Learning Deep Representations of Fine-Grained Visual Descriptions

Scott Reed, Zeynep Akata, Honglak Lee and Bernt Schiele

CVPR 2016

Park, MinKyu

2019.05.30

Dongguk University

Artificial Intelligence Laboratory

Contents

Generating Visual Explanations - Related Work

Fine-grained Classification. Object classification, and fine-grained classification in particular, is attractive to demonstrate explanation systems because describing image content is not sufficient for an explanation. Explanation models must focus on aspects that are both class-specific and depicted in the image.

Most fine-grained zero-shot and few-shot image classification systems use attributes [29] as auxiliary information that can support visual information. Attributes can be thought of as a means to discretize a high dimensional feature space into a series of simple and readily interpretable decision statements that can act as an explanation. However, attributes have several disadvantages. They require fine-grained object experts for annotation which is costly. For each additional class, the list of attributes needs to be revised to ensure discriminativeness so attributes are not generalizable. Finally, though a list of image attributes could help explain a fine-grained classification, attributes do not provide a natural language explanation like the user expects. We therefore, use natural language descriptions collected in [30] which achieved superior performance on zero-shot learning compared to attributes.

Abstract

- zero-shot visual recognition

  a joint embedding problem of images and side information **(attributes)**

- Despite good performance, attributes have **limitations**:

  (1) finer-grained recognition requires commensurately more attributes

  (2) attributes do not provide a natural language interface.

- We propose to overcome these limitations by training neural language models from scratch; i.e. without pre-training and only consuming words and characters.

Introduction

- the problem of relating images and text is still far from solved.
- previous zero-shot learning approaches [13, 2, 3]

  human-encoded attributes [24], or simplified language models such as bag-of-words [16], WordNet-hierarchy-derived features [29], and neural word embeddings such as Word2Vec [28] and GloVE [37].
- Previous text corpora used for fine-grained label embedding were **either very large but not visually focused**,

  e.g. the entire wikipedia, or somewhat visually relevant but very short,

  e.g. the subset of wikipedia articles that are related to birds.

  Furthermore, these wikis do not provide enough aligned images and text to train a high-capacity sentence encoder.

  Given the data limitations, **previous text embedding methods work** surprisingly well for zero-shot visual recognition,

  but there remains a large gap between the text embedding methods and human-annotated attributes (28.4% vs 50.1% average top-1 per-class accuracy on CUB [2]).
- we hypothesize that **higher-capacity text models** are required.

- Our contributions

  First, we **collected two datasets** of fine-grained visual descriptions: one for the Caltech-UCSD birds dataset, and another for the Oxford-102 flowers dataset [32].

  Second, we propose a novel extension of structured joint embedding [2], and show that it can be used for end-to-end training of deep neural language models.

  Third, we evaluate several variants of word- and character-based neural language models, including our novel hybrids of convolutional and recurrent networks for text modeling.

  We demonstrate significant improvements over the state-of-the-art on CUB and Flowers datasets in both zero-shot recognition and retrieval.

Related work

- a remaining challenge is fine-grained image classification [46, 10, 7, 51], i.e. classifying objects of many visually similar classes.
- The setting we study in this work is both *fine-grained* and *zero-shot*, *e.g.* **we want to do fine-grained classification of previously unseen categories of birds and flowers.**
- Zero-shot retrieval and detection have also been studied in [5, 15, 48, 21],

  but no other work has studied zero-shot text-based retrieval in the fine-grained context of CUB and flowers.
- Deep multi-modal representation learning

  In [31], audio and video signals were combined in an autoencoder framework, yielding improved speech signal classification for noisy inputs, and learning a shared representation across modalities.

  In [43], a deep Boltzmann machine architecture was used for multimodal learning on Flickr images and text tags.
- Recent image and video captioning models [26, 45, 20, 49, 8] go beyond tags to generate natural language descriptions.

- Convolutional and recurrent components (CNN-RNN) end-to-end

  for encoding spatial dependencies in segmentation [53] and video classification [30].

  Here **we extend CNN-RNN to learn a visual semantic embedding "from scratch" at the character level**, yielding competitive performance, robustness to typos, and scalability to large vocabulary

- to improve label embeddings for image classification [4, 47, 12, 1, 33].

  Embedding labels in an Euclidean space is an effective way to model latent relationships between classes [4, 47]

  For zero-shot learning, DeViSE [12] and ALE [1] employ two variants of a ranking formulation to learn a compatibility between images and textual side-information.

