Improving Visually Grounded Sentence Representations with Self-Attention

Kang Min Yoo, Youhyun Shin, Sang-goo Lee
Deparment of Computer Science
Seoul National University
{kangminyoo, shinu89, sglee}@europa.snu.ac.kr

Abstract

Sentence representation models trained only on language could potentially suffer from the grounding problem. Recent work has shown promising results in improving the qualities of sentence representations by jointly training them with associated image features. However, the grounding capability is limited due to distant connection between input sentences and image features by the design of the architecture. In order to further close the gap, we propose applying self-attention mechanism to the sentence encoder to deepen the grounding effect. Our results on transfer tasks show that self-attentive encoders are better for visual grounding, as they exploit specific words with strong visual associations.

1 Introduction

Recent NLP studies have thrived on distributional hypothesis. More recently, there have been efforts in applying the intuition to larger semantic units, such as sentences, or documents. However, approaches based on distributional semantics are limited by the *grounding problem* [7], which calls for techniques to ground certain conceptual knowledge in perceptual information.

Both NLP and vision communities have proposed various multi-modal learning methods to bridge the gap between language and vision. However, how general sentence representations can be benefited from visual grounding has not been fully explored yet. Very recently, [8] proposed a multi-modal encoder-decoder framework that, given an image caption, jointly predicts another caption and the features of associated image. The work showed promising results for further improving general sentence representations by grounding them visually. However, according to the model, visual association only occurs at the final hidden state of the encoder, potentially limiting the effect of visual grounding.

Attention mechanism helps neural networks to focus on specific input features relevant to output. In the case of visually grounded multi-modal framework, applying such attention mechanism could help the encoder to identify visually significant words or phrases. We hypothesize that a language-attentive multi-modal framework has an intuitive basis on how humans mentally visualize certain concepts in sentences during language comprehension.

In this paper, we propose an enhanced multi-modal encoder-decoder model, in which the encoder attends to the input sentence and the decoders predict image features and the target sentence. We train the model on images and respective captions from COCO5K dataset [9]. We augment the state-of-the-art sentence representations with those produced by our model and conduct a series of experiments on transfer tasks to test the quality of sentence representations. Through detailed analysis, we confirm our hypothesis that self-attention help our model produce more feature-rich visually grounded sentence representations.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

2 Related Work

Sentence Representations. Since the inception of word embeddings [1], extensive work have emerged for larger semantic units, such as sentences and paragraphs. These works range from deep neural models [4] to log-bilinear models [2, 3]. A recent work proposed using supervised learning of a specific task as a leverage to obtain general sentence representation [10].

Joint Learning of Language and Vision. Convergence between computer vision and NLP researches have increasingly become common. Image captioning [11–14] and image synthesis [15] are two common tasks. There have been significant studies focusing on improving word embeddings [16, 17], phrase embeddings [18], sentence embeddings [8, 19], language models [20] through multi-modal learning of vision and language. Among all studies, [8] is the first to apply skip-gram-like intuition (predicting multiple modalities from langauge) to joint learning of language and vision in the perspective of general sentence representations.

Attention Mechanism in Multi-Modal Semantics. Attention mechanism was first introduced in [21] for neural machine translation. Similar intuitions have been applied to various NLP [5, 22, 23] and vision tasks [11]. [11] applied attention mechanism to images to bind specific visual features to language. Recently, self-attention mechanism [5] has been proposed for situations where there are no extra source of information to "guide the extraction of sentence embedding". In this work, we propose a novel sentence encoder for the multi-modal encoder-decoder framework that leverages the self-attention mechanism. To the best of our knowledge, such attempt is the first among studies on joint learning of language and vision.

3 Proposed Method

Given a data sample $(\mathbf{X}, \mathbf{Y}, \mathbf{h}_I) \in \mathcal{D}$, where $\mathbf{X}$ is the source caption, $\mathbf{Y}$ is the target caption, and $\mathbf{h}_I$ is the hidden representation of the image, our goal is to predict $\mathbf{Y}$ and $\mathbf{h}_I$ with $\mathbf{X}$, and the hidden representation in the middle serves as the general sentence representation.

3.1 Visually Grounded Encoder-Decoder Framework

We base our model on the encoder-decoder framework introduced in [8]. A bidirectional Long Short-Term Memory (LSTM) [24] encodes an input sentence and produces a sentence representation for the input. A pair of LSTM cells encodes the input sequence in both directions and produce two final hidden states: $\vec{\mathbf{h}}_t$ and $\overleftarrow{\mathbf{h}}_t$. The hidden representation of the entire sequence is produced by selecting maximum elements between the two hidden states: $\mathbf{h}_S = \max(\vec{\mathbf{h}}_t, \overleftarrow{\mathbf{h}}_t)$.

