SPATIAL-SEMANTIC ATTENTION FOR GROUNDED IMAGE CAPTIONING

Wenzhe Hu, Lanxiao Wang, Linfeng Xu[⋆]

University of Electronic Science and Technology of China, Chengdu, China

ABSTRACT

Grounded image captioning models usually process high-dimensional vectors from the feature extractor to generate descriptions. However, mere vectors do not provide adequate information. The model needs more explicit information for grounded image captioning. Besides high dimensional vectors, the feature extractor also predicts the locations and categories of the objects, which contains low-level spatial information and high-level semantic information. To this end, we propose a new attention module called Spatial-Semantic (SS) Attention, which utilizes the predictions from the backbone network to help the model attend to the correct objects. Specifically, the SS attention module collects the position of proposals and the class probabilities from the feature extractor as spatial and semantic information to assist attention weighting. In addition, we propose a grounding loss to supervise the SS attention. Our method achieves high performance on captioning and grounding metrics and outperforms some powerful previous models on the Flickr30k Entities dataset.

Index Terms— Grounded Image Captioning, Visual Grounding, Multimodal

1. INTRODUCTION

Image captioning [1, 2, 3, 4] requires the model to understand both visual and textual information and align them. The conventional image captioning task only focuses on generated sentences, whereas such descriptions are sometimes not well grounded. Therefore, the grounded image captioning task has emerged [5]. It requires the model not only to generate descriptions but also to ground the object words in the image. In other words, when generating an object word, the model should find the corresponding object in the image. With such grounding, the generated description could be more reliable and authentic.

Most image captioning models follow the encoder-decoder paradigm [2], using CNN to extract visual features of images and encode them, then RNN to decode the features to generate sentences. However, this plausible structure has a drawback: those features do not provide enough information for the decoder, the model needs more explicit information to generate grounded captions. Previous works adopt various techniques to acquire more information for decoding. NBT [6] proposes *Slotted Caption Template Generation* to force sentences into the designated structure, whereas such a template makes the model less intelligent. GCN-LSTM [7] utilizes a GCN to explore the spatial and semantic relationships between features. But this information is implicitly included in high-dimensional vectors, which is difficult for networks to extract. LSTM-A [8] acquires attribute information from features. But from a holistic perspective, this type of information is ”external” information, introduced into the whole encoder-decoder structure. Thus, this may impair the generalization of the model to some extent. Recently, Zhou et al. [9] distill knowledge from an image-text retrieval model, guiding the attention weights. Chen et al. [10] propose the distributed attention model, designing several attention branches and merging the proposals.

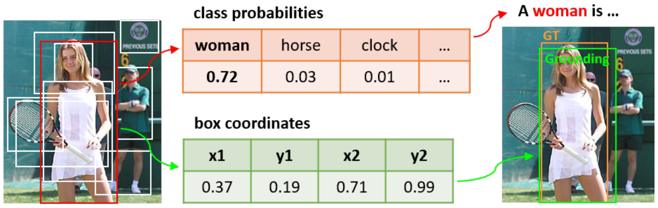


Fig. 1. The effect of semantic and spatial information. Class probabilities indicate the object word, while box coordinates show the position of the corresponding object in the image.

We notice that the outputs of the backbone network are more than high-dimensional vectors. The feature extractor also predicts the location and probabilities of objects. In this paper, we make use of this information from backbone. As shown in Fig. 1, the class probabilities indicate the category of the object, helping the model predict the correct object word; the box coordinates directly show the location of the corresponding object, assisting the model ground the object in the image. These two types of information are conducive to captioning and grounding, so we introduce them into the attention module. Specifically, we train some Fully Connected Networks (FCNs) to predict a distribution based on spatial and semantic predictions. Intuitively, the FCNs estimate the importance of an object based on *where it is* and *what it is*. Then we add the scores to basic attention weights so that