  ConSe [33] uses the probabilities of a softmax-output layer to weigh the semantic vectors of all the classes.

  Akata et al. [2] showed a large performance gap in zero-shot classification between attributes and unsupervised word embeddings.

- Our contribution builds on previous work on **character-level language models [52] and fine-grained zero-shot learning [1] to train high capacity text encoders from scratch** to jointly embed fine-grained visual descriptions and images.

Deep Structured Joint Embedding

- our approach to jointly embedding images and fine-grained visual descriptions
- As in previous multimodal structured learning methods [1, 2],

  we learn a compatibility function of images and text.

  However, instead of using a bilinear compatibility function we use the inner product of features generated by deep neural encoders.

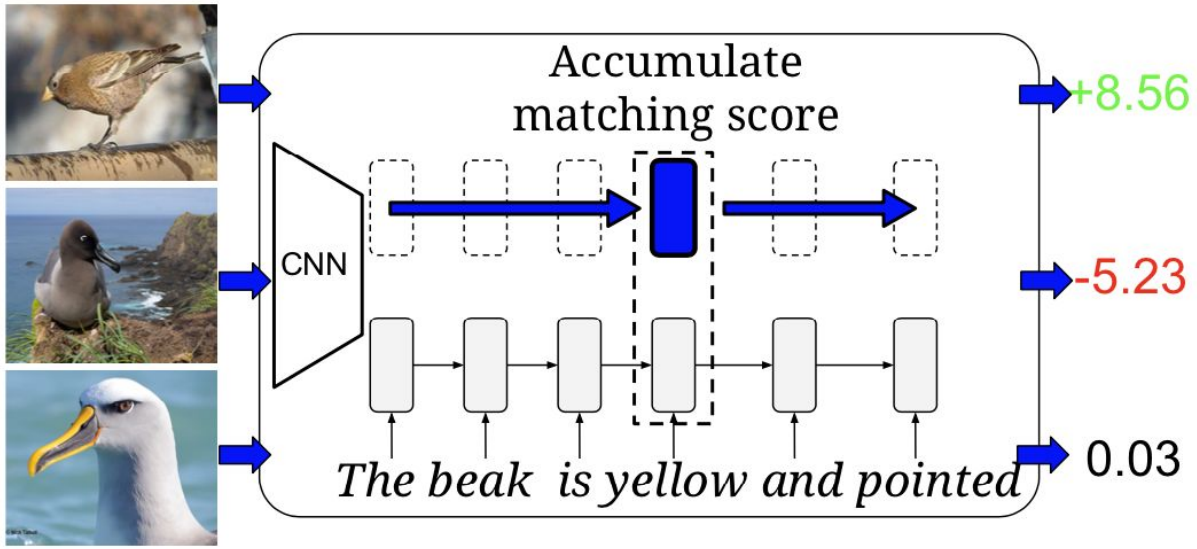


Figure 1: Our model learns a scoring function between images and text descriptions. A word-based LSTM is shown here, but we also evaluate several alternative models.

Objective

Objective. Given data $\mathcal{S} = \{(v_n, t_n, y_n), n = 1, ..., N\}$ containing visual information $v \in \mathcal{V}$, text descriptions $t \in \mathcal{T}$ and class labels $y \in \mathcal{Y}$, we seek to learn functions $f_v : \mathcal{V} \rightarrow \mathcal{Y}$ and $f_t : \mathcal{T} \rightarrow \mathcal{Y}$ that minimize the empirical risk

$$\frac{1}{N} \sum_{n=1}^{N} \Delta(y_n, f_v(v_n)) + \Delta(y_n, f_t(t_n)) \qquad (1)$$

where $\Delta : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ is the 0-1 loss. Note that N is the number of image and text *pairs* in the training set, and so a given image can have multiple corresponding captions.

- deep symmetric structured joint embedding (DS-SJE)
- deep asymmetric structured joint embedding (DA-SJE) : only image encoder *fv* is trained

Inference

Inference. We define a compatibility function $F : \mathcal{V} \times \mathcal{T} \rightarrow \mathbb{R}$ that uses features from learnable encoder functions $\theta(v)$ for images and $\varphi(t)$ for text:

$$F(v, t) = \theta(v)^T \varphi(t) \tag{2}$$

We then formulate image and text classifiers as follows:

$$f_v(v) = \arg\max_{y \in \mathcal{Y}} \mathbb{E}_{t \sim \mathcal{T}(y)}[F(v, t)] \tag{3}$$

$$f_t(t) = \arg\max_{y \in \mathcal{Y}} \mathbb{E}_{v \sim \mathcal{V}(y)}[F(v, t)] \tag{4}$$

- From the perspective of the text encoder, this means that text features must produce a higher compatibility score to a matching image compared to both 1) the score of that image with any mismatching text, and 2) the score of that text with any mismatching image.

