 VieCap4H - VLSP 2021: ObjectAoA - Enhancing performance of Object Relation Transformer with Attention on Attention for Vietnamese image captioning

Nghia Hieu Nguyen, Duong T.D. Vo, Minh-Quan Ha

1 University of Information Technology, Ho Chi Minh City, Vietnam
2 Vietnam National University, Ho Chi Minh City, Vietnam

Abstract
Image captioning is currently a challenging task that requires the ability to both understand visual information and use human language to describe this visual information in the image. In this paper, we propose an efficient way to improve the image understanding ability of transformer-based method by extending Object Relation Transformer architecture with Attention on Attention mechanism. Experiments on the VieCap4H dataset show that our proposed method significantly outperforms its original structure on both the public test and private test of the Image Captioning shared task held by VLSP.

Keywords: Vietnamese image captioning, Transformer

1. Introduction
Image captioning is first introduced with the release of Flickr8k dataset. This task has many practical applications: it can be used as a feature for automatically generating alternative text for images in social media applications; or can be used as a function in support devices for visually impaired people. Moreover, results from image captioning researches can be inherited to propose novel methods for visual question answering tasks. Regardless of its helpful application, image captioning is still challenging as it requires the model to understand visual features provided by the images and exploit the linguistic features of human language to describe the content of images.

Some early approaches tackled this task by using the combination of Convolutional Neural Network (CNN) to learn visual features from images and Recurrent Neural Network (RNN) to learn linguistic features from captions. With the advance of attention mechanisms, deep learning models can generate more accurate captions for images by adding attention features in their RNN structure or using Transformer structure instead of RNN structure.

Although the improvement in linguistic aspect of captions generated from recently
novel state-of-the-art models, these captions still have some drawbacks. Their meanings are somewhat general, in which a class of similar images can be given exactly the same caption. Moreover, considering more about captions, most of them describe the overall context in images, not go deeper into the details. These drawbacks draw a large challenge for image captioning models, requiring them to focus their attention on main objects in images and exploit more the human linguistic aspect to make better captions. To this end, Bottom-up Attention mechanism [9] is proposed to make deep learning model focus better on meaningful objects in the images, leaving the research community the opportunity to explore and propose the structure which can effectively use these attention-worthy visual features to learn how to make better captions.

Based on the work of [9], many attention architectures were proposed to enhance the deep learning model to generate more meaningful captions [7, 6, 8]. These novel methods respectively achieved state-of-the-art (SOTA) results on the MS COCO dataset [10]. Although there are many mechanisms proposed in encoder layers of transformer-based methods to enhance image understanding ability, there is no work exploring how we can use these mechanisms together effectively. Hence in this work, we introduce a novel architecture which effectively uses together the two SOTA approaches proposed by [7] and [5].

2. Related works

Vinyals et al. proposed the early approach, Show and Tell [3], for image captioning task by using CNN for learning features from images, then using RNN to learn how to make captions from learned visual features. Xu et al. improved the Show and Tell model [2] by expanding its RNN structure with attention mechanisms [11, 12]. After analysing visual aspect of captioning models, [9] proposed another attention stage at visual extraction module of image caption models, called Bottom-up attention instead of using grid features. Recently, Vaswani et al. [4] proposed Transformer architecture which have become the new research trend for the natural language processing (NLP) community. To take into account the advantage of Transformer architecture for linguistic aspect of captioning models, he et al. [6] proposed replacing RNN structure and its attention with transformer-based architecture. Huang et al. [7] discovered that the information returned from attention layer of transformer lack correlation with the information from the input, so they proposed another attention stage for Transformer architecture, called Attention on Attention (AoA). Herdade et al. [5] enhanced the attention module in transformer structure by proposing geometric attention mechanisms to make better understanding of visual features, so introduce a novel method called Object Relation Transformer. All the three transformer-based approaches, Image Transformer, AoA, and Object Relation Transformer, respectively achieved SOTA results on MS COCO dataset [10].

For image captioning tasks in Vietnamese, there are quite limited works and experiments on Vietnamese. Recently, the first qualified image captioning dataset for Vietnamese
UIT-ViIC\cite{13} has been introduced to the research community. At the same time,\cite{14} published the first experiment for image captioning task in Vietnamese, but their approach was too trivial. They used machine translating models to translate caption from MS COCO dataset into Vietnamese then evaluated image captioning models on these captions, which have many drawbacks as analysed by\cite{13}.

