The 26th Meeting on Image Recognition and Understanding

Leverage Semantic Alignment of Object Relations for Image Captioning

Da Huo[1,a)] Marc A. Kastner[2,b)] Takatsugu Hirayama[3,c)]
Takahiro Komamizu[1,d)] Ichiro Ide[1,e)]

Abstract

Image captioning is a popular task in vision and language, which aims to generate proper textual descriptions of images. Recently, some works use objects to ease image and text alignment for learning better cross-modal representation, resulting in good performance in this task. In this paper, we consider relation is also important for learning semantics, here we use relations between objects to explore if relations as a prior can also improve performance. First, we consider the annotated relations between objects, and use them as tags in an image captioning model for aligning the image and text. Moreover, we also aim at integrating relationships between text to image features. For this, we focus on the masking strategy and change the strategy from random masking to relation masking to further study the training strategy for enhancing semantic alignment of object relations. In the experiments, we found that considering object relations improved the captioning performance in common metrics. Further, when changing the masking strategy for focusing on a specific part in caption to be masked when training, we found that it could lead to capturing more object relations of an image, while it destroyed the randomness when training, the performance decreases and the relations appear to be not compatible with the image contents.

1. Introduction

With the rapid development of deep learning, many tasks in multimedia processing have received a lot of attentions and made remarkable progresses [1], such as object detection, semantic segmentation, image captioning, and so on. In the image captioning tasks, some methods introduce the self-attention mechanism using Transformer [2], which can capture the long-distance relationship between the text and the input image leading to good performance. Most existing methods only use image features and texts to perform image captioning, but recently, a Vision and Language Pretraining (VLP) model called Oscar [3] has appeared. It uses object tags to learn the semantic alignment between the two modalities and has shown better performance. However, the object tags only show the object details, so it ignores the relationships between objects. In previous work [5], we found that the use of the action part from text as tags also can be used for semantic alignment. But the action used as tags consisting of a list into captioning model, it loses the connection with its corresponding objects, which is the limitation in learning semantic information.

In this paper, we extend this idea by using relations of objects as tags as well as their corresponding object location information that the previous work [5] does not have for aligning elements within two modalities. Different from action tags which comes from texts, here we focus on the image and explore if the object relations in an image can align the semantics of image and text and catch more relations.

The main contributions of our work can be summarized as follows: (i) We consider the relation between objects as anchor points to learn cross-modal representations in the image captioning task. (ii) We show that the model adding relations improves performance. (iii) We also explore the impact of masking strategies while training. The experiments show that by additionally masking an relation in an caption, although it can lead the model to capture more relations in the generated captions, due to the losing of randomness, the captioning model will not learn object relation semantics well.

2. Related work

2.1 Image captioning

With the rapid development of deep learning, various models have been proposed, such the Transformer model [2] which used attention mechanism to capture long-distance dependency in a sequence. It has obtained substantial enhancement in performance across both vision and language tasks. Masked language model has been proposed in Bidirectional Encoder Representations from Transformers (BERT) [4] architecture, where the idea is to randomly mask a small number of input tokens and training the model to predict the masked tokens according to the rest of the tokens. The BERT architecture that fuses visual and textual

[1] Nagoya University
[2] Kyoto University
[3] University of Human Environments
[a)] huod@cs.is.i.nagoya-u.ac.jp
[b)] mkastner@i.kyoto-u.ac.jp
[c)] t-hirayama@uhe.ac.jp
[d)] taka-coma@acm.org
[e)] ide@i.nagoya-u.ac.jp