The decoder calculates the probability of a target word $\mathbf{y}_t$ at each time step t, conditional to the sentence representation $\mathbf{h}_S$ and all target words before t. $P\left(\mathbf{y}_t \mid \mathbf{y}_{<t}, \mathbf{h}_S\right)$.

The objective of the basic encoder-decoder model is thus the negative log-likelihood of the target sentence given all model parameters: $\mathcal{L}_C = -\sum_{\mathbf{X},\mathbf{Y}\in\mathcal{D}} \sum_{\mathbf{y}_t\in\mathbf{Y}} \log P\left(\mathbf{y}_t \mid \mathbf{y}_{<t}, \mathbf{X}, \Theta\right)$.

3.2 Visual Grounding

Given the source caption representation $\mathbf{h}_S$ and the relevant image representation $\mathbf{h}_I$, we associate the two representations by projecting $\mathbf{h}_S$ into image feature space. We train the model to rank the similarity between predicted image features $\tilde{\mathbf{h}}_I$ and the target image features $\mathbf{h}_I$ higher than other pairs, which is achieved by ranking loss functions. Although margin ranking loss has been the dominant choice for training cross-modal feature matching [8, 20, 25], we find that *log-exp-sum pairwise ranking* [26] yields better results in terms of evaluation performance and efficiency. Thus, the objective for ranking

$$\mathcal{L}_{VG} = \log\left(1 + \sum_{\tilde{\mathbf{h}}_I, \mathbf{h}_I} \sum_{\left(\mathbf{h}'_I, \tilde{\mathbf{h}}'_I\right)\in\mathcal{N}} \exp\left(sim\left(\tilde{\mathbf{h}}'_I, \mathbf{h}'_I\right) - sim\left(\tilde{\mathbf{h}}_I, \mathbf{h}_I\right)\right)\right) \tag{1}$$

where $\mathcal{N}$ is the set of negative examples and sim is cosine similarity.

3.3 Visual Grounding with Self-Attention

Let $h_t \in \mathbb{R}^{d_h}$ be the encoder hidden state at timestep t concatenated from two opposite directional LSTMs (d_h is the dimensionality of sentence representations). Let $H \in \mathbb{R}^{d_h \times T}$ be the hidden state matrix where t-th column of H is h_t. The self-attention mechanism aims to learn attention weight α_t, i.e. how much attention must be paid to hidden state $\mathbf{h}_t$, based on all hidden states H. Since there could be multiple ways to attend depending on desired features, we allow multiple attention vectors to be learned. Attention matrix $\mathbf{A} \in \mathbb{R}^{n_a \times T}$ is a stack of n_a attention vectors, obtained through attention layers: $\mathbf{A} = softmax\left(\mathbf{W}_{a2} \tanh\left(\mathbf{W}_{a1} H\right)\right)$. $\mathbf{W}_{a1} \in \mathbb{R}^{d_a \times d_h}$ and $\mathbf{W}_{a2} \in \mathbb{R}^{n_a \times d_a}$ are attention parameters and d_a is a hyperparameter. The context matrix $\mathbf{C} \in \mathbb{R}^{n_a \times d_h}$ is obtained by $\mathbf{C} = \mathbf{AH}$. Finally, we compress the context matrix into a fixed size representation $\mathbf{h}_A$ by max-pooling all context vectors: $\mathbf{h}_A = \max\left(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{n_a}\right)$. Attended representation $\mathbf{h}_A$ and encoder-decoder representation $\mathbf{h}_S$ are concatenated into the final self-attentive sentence representation $\mathbf{h}$. This hybrid representation replaces $\mathbf{h}_S$ and is used to predict image features (Section 3.2) and target caption (Section 3.1).

3.4 Learning Objectives

Following the experimental design of [8], we conduct experiments on three different learning objectives: CAP2ALL, CAP2CAP, CAP2IMG. Under CAP2ALL, the model is trained to predict both the target caption and the associated image: $\mathcal{L} = \mathcal{L}_C + \mathcal{L}_{VG}$. Under CAP2CAP, the model is trained to predict only the target caption ($\mathcal{L} = \mathcal{L}_C$) and, under CAP2IMG, only the associated image ($\mathcal{L} = \mathcal{L}_{VG}$).

4 Experiments

4.1 Implementation Details

Word embeddings $\mathbf{W}_E$ are initialized with GloVe [27]. The hidden dimension of each encoder and decoder LSTM cell (d_h) is 1024[1]. We use Adam optimizer [28] and clip the gradients to between -5 and 5. Number of layers, dropout, and non-linearity for image feature prediction layers are 4, 0.3 and ReLU [29] respectively. Dimensionality of hidden attention layers (d_a) is 350 and number of attentions (n_a) is 30. We employ orthogonal initialization [30] for recurrent weights and xavier initialization [31] for all others. For the datasets, we use Karpathy and Fei-Fei's split for MS-COCO dataset [13]. Image features are prepared by extracting hidden representations at the final layer of ResNet-101 [32]. We evaluate sentence representation quality using `SentEval`[2] [8, 10] scripts. Mini-batch size is 128 and negative samples are prepared from remaining data samples in the same mini-batch.

