A Distributed Representation Based Query Expansion Approach for Image Captioning

Semih Yagcioglu[1] **Erkut Erdem**[1] **Aykut Erdem**[1] **Ruket Çakıcı**[2]
[1] Hacettepe University Computer Vision Lab (HUCVL)
Dept. of Computer Engineering, Hacettepe University, Ankara, TURKEY
semih.yagcioglu@hacettepe.edu.tr, {erkut,aykut}@cs.hacettepe.edu.tr
[2] Dept. of Computer Engineering, Middle East Technical University, Ankara, TURKEY
ruken@ceng.metu.edu.tr

Abstract

In this paper, we propose a novel query expansion approach for improving transfer-based automatic image captioning. The core idea of our method is to translate the given visual query into a distributional semantics based form, which is generated by the average of the sentence vectors extracted from the captions of images visually similar to the input image. Using three image captioning benchmark datasets, we show that our approach provides more accurate results compared to the state-of-the-art data-driven methods in terms of both automatic metrics and subjective evaluation.

1 Introduction

Automatic image captioning is a fast growing area of research which lies at the intersection of computer vision and natural language processing and refers to the problem of generating natural language descriptions from images. In the literature, there are a variety of image captioning models that can be categorized into three main groups as summarized below.

The first line of approaches attempts to generate novel captions directly from images (Farhadi et al., 2010; Kulkarni et al., 2011; Mitchell et al., 2012). Specifically, they borrow techniques from computer vision such as object detectors and scene/attribute classifiers, exploit their outputs to extract the visual content of the input image and then generate the caption through surface realization. More recently, a particular set of generative approaches have emerged over the last few years, which depends on deep neural networks (Chen and Zitnick., 2015; Karpathy and Fei-Fei, 2015; Xu et al., 2015; Vinyals et al., 2015). In general, these studies combine convolutional neural networks (CNNs) with recurrent neural networks (RNNs) to generate a description for a given image.

The studies in the second group aim at learning joint representations of images and captions (Hodosh et al., 2013; Socher et al., 2014; Karpathy et al., 2014). They employ certain machine learning techniques to form a common embedding space for the visual and textual data, and perform cross-modal (image-sentence) retrieval in that intermediate space to accordingly score and rank the pool of captions to find the most proper caption for a given image.

The last group of works, on the other hand, follows a data-driven approach and treats image captioning as a caption transfer problem (Ordonez et al., 2011; Kuznetsova et al., 2012; Patterson et al., 2014; Mason and Charniak, 2014). For a given image, these methods first search for visually similar images and then use the captions of the retrieved images to provide a description, which makes them much easier to implement compared to the other two classes of approaches.

The success of these data-driven approaches depends directly on the amount of data available and the quality of the retrieval set. Clearly, the image features and the corresponding similarity measures used in retrieval play a significant role here but, as investigated in (Berg et al., 2012), what makes this particularly difficult is that while describing an image humans do not explicitly mention every detail. That is, some parts of an image are more salient than the others. Hence, one also needs to bridge the semantic gap between what is there in the image and what people say when describing it.

As a step towards achieving this goal, in this paper, we introduce a novel automatic query expansion approach for image captioning to retrieve semantically more relevant captions. As illustrated in Fig. 1, we swap modalities at our query expan-

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 106–111,
Beijing, China, July 26-31, 2015.

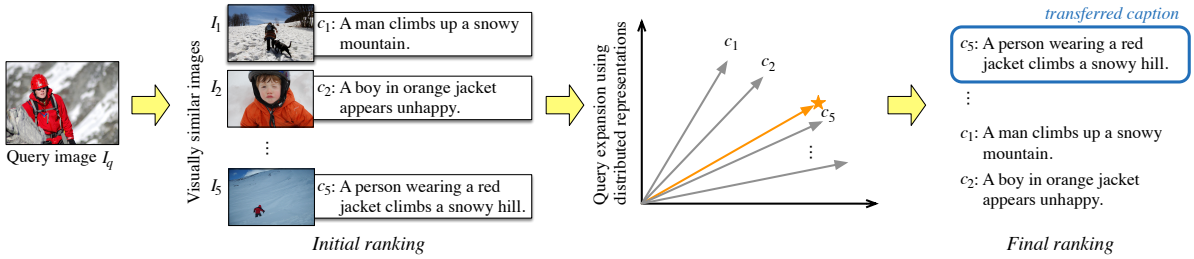

Figure 1: A system overview of the proposed query expansion approach for image captioning.