Learning

Learning. Since the 0-1 loss is discontinuous, we instead optimize a surrogate objective function (related to equation 1) that is continuous and convex:

$$\frac{1}{N}\sum_{n=1}^{N}\ell_v(v_n,t_n,y_n)+\ell_t(v_n,t_n,y_n) \tag{5}$$

where the misclassification losses are written as:

$$\ell_v(v_n,t_n,y_n) = \max_{y\in\mathcal{Y}}(0,\Delta(y_n,y)+\mathbb{E}_{t\sim\mathcal{T}(y)}[F(v_n,t)-F(v_n,t_n)]) \tag{6}$$

$$\ell_t(v_n,t_n,y_n) = \max_{y\in\mathcal{Y}}(0,\Delta(y_n,y)+\mathbb{E}_{v\sim\mathcal{V}(y)}[F(v,t_n)-F(v_n,t_n)]) \tag{7}$$

- For the image encoder, we keep the network weights fixed to the original GoogLeNet.

Text encoder models

- Text-based ConvNet (CNN)

  The text-based CNN can be viewed as a standard CNN for images, except that the image width is 1 pixel and the number of channels is equal to the alphabet size.

  the Char-CNN, The Word-CNN

- Convolutional Recurrent Net (CNN-RNN)

  To get the benefits of both recurrent models and CNNs, we propose to stack a recurrent network on top of a mid-level temporal CNN hidden layer.

- Long Short-Term Memory (LSTM)

- Baseline representations

  BoW, word2vec, attributes

  The CUB dataset also has per-image attributes, **but we found that using these does not improve performance compared to using a single averaged attribute vector per class.**

