Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention

Dat Huynh
Northeastern University
huynh.dat@husky.neu.edu

Ehsan Elhamifar
Northeastern University
eelhami@ccs.neu.edu

Abstract

We address the problem of fine-grained generalized zero-shot recognition of visually similar classes without training images for some classes. We propose a dense attribute-based attention mechanism that for each attribute focuses on the most relevant image regions, obtaining attribute-based features. Instead of aligning a global feature vector of an image with its associated class semantic vector, we propose an attribute embedding technique that aligns each attribute-based feature with its attribute semantic vector. Hence, we compute a vector of attribute scores, for the presence of each attribute in an image, whose similarity with the true class semantic vector is maximized. To tackle the challenge of bias towards seen classes during testing, we propose a new self-calibration loss that adjusts the probability of unseen classes to account for the training bias. We conduct experiments on three popular datasets of CUB, SUN and AWA2. We show that our model significantly improves the state of the art.

This is a short version of a full paper that is accepted for presentation in CVPR 2020.

1. Introduction

Fine-grained zero-shot recognition, which is to classify categories without training data in addition to being visually very similar, is an important yet challenging task with a wide range of applications from fashion industry, e.g., recognition of different types of shoe or cloth, to face recognition and environmental conservation, e.g., recognizing endangered species of birds or plants. However, training fine-grained classification systems is challenging, as collecting training samples from every class requires costly annotations by domain experts to distinguish between similar classes. As a result, training samples often follow a long-tail distribution, where many classes have few or no training samples. In this work, we aim to generalize fine-grained recognition to new classes without training samples via capturing and transferring fine-grained knowledge from seen to unseen classes without overfitting on seen classes.

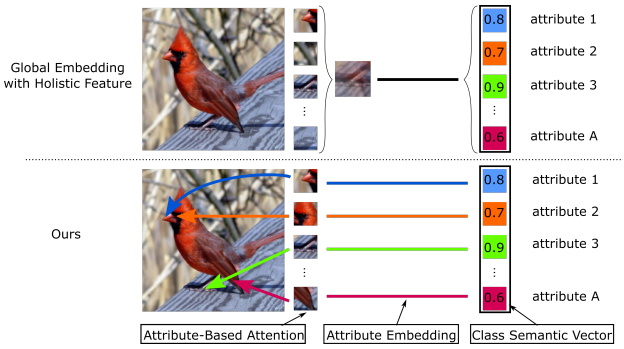

Figure 1: Traditional zero-shot classification (top) compresses visual features to perform global embedding with class semantic descriptions, hence, not efficiently capturing fine-grained discriminative visual information. Our method (bottom) finds local discriminative regions through dense attribute-based attention and individually embeds each attribute-based feature with the attribute semantic description, allowing for knowledge transfer to unseen classes while preserving all fine-grained details.

We propose a *dense attribute-based attention mechanism* that for each attribute focuses on the most relevant image regions, obtaining attribute-based features. Our attribute-based attention model is guided by each *attribute semantic vector*, hence, building the same number of feature vectors as the number of attributes. Instead of aligning a combination of all attribute-based features with the true class semantic vector, we propose an *attribute embedding technique* that aligns each attribute-based feature with its attribute semantic vector, see Figure 1. Hence, we compute a vector of attribute scores, for the presence of each attribute in an image, whose similarity with the true *class semantic vector* is maximized. To tackle the challenge of bias towards seen classes during testing, we propose a new self-calibration loss that adjusts the probability of unseen classes to account for the training bias.

We conduct experiments on three popular datasets of CUB, SUN and AWA2. We show that our model signifi-

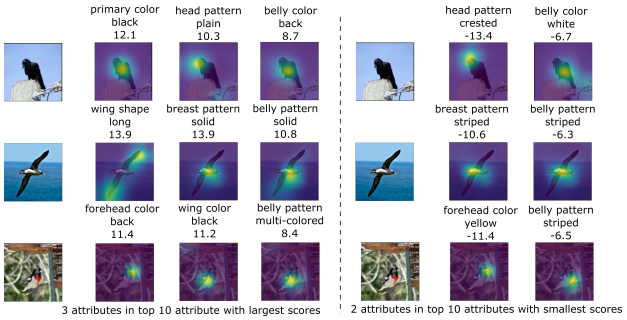

Figure 2: Visualization of attention maps from top positive/negative attributes in CUB dataset

cantly improves the state-of-the-art performance in the challenging generalized zero-shot setting where the model must recognize both seen and unseen classes.