3. Transformer-based methods

In this section, we briefly review the two SOTA transformer-based methods: Attention on Attention Network and Object Relation Transformer. Relied on the ideas of these methods, we then propose the novel approach (section 4) and analyse our proposed methods compared with its original architecture on both the UIT-ViIC dataset\cite{13} and the vieCap4H dataset\cite{15} (section 5).

3.1. Attention on Attention (AoA)

Original cross-attention mechanism in transformer \cite{4} (figure\ref{fig:cross-attention}) forms the attention weight matrix as:

$$\Omega_A = \frac{QK^T}{\sqrt{d_k}}$$

then the output of the attention layer is

$$f_{att}(Q, K, V) = \text{Softmax}(\Omega_A)V$$

where $Q \in \mathbb{R}^{\text{seq} \times d_q}, K \in \mathbb{R}^{\text{seq} \times d_k}, V \in \mathbb{R}^{\text{seq} \times d_v}$ are the three input vectors of attention layer. Huang et al.\cite{7} observed that the output weights can have low correlation with the input vectors as not all parts from captions have its correspondences in images. Therefore, we do not need to pay any attention at visual features for such these parts. To this end, they proposed to use additional linear layers as extended attention stage to ensure strong relation between the output of attention layer and the input vectors, reformulated the output of the attention layer as:

$$\text{AoA}(Q, K, V) = \sigma(G(Q, K, V)) \odot I(Q, K, V)$$

where

$$G(Q, K, V) = W_q^gQ + W_v^g f_{att}(Q, K, V) + b^g$$

$$I(Q, K, V) = W_q^iQ + W_v^i f_{att}(Q, K, V) + b^i$$

3.2. Object Relation Transformer (ORT)

Inherited the advantage of Bottom-up attention \cite{9,5} proposed to use additionally box coordinates achieved from RPN layer of FasterRCNN\cite{16} to enhance the visual information from images. Given $Q \in \mathbb{R}^{\text{seq} \times d_q}, K \in \mathbb{R}^{\text{seq} \times d_k}, V \in \mathbb{R}^{\text{seq} \times d_v}$, the three input vectors of attention layer, the attention weights are calculated as:

$$\Omega_A = \frac{QK^T}{\sqrt{d_k}}$$
The relative position of object $m$ to object $n$ in an image is represented as a 4-dimension vector:

$$\lambda = \lambda(m, n) = (\log(t_x), \log(t_y), t_w, t_h)$$

where:

$$t_x = \frac{|x_m - x_n|}{w_m}$$
$$t_y = \frac{|y_m - y_n|}{h_m}$$
$$t_w = \log\left(\frac{w_m}{w_n}\right)$$
$$t_h = \log\left(\frac{h_m}{h_n}\right)$$

Then the attention weight for the pair of object $m, n$ is given as:

$$w_{m,n} = ReLU(Emb(\lambda)W_G)$$

The final attention weight is calculated as:

$$\Omega = \Omega_A \circ \exp(\Omega_W)$$

where $\Omega_W$ is the matrix formed by $w_{m,n}$.

4. Our proposal

Huang et al. [7] analysed the correlation between input information and attended information of self-attention module in transformer. But they equipped LSTM structure for their decoder layer which limited the ability to catch and understand linguistic features, potentially leading to lower performance. Herdade et al. [5] proposed to use geometric features of detected objects but they used the coordinates instead of area of bounding boxes. Moreover, the decoder of their model uses cross-attention mechanism to capture the human language rules, which can achieve better performance compared to the above methods.

To this end, we propose to extend the utilization of geometric features with the idea of AoA to combine the advantages of these SOTA methods together. Specifically, from the attention weights $\Omega$ of geometric attention layer in Object Relation Transformer, the output of attention layer in our proposal is reformulated as:

$$\hat{V} = \sigma(G(Q, K, V)) \odot I(Q, K, V)$$

where

$$G(Q, K, V) = W^q_Q Q + W^v_{f_{att}}(Q, K, V) + b^q$$
$$I(Q, K, V) = W^t_Q Q + W^f_{f_{att}}(Q, K, V) + b^t$$
$$f_{att}(Q, K, V) = \text{softmax}(\Omega)V$$

