84.6	66.7
0.0	0.5	0.0	39.3	57.5	46.7	54.0	78.1	63.8
✓	0.5	0.0	53.8	49.6	51.6	71.7	67.2	69.4
✓	0.0	0.5	47.4	36.6	41.3	51.5	60.5	55.6
✓	1.0	0.5	53.9	53.8	53.8	69.5	73.9	71.6
✓	0.5	0.5	59.3	52.6	**55.8**	73.0	73.6	**73.3**

Architecture and training details. see Appendix D.

Our key empirical results are shown in Tables 1 and 2. Our results are based on the settings of $\lambda = 0.5$, $\kappa = 0.5$ and α selected on the validation sets of CUB and FLOWERS datasets. Clearly, the combination of the proposed training method and the rebalancing of the metric space results in very impressive performance, especially taking into account the simplicity of our method. In the rest of the section we further analyze the stability with respect to the choices of λ and κ and provide more details on the selection of α.

The seen/unseen accuracy balancing. Fig. 2 confirms that **H** exhibits inverted U-shape behavior as a function of α on the validation sets of CUB and FLOWERS datasets, as expected based on results of Section 2.1. Once the value of α is determined by maximizing **H** on validation set, we train the representation on the full train+val subset and report results on the test split (the usual practice in GZSL). Validation set construction is detailed in Appendix C.

Ablation studies. Fig. 3 studies the importance of image and text retrieval losses. We see that all Top-1 accuracies (**H**, **s**, **u**) are stable in the range $\lambda \in [0.2, 0.9]$. Removing text retrieval loss ($\lambda = 0$) results in the most significant drop. Indeed, at the batch level, retrieving the correct text given an image is related to identifying the correct class encoded by a text prototype during GZSL inference step. Fig. 4 studies the interplay between the retrieval and the classification losses. We again observe that there exists a reasonably stable range of $\kappa \in [0.2, 0.6]$. $\kappa = 1$ results in the catastrophic performance drop: the classification losses alone do not enforce the necessary modality alignment.

Table 3 further studies the effects of different loss terms. The best result is achieved when all loss terms are active and when the metric rescaling is on (the last line in the table: $\lambda, \kappa = 0.5$ and α checked). Comparing this to the case with no metric rescaling (first line, $\alpha = 0$), we see that the rescaling helps to greatly decrease the gap between seen and unseen classification accuracy, both on CUB and FLOWERS. Interestingly, we only use images and texts from the training set to achieve it. Going to the second line in the table (the image/text classification loss is inactive, $\kappa = 0$) and comparing it to the first one, we assess the effect of the image/text classification loss. It barely affects the performance on unseen set, but it significantly boosts the classification accuracy on the seen set (around 8% on both datasets). However, it improves GZSL accuracy only when applied together with metric rescaling (please refer to lines 1 and 6 in Table 3). Our interpretation is that the image/text classifier loss reduces the intraclass variability and enforces tighter embedding clustering. Yet, this also leads to overfit on classification task. This is accounted for by metric rescaling that enables the learnings from the image/text classification task be transferred effectively into the GZSL task. Finally, an interesting observation can be made by comparing line 3 of Table 3 with performance of algorithms in Table 1. In this case our training relies only on retrieval losses computed without class labels solely based on the pairwise relationships between texts and images. The learned representation is competitive against the latest GAN/VAE based approaches on CUB and is state-of-the-art on FLOWERS. We conclude that when very fine-grained modality outputs are available (image and text pairs being a very prominent example), the high-quality representations may be learned without relying on manually supplied class labels.

5. Conclusions

We propose and empirically validate two contributions for learning fine-grained cross-modal representations. First, we confirm the hypothesis that in the context of paired images and texts, a deep metric learning approach can be driven by an instance-based retrieval loss resulting in impressive GZSL classification results. This demonstrates that high-quality deep representations can be trained relying largely on pairwise modality relationships. Second, we mathematically analyze and empirically validate a simple method of balancing seen/unseen accuracy in the GZSL task.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *TPAMI*, 2016. 3

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015. 3

[3] Yuval Atzmon and Gal Chechik. Adaptive confidence smoothing for generalized zero-shot learning. In *CVPR*, 2019. 3

[4] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV (2)*, pages 52–68, 2016. 1, 3

[5] Debasmit Das and C Lee. Zero-shot image recognition using relational matching, adaptation and calibration. In *International Joint Conference on Neural Networks*, 2019. 1, 3

[6] Rafael Felix, Vijay Kumar B G, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, 2018. 3

[7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013. 3

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 3

[9] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *NIPS*, 2007. 3

[10] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Workshop*, 2015. 3

[11] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. In *CVPR*, 2017. 3

[12] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 3

[13] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 2014. 3

[14] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, 2008. 1, 3

[15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 2015. 1

[16] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized zero-shot learning with deep calibration network. In *NIPS*, 2018. 1, 3

[17] M. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, 2008. 3

[18] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009. 3

[19] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *IJCV*, 2014. 3

[20] Scott E. Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016. 1, 2, 3, 8

[21] Edgar Schönfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. *CVPR*, 2019. 3

[22] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013. 3

[23] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, 2018. 3

[24] Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Trans. Intell. Syst. Technol.*, 2019. 1

[25] Yaqing Wang and Quanming Yao. Few-shot learning: A survey. In *arXiv*, 2019. 1

[26] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical report, California Institute of Technology, 2010. 3

[27] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh N. Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016. 3

[28] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Pretrained CUB features, 2018. 8

[29] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Pretrained FLOWERS features, 2018. 8

[30] Yongqin Xian, H. Christoph Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning: A comprehensive evaluation of the good, the bad and the ugly. *TPAMI*, 2018. 1, 2, 3, 8

[31] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 3

[32] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. f-vaegan-d2: A feature generating framework for any-shot learning. *CVPR*, 2019. 3