2. Dense Attribute-Based Attention for Fine-Grained Generalized Zero-Shot Learning

2.1. Attribute Localization via Dense Attention

As the first component of our method, we propose an attribute-based spatial attention model, where for each attribute, we localize the most relevant image regions to the attribute to extract an *attribute-based attention feature* from a given image. Let $\{\boldsymbol{v}_a\}_{a=1}^{A}$ is the set of attribute semantic vectors, where $\boldsymbol{v}_a$ denotes the pretrained GloVe representation of the a-th attribute such as'yellow beak', and $\{\boldsymbol{f}_i^r\}_{r=1}^{R}$ denotes the region features of the image i divided into R equal regions. For the a-th attribute, we define its attention weights of focusing on different regions of image i as

$$\alpha(\boldsymbol{f}_i^r, \boldsymbol{v}_a) \triangleq \frac{\exp(\boldsymbol{v}_a^T \boldsymbol{W}_\alpha \boldsymbol{f}_i^r)}{\sum_{r'} \exp(\boldsymbol{v}_a^T \boldsymbol{W}_\alpha \boldsymbol{f}_i^{r'})}, \tag{1}$$

where $\boldsymbol{W}_\alpha$ denotes a learnable matrix to measure the compatibility between each attribute semantic vector and the visual feature of each region. Using the set of attention weights $\{\alpha(\boldsymbol{f}_i^r, \boldsymbol{v}_a)\}_{r=1}^{R}$, we compute the *attribute-based attention feature* for the a-th attribute as

$$\boldsymbol{h}_i^a \triangleq \sum_{r=1}^{R} \alpha(\boldsymbol{f}_i^r, \boldsymbol{v}_a) \boldsymbol{f}_i^r. \tag{2}$$

Thus, $\boldsymbol{h}_i^a$ represents the visual feature of the image i that is relevant to the a-th attribute according to the semantic vector $\boldsymbol{v}_a$. Notice that when an attribute is absent in the image, $\boldsymbol{h}_i^a$ captures the visual evidence used to reject the attribute in the image. For instance, the model could focus on a 'back belly' (and later assigns a negative score to it) to indicate the absence of a 'white belly' as in Figure 2.

2.2. Attribute Embedding

Given the set of attribute-based attention features $\{\boldsymbol{h}_i^a\}_{a=1}^{A}$ for each training image i, our goal is to compute the *class score* s_i^c of the image i belonging to a class c. During training, the class score would be optimized to be large for the ground-truth class $c = y_i$ and small for other classes $c \neq y_i$. To do so, we define A *attribute scores*, where each score measures the strength of having each attribute in the image (where A is the number of attributes). To compute the class score, we fuse these attribute scores using each class semantic vector which encodes the absence/presence of attributes in each class.

More specifically, we define the attribute score e_i^a, as the confidence of having the a-th attribute in the image i, by matching the attribute attention feature $\boldsymbol{h}_i^a$ with the attribute semantic vector $\boldsymbol{v}_a$,

$$e_i^a \triangleq \boldsymbol{v}_a^T \boldsymbol{W}_e \boldsymbol{h}_i^a, \tag{3}$$

where $\boldsymbol{W}_e$ is an embedding matrix that embeds the attribute feature $\boldsymbol{h}_i^a$ to the a-th attribute semantic space. In fact, when the attribute is visually present in an image, the associated image feature would be projected near its attribute semantic vector. We compute the class score s_i^c as the sum of products between each attribute score e_i^a and the strength of having the attribute a in class c, i.e, z_a^c, as

$$s_i^c = \sum_{a=1}^{A} e_i^a \times z_a^c. \tag{4}$$

As a result, when a class c has attribute a, i.e., $z_c^a > 0$, we would maximize the attribute score e_i^a.