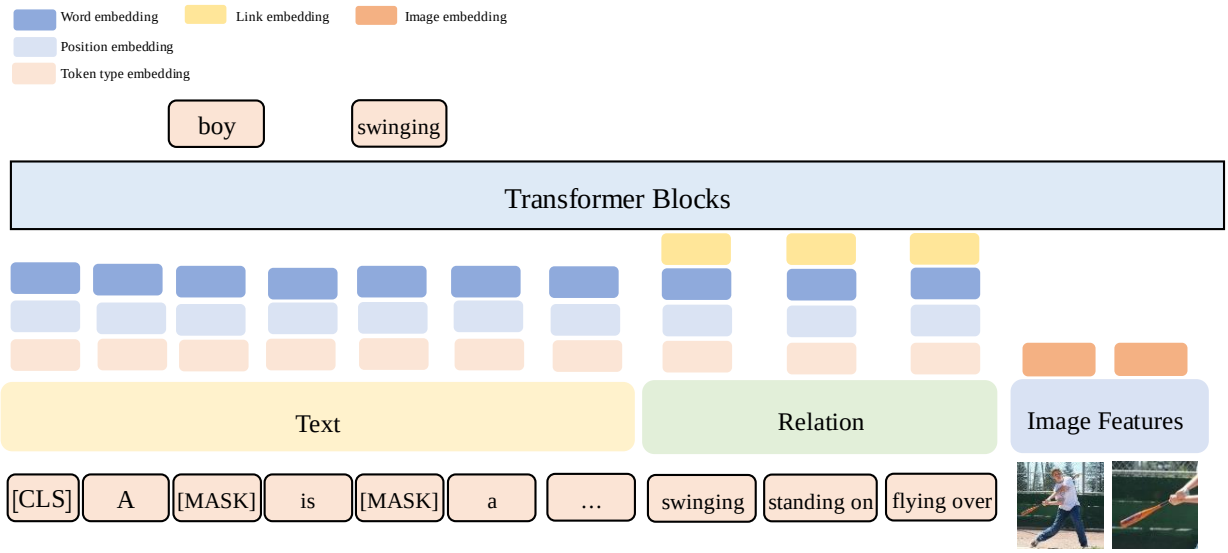

Fig. 1 Proposed architecture of the captioning model.

modalities is proposed for image captioning .

2.2 Alignment based captioning

Recently some works, such as Oscar [3], use not only the image-text pair as input, but in addition, introduce objects detected from image as an input to the captioning model. In their work, they introduce the objects as anchor points when training the model to ease the learning of image-text alignment. These objects serve as anchors in training, and make the model capture the representation of channel invariant (or modal invariant) information. Benefiting from that, it allows the model to generate captions with more objects and have better quality. Following this idea, another work [5] considering the action of text and using action tags for semantic alignments for image and text.

In this work, different from previous research by focusing on action in text, here we focus on the images and using the

ing have two objects people and bat. In the text part the verb shows similar semantic and easy for alignment, however, the semantic alignment for the image part shows limitation because the relation don not have entity and difficult for alignment in the image part. Due to the objects have its own region features with the bounding box information, so we enhance the connections for relations to their object features using the bounding box information into relations into the model, and we create a linking embedding to enhance the image part semantic alignment.

To do this, we use the bounding boxes of image region features $[x_1, x_2, y_1, y_2, w, h]$ for producing the 6-dim for each objects, and each relation has two corresponding objects, we concatenate them for a relation, in which x_1, x_2, y_1, y_2 are objects regions and the w and h are width and height of the image. The examples of the creation of vector used for linking embedding as shown in the Fig. 2. In this example, 'people-swinging-bat' is a relation triple pair, and we use the bounding boxes of people and tennis for concatenation to create a 12-dim vector, and then produce the location information for a relation swinging.

We create a new embedding named linking embedding to enhance the relationship of the relation part to the corresponding image features. Similar to positional embedding, this linking embedding is added into the model as shown in Fig. 1.

Fig. 2 In this example, we show the relation *swinging* between the object *people* and *bat*. The use of bounding boxes of objects to create location information for the relation and by linking embedding to better learn the semantic in both image and text.

4. Experiments

4.1 Model training

For the captioning model, we choose the recent model Oscar [3] as our baseline, directly use their pre-trained base model and fine-tune on it. We focus on the fine-tuning process and introduce relations for captioning.

The architecture is shown in Fig. 1, in our experiments, the input consists of image features of different regions, relations, and a caption. While training, there are 15% of tokens in the text being chosen for masking. Next, we train the model to predict the masked tokens. The training and interface processes are the same with previous work [3].

Here, we use the intersection of PSG dataset [6] and the MS COCO dataset [7] for image captioning and find the its impact on semantic alignment. In details, the data we used contains training set for 42,000 images and validation set for 1,800 images with the annotated relations.

Table 1 Performance on part MS COCO [7] validation set trained with (a) no relation, (b) relation and (c) relation with linking.

Relation information	B4	M	C	S
None	31.5	26.8	106.4	19.8
Relation	34.3	28.4	115.6	21.1
Relation with Linking	**35.3**	**28.7**	**118.1**	**21.6**

Table 2 Performance on part MS COCO [7] validation set with (a) random masking and (b) relation masking strategies.

Masking strategy	B4	M	C	S
Random masking	**35.3**	**28.7**	**118.1**	**21.6**
Relation masking	25.6	25.6	91.7	18.4

4.2 Main results

In this section, we explore the performance of the relation for semantic alignment and how it led the performance improvement. We perform the experiments with training the image captioning model in circumstances with no relation, relation, and relation with linking. The results are shown in Table 1. All the cases are evaluated on the common metrics for image captioning as BLEU-4 [9], METEOR [10], CIDEr [11] and SPICE [12].

