GENERALIZATION TO NEW COMPOSITIONS OF KNOWN ENTITIES IN IMAGE UNDERSTANDING

Yuval Atzmon
Bar Ilan University, Israel
yuval.atzmon@biu.ac.il

Jonathan Berant & Amir Globerson
Tel Aviv University, Israel

Vahid Kazemi & Gal Chechik
Google Research,
Mountain View CA, USA

ABSTRACT

Recurrent neural networks can be trained to describe images with natural language, but it has been observed that they generalize poorly to new scenes at test time. Here we provide an experimental framework to quantify their generalization to unseen compositions. By describing images using short structured representations, we tease apart and evaluate separately two types of generalization: (1) generalization to new images of similar scenes, and (2) generalization to unseen compositions of known entities. We quantify these two types of generalization by a large-scale experiment on the MS-COCO dataset with a state-of-the-art recurrent network, and compare to a baseline structured prediction model on top of a deep network. We find that a state-of-the-art image captioning approach is largely "blind" to new combinations of known entities (∼2.3% precision@1), and achieves statistically similar precision@1 to that of a considerably simpler structured-prediction model with much smaller capacity. We therefore advocate using compositional generalization metrics to evaluate vision and language models, since generalizing to new combinations of known entities is key for understanding complex real data.

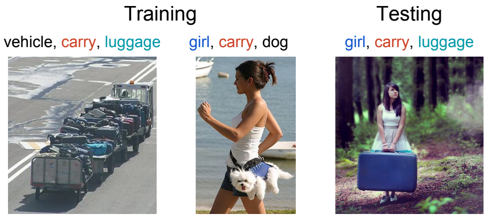
1 INTRODUCTION

Recently, deep neural networks were successfully used for training models that describe images with natural language. While the results were both inspiring and impressive, it became clear in the aftermath of analyzing the results, that current approaches suffer from two fundamental issues. First, generalization was poor for images describing scenarios not seen at training time. Second, evaluating descriptions was challenging, because strong language models can generate sensible descriptions that are missing essential components in the image (Cui et al., 2015; Anderson et al., 2016). Here we propose to address these issues by using structured representations for image descriptions. As a first step, we use a simple representation consisting of subject-relation-object (SRO) triplets (Farhadi et al., 2010; Gupta et al., 2012). By reducing sentences to an SRO representation, we focus on the composition of entities in an image. This allows to partition the data such that the model is tested only on *new unseen compositions*, which are not included in the training set. Our key observation is that one should separate two kinds of generalization when generating image descriptions. The first, generalizing to new images of the same class, is routinely being evaluated in the current data split of the MS-COCO challenge (Lin et al., 2014). The second type, which we focus on, is concerned with generalizing to new scenarios, akin to *transfer* or *zero-shot* learning (Fei-Fei et al., 2006), where learning is extended to semantically-similar classes. Importantly, this generalization is the crux of learning in scenes, since both language and visual scenes are compositional, resulting in an exponentially large set of possible descriptions. Hence, a key component of image captioning is to properly quantify generalization to new combinations of known entities and relations. In practice, we first map image descriptions to short open-IE style phrases of the form subject-relation-object (termed *SRO triplets*). We then partition the examples such that the test and training sets share no common images or SRO triplets, (Fig. 1).

2 RELATED WORK

There is already substantial body of work on image captioning, which is not reviewed here. Relevant to the currnet work, it has been shown that spatial attention can improve captioning quality (Xu et al., 2015; Lu et al., 2016; Rennie et al., 2016), and that further improvement can be achieved using

Figure 1: Learning to generalize to new compositions of entities in images, reflected in their descriptions. Each image is represented with subject-relation-object (SRO) tuple. In a compositional split, testing is performed over novel compositions of entities observed during training.



Visual-Concepts (VC) predictions (Fang et al., 2015; You et al., 2016; Chen et al., 2016). Compositional aspects of language and images were explored using both DNNs and structured models. Andreas et al. (2015) addressed a visual QA task by breaking questions into substructures, and composing modular networks. Socher et al. (2011) and (Lin et al., 2016) aggregated objects into parse trees using recursive NNs. Out-of-vocabulary terms are often handled by "smearing" to semantically-related entities using external language priors (Hendricks et al., 2016; Frome et al., 2013; Xian et al., 2016), and smearing also improves subject-relation-object prediction (Lu et al., 2016). Several authors used structured CRF models: Kulkarni et al. (2011) learned spatial relations for generating descriptions based on templates; Zitnick et al. (2013) generated synthetic scenes. Yatskar et al. (2016) modelled combinations of entities using CRFs. Closely related to our structured approach, Farhadi et al. (2010) matched sentences and images, through a space of meaning parametrized by subject-verb-object triplets, and Johnson et al. (2015) combined subjects, objects and relationships in a graph structure for image retrieval.