As described above, our attention formulation is now equipped with an extra attention stage. We assume that the geometric attention exploits the geometric relationships between objects in images beside the internal relationships learned from visual feature vectors of these objects. However, not all the contexts in images contain the geometric relationship. Therefore, additional information about the locations of objects in images may somewhat cause confusion to the model when it learns how to use attention weights from the encoder layers in order to generate appropriate captions. This is when the Attention on Attention mechanisms come in handy. As analysed by [7], Attention on Attention can increase the correlation between input information and output vectors from attention layers to create relevantly attended
information. By applying Attention on Attention together with geometric attention, we can avoid uncorrelated output vectors of attention layers from the query vectors, making the model clearer when learning how to combine the information from visual and linguistic features, leading the model to generate better captions as a result. Moreover, to make the model better exploiting linguistic features, we use transformer as its decoder. For visual features, we inherit the advantage of region-based features yielded from Bottom-up Attention mechanisms [9] to let the model have the ability to focus on which objects it needs to describe. Finally, we named this method ObjectAoA1 (figure 2).

5. Experiments

5.1. Datasets

In our experiments, we have evaluated the ORT method and our proposed method ObjectAoA on the UIT-ViIC dataset. Consequently, relying on the improvement on this dataset, we continue using the previous methods to train and evaluate on the VieCap4H dataset [15] which was published in Image Captioning Shared Task by VLSP Challenge 2021.

5.2. Metrics

The popular metrics used for evaluating image captioning methods are BLEU [17], METEOR [18], ROUGE [19], CIDEr [20]. In our experiments, these four metrics are utilized to evaluate the two transformer-based models on the UIT-ViIC. For the VieCap4H datasets [15], we use the average of the

1The implementation of ObjectAoA can be found in this repository https://github.com/hieunghia-pat/UIT-ObjectAoA

2Our results on the VieCap4H dataset can be found in the competition leaderboard at https://aihub.vn/competitions/68#results
four BLEU scores to evaluate ORT and our proposal.

5.3. Training scheme

We follow the work from Rennie et al. [21] where the training scheme is divided into two stages: in the first stage the model is trained using cross-entropy loss and in the second stage the model is refined using self-critical sequential training [21]. The self-critical sequential training scheme is a reinforcement learning strategy that consider the model as the "agent" and features from images and training captions are "signals" from the "environment". The rewards are calculated based on the CIDEr score in which the model will receive higher rewards if it can generate captions closer to the training captions.

5.4. Configuration

For first training training stage on the UIT-ViIC dataset, we trained the methods on the training set, used dev set to select the best models then evaluated them on the test set. All methods were trained on the GPU NVIDIA P100, using Adam [22] optimizer with the learning rate adjusted follow the learning rate scheduler from Vaswani et al. [4]. The best models on dev set were chosen based on the CIDEr metric. We also set the batch size to 32 and trained the models until the CIDEr score decreases.

For the self-critical training stage on both VieCap4H dataset and UIT-ViIC dataset, we fixed the learning rate to $5 \cdot 10^{-6}$ and trained the two models until the CIDEr score decreases.

5.5. Experimental results

We first use the UIT-ViIC dataset [13] to analyse the performance of our proposed method and ORT method. Results are showed in Table 1.

Results in Table 1 indicate that ObjectAoA designed based on the Object Relation Transformer has better performance to its original architecture on all of the seven metrics.

In our experiments on the UIT-ViIC dataset, we showed that our proposal yielded better results on the BLEU metric to other methods, therefore we used ObjectAoA as the main method to compete in the VLSP image captioning shared task. We concurrently conducted experiments for Object Relation Transformer on the VieCap4H dataset [15] to compare with our proposal.

As Table 2 indicates, the ObjectAoA Transformer outperforms the Object Relation Transformer in both dev set and test set. Relyed on these results, we used ObjectAoA as our best method to submit to the private test phase.

The example captions generated by ORT and ObjectAoA are shown in figures A.3.
Table 1. Results of the transformer-based methods on the UIT-ViIC dataset.

| Methods      | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | Meteor | Rouge | CIDEr |
|--------------|--------|--------|--------|--------|--------|-------|-------|
| ORT          | 0.744  | 0.634  | 0.544  | 0.47   | 0.364  | 0.638 | 1.362 |
| ObjectAoA    | 0.769  | 0.66   | 0.569  | 0.497  | 0.373  | 0.648 | 1.488 |

Table 2. Results of Object Relation Transformer and ObjectAoA Transformer on the VieCap4H dataset.

| Model                     | dev     | public test | private test |
|---------------------------|---------|-------------|--------------|
| Object Relation Transformer| 0.3229  | 0.2645      | 0.2741       |
| ObjectAoA Transformer     | 0.3278  | 0.295912    | 0.285222     |

For each example image, there is a comparison between two corresponding captions given by two models. From the examples, it is clear that with the usage of AoA, the model can focus more precisely and describe the object correctly with better interpretation.