sion step and synthesize a new query, based on distributional representations (Baroni and Lenci, 2010; Turney and Pantel, 2010; Mikolov et al., 2013; Pennington et al., 2014) of the captions of the images visually similar to the input image. Through comprehensive experiments over three benchmark datasets, we show that our model improves upon existing methods and produces captions more appropriate to the query image.

2 Related Work

As mentioned earlier, a number of studies pose image captioning as a caption transfer problem by relying on the assumption that visually similar images generally contain very similar captions. The pioneering work in this category is the im2text model by Ordonez et al. (2011), which suggests a two-step retrieval process to transfer a caption to a given query image. The first step, which provides a baseline for the follow-up caption transfer approaches, is to find visually similar images in terms of some global image features. In the second step, according to the retrieved captions, specific detectors and classifiers are applied to images to construct a semantic representation, which is then used to re-rank the associated captions.

Kuznetsova et al. (2012) proposed performing multiple retrievals for each detected visual element in the query image and then combining the relevant parts of the retrieved captions to generate the output caption. Patterson et al. (2014) extended the baseline model by replacing global features with automatically extracted scene attributes, and showed the importance of scene information in caption transfer. Mason and Charniak (2014) formulated caption transfer as an extractive summarization problem and proposed to perform the re-ranking step by means of a word frequency-based representations of captions. More recently, Mitchell et al. (2015) proposed to select the caption that best describes the remaining descriptions of the retrieved similar images wrt an n-gram overlap-based sentence similarity measure.

In this paper, we take a new perspective to data-driven image captioning by proposing a novel query expansion step based on compositional distributed semantics to improve the results. Our approach uses the weighted average of the distributed representations of retrieved captions to expand the original query in order to obtain captions that are semantically more related to the visual content of the input image.

3 Our Approach

In this section, we describe the steps of the proposed method in more detail.

3.1 Retrieving Visually Similar Images

Representing Images. Data-driven approaches such as ours rely heavily on the quality of the initial retrieval, which makes having a good visual feature of utmost importance. In our study, we use the recently proposed Caffe deep learning features (Jia et al., 2014), trained on ImageNet, which have been proven to be effective in many computer vision problems. Specifically, we use the activations from the seventh hidden layer (fc7), resulting in a 4096-dimensional feature vector.

Adaptive Neighborhood Selection. We create our expanded query by using the distributed representations of the captions associated with the retrieved images, and thus, having no outliers is also an important factor for the effectiveness of the approach. For this, instead of using a fixed neighborhood, we adopt an adaptive strategy to select the initial candidate set of image-caption pairs $\{(I_i, c_i)\}$.

For a query image I_q, we utilize a ratio test and only consider the candidates that fall within a radius defined by the distance score of the query im-

age to the nearest training image $I_{closest}$, as

$$\mathcal{N}(I_q) = \{(I_i, c_i) \mid dist(I_q, I_i) \leq (1+\epsilon) dist(I_q, I_{closest}), \\ I_{closest} = \arg\min\ dist(I_q, I_i), I_i \in \mathcal{T}\} \quad (1)$$

where $dist$ denotes the Euclidean distance between two feature vectors, $\mathcal{N}$ represents the candidate set based on the adaptive neighborhood, $\mathcal{T}$ is the training set, and ϵ is a positive scalar value[1].

