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

Ablation Studies on Other Datasets. In the main paper, we showed ablation studies on the DeepFashion dataset. In this section, we report the ablation study results on CUB, SUN and AWA2 datasets in order to demonstrate the effectiveness of the different parts of our framework as well as its limitation.

As Table 1 shows, on CUB and AWA2, using Dense Attention significantly improves the performance compared to No Attention in both settings of without/with Self-Calibration. Notice that in the No Attention variant, the model still learns the Attribute Embedding matrix $\boldsymbol{W}_e$ and refines attribute semantic representation $\{\boldsymbol{v}_a\}_{a=1}^{\boldsymbol{A}}$. On SUN dataset, we observe that not having any attention module achieves the best seen and unseen accuracies. This demonstrates that due to the small number of training images per class, the attention modules overfit to the training set and degrade the performance compared to No Attention variant, which has the least number of parameters. On all datasets, having Self-Calibration prevents bias towards seen classes by trading off accuracy on seen classes for accuracy on unseen classes, which improves the harmonic mean score.

We also observe that attention on attribute has less effect on these three datasets compared to DeepFashion dataset in the main paper. This is because the attributes in these datasets are well-constructed by human annotators, thus attributes are independent from each other and necessary for the recognition of all classes.

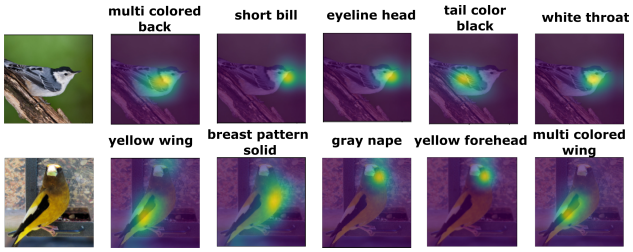

Figure 1: Visualization of our dense attention on two images from two classes in CUB for five attributes with largest attribute attention scores. Each attribute-based spatial attention feature successfully focuses on the relevant region.

In this case, the model learns the necessity of all attributes and produces high attention score for all of them, i.e., $\beta(\boldsymbol{h}_i^a, \boldsymbol{v}_a) \approx 1$. Attributes in DeepFashion, on the other hand, are extracted automatically from the web, thus they are noisy and redundant which requires attention on attributes to augment prediction.

Qualitative Results: Figure 1 shows the results of our dense attention on two unseen classes from CUB. Attributes are ranked according to their attention scores for each image. Notice that our model effectively localizes fine-grained details given weak supervision, i.e., only image labels. Moreover, the model learns to assign the largest attribute attention score to the most visually descriptive attribute of each class, e.g., 'multi-colored back' or 'yellow wing'. This demonstrates the effectiveness of both our dense attribute-based spatial attention and attention on attributes.

Method	Self Calibration	Attention on Attribute	CUB			SUN			AWA2		
			acc_s	acc_u	H	acc_s	acc_u	H	acc_s	acc_u	H
No Attention	No $\mathcal{L}_{ce}$ (seen classes)	No	57.2	25.1	34.9	**34.8**	24.2	28.6	**88.5**	20.9	33.8
Dense Attention		No	65.1	41.4	50.6	32.6	23.9	27.6	83.7	22.6	35.6
Dense Attention		Yes	**65.3**	42.0	51.1	31.9	21.7	25.8	82.5	25.7	39.2
No Attention	Yes $\mathcal{L}_{ce} + \mathcal{L}_{cal}$ (all classes)	No	48.2	43.6	45.8	25.4	**53.8**	34.5	69.9	58.5	63.7
Dense Attention		No	59.6	**56.7**	**58.1**	25.9	48.3	33.7	76.2	**60.5**	**67.5**
Dense Attention		Yes	59.6	**56.7**	**58.1**	24.3	52.3	33.2	75.7	60.3	67.1

Table 1: Ablation study for generalized zero-shot learning on CUB, SUN and AWA2 datasets.