4.2 Evaluation

Adhering to the experimental settings of [8], we concatenate sentence representations produced from our model with those obtained from the state-of-the-art unsupervised learning model (Layer Normalized Skip-Thoughts, ST-LN) [33]. We evaluate the quality of sentence representations produced from different variants of our encoders on well-known transfer tasks: movie review sentiment (MR) [34], customer reviews (CR) [35], subjectivity (SUBJ) [36], opinion polarity (MPQA) [37], paraphrase identification (MSRP) [38], binary sentiment classification (SST) [39], SICK entailment and SICK relatedness [40].

4.3 Results

Results are shown in Table 1. Results show that incorporating self-attention mechanism in the encoder is beneficial for most tasks. However, original models were better in some tasks (CR, MPQA, MRPC), suggesting that self-attention mechanism could sometimes introduce noise in sentence features. Overall, utilizing self-attentive sentence representation further improves performances in

[1] However, for baseline models (without self-attention), we use $d_h = 2048$ to match the dimensionality (2048) of sentence representations produced by our proposed models.

[2] https://github.com/facebookresearch/SentEval

5 out of 8 tasks. Considering that models with self-attention employ smaller LSTM cells (1024) than those without (2048) (Section 4.1), the performance improvements are significant. Results on COCO5K image and caption retrieval tasks (not included in the paper due to limited space) show comparable performances to other more specialized methods [13, 41].

Table 1: Classification performance on transfer tasks. We report F1-score for MRPC, Pearson coefficient for SICK-R and accuracy for most others. All sentence representations have been concatenated with ST-LN embeddings. Note that the discrepancy between results reported in this paper and the referenced paper is likely due to differences in minor implementation details and experimental environment. Our models are denoted by †.

Method	MR	CR	SUBJ	MPQA	MRPC	SST	SICK-E	SICK-R
ST-LN [33]	75.46	76.98	92.60	86.46	82.23	82.26	80.76	84.39
CAP2CAP [8]	75.45	**77.85**	92.84	**87.45**	81.92	82.54	80.98	83.62
CAP2IMG [8]	75.81	77.35	92.59	86.99	73.16	82.43	81.25	81.59
CAP2ALL [8]	75.92	77.46	92.86	87.04	**82.26**	81.16	81.59	84.37
Att. CAP2IMG †	75.88	77.16	**92.91**	86.57	82.16	**83.03**	**81.69**	83.95
Att. CAP2ALL †	**75.98**	77.43	92.44	86.33	81.88	81.60	81.08	**84.54**

4.4 Attention Mechanism at Work

In order to study the effects of incorporating self-attention mechanism in joint prediction of image and language features, we examine attention vectors for selected samples from MS-COCO dataset and compare them to associated images (Figure 1). For example, given the sentence "man in black shirt is playing guitar", our model identifies words that have association with strong visual imagery, such as "man", "black" and "guitar". Given the second sentence, our model learned to attend to visually significant words such as "cat" and "bowl". These findings show that visually grounding self-attended sentence representations helps to expose word-level visual features onto sentence representations [8].

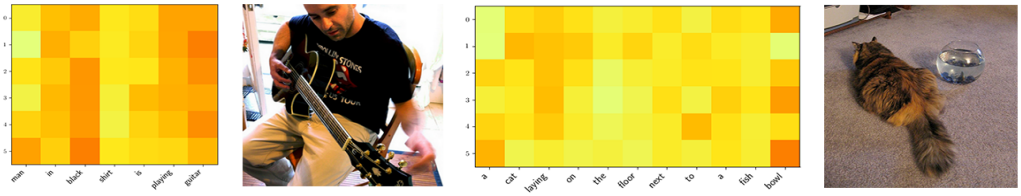

Figure 1: Activated attention weights on two samples from MS-COCO dataset. Vertical axis shows attention vectors learned by our model (compressed due to space limit). Note how the sentence encoder learned to identify words with strong visual associations.

5 Conclusion and Future Work

In this paper, we proposed a novel encoder that exploits self-attention mechanism. We trained the model using MS-COCO dataset and evaluated sentence representations produced by our model (combined with universal sentence representations) on several transfer tasks. Results show that the self-attention mechanism not only improves the qualities of general sentence representations but also guides the encoder to emphasize certain visually associable words, which helps to make visual features more prominent in the sentence representations. As future work, we intend to explore cross-modal attention mechanism to further intertwine language and visual information for the purpose of improving sentence representation quality.