[⋆]Corresponding Author, {lfxu@uestc.edu.cn}

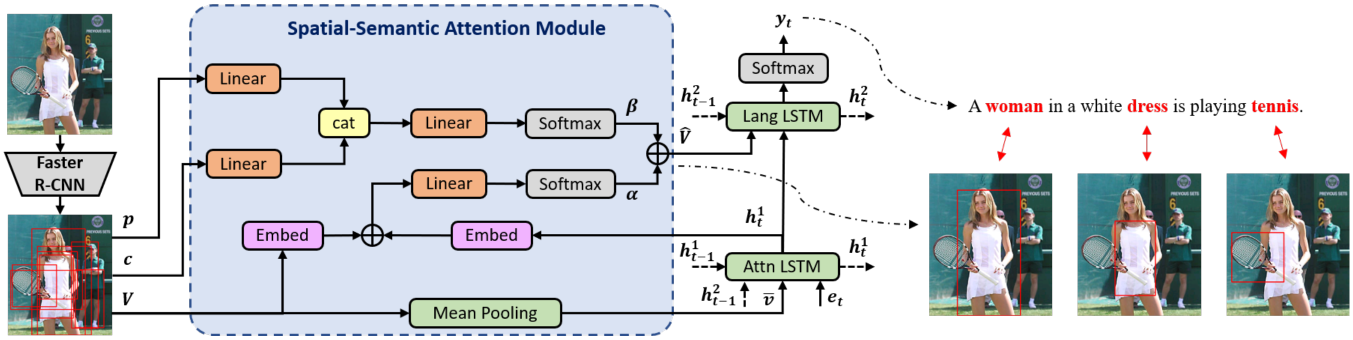


Fig. 2. The overall architecture of our model. The SS attention module utilizes the spatial vector p and semantic vector c to predict a distribution β and adds it to basic attention weights α. The object words and corresponding regions are marked in red.

the decoder can attend to those salient objects more. To supervise the FCNs, we design a new grounding loss onto the distribution, which helps the network assign larger weights on salient objects. We name the whole attention module as Spatial-Semantic (SS) Attention.

In summary, the main contributions of this paper are as follows:

- We propose the Spatial-Semantic Attention module to acquire and incorporate explicit spatial and semantic information of objects and use them to assist attention weights assignment.
- We employ a new grounding loss to supervise the SS attention module, which forces the model to attend to correct objects.

2. APPROACH

In this section, we elaborate on the proposed spatial-semantic attention module. We first introduce the encoder of our model in Section 2.1. Then in Section 2.2, we describe how the SS attention mechanism leverages the spatial and semantic information to assign weights to objects. In Section 2.3, we give the details of the grounding loss. The overall framework of our model is illustrated in Fig. 2.

2.1. Encoder

We choose the commonly used Bottom-Up Attention model [3] as our encoder, which generates *K* proposals with their visual features $V = \{v_1, v_2, ..., v_K\}$. For each region *i*, the location is represented by a 5-D vector $p_i = [x_1, x_2, y_1, y_2, s]^{\mathrm{T}}$, where (x_1, y_1) and (x_2, y_2) are coordinates of the upper-left and lower-right points, s is the area of region normalized by area of the image. The overall location vector is $p = \{p_1, p_2, ..., p_K\}$. In addition, the encoder also gives the class probabilities $c = \{c_1, c_2, ..., c_K\}$. We feed V, p, c into our attention module and decoder.

2.2. Spatial-Semantic Attention

We take the Top-Down captioning model [3] as our baseline, and modify its attention module to implement our spatial-semantic attention.

We first revisit the basic top-down decoder. The attention LSTM layer takes the global feature, previous hidden state, and word embedding vector as the inputs, then gets the current hidden state vector. At time step t, the model assigns an attention weight for each region i according to current hidden state h_t and the feature v_i:

$$a_{i,t} = W_a \mathrm{ReLU}(W_h h_t + W_v v_i)$$

$$\alpha_t = \mathrm{softmax}(a_t)$$

where W_a, W_h, W_v are learned parameters.