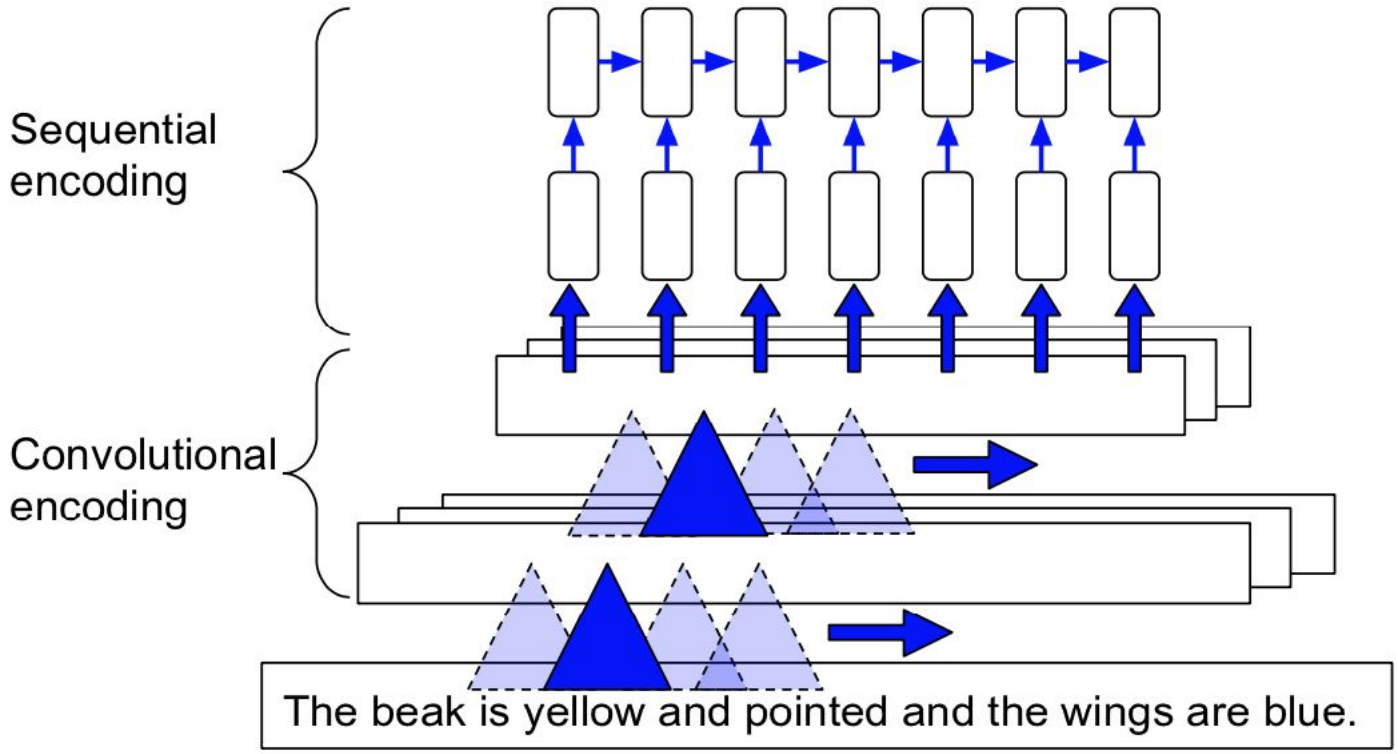


Figure 2: Our proposed convolutional-recurrent net.

Experimental results

- Caltech-UCSD Birds dataset (CUB) and Oxford Flowers 102 (Flowers) dataset

  CUB contains 11,788 bird images from 200 different categories. Flowers contains 8189 flower images from 102 different categories.

  we extracted 1,024-dimensional pooling units from GoogLeNet [44] with batch normalization [19] implemented in Torch2

  The CNN input size (sequence length) was set to 30 for word-level and 201 for character-level models

- Collecting fine-grained visual descriptions

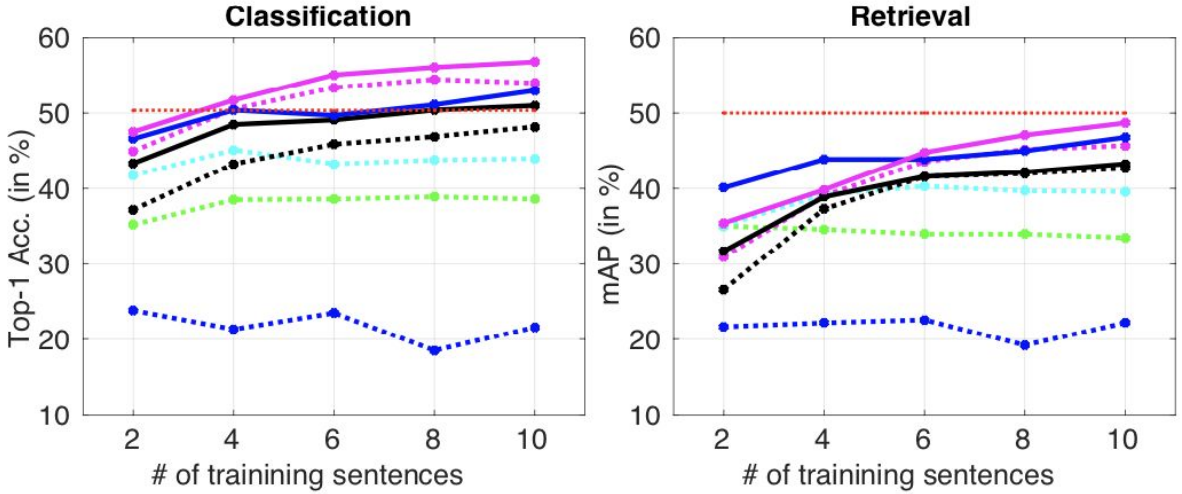
  the Amazon Mechanical Turk (AMT) platform for data collection, using non-“Master” certified workers situated in the US with average work approval rating above 95%

  We asked workers to describe only visual appearance in at least 10 words, to avoid figures of speech, to avoid naming the species even if they knew it, and not to describe the background or any actions being taken.