The experiments shows the performance of introducing the relation in the Table 1, we can see after adding the relations, it leads to the performance increases in all captioning metrics and, further adding bounding box information, the performance also increases in some extent. All the performance are reported on part of the validation set [7], the data used for both training and validation are less than half of that in other works, therefore the performance is not as good as others [5], in this paper we mainly concern the ablation study with different settings.

4.3 Impact of masking strategy

Using the relations and bounding box information for enhancing the semantic alignment, we found the performance of the model increases. Moreover, to further enhance the

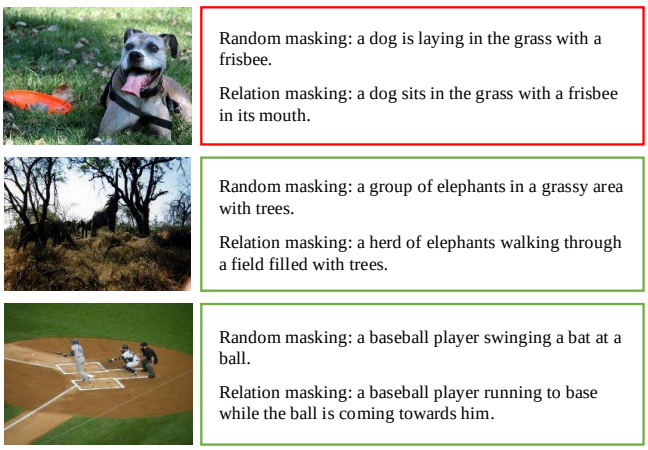

Fig. 3 The examples of captions for the training strategy with default random masking, comparing with that without randomness. The box colors green is the proper captions for image content, but the red one is not compatible with semantics in that image.

alignment between the relation part and the text part, we change the training strategy, which destroy the randomness for masking and study the impact of the learning semantics.

Here, we change the training strategy from random masking to masking the relation part of a text, for further finetuning based on our random training one. The performance are shown in the Table 2. We can see the all metrics drops a lot in this masking strategy. Experiments shows that this training strategy is not compatible with semantic alignment because of it eliminates the randomness, which is essential for learning semantics.

On the one hand, the random masking contributing to good performance, while losing the randomness lead to limitation when learning semantic. We found that introducing training strategy without randomness, the performance drops a lot and then we further analyze why it is dropped a lot. Here we show some generated caption examples without random masking strategy. For example, in the image with a dog, the captions with random masking strategy is correct for describe the content, while after destroy the randomness when training, the captions appears to over reasoning, which is shows the frisbee in mouth but actually is not, the dog is only open the mouth.

The caption examples are shown in the Fig. 3, it shows that part of some generated captions changed to catch more relations that seems good for semantic alignment of relation, but we also found lots of the captions like above are not correct to describe the image content, which is the reason all the metrics dropped a lot.

5. Conclusion

We proposed an method for introducing relations of objects to align the image and text. Our contributions are as follows: 1) We use relations between objects for aligning the semantic information for image and text and find how it contributing to semantic alignment. 2) For the current experiments, we found that using the relation for alignment it can make better performance than the one not used it. 3) We also try to use different strategy for training, experimental results showed that when changing the random masking strategy, the performance dropped a lot and the randomness of training strategy is significant for learning semantics.

Acknowledgments

Part of the work presented in this paper was supported by the CSC / MEXT scholarship and MEXT Grant-in-aid for Scientific Research (22H03612).

References

[1] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietik¨ainen, Deep learning for generic object detection: A survey., Int. J. Comput. Vis., 2020, pp. 261–318

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", NeurIPS, vol. 30, pp. 1–11, 2017.

[3] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. "Oscar: Object-semantics aligned pre-training for vision-language tasks", ECCV, pp. 121–137, 2020.

[4] J. Devlin, M. Chang and K. Lee, K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", Proc. 16th NAACL-HLT, pp. 4171–4186, 2018.

[5] D. Huo, M. A. Kastner, T. Komamizu, I. Ide, "Action Semantic Alignment for Image Captioning", MIPR, 2022.

[6] J. Yang, Y. Z, Ang, Z. Guo, K. Zhou, W. Zhang, Z. Liu, "Panoptic scene graph generation", ECCV, pp. 178–196 2022.

[7] 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.", ECCV, pp. 740– 755, 2014.

[8] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma et al., "Visual Genome: Connecting language and vision using crowdsourced dense image annotations.", Int. J. Comput. Vis., 2017, vol. 123, pp. 32–73.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, ACL, 2002.

[10] S. Banerjee and A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgments., ACL Workshops, 2005.

[11] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," in CVPR, 2015.

[12] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic Propositional Image Caption Evaluation," in ECCV, 2016.