Learning Deep Representations of Fine-Grained Visual Descriptions

Liangjie Cao

25 May 2018

1. Introduction

State-of-the-art methods for zero-shot visual recognition formulate learning as a joint embedding problem of images and side information. Their proposed models train end-to-end to align with the fine-grained and category-specific content of images. Actually natural language provides a flexible and compact way of encoding only the salient visual aspects for distinguishing categories. By training on raw text, our model can do inference on raw text as well, providing humans a familiar mode both for annotation and retrieval. A key challenge in image understanding is to correctly relate natural language concepts to the visual content of images. In recent years there has been significant progress in learning visual-semantic embeddings, *e.g.* for zero-shot learning [1, 7, 12, 13, 16–18], and automatically generating image captions for general web images. [4, 9, 11, 15, 20]

The authors contributions in this work are as follows. First, they collected two datasets of negrained visual descriptions: one for the Caltech-UCSD birds dataset, and another for the Oxford-102 flowers dataset [8]. Both data and code will be made available. Second, they propose a novel extension of structured joint embedding [1], and show that it can be used for end-to-end training of deep neural language models. It also dramatically improves zero-shot retrieval performance for all models. Third, they evaluate several variants of word- and character-based neural language models, including our novel hybrids of convolutional and recurrent networks for text modeling. They demonstrate signicant improvements over the state-of-the-art on CUB and Flowers datasets in both zero-shot recognition and retrieval.

2. Related work

In the past few years, advances in deep convolutional networks [5, 10, 19] have driven rapid progress in general-purpose visual recognition on large-scale benchmarks such as ImageNet. [2] The learned features of these networks have proven transferable to many other problems [14]. However, a remaining challenge is finegrained image classifcation [3, 6, 21, 22], *i.e.* classifying objects of many visually similar classes. The difficulty is increased by the lack of extensive labeled images, which for fine-grained data sets may even require annotation by human experts. They demonstrate that with suf?cient training data, text-based label embeddings can outperform the previous attributes-

based state-of-the art for zero-shot recognition on CUB (at both word and character level). It is also possible to build an asymmetric model in the opposite direction, *i.e.* only train ft in order to perform zero-shot image retrieval, although we are not aware of previous works doing this. From a practical perspective it is clearly better to have a single model that does both tasks well. Thus in the experiments they compare DS-SJE with DA-SJE (training only fv) for zero-shot classi?cation.

1	Translation language
2	Control robot
3	Image analysis
4	Image analysis

Table 1. **What can lstm - based system do**

3. Deep Structured Joint Embedding

As in previous multimodal structured learning methods, the authors 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. An instantiation of our model using a word-level LSTM(Table 1) is illustrated in Figure 1. Since their text encoder models are all differentiable, their backpropagate (sub)-gradients through all text network parameters for end-to-end training. For the image encoder, they keep the network weights ?xed to the original GoogLeNet. I will continue learning the following days.

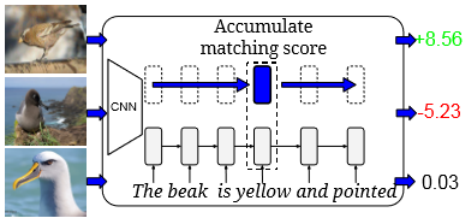

Figure 1. **RON object detection overview**

References

[1] Zeynep Akata, Scott E Reed, Daniel J Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. *computer vision and pattern recognition*, pages 2927–2936, 2015.

[2] Jia Deng, Wei Dong, Richard Socher, Lijia Li, Kai Li, and Li Feifei. Imagenet: A large-scale hierarchical image database. pages 248–255, 2009.

[3] Jia Deng, Jonathan Krause, and Li Feifei. Fine-grained crowdsourcing for fine-grained recognition. pages 580–587, 2013.

[4] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. *computer vision and pattern recognition*, pages 2625–2634, 2015.

[5] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. *international conference on machine learning*, pages 647–655, 2014.

[6] Kun Duan, Devi Parikh, David J Crandall, and Kristen Grauman. Discovering localized attributes for fine-grained recognition. pages 3474–3481, 2012.

[7] Andrea Frome, Gregory S Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marcaurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. pages 2121–2129, 2013.

[8] Yuli Gao and Jianping Fan. Automatic function selection for large scale salient object detection. pages 97–100, 2006.

[9] Andrej Karpathy and Li Feifei. Deep visual-semantic alignments for generating image descriptions. *computer vision and pattern recognition*, pages 3128–3137, 2015.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. pages 1097–1105, 2012.

[11] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating simple image descriptions. pages 1601–1608, 2011.

[12] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[13] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *international conference on learning representations*, 2014.

[14] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. pages 1717–1724, 2014.

[15] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. pages 1143–1151, 2011.

[16] Mark Palatucci, Dean A Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. 22:1410–1418, 2009.

[17] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. pages 1641–1648, 2011.

[18] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Y Ng. Zero-shot learning through cross-modal transfer. *neural information processing systems*, pages 935–943, 2013.

[19] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *computer vision and pattern recognition*, pages 1–9, 2015.

[20] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tel-

l: A neural image caption generator. *computer vision and pattern recognition*, pages 3156–3164, 2015.

[21] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. *California Institute of Technology*, 2010.

[22] Ning Zhang, Jeff Donahue, Ross B Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. *european conference on computer vision*, pages 834–849, 2014.