3.2 Query Expansion Based on Distributed Representations

Representing Words and Captions. In this study, we build our query expansion model on the distributional models of semantics where the meanings of words are represented with vectors that characterize the set of contexts they occur in a corpus. Existing approaches to distributional semantics can be grouped into two, as count and predict-based models (Baroni et al., 2014). In our experiments, we tested our approach using two recent models, namely *word2vec* (Mikolov et al., 2013) and *GloVe* (Pennington et al., 2014), and found out that the predict-based model of Mikolov et al. (2013) performs better in our case.

To move from word level to caption level, we take the simple addition based compositional model described in (Blacoe and Lapata, 2012) and form the vector representation of a caption as the sum of the vectors of its constituent words. Note that here we only use the non-stop words in the caption.

Query Expansion. For a query image I_q, we first retrieve visually similar images from a large dataset of captioned images. In our query expansion step, we swap modalities and construct a new query based on the distributed representations of captions. In particular, we expand the original visual query with a new textual query based on the weighted average of the vectors of the retrieved captions as follows:

$$q = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} sim(I_q, I_i) \cdot {c_i}^j \quad (2)$$

where N and M respectively denote the total number of image-caption pairs in the candidate set $\mathcal{N}$ and the number of reference captions associated with each training image, and $sim(I_q, I_i)$ refers to the visual similarity score of the image I_i to the query image ${I_q}$[2] which is used to give more importance to the captions of images visually more close to the query image.

Then, we re-rank the candidate captions by estimating the cosine distance between the distributed representation of the captions and the expanded query vector q, and finally transfer the closest caption as the description of the input image.

4 Experimental Setup and Evaluation

In the following, we give the details about our experimental setup.

Corpus. We estimated the distributed representation of words based on the captions of the MS COCO (Lin et al., 2014) dataset, containing 620K captions. As a preprocessing step, all captions in the corpus were lowercased, and stripped from punctuation.

In the training of word vectors, we used 500 dimensional vectors obtained with both *GloVe* (Pennington et al., 2014) and *word2vec* (Mikolov et al., 2013) models. The minimum word count was set to 5, and the window size was set to 10. Although these two methods seem to produce comparable results, we found out that *word2vec* gives better results in our case, and thus we only report our results with *word2vec* model.

Datasets. In our experiments, we used the popular Flickr8K (Hodosh et al., 2013), Flickr30K (Young et al., 2014), MS COCO (Lin et al., 2014) datasets, containing 8K, 30K and 123K images, respectively. Each image in these datasets comes with 5 captions annotated by different people. For each dataset, we utilized the corresponding validation split to optimize the parameters of our method, and used the test split for evaluation where we considered all the image-caption pairs in the training and the validation splits as our knowledge base.

Although Flickr8K, and Flickr30K datasets have been in use for a while, MS COCO dataset is under active development and might be subject to change. Here, we report our results with version 1.0 of MS COCO dataset where we used the train, validation and test splits provided by (Karpathy et al., 2014).

We compared our proposed approach against the adapted baseline model (VC) of *im2text* (Ordonez et al., 2011) which corresponds to using the caption of the nearest visually similar im-

[1] The adaptive neighborhood parameter ϵ was emprically set to 0.15.

[2] We define $sim(I_q, I_i) = 1 - dist(I_q, I_i)/Z$ where Z is a normalization constant.

MC-KL	a black and white dog is playing or fighting with a brown dog in grass	a man is sitting on a blue bench with a blue blanket covering his face	a man in a white shirt and sunglasses is holding hands with a woman wearing a red shirt outside	one brown and black pigmented bird sitting on a tree branch
MC-SB	a dog looks behind itself	a girl looks at a woman s face	a woman and her two dogs are walking down the street	a tree with many leaves around it
VC	a brown and white dog jumping over a red yellow and white pole	a father feeding his child on the street	a girl is skipping across the road in front of a white truck	a black bear climbing a tree in forest area
OURS	a brown and white dog jumps over a dog hurdle	a man in a black shirt and his little girl wearing orange are sharing a treat	a girl jumps rope in a parking lot	a bird standing on a tree branch in a wooded area
HUMAN	a brown and white sheltie leaping over a rail	a man and a girl sit on the ground and eat	a girl is in a parking lot jumping rope	a painted sign of a blue bird in a tree in the woods

Figure 2: Some example input images and the generated descriptions.