5.6. Ablation study

As depicted in figure 2, our proposed method ObjectAoA and the original ORT method are mostly the same, except for the AoA module in encoder layers. Therefore, the results on the UIT-ViIC dataset (table 1) and on the VieCap4H dataset (table 2) are the ablation study results for the effectiveness of the additional AoA module equipped at the encoder layers.

5.7. Results analysis

As depicted in Appendix A, most of the wrong cases of ORT method is catching wrong objects in the images (Fig A.4 #4, Fig A.5 #6, Fig A.12). Moreover, the ORT model failed to understand the context round objects in the images, which is a significant drawback causing ORT struggle to describe objects in images more specifically (Fig A.3 #1 and #2, Fig A.6). We can realise this conclusion through the examples from the UIT-ViIC dataset, where captions generated from ORT model have general meaning compared to captions generated from ObjectAoA model which can describe both action and intention of objects in images. With the additional AoA module, the attention layer of ORT model is improved to enhance the correlation of the output features with the input features, which makes the ObjectAoA model can understand images better.

6. Conclusion and future work

In this paper, we have analysed and proposed a novel transformer architecture based on the two recent SOTA methods AoA [7] and ORT [6] for image captioning task. In the future, we will conduct more experiments and perform more deeper analysis on linguistic aspect of the ObjectAoA Transformer as well as extend other recent SOTA transformer-based models such as Meshed-Memory Transformer [8] with Attention on Attention mechanism to find out the better method for image captioning.
in Vietnamese. Moreover, we will develop the UIT-ViIC dataset to different domains and continue investigating image captioning architectures to finally conduct a cross-domain method for Vietnamese image captioning task.

References

[1] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics, Journal of Artificial Intelligence Research 47 (2013) 853–899.
[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: International conference on machine learning, PMLR, 2015, pp. 2048–2057.
[3] O. Vinyals, A. Toshev, S. Bengio, D. 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.
[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv preprint arXiv:1706.03762.
[5] S. Herdade, A. Kappeler, K. Boakye, J. Soares, Image captioning: Transforming objects into words, arXiv preprint arXiv:1906.05963.
[6] S. He, W. Liao, H. R. Tavakoli, M. Yang, B. Rosenhahn, N. Pugeault, Image captioning through image transformer, in: Proceedings of the Asian Conference on Computer Vision, 2020.
[7] L. Huang, W. Wang, J. Chen, X.-Y. Wei, Attention on attention for image captioning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4634–4643.
[8] M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, Meshed-memory transformer for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10578–10587.
[9] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. 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.
[10] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.
[11] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473.
[12] M.-T. Luong, H. Pham, C. D. Manning, Effective approaches to attention-based neural machine translation, arXiv preprint arXiv:1508.04025.
[13] Q. H. Lam, Q. D. Le, V. K. Nguyen, N. L.-T. Nguyen, Uit-viic: A dataset for the first evaluation on vietnamese image captioning, in: International Conference on Computational Collective Intelligence, Springer, 2020, pp. 730–742.
[14] H. N. Tien, T.-H. Do, V.-A. Nguyen, Image captioning in vietnamese language based on deep learning network, in: M. Hernes, K. Wojtkiewicz, E. Szczerbicki (Eds.), Advances in Computational Collective Intelligence, Springer International Publishing, Cham, 2020, pp. 789–800.
[15] L. T. Minh, D. L. Hoang, N. Thanh-Son, N. T. M. Huyen, V. Xuan-Son, ViLsp 2021 - vicap4h challenge: Automatic image caption generation for healthcare domain in vietnamese, VNU Journal of Science: Computer Science and Communication Engineering 38 (1).
[16] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Advances in neural information processing systems 28 (2015) 91–99.
[17] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318. doi:10.3115/1073083.1073135.
URL https://aclanthology.org/P02-1040
[18] S. Banerjee, A. 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, Association for Computational Linguistics, Ann Arbor, Michigan, 2005, pp. 65–72. URL https://aclanthology.org/W05-0909

[19] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL https://aclanthology.org/W04-1013

[20] R. Vedantam, C. Lawrence Zitnick, D. Parikh, Cider: Consensus-based image description evaluation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[21] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, V. Goel, Self-critical sequence training for image captioning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7008–7024.