With such attention, the model can attend to the most related region when generating the current word. However, we argue that the high-dimensional vector v_i contains limited and implicit information, so more explicit knowledge should be introduced. Therefore, we add our spatial-semantic attention to the basic one. Notice that the predictions of the locations p and categories of the objects c can be regarded as spatial and semantic information. We train an FCN to predict a distribution based on such information. Specifically, given the position matrix p and class probability matrix c, the weights β can be formulated as:

$$b_i = W_b \mathrm{ReLU}([W_c c_i, W_p p_i])$$

$$\beta = \mathrm{softmax}(b)$$

where W_b, W_p, W_c are learned parameters. Note that β is a constant with respect to time t. The distribution β provides the model a clue of which objects are salient.

The overall attention weights γ are the weighted sum of the above two weights.

$$\gamma_t = \alpha_t + \lambda\beta$$

$$\hat{v_t} = \sum_{i=1}^{K} \gamma_{i,t} v_i$$

Table 1. Experiment results on Flickr30k dataset. † denotes the model is reimplemented in this paper.

Method	B@1	B@4	M	C	S	$F1_{all}$	$F1_{loc}$
NBT[6]	69.0	27.1	21.7	57.5	15.6	-	-
GVD[5]	69.9	27.3	22.5	62.3	16.5	7.55	**22.2**
Prophet[11]	-	27.2	22.3	60.8	16.3	5.45	15.30
Distributed[10]	69.2	27.2	22.5	62.5	16.5	**7.91**	21.54
UpDown†[3](Baseline)	71.0	29.5	22.2	63.4	16.7	4.22	12.45
SS Attn.(w/o gnd loss)	71.2	29.6	22.4	64.9	16.7	4.75	13.15
SS Attn.	**72.0**	**30.8**	**22.7**	**65.6**	**17.1**	5.52	15.69

where λ is a hyper-parameter. We set $\lambda = 0.5$ in practice.

Finally, the language LSTM layer receives the attended image feature and predicts the next word.

$$p(y_t|y_{1:t-1}) = \text{softmax}(W_y h_t^2)$$

where h_t^2 denotes the hidden state of the language LSTM, $p(y_t|y_{1:t-1})$ represents the conditional probability of the word y_t.

2.3. Objective

On word prediction, given a ground-truth sentence $y^*_{1:T}$, we adopt the following cross-entropy loss:

$$L_{ce} = -\sum_{t=1}^{T} \log(p_\theta(y_t^*|y^*_{1:t-1}))$$

where θ is the model parameters.

Moreover, we employ a new grounding loss to supervise our proposed SS attention. Denote the predicted object regions as $R = \{r_i\}$. We collect the ground-truth (GT) bounding boxes $G = \{g_j\}$ of the objects mentioned in GT sentences. Similar to attention supervision in GVD [5], for each GT box g_j, we choose the predicted region which has the largest IoU with g_j as the positive region. Define an indicator σ_i:

$$\sigma_i = \begin{cases} 1, \arg\max\limits_{r_i \in R} \text{IoU}(r_i, g_j) \\ 0, \text{otherwise} \end{cases}$$

For regions R, we get the corresponding indicator vector $\sigma = \{\sigma_i\}$. Then we utilize cross-entropy loss to regress SS attention β to σ. The grounding loss is defined as:

$$L_{gnd} = -\sum_{i=1}^{N} \sigma_i \log \beta_i$$

With such grounding loss, the SS attention model can be encouraged to attend to the objects mentioned in the sentence. We formulate the total loss as a sum of the above two losses:

$$L = L_{ce} + \mu L_{gnd}$$

where μ is a hyper-parameter which set to 1 in practice.