	Flickr8K			Flickr30K			MS COCO		
	BLEU	METEOR	CIDEr	BLEU	METEOR	CIDEr	BLEU	METEOR	CIDEr
OURS	**3.78**	**11.57**	**0.31**	**3.22**	**10.06**	0.20	**5.36**	**13.17**	**0.58**
MC-KL	2.71	10.95	0.15	2.02	9.92	0.07	4.04	12.56	0.37
MC-SB	3.08	9.06	0.27	2.76	8.06	**0.20**	5.02	11.78	0.56
VC	2.79	8.91	0.19	2.33	7.53	0.14	3.71	10.07	0.35
HUMAN	7.59	17.72	2.67	6.52	15.70	2.53	7.42	16.73	2.70

Table 1: Comparison of the methods on the benchmark datasets based on automatic evaluation metrics.

age, and the word frequency-based approaches of Mason and Charniak (2014) (MC-SB and MC-KL). We also provide the human agreement results (HUMAN) by comparing one groundtruth caption against the rest.

For a fair comparison with the MC-SB and MC-KL models (Mason and Charniak, 2014) and the baseline approach VC, we used the same image similarity metric and training splits in retrieving visually similar images for all models. For human agreement, we had five groundtruth image captions, thus we determine the human agreement score by following a leave-one-out strategy. For display purposes, we selected one description randomly from the available five groundtruth captions in the figures.

Automatic Evaluation. We evaluated our approach with a range of existing metrics, which are thoroughly discussed in (Elliott and Keller, 2014; Vedantam et al., 2015). We used smoothed BLEU (Papineni et al., 2002) for benchmarking purposes. We also provided the scores of METEOR (Denkowski and Lavie, 2014) and the recently proposed CIDEr metric (Vedantam et al., 2015), which has been shown to correlate well with the human judgments in (Elliott and Keller, 2014) and (Vedantam et al., 2015), respectively[3].

Human Evaluation. We designed a subjective experiment to measure how relevant the transferred caption is to a given image using a setup similar to those of (Kuznetsova et al., 2012; Mason and Charniak, 2014)[4]. In this experiment, we provided human annotators an image and a candidate description where it is rated according to a scale of 1 to 5 (5: perfect, 4: almost perfect, 3: 70-80% good, 2: 50-70% good, 1: totally bad) for its relevancy. We experimented on a randomly selected set of 100 images from our test set and evaluated our captions as well as those of the competing approaches.

[3]We collected METEOR and BLEU scores via MultEval (Clark et al., 2011) and for CIDEr scores we used the authors' publicly available code.

[4]We used CrowdFlower and at least 5 different human annotators for each question.

	Rate	Variance
OURS	**2.73**	0.65
MC-SB	2.38	0.58
VC	2.27	0.66
MC-KL	2.03	0.62
HUMAN	4.84	0.26

Table 2: Human judgment scores on a scale of 1 to 5.

5 Results and Discussion

In Figure 2, we present sample results obtained with our framework, MC-SB, MC-KL and VC models along with the groundtruth caption. We provide the quantitative results based on automatic evaluation measures and human judgment scores in Table 1 and Table 2, respectively.

Our findings indicate that our query expansion approach which is based on distributed representations of captions gives results better than those of VC, MC-SB and MC-KL models. Although our method makes a modest improvement compared to the human scores we believe that there is still a big gap between the human baseline, which align well with the recently held MS COCO 2015 Captioning Challenge results.

One limitation in this work is the Out-of-Vocabulary (OOV) words, which is around 1% on average for the benchmark datasets. We omit them in our calculations, since there is no practical way to map word vectors for the OOV words, as they are not included in the training of the word embeddings. Another limitation is that this approach currently does not incorporate the syntactic structures in captions, therefore the position of a word in a caption does not make any difference in the representation, i.e. *"a man with a hat is holding a dog"* and *"a man is holding a dog with a hat"* are represented with the same vector. This limitation is illustrated in Fig. 3, where the closest caption from retrieval set contains similar scene elements but does not depict the scene well.