3 Experiments

3.1 The Data

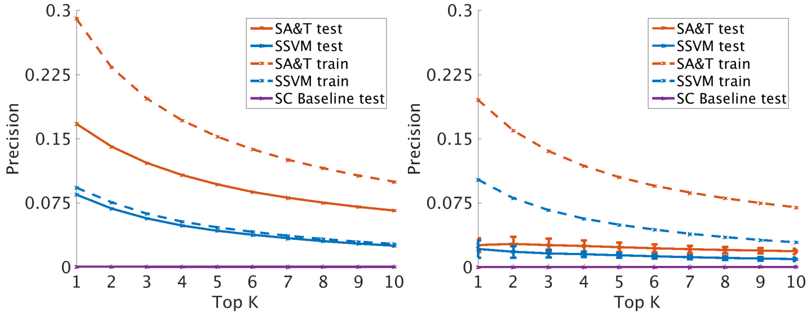
We evaluated image captioning on the MS-COCO data (Lin et al., 2014), currently the standard benchmark for evaluating image captioning models (122K images, ≤ 5 textual descriptions per image). We parsed MS-COCO descriptions into SRO triplets by first constructing dependency parse trees for each description (Andor et al., 2016), and then using manually-constructed patterns to extract triplets from each description. Finally, relation words were stemmed. Removing descriptions without SROs (due to noun phrases, rare prepositions, or parsing errors), yielded 444K (image, SRO) pairs. We will provide the triplets and splits online. We limited the vocabulary to the 800 most frequent nouns (covering 89%) and 100 stemmed relations (90%). Taking only SRO labels that jointly participate at the vocabulary and removing duplicate SROs per image yielded 268K unique pairs of (image, SRO) over 105K unique images. This dataset was split in two ways: by intersecting with the COCO benchmark split, and in a compositional way as described in Section 3.3.

3.2 Compared Methods

We compare four methods and baselines: **(1) SSVM/Conv**. Jointly predict the subject, relation and object of an SRO triplet, by training a structured-prediction model on top of Visual-Concepts (VC) convolutional network (Fang et al., 2015). VC optimizes a multiple-instance-learning (MIL) objective, yielding probabilities of $P(S|Image)$, $P(R|Image)$ and $P(O|Image)$. On top of VC, we trained a *structured SVM* (SSVM) (Tsochantaridis et al., 2005), minimizing the hinge loss between the predicted and ground-truth SRO triplets. The model learns a score function over SRO triplets, decomposed as: $f(s, r, o) = w_S f_S(s) + w_O f_O(o) + w_R f_R(r) + w_{SR} f_{SR}(s, r) + w_{RO} f_{RO}(r, o)$, where $w_S, w_O, w_R, w_{SR}, w_{RO}$ are scalar weights learned by the algorithm. Node scores $f_S(s)$, $f_O(o)$, $f_R(r)$ were set as the VC predictions of $P(S|Image)$, $P(R|Image)$ and $P(O|Image)$. The binary features $f_{SR}(s, r)$, $f_{RO}(r, o)$ were set as the bigram probability of (s, r) and (r, o) in the training SRO data. **(2) Show-Attend-and-Tell (SA&T)**. An LSTM with attention model for caption generation (Xu et al., 2015). For a fair comparison with the structured model we replaced the VGG 512x14x14 convolutional feature maps with the 900x12x12 VC spatial response maps, where each 12x12 response map represents a single visual concept.[1] We re-trained the decoder layers to predict SRO triplets with soft-attention. Hyper-parameters were tuned to minimize perplexity on a validation set, learning rate in $2 \cdot (10^{-3}, 10^{-4}, \ldots 10^{-6})$ and weight decay in $(0, 10^{-6}, \ldots, 10^{-3})$. Importantly, we also controlled for model capacity by tuning the embedding dimensionality $(128, 512, 1024)$ and the LSTM dimensionality $(200, 600, 1800, 3600)$. The remaining parameters were set as in Xu et al. (2015). **(3) Stochastic conditional (SC)**. Draw R based on

[1] We found that replacing VGG convolutional feature maps by VC spatial response maps also improves performance for SA&T.

the training distribution, then draw S and O based on $p_{train}(S|R)$, $p_{train}(O|R)$. This baseline is designed to capture the gain that can be attributed to bigram statistics. **(4) Most frequent triplet (MF)**. Predict an SRO consisting of the most frequent subject, most frequent relation, and most frequent object, based on the training set. By construction, for the compositional split, the most frequent SRO in the train set can not appear in the test set.



(a) **COCO split** (b) **Compositional split**

Figure 2: Comparing precision@k for SA&T vs. SSVM/conv. Error bars denote the standard deviation of the precision across five compositional folds. Both approaches are largely "blind" to unseen compositions and achieve statistically similar $p@1$.