Acknowledgments

This work was supported by BK21 Plus for Pioneers in Innovative Computing(Dept. of Computer Science and Engineering, SNU) funded by National Research Foundation of Korea(NRF)

(21A20151113068). Also, this work would not be possible without invaluable discussions with knowledgeable and helpful colleagues.

References

[1] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.

[2] F. Hill, K. Cho, and A. Korhonen, "Learning distributed representations of sentences from unlabelled data," in *Proceedings of NAACL-HLT*, pp. 1367–1377, 2016.

[3] M. Chen, "Efficient vector representation for documents through corruption," *arXiv preprint arXiv:1707.02377*, 2017.

[4] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-thought vectors," in *Advances in neural information processing systems*, pp. 3294–3302, 2015.

[5] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, "A structured self-attentive sentence embedding," *arXiv preprint arXiv:1703.03130*, 2017.

[6] Y. Zhang, D. Shen, G. Wang, Z. Gan, R. Henao, and L. Carin, "Deconvolutional paragraph representation learning," *arXiv preprint arXiv:1708.04729*, 2017.

[7] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1-3, pp. 335–346, 1990.

[8] D. Kiela, A. Conneau, A. Jabri, and M. Nickel, "Learning visually grounded sentence representations," *arXiv preprint arXiv:1707.06320*, 2017.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[10] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[11] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, pp. 2048–2057, 2015.

[12] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

[13] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137, 2015.

[14] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 207–218, 2014.

[15] E. Mansimov, E. Parisotto, J. L. Ba, and R. Salakhutdinov, "Generating images from captions with attention," *arXiv preprint arXiv:1511.02793*, 2015.

[16] A. Lazaridou, N. T. Pham, and M. Baroni, "Combining language and vision with a multimodal skip-gram model," *arXiv preprint arXiv:1501.02598*, 2015.

[17] D. Kiela and L. Bottou, "Learning image embeddings using convolutional neural networks for improved multi-modal semantics.," in *EMNLP*, pp. 36–45, 2014.

[18] J. Krishnamurthy and T. Kollar, "Jointly learning to parse and perceive: Connecting natural language to the physical world," *Transactions of the Association for Computational Linguistics*, vol. 1, pp. 193–206, 2013.

[19] G. Chrupała, A. Kádár, and A. Alishahi, "Learning language through pictures," *arXiv preprint arXiv:1506.03694*, 2015.

[20] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[22] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," *arXiv preprint arXiv:1611.01603*, 2016.

[23] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, pp. 577–585, 2015.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] J. Lee, K. Cho, J. Weston, and D. Kiela, "Emergent translation in multi-agent communication," *arXiv preprint arXiv:1710.06922*, 2017.

[26] Y. Li, Y. Song, and J. Luo, "Improving pairwise ranking for multi-label image classification," *arXiv preprint arXiv:1704.03135*, 2017.

[27] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.

[30] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[31] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[33] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[34] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp. 115–124, Association for Computational Linguistics, 2005.

[35] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 168–177, ACM, 2004.

[36] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, p. 271, Association for Computational Linguistics, 2004.

[37] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language resources and evaluation*, vol. 39, no. 2, pp. 165–210, 2005.

[38] B. Dolan, C. Quirk, and C. Brockett, "Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources," in *Proceedings of the 20th international conference on Computational Linguistics*, p. 350, Association for Computational Linguistics, 2004.

[39] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

[40] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, and R. Zamparelli, "Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment.," in *SemEval@ COLING*, pp. 1–8, 2014.

[41] B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4437–4446, 2015.

[42] S. Tang, H. Jin, C. Fang, Z. Wang, and V. R. de Sa, "Rethinking skip-thought: A neighborhood based approach," *arXiv preprint arXiv:1706.03146*, 2017.

[43] D. Kiela and S. Clark, "Multi-and cross-modal semantics beyond vision: Grounding in auditory perception.," in *EMNLP*, pp. 2461–2470, 2015.

[44] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, and C. D. Manning, "Generating semantically precise scene graphs from textual descriptions for improved image retrieval," in *Proceedings of the fourth workshop on vision and language*, vol. 2, 2015.

[45] D. K. Roy, "Learning visually grounded words and syntax for a scene description task," *Computer speech & language*, vol. 16, no. 3, pp. 353–385, 2002.

[46] S. Kottur, R. Vedantam, J. M. Moura, and D. Parikh, "Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4985–4994, 2016.

[47] A. K. Vijayakumar, R. Vedantam, and D. Parikh, "Sound-word2vec: Learning word representations grounded in sounds," *arXiv preprint arXiv:1703.01720*, 2017.

[48] D. Kiela, L. Bulat, and S. Clark, "Grounding semantics in olfactory perception.," in *ACL (2)*, pp. 231–236, 2015.

[49] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, *et al.*, "Devise: A deep visual-semantic embedding model," in *Advances in neural information processing systems*, pp. 2121–2129, 2013.