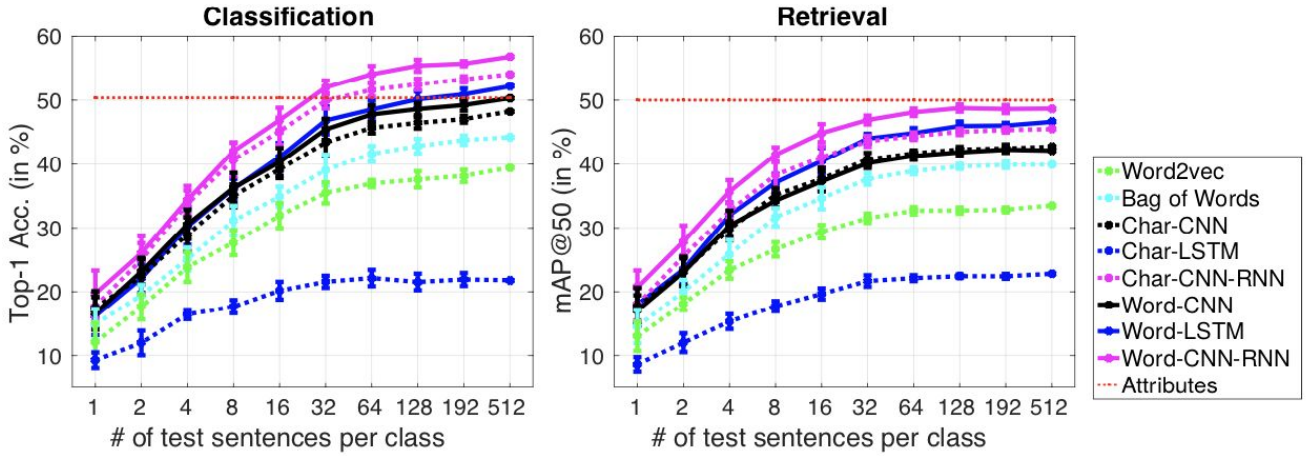
CUB zero-shot recognition and retrieval

	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. "DS-SJE" and "DA-SJE" refer to symmetric and asymmetric forms of our joint embedding objective, respectively.



(a) Increasing number of training sentences



(b) Increasing number of test sentences

Zero-shot image classification 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.

Qualitative results

- zero-shot retrieval results using a single text description

  BoW achieves 14.6% AP@50 with a single query compared to 18.0% with word-LSTM and 20.7% with Word-CNN-RNN