3. EXPERIMENTS

3.1. Datasets and Evaluation Metrics

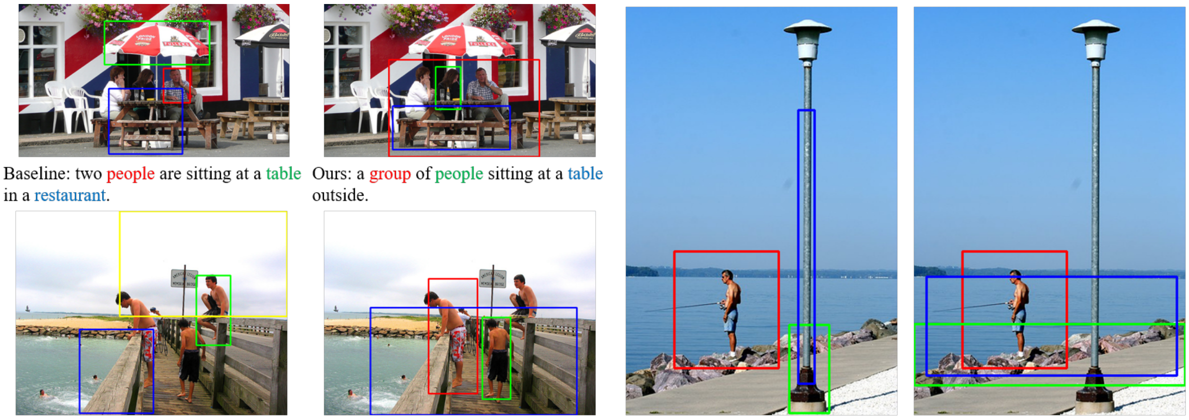
We conduct all the experiments on the widely used Flickr30k Entities dataset [12]. This dataset has more than 31k images and 158k captions in total. Besides five captions per image, it contains 276k bounding boxes linked to object words in captions. We follow the Karpathy split [1] for evaluation, which assigns 29k images for training, 1k for validation, and 1k for testing. We adopt the same vocabulary as GVD [5] for compatibility. The evaluation of the model should be from two aspects: captioning and grounding. For captioning, we use the standard caption metrics BLEU [13], METEOR [14], CIDEr [15], and SPICE [16] for captioning evaluation. For grounding, we evaluate our model on $F1_{all}$ and $F1_{loc}$ defined by GVD.

3.2. Implementation Details

We exploit the off-the-shelf Bottom-Up [3] model based on Faster R-CNN [17] to extract region features. We extract 36 regions for each image, and the dimension of visual features is 2048. In the attention module, the hidden state and visual features are projected into a 512-dimensional space, while the spatial and semantic vectors are both transformed to a 1-dimensional vector, i.e. a value. Our whole model is implemented with X-modaler [18] toolbox based on PyTorch. At the training stage, we use Adam [19] optimizer, setting the mini-batch size to 8. The initial learning rate is set to 4×10^{-4} and decayed every 3 epochs by a factor of 0.8. We train our model for 50 epochs.

3.3. Performance Comparison

We compare our model with some previous grounded image captioning methods, including NBT [6], GVD [5], Prophet [11], and Distributed [10], as shown in Table 1. Our method outperforms all the above approaches on captioning metrics, especially CIDEr. The full Spatial-Semantic Attention model reaches 65.2 on the CIDEr score, making nearly 3 points improvement. As for grounding metrics, our SS attention



Baseline: two people are sitting at a table in a restaurant.

Ours: a group of people sitting at a table outside.

Baseline: a group of men are standing on a dock in the water.

Ours: a group of men are standing on a pier.

Baseline: a man standing on a rock overlooking the water.

Ours: a man is standing on a rock overlooking the water.

Fig. 3. Comparison of results of baseline and our model. Generated captions and grounding regions are both shown.