3.3 Evaluation Procedure

We experimented with two cross-validation procedures. **(1) COCO split**: The split provided by ms-coco, restricted to the set of images with SRO triplets. Since captions are not provided for the COCO test set, we used the COCO validation set for evaluations. **(2) Compositional split**: We first split unique SRO triplets to training, validation and test sets (80%/10%/10%). To determines how images are split, note that an image may have more than one caption, and as a result may have SRO triplets on more than one split set. We remove images from the training set that have SRO triplets of the test set, and then removed images from the test set that have SRO triplets from the train set (same for the validation set). Repeating this process with 5 randomizations, yielded an average of 140K training triplets over 61K unique images and an average of 5.1K testing triplets (4K images). Since images have up to 5 descriptions, they may have more than one ground-truth SRO. For SSVM, we computed precision@k by ranking candidate SRO triplets by their scores and comparing against the set of ground-truth SRO triplets.

3.4 Results

Compositional vs. within-class generalization: We find that (1) SA&T is largely "blind" to new combinations of known entities, and (2) it achieves statistically similar precision@1 to that of a considerably simpler structured-prediction model with much smaller capacity.

Figure 2 shows the average precision@k across images, comparing SSVM to SA&T for both their test and training performance. In standard **COCO split** (left panel) SA&T model (red) wins with test precision of $p@1 = 16.7\%$ and the SSVM/Conv model (blue) achieves $p@1 = 8.4\%$. Test precision of the baselines was $p@1 = 0.026\%$ for SC, and 0% for MF. Both approaches mostly predicted compositions that appeared on the train set (LSTM 94%, SSVM 91%). In the **compositional split** (right panel), SSVM test precision (solid blue) was $p@1 = 2.1\% \pm 1\%$, which is statistically similar to LSTM (red) $p@1 = 2.5\% \pm 0.8\%$. Test precision of the baselines (purple) was $p@1 = 0\%$ for SC. The most frequent S, R and O in the dataset were *man*, *with* and *field*, but the triplet (man with field) did not appear at all in the data, yielding 0% MF accuracy. .

Model capacity: The large generalization gap, manifested by the precision difference between training and test set, could be due to over-fitting. We tested SA&T with different capacities, varying the number of parameters (word dimensionality and LSTM hidden state dimensionality), but found that the generalization gap was always very large. This shows that merely reducing the capacity of the SA&T model was not sufficiently effective to control overfitting for the compositional case.

3.5 Conclusion

This paper highlights the role of generalization to new combinations of known objects in vision-to-language problems, and proposes an experimental framework to measure compositional generalization. In a large-scale experiment, we found that existing state-of-the-art image captioning models generalize poorly to new combinations of known entities, and achieve statistically similar precision@1 to that of a structured-prediction model with much smaller capacity.

REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.

Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins. Globally normalized transition-based neural networks. *arXiv preprint arXiv:1603.06042*, 2016.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Deep compositional question answering with neural module networks. *arXiv preprint arXiv:1511.02799*, 2015.

Wenhu Chen, Aurelien Lucchi, and Thomas Hofmann. Bootstrap, review, decode: Using out-of-domain textual data to improve image captioning. *arXiv preprint arXiv:1611.05321*, 2016.

Yin Cui, Matteo Ruggero Ronchi, and Tsung-Yi Lin. 1st captioning challenge. In *Large-scale Scene UNderstanding Workshop, CVPR*, 2015.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1473–1482, 2015.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pp. 15–29. Springer, 2010.

Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):594–611, 2006.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pp. 2121–2129, 2013.

Ankush Gupta, Yashaswi Verma, and C. Jawahar. Choosing linguistics over vision to describe images. In *AAAI Conference on Artificial Intelligence*, 2012.

Lisa Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A Shamma, Michael S Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pp. 3668–3678. IEEE, 2015.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*, 2011.

Liang Lin, Guangrun Wang, Rui Zhang, Ruimao Zhang, Xiaodan Liang, and Wangmeng Zuo. Deep structured scene parsing by learning with image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pp. 740–755. Springer, 2014.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016.

J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. *arXiv preprint arXiv:1612.01887*, 2016.

S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical Sequence Training for Image Captioning. *arXiv preprint arXiv:1612.00563*, 2016.

Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 129–136, 2011.

Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. In *Journal of Machine Learning Research*, pp. 1453–1484, 2005.

Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. *arXiv preprint arXiv:1603.08895*, 2016.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2015. https://github.com/kelvinxu/arctic-captions.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2016.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Lawrence Zitnick, Devi Parikh, and Lucy Vanderwende. Learning the visual interpretation of sentences. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1681–1688, 2013.