Remark 1 *Notice that instead of computing the class compatibility score between a class semantic vector and a global image feature, we first compute A compatibility scores between each attribute-based attention feature and each attribute semantic vector to form the class compatibility score. This gives our model the ability to use a rich set of features for classification based on localization of each attribute in an image.*

2.3. Loss Function with Self-Calibration Component

In order to find the parameters of our model, we need to optimize the cross-entropy loss between the model prediction and the ground-truth label y_i over training images, i.e.,

$$\mathcal{L}_{ce}\big(\{s_i^c\}_{c\in\mathcal{C}}\big) = -\sum_i \log p(s_i^{y_i}). \tag{5}$$

Here, $p(s_i^c)$ is the probability that image i belongs to class $c \in \mathcal{C}$, where $\mathcal{C} = \mathcal{C}_s \cup \mathcal{C}_u$ is the set of both seen classes $\mathcal{C}_s$ and unseen classes $\mathcal{C}_u$. $p(s_i^c)$ is computed by applying softmax to class scores $\{s_i^c\}_{c\in\mathcal{C}}$,

$$p(s_i^c) \triangleq \frac{\exp(s_i^c)}{\sum_{c'\in\mathcal{C}} \exp(s_i^{c'})}. \tag{6}$$

However, optimizing the cross-entropy loss on training images that consist of only seen classes is prone to bias towards seen classes.

To overcome this challenge, we start by considering a calibration loss that allows to shift some of the prediction probabilities from seen to unseen classes during training. More specifically, we define

$$\mathcal{L}_{cal}\big(\{s_i^c\}\big) \triangleq -\sum_i \log\left(\sum_{u\in\mathcal{C}_u} p(s_i^u)\right), \tag{7}$$

where for brevity of notation, we have dropped the subscripts in $\{s_i^c\}_{c\in\mathcal{C}}$, which can be inferred from the context. Thus, minimization of $\mathcal{L}_{cal}$ in conjunction with the cross-entropy loss, promotes to put nonzero probability on the unseen classes during training. Hence, at testing time, for an image from an unseen class, the model can produce a (large) non-zero probability for the true unseen class. However, the drawback of using (7), as it is, is that it reduces the scores of seen classes and increases the scores of unseen classes during training on images from seen classes, which is not desired. Thus, to allow for nonzero prediction probability mass in unseen classes during training while keeping the scores of unseen classes low, we propose to augment the unseen scores using a margin (here, set to one), and use

$$\mathcal{L}_{cal}\big(\{s_i^c + \mathbb{1}_{\mathcal{C}_u}(c)\}\big). \tag{8}$$

where $\mathbb{1}_{\mathcal{C}_u}(\cdot)$ is a margin having the value of 1 when $c \in \mathcal{C}_u$ and -1 otherwise. The margin artifically reduces the effect of seen score on the calibration loss. Thus, this formulation allows high seen prediction on training data without incurring high loss value as long as unseen scores are sufficiently large.

Final Loss Function. Combining the cross-entropy and the self-calibration loss functions, we propose to minimize

$$\min_{\boldsymbol{W}_\alpha, \boldsymbol{W}_e, \{\boldsymbol{v}_a\}_{a=1}^A} \mathcal{L}_{ce}\big(\{s_i^c\}\big) + \lambda\mathcal{L}_{cal}\big(\{s_i^c + \mathbb{1}_{\mathcal{C}_u}(c)\}\big), \tag{9}$$

over the parameters of the attention models, attribute-image embedding and the attribute semantic vectors.

Remark 2 *Notice that in our method, we are optimizing over attribute semantic vectors $\{\boldsymbol{v}_a\}_{a=1}^A$, which results in visual grounding of each attribute meaning to the visual feature of training images. Also, by sharing $\{\boldsymbol{v}_a\}_{a=1}^A$ among all classes, we effectively allow transfering fine-grained knowledge from seen to unseen classes.*

Finally, at inference time, we predict the class of a test image as the class that has the maximum augmented score,

$$c^* = \text{argmax}_{c\in\mathcal{C}}\, s_i^c + \mathbb{1}_{\mathcal{C}_u}(c), \tag{10}$$

Thus, we effectively make prediction based on the augmented class scores, which we have explicitly calibrated to be sensitive toward unseen classes.