6 Conclusion

In this paper, we present a novel query expansion approach for image captioning, in which we utilize a distributional model of meaning for sentences. Extensive experimental results on three well-established benchmark datasets have demonstrated that our approach outperforms the state-of-the art data-driven approaches. Our future plans focus on incorporating other cues in images, and considering the syntactic structures in image descriptions.

a man wearing a santa hat holding a dog posing for a picture

a boy is holding a dog that is wearing a hat

Figure 3: Limitation. A query image on the left and its actual caption, a proposed caption on the right along with its actual image.

Acknowledgments

This study was supported in part by The Scientific and Technological Research Council of Turkey (TUBITAK), with award no 113E116.

References

Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):673–721. 2

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proc. of ACL*. 3

Alexander C Berg, Tamara L Berg, Hal Daume, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, et al. 2012. Understanding and Predicting Importance in Images. In *Proc. of CVPR*. 1

William Blacoe and Mirella Lapata. 2012. A Comparison of Vector-based Representations for Semantic Composition. In *Proc. of EMNLP-CoNLL*. 3

Xinlei Chen and C. Lawrence Zitnick. 2015. Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. In *Proc. of CVPR*. 1

Jonathan H Clark, Chris Dyer, Alon Lavie, and Noah A Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proc. of ACL*. 4

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proc. of EACL Workshop on Statistical Machine Translation*. 4

Desmond Elliott and Frank Keller. 2014. Comparing Automatic Evaluation Measures for Image Description. In *Proc. of ACL*. 4

Ali Farhadi, M Hejrati, Mohammad Amin Sadeghi, P Young, C Rashtchian, J Hockenmaier, and David Forsyth. 2010. Every Picture Tells a Story: Generating Sentences from Images. In *Proc. of ECCV*. 1

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research*. 1, 3

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In *Proc. of ACM MM*. 2

Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-semantic Alignments for Generating Image Descriptions. In *Proc. of CVPR*. 1

Andrej Karpathy, Armand Joulin, and Li Fei-Fei. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In *Proc. of NIPS*. 1, 3

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby Talk: Understanding and Generating Simple Image Descriptions. In *Proc. of CVPR*. 1

Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. 2012. Collective Generation of Natural Image Descriptions. In *Proc. of ACL*. 1, 2, 4

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proc. of ECCV*. 3

Rebecca Mason and Eugene Charniak. 2014. Nonparametric Method for Data-driven Image Captioning. In *Proc. of ACL*. 1, 2, 4

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. of NIPS*. 2, 3

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating Image Descriptions from Computer Vision Detections. In *Proc. of EACL*. 1

Margaret Mitchell, Hao Fang, Hao Cheng, Saurabh Gupta, Jacob Devlin, and Geoffrey Zweig. 2015. Language Models for Image Captioning: The Quirks and What Works. In *Proc. of ACL*. 2

Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2text: Describing Images using 1 Million Captioned Photographs. In *Proc. of NIPS*. 1, 2, 3

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. of ACL*. 4

Genevieve Patterson, Chen Xu, Hang Su, and James Hays. 2014. The SUN Attribute Database: Beyond Categories for Deeper Scene Understanding. *International Journal of Computer Vision*, 108(1-2):59–81. 1, 2

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. *Proc. of EMNLP*. 2, 3

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded Compositional Semantics for Finding and Describing Images with Sentences. *Transactions of the Association for Computational Linguistics*. 1

Peter Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*. 2

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *Proc. of CVPR*. 4

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *Proc. of CVPR*. 1

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual attention. In *Proc. of ICML*. 1

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From Image Descriptions to Visual Denotations: New similarity Metrics for Semantic Inference over Event Descriptions. *Transactions of the Association for Computational Linguistics*. 3