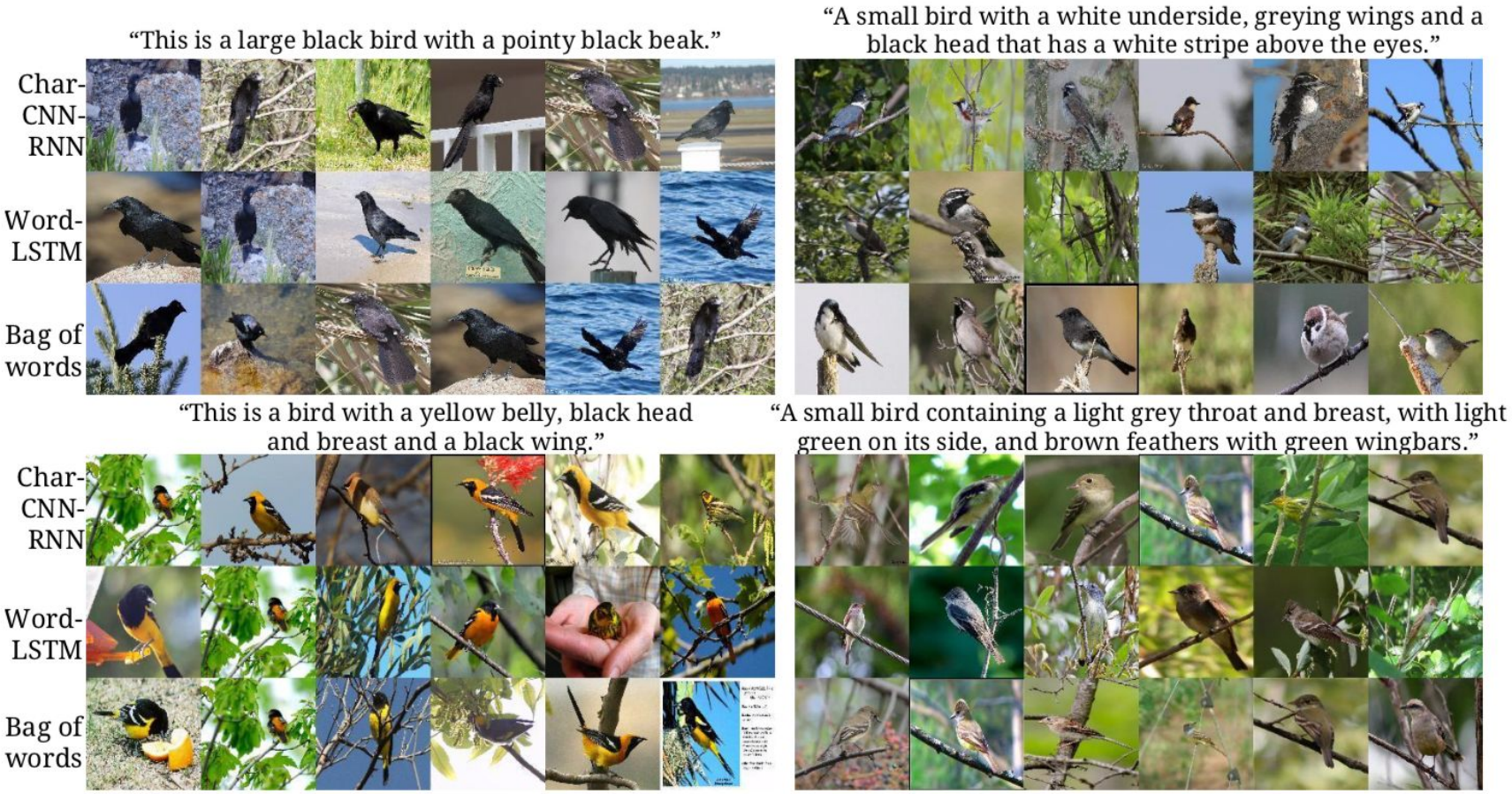


Figure 5: Zero-shot retrieval given a single query sentence. Each row corresponds to a different text encoder.

Comparison to the state-of-the-art

- CSHAP$_H$ [18] uses 4K-dim features from the Oxford VGG net [40] and also attributes to learn a hypergraph on the attribute space.
- AHLE [1] uses Fisher vector image features and attribute embeddings
- TMV-HLP [14] builds a hypergraph on a multiview embedding space 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 [44] features and various other embeddings including attributes.

Approach	CUB	Flowers
CSHAP$_H$ [18]	17.5	–
AHLE [1]	27.3	–
TMV-HLP [14]	47.9	–
SJE [2]	50.1	–
DA-SJE (ours)	54.3	62.3
DS-SJE (ours)	**56.8**	**65.6**

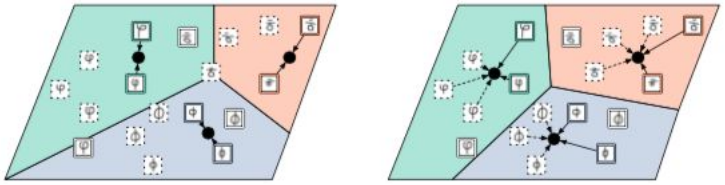
Table 3: Summary of zero-shot % classification accuracies. Note that different features are used in each work, although [1] uses the same features as in this work.

Discussion

- We developed a deep symmetric joint embedding model, collected a high-quality dataset of fine-grained visual descriptions, and evaluated several deep neural text encoders.
- We showed that a text encoder trained from scratch on characters or words can achieve state-of-the-art zero-shot recognition accuracy on CUB, outperforming attributes.
- Our 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.
- our contributions (data, objective and text encoders) improve performance at multiple operating points of training text size.

Opinion

- Visual Descriptions □ Interpretable explanations + Multimodal explanations
- CUB + Flower □ ? : Transfer Learning, Generalized Zero-shot Learning
- Visual Descriptions to Attribute extraction, Attribute Transfer ?
- Prototypical Networks 적용
- zero-shot text-based image retrieval 연구
- 0-1 Loss 연구 □ 비교 분석?



Before Refinement　　After Refinement

Figure 3: Left: The prototypes are initialized based on the mean location of the examples of the corresponding class, as in ordinary Prototypical Networks. Support, unlabeled, and query examples have solid, dashed, and white colored borders respectively. Right: The refined prototypes obtained by incorporating the unlabeled examples, which classifies all query examples correctly.

Learning with Average Top-k Loss

Yanbo Fan[3,4,1] **, Siwei Lyu**[1]*, **Yiming Ying**[2] **, Bao-Gang Hu**[3,4]
[1]Department of Computer Science, University at Albany, SUNY
[2]Department of Mathematics and Statistics, University at Albany, SUNY
[3]National Laboratory of Pattern Recognition, CASIA
[4]University of Chinese Academy of Sciences (UCAS)
{yanbo.fan,hubg}@nlpr.ia.ac.cn, slyu@albany.edu, yying@albany.edu