Table 2. Ablation study on the effects of spatial information p and semantic information c.

p	c	B@4	M	C	S	$F1_{all}$	$F1_{loc}$
×	×	29.6	22.2	62.3	16.5	2.87	6.91
✓	×	29.0	22.3	62.8	16.6	5.68	16.26
×	✓	29.9	22.5	62.7	16.8	2.76	7.68
✓	✓	30.8	22.7	65.6	17.1	5.52	15.69

model performs better than those methods above except for GVD and Distributed. This is because GVD employs multiple losses to supervise the attention at every time step, thus promoting its grounding performance. But our model only uses a simple loss to restrict the time-invariant distribution β. Distributed Attention improves grounding performance by merging proposals, but their captioning metrics do not increase. However, our captioning scores are higher than theirs, illustrating that our model can help generate more grounded captions.

3.4. Ablation Study

We conduct ablation studies to verify the effect of our new spatial-semantic module and grounding loss. The results are shown in the lower part of Table 1. We first reimplement the baseline Up-Down model. To prove the effectiveness of our spatial-semantic distribution, we train our SS attention model only with cross-entropy loss, and the scores are improved. Then we add the grounding loss to train the whole model, and the performance is further improved. The above experiments demonstrate that the proposed spatial-semantic attention module and grounding loss both can promote the performance either on captioning or on grounding.

In Table 2, we test the effect of spatial and semantic information. In order to keep the structure unchanged, we replace p or c by visual features V rather than remove networks. The results show that the spatial information p brings a large progress on grounding, and the semantic information improves the captioning performance in turn. Incorporating such two types of information further boost the performance.

3.5. Qualitative Results

We visualize the captioning and grounding results and compare them with baseline in Fig. 3. It proves that our model outperforms the baseline both on description accuracy and grounding correctness. The images on the right show that sometimes our model generates the same caption as the baseline, but the grounded regions are much more accurate.

4. CONCLUSION

In this paper, we propose the Spatial-Semantic Attention model, which exploits the position and category information of bounding boxes for grounded image captioning. This spatial and semantic information is from the feature extractor, providing some explicit information for the model to generate captions. We train FCNs to explore and incorporate such information. We also propose a new grounding loss to supervise our SS attention module. Experiments prove that our model achieves higher performance on all captioning and grounding metrics than some existing powerful methods.

5. ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 62071086 and Sichuan Science and Technology Program under Grant 2021YFG0296 and science and technology innovation (seedling project) cultivation and invention creation project in Sichuan province No.2021015.

6. REFERENCES

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

[2] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[4] Haishun Chen, Ying Wang, Xin Yang, and Jie Li, "Captioning transformer with scene graph guiding," in *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 2021, pp. 2538–2542.

[5] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach, "Grounded video description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6578–6587.

[6] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh, "Neural baby talk," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7219–7228.

[7] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 684–699.

[8] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei, "Boosting image captioning with attributes," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4894–4902.

[9] Yuanen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang, "More grounded image captioning by distilling image-text matching model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4777–4786.

[10] Nenglun Chen, Xingjia Pan, Runnan Chen, Lei Yang, Zhiwen Lin, Yuqiang Ren, Haolei Yuan, Xiaowei Guo, Feiyue Huang, and Wenping Wang, "Distributed attention for grounded image captioning," in *Proceedings of the ACM international conference on Multimedia*, 2021, pp. 1966–1975.

[11] Fenglin Liu, Xuancheng Ren, Xian Wu, Shen Ge, Wei Fan, Yuexian Zou, and Xu Sun, "Prophet attention: Predicting attention with future attention.," in *Advances in Neural Information Processing Systems*, 2020.

[12] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2641–2649.

[13] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[14] Satanjeev Banerjee and Alon Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[15] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.

[16] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, "Spice: Semantic propositional image caption evaluation," in *Proceedings of the European Conference on Computer Vision*. Springer, 2016, pp. 382–398.

[17] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[18] Yehao Li, Yingwei Pan, Jingwen Chen, Ting Yao, and Tao Mei, "X-modaler: A versatile and high-performance codebase for cross-modal analytics," in *Proceedings of the ACM international conference on Multimedia*, 2021, pp. 3799–3802.

[19] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.