3. Experiments

3.1. Experimental Setup

Datasets. Following [1], we conduct experiments on the three popular datasets of CUB, SUN and AWA2. CUB [2] contains images from fine-grained bird-species with 150 seen and 50 unseen classes. SUN [3] is a dataset of visual scenes having 645 seen and 72 unseen classes and has the largest number of classes among the datasets. However, it only contains 16 training images per class due to its small overall training set. AWA2 [1] has been proposed for animal classification with 40 seen and 10 unseen classes and has a medium size of 37,322 samples in total. For CUB, SUN, AWA2, we follow the proposed training, validation and testing splits in [1].

Evaluation Metrics. Following [1], we measure the top-1 accuracy on two settings: i) traditional zero-shot learning, where test images are only from unseen classes, thus all predictions are constrained to be from unseen classes; ii) generalized zero-shot learning, where test images are from seen and unseen classes. In the latter case, we report the accuracy on testing images from seen classes, acc_s, and from unseen classes, acc_u. Also, to capture the trade-off between seen and unseen performance, we compute the harmonic mean, H, between seen and unseen accuracy.

Implementation Details. Following the canonical setting in [1], we use pretrained ResNet-101 with the input size of 224×224 to extract feature map at the last convolutional layer ($7 \times 7 \times 2048$) as the image region features. We implement all methods in PyTorch and optimize with the default setting of RMSprop with the learning rate 0.0001 and batch size of 50. We train all models on an NVIDIA V100 GPU for at most 20 epochs on CUB, AWA2, SUN. *In our method, we fix $\lambda = 0.1$ on all datasets*, which also shows that our self-calibration loss works on different datasets without the need for heavy hyper-parameter tuning.

3.2. Experimental Results

Fine-Grained Zero-Shot Learning. We measure the fine-grained zero-shot performance on the CUB dataset that contains different bird species with small visual differences, hence, demands the ability to focus on discriminative regions for classification. We report the traditional zero-shot performance given unseen class semantic vectors. Due to differences in the experiment settings among previous works, we conduct experiments on both the standard split (SS) and on the proposed split (PS) in [1] for comparison with the state-of-the-art methods. Table 2 shows the accuracies of different methods on the two splits of the CUB. Notice that on SS, we achieve at least 5.8% improvement over methods trained on holistic image features, while we have comparable performance (within 1% difference) to methods

Model	Approach	CUB			SUN			AWA2		
		acc_s	acc_u	H	acc_s	acc_u	H	acc_s	acc_u	H
SYNC [4]	Holistic Seen Feature	**70.9**	11.5	19.8	**43.3**	7.9	13.4	90.5	10.0	18.0
RNet [5]		61.1	38.1	47.0	–	–	–	93.4	30.0	45.3
TCN [6]		52.0	52.6	52.3	37.3	31.2	34.0	65.8	61.2	63.4
f-VAEGAN-D2 [7]	Holistic Feature Generation	60.1	48.4	53.6	38.0	45.1	41.3	70.6	57.6	63.5
CADA-VAE [8]		53.5	51.6	52.4	35.7	47.2	**40.6**	75.0	55.8	63.9
Ours ($\mathcal{L}_{ce}$)	Dense Attention	69.4	27.4	39.3	28.3	46.5	35.2	**91.1**	13.2	23.0
Ours ($\mathcal{L}_{ce} + \mathcal{L}_{cal}$)		59.6	**56.7**	**58.1**	25.9	**48.3**	33.7	76.2	**60.5**	**67.5**

Table 1: Generalized zero-shot classification performance on CUB, SUN and AWA2. We report accuracy per seen class, acc_s, and accuracy per unseen class, acc_u, as well as their harmonic mean, H.

