Learning Deep Representations of Fine-Grained Visual Descriptions

Liangjie Cao
29 May 2018

1. Collecting fine-grained visual descriptions

In this section the authors describe the collection of our new dataset of fine-grained visual descriptions. For each image in CUB and Flowers, they collected ten single-sentence visual descriptions. They used the Amazon Mechanical Turk (AMT) platform for data collection, using non-“Master” certified workers situated in the US with average work approval rating above 95%. Figure 1 shows several representative examples of the results from our data collection. The descriptions almost always accurately describe the image, to varying degrees of comprehensiveness. Thus, in some cases multiple captions might be needed to fully disambiguate the species of bird category. However, as the authors show subsequently, the data is descriptive and large enough to support training high-capacity text models and greatly improve the performance of textbased embeddings for zero-shot learning.

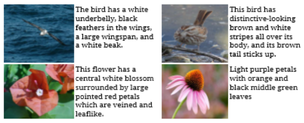
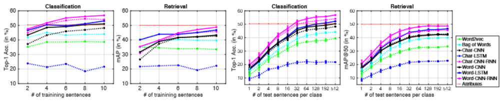

The bird has a white underbelly, black feathers in the wings, a large wingspan, and a white beak.
This bird has distinctive-looking brown and white stripes all over its body, and its brown tail sticks up.
This flower has a central white blossom surrounded by large pointed red petals which are veined and leaflike.
Light purple petals with orange and black middle green leaves

Figure 1. **Example annotations of birds and flowers.**

2. CUB zero-shot recognition and retrieval

In this section the authors describe the protocol and results for our zero-shot tasks. For both recognition and retrieval, they first extract text encodings from test captions and average them per-class. In this experiment they use all test captions and in a later section they vary this number, including using a single caption per class. Table 1 summarizes their results. Both in the classification (first two columns) and for retrieval (last two columns) settings, the symmetric (DS-SJE) formulation of our model improves over the asymmetric (DA-SJE) formulation. Especially for retrieval, DS-SJE performs much better than DA-SJE consistently for all the text embedding variants. It makes the difference between working very well and failing, particularly for the high-capacity models which likely overfit to the classification task in the asymmetric setting. In the classi?cation setting there are notable differences between the language models. For DA-SJE (first column), Char-CNN-RNN (54.0% Top-1 Acc) and Word-CNN-RNN (54.3%) outperform the attributes-based state-of-the-art [1] for zero-shot classification (50.1%). In fact they replicated the attribute-based model in [1] and got slightly better results (50.9%, also reported in Table 1), probably due to training on 10 image crops instead of a single crop. Similar observations hold for DS-SJE (second column). Notably for DS-SJE, Char-CNN-RNN (54.0%), Word-CNN (51.0%), Word-LSTM (53.0%) and Word-CNN-RNN (56.8%) outperform the attributes. [2] In the case of retrieval and DS-SJE (last column), attributes still performs the best (50.0% AP), but Word-CNN-RNN (48.7%) approaches this result.



(a) Increasing number of training sentences

(b) Increasing number of test sentences

Figure 2. **Zero-shot image classfication and retrieval accuracy versus number of sentences per-image used in training and number of sentences in total used for testing. Results reported on CUB.**

3. Effect of visual description training set size

The authors show the performance of several text encoding models in Fig 2. In zero-shot classification, attributes are competitive when two captions per-image are available, but with more training captions the deep network models win. For retrieval, the crossover point might happen with more than ten captions per image as the results seem to be increasing. The baseline word2vec and BoW encodings do not gain much from more data. The results suggests that given a moderate number of sentences, i.e. four per image, neural text encoders improve the performance over the state-ofthe-art attribute-based methods significantly. Figure 2shows the averaged results for zero-shot classfication and for zero-shot retrieval. Both figures include error bars to ¡□1 standard deviation. Note that the error bars are larger towards the left side of both figures because in the few-text case, especially discriminative or especially vague (or wrong) descriptions can have a relatively larger impact on the text embedding quality. BoW again shows a surprisingly good performance, significantly better than word2vec and competitive with Char-CNN. However, the word-level neural text encoders outperform word2vec and BoW at all operating points.

	Top-1 Acc (%)		AP@50 (%)	
Embedding	DA-SJE	DS-SJE	DA-SJE	DS-SJE
ATTRIBUTES	50.9	50.4	20.4	50.0
WORD2VEC	38.7	38.6	7.5	33.5
BAG-OF-WORDS	43.4	44.1	24.6	39.6
CHAR CNN	47.2	48.2	2.9	42.7
CHAR LSTM	22.6	21.6	11.6	22.3
CHAR CNN-RNN	54.0	54.0	6.9	45.6
WORD CNN	50.5	51.0	3.4	43.3
WORD LSTM	52.2	53.0	36.8	46.8
WORD CNN-RNN	54.3	56.8	4.8	48.7

Table 1. **Zero-shot recognition and retrieval on CUB**

References

[1] 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.

[2] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In CVPR, 2016.