[22] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

Appendix A. Example captions by ORT and ObjectAoA

We note that as the captions of public test set and private test set of VieCap4H dataset [15] are keep privately, we can only present the captions generated from ORT and ObjectAoA to compare the qualification of their results on this dataset.
Fig. A.3. Example captions by ORT and ObjectAoA on the VieCap4H dataset #1 and #2

ORT: hô biến bút và đục ở trên tấm gỗ (medical masks are placed on the wooden panel).
ObjectAoA: có nhiều hỏng trong được đặt trên bàn gỗ (there are lots of medical masks placed on the wooden table).

ORT: một tấm biển thương bao đằng được đặt trên một tấm biển (a sign is placed on another sign).
ObjectAoA: một tấm biển thương bao đằng được đặt trên cửa ra vào (a sign is placed near a door).

Fig. A.4. Example captions by ORT and ObjectAoA on the VieCap4H dataset #3 and #4

ORT: hai người mặc đồ bảo hộ màu trắng đang dùng chăn xe (two people wearing white protected gear are standing nearby the car).
ObjectAoA: hai nhân viên y tế mặc đồ bảo hộ màu trắng đang dùng chăn chiếc xe cứu thương (two medical staffs wearing white protected gear are standing nearby the ambulance).

ORT: một người mặc trang phục bảo hộ màu trắng và đeo kính chắn gió ở bên (a man wearing a white protected gear and protected glasses).
ObjectAoA: một nhân viên y tế mặc đồ bảo hộ màu trắng và mũ trùm (a medical staff wearing a white protected gear and a hood).
Fig. A.5. Example captions by ORT and ObjectAoA on the VieCap4H dataset #5 and #6

ORT: một người đàn ông đong ngoài cạnh một người đàn ông đong ngoài ghe (a man sitting nearby another man)
ObjectAoA: bác sĩ đang khám mắt cho một người đàn ông (the doctor is taking care of the eyes of a man)

ORT: một người mặc đồ bảo hộ màu xanh đang làm việc (a person wearing a blue protector gear is working)
ObjectAoA: một nhân viên y tế mặc đồ bảo hộ màu trắng đang cầm một ông nghiệm (a medical staff wearing the white protected gear is holding a test tube)

Fig. A.6. Example captions by ORT and ObjectAoA on the VieCap4H dataset #7 and #8

ORT: hai người phụ nữ đeo khẩu trang đang bế một đứa bé sơ sinh (two women wearing masks are holding a baby)
ObjectAoA: một người phụ nữ đeo khẩu trang màu xanh dương đang nhìn về phía trước (a woman wearing a blue mask is looking towards)

ORT: một người đàn ông đeo khẩu trang trắng đang cười (a man wearing a white mask is laughing)
ObjectAoA: một người đàn ông mặc áo vest đeo khẩu trang màu đen (a man with a vest is wearing a black mask)
Fig. A.7. Example captions by ORT and ObjectAoA on the VieCap4H dataset #9 and #10

Fig. A.8. Example captions by ORT and ObjectAoA on the UIT-ViIC dataset #1
Fig. A.9. Example captions by ORT and ObjectAoA on the UIT-ViIC dataset #2

**GT1**: Một người phụ nữ đang chuẩn bị dùng vợt để đánh quả bóng tennis (a woman is preparing to use the racket to hit the tennis ball)

**GT2**: Một người phụ nữ đang chuẩn bị dùng vợt để đánh quả bóng tennis (a woman is about to use a racket to hit a tennis ball)

**GT3**: Một người phụ nữ đang chuẩn bị dùng vợt để đánh quả bóng tennis (a woman is about to use a racket to hit a tennis ball)

**GT4**: Một người phụ nữ đang đứng trước để đánh quả bóng tennis (a woman is standing in front to hit a tennis ball)

**GT5**: Một người phụ nữ đang đứng trước để đánh quả bóng tennis (a woman is standing in front to hit a tennis ball)

Fig. A.10. Example captions by ORT and ObjectAoA on the UIT-ViIC dataset #3

**ORT**: Một đội viên quần vợt đang cầm vợt thi đấu trên sân (the tennis athlete is holding the racket and competing on the ground)

**ObjectAoA**: Một người đàn ông đang dùng vợt để đánh quả bóng tennis (a man is using the racket to hit the ball)
Fig. A.11. Example captions by ORT and ObjectAoA on the UIT-ViIC dataset #4

**Fig. A.12. Example captions by ORT and ObjectAoA on the UIT-ViIC dataset #5**