

Learning Deep Representations of Fine-Grained Visual Descriptions

Liangjie Cao

31 May 2018

1. Flowers zero-shot recognition and retrieval

To demonstrate that their results generalize beyond the case of bird images, they report the same set of experiments on the Flowers dataset. All neural text model architectures are the same as they used for CUB, and they used the same hyperparameters from crossvalidation on CUB. Table 2 summarizes their results. From the Table, Char CNN-RNN achieves competitive results to wordlevel models both for DA-SJE and DS-SJE. The wordlevel models achieve the best result, significantly better than both the shallow embeddings and character-level models. Among different models, Word LSTM is the winner for DASJE both in classification and retrieval. On the other hand, Word CNN-RNN is the winner for DS-SJE for the same. As in the case for CUB, DS-SJE achieves strong retrieval performance, and DA-SJE often fails in comparison.

2. Qualitative results

Figure 1 shows several example zero-shot retrieval results using a single text description. Both the text queries and images are real data points drawn from the test set. The authors observe that having trained on our dataset of visual descriptions,

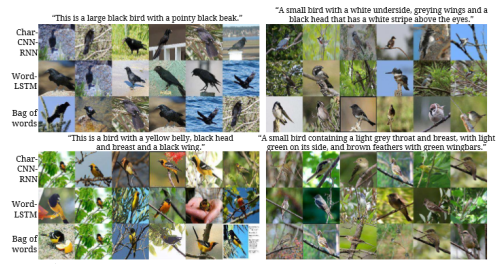


Figure 1. Zero-shot retrieval given a single query sentence. Each row corresponds to a different text encoder.

our proposed method returns results that accurately reflect the text, even when using only a single caption. Quantitatively, BoW achieves 14.6% AP@50 with a single query compared to 18.0% with word-LSTM and 20.7% with Word-CNN-RNN. Note that although almost all retrieved images match the text query well, the actual class of that image can still be incorrect. This is why the average precision may seem low compared to the generally good qualitative results. The performance appears to degrade gracefully; our model at least returns visually-consistent results if not of the correct class. And they show a t-SNE embedding of test-set description embeddings in Figure 2, successfully clustering according to visual similarities (i.e. color, shape). Additional examples from test images and

queries are included in the supplementary material.



Figure 2. t-SNE embedding of test class description embeddings from Oxford-102 (left) and CUB (right), marked with corresponding images. Best viewed with zoom.

Approach	CUB	Flowers
CSHAPH [4]	17.5	-
AHLE [1]	27.3	-
TMV-HLP [3]	47.9	-
SJE [2]	50.1	-
DA-SJE (ours)	54.3	62.3
DS-SJE (ours)	56.8	65.6

Table 1. Summary of zero-shot % classification accuracies

3. Comparison to the state-of-the-art

In this section we compare to the previously published results on CUB, including results that use the same zeroshot split. CSHAPH [4] uses 4K-dim features from the Oxford VGG net [5] and also attributes to learn a hypergraph on the attribute space. AHLE [1] uses Fisher vector image features and attribute embeddings to learn a bilinear compatibility function between these embeddings. TMVHLP [3] builds a hypergraph on a multiview embedding s-

Embedding	Top-1 Acc (%)		AP@50 (%)	
	DA-SJE	DS-SJE	DA-SJE	DS-SJE
WORD2VEC	54.6	54.2	16.3	52.1
BAG-OF-WORDS	56.7	57.7	28.2	57.3
CHAR CNN	51.1	47.3	8.3	46.1
CHAR LSTM	29.1	25.8	19.3	27.0
CHAR CNN-RNN	61.7	63.7	13.6	57.3
WORD CNN	60.2	60.7	8.7	56.3
WORD LSTM	62.3	64.5	45.9	52.3
WORD CNN-RNN	60.9	65.6	7.6	59.6

Table 2. Zero-shot % recognition accuracy and retrieval average precision on Flowers

pace learned via CCA which uses deep image features and attributes. In SJE [2] as in AHLE [1] a compatibility function is learned, in this case between 1K-dim GoogleNet [6] features and various other embeddings including attributes. Overall, the results in Table 1 demonstrate that state-of-the-art zero-shot prediction performance can be achieved directly from text descriptions. This does not require access to any form of test label embeddings. Although attributes are richer and more compact than text descriptions, attributes alone form a very small training set.

4. Discussion

Their visual descriptions data also improved the zero shot accuracy using BoW and word2vec encoders. While these win in the smaller data regime, higher capacity encoders dominate when enough data is available. Thus our contributions (data, objective and text encoders) improve performance at multiple operating points of training text

size.

References

- [1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *IEEE TPAMI*, 38(7):1425–1438, 2016.
- [2] Z. Akata, S. E. Reed, D. J. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In CVPR, 2015.
- [3] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE TPAMI*, 37(11):2332–2345, 2015.
- [4] S. Huang, M. Elhoseiny, A. M. Elgammal, and D. Yang. Learning hypergraph-regularized attribute predictors. In CVPR, 2015.
- [5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In CVPR, 2015.