Method	Bounding Box Annotations	Accuracy SS	Accuracy PS
RNet [5]	Not Required	62.0	55.6
TCN [6]		-	59.5
f-VAEGAN-D2 [7]		-	61.0
S^2GA (one-attention-layer) [9]	Required	67.1	-
S^2GA (two-attention-layer) [9]		**68.9**	-
Ours ($\mathcal{L}_{ce}$)	Not Required	67.1	60.0
Ours ($\mathcal{L}_{ce} + \mathcal{L}_{cal}$)		67.8	**65.9**

Table 2: Zero-shot classification performance on CUB dataset.

that use ground-truth bounding box annotations of the discriminative parts during training (we do not use this information). In fact, this shows the effectiveness of our dense attribute-based attention on capturing fine-grained details, achieving similar performance to S^2GA without the need for the costly annotations of the discriminative parts locations. On the other hand, on PS, we outperform other methods with at least 4.9% improvement, in particular, with respect to the state-of-the-art generative methods, which lack the ability to synthesize local discriminative regions of images. Also, notice that having the self-calibration loss facilitates knowledge transfer from seen to unseen classes, boosting the accuracy on PS by 5.9% compared to not using it.

Fine-Grained Generalized Zero-Shot Learning. Table 1 shows the performance of different methods for generalized zero-shot learning, where both seen and unseen classes appear at the test time. Notice that compared to SYNC [4], which achieves the best seen accuracy, our method ($\mathcal{L}_{ce}$) generalizes better to unseen classes with similar seen accuracy. This shows the effectiveness of our dense attention mechanism in generalization to unseen classes by only focusing on transferable attribute features instead of holistic visual appearance features, which often contain irrelevant background information. However, without the self-calibration loss, our method has lower unseen accuracy especially compared to feature generation techniques, which simulate the inference distribution by augmenting training samples with synthesized features from unseen classes.

On the other hand, using the calibration loss, $\mathcal{L}_{ce} + \mathcal{L}_{cal}$, our method significantly outperforms other algorithms on unseen accuracy, in particular, improves over the state-of-the-art generative model CADA-VAE [8] on unseen accuracy by 5.1%, 1.1% and 4.7% on CUB, SUN and AWA2, respectively. In addition, our method improves the harmonic mean score by 5.7% and 3.6%, respectively, on CUB and AWA2. However, it does not achieve the best harmonic mean on SUN. We believe this is due to having only 16 training samples for all seen classes, which does not allow to effectively train our dense attention model and results in even low seen performance compared to SYNC [4].

Qualitative Results: Figure 2 visualizes the results of our dense attention on CUB dataset. We observe that our model is able to localize fine-grained details given weak supervision, i.e., only image labels. Our model is also capable of learning different levels of abstraction/granularity thanks to the hierarchical structure of $\boldsymbol{W}_\alpha$ and $\boldsymbol{W}_e$ where the input of Eq 3 depends on the output of Eq 2. As Figure 2 shows, $\boldsymbol{W}_\alpha$ well localizes different parts (see spatial attention on each image) and $\boldsymbol{W}_e$ assigns appropriate attribute scores (see the score on top of each image) such as 'pattern' and 'color' for these parts.

References

[1] Y. Xian et al., "Zero-shot learninga comprehensive evaluation of the good, the bad and the ugly," *IPAMI*, 2017.

[2] P. Welinder et al., "Caltech-UCSD Birds 200," California Institute of Technology, Tech. Rep. 2010.

[3] G. Patterson et al., "Sun attribute database: Discovering, annotating, and recognizing scene attributes," *CVPR*, 2012.

[4] S. Changpinyo et al., "Synthesized classifiers for zero-shot learning," *CVPR*, 2016.

[5] F. Sung et al., "Learning to compare: Relation network for few-shot learning," *CVPR*, 2017.

[6] H. Jiang et al., "Transferable contrastive network for generalized zero-shot learning," *ICCV*, 2019.

[7] Y. Xian et al., "f-vaegan-d2: A feature generating framework for any-shot learning," *CVPR*, 2019.

[8] E. Schönfeld et al., "Generalized zero- and few-shot learning via aligned variational autoencoders," *CVPR*, 2019.

[9] Y. Yu et al., "Stacked semantics-guided attention model for fine-grained zero-shot learning," *neurIPS*, 2018.