Image Captioning with Transformer-based models on UIT-ViIC corpus

Khang Pham Minh[1,2*], Khanh Truong Phuoc Bao[1,2] and Kiet Nguyen Van[1,2]

[1]Viet Nam National University, HCM City, Viet Nam.
[2]University of Information Technology, HCM City, Viet Nam.

*Corresponding author(s). E-mail(s): 19520122@gm.uit.edu.vn;
Contributing authors: 20520579@gm.uit.edu.vn;
kietnv@uit.edu.vn;

Abstract

Image captioning is a difficult task that requires the ability to understand visual information as well as use human language to describe the visual information in the image. In this paper, we conduct experiments to address the issue of Vietnamese Image Captioning by combining Bottom-up and Top-down architecture at the encoder layer with Transformers architecture at the decoder layer. When it comes to Computer Vision, Top-down and Bottom-up architectures assist in extracting features from images that include certain items, while Transformer can generate texts from image features.

Keywords: Deep Learning, Image Captioning, Transfomer Models

1 Introduction

At the nexus of computer vision and natural language processing, image captioning is the problem of producing a natural language description of the material within a picture. Progress in picture captioning has logically followed pace given how busy and recently advanced each of these fields of study are. Improved object detection and convolutional neural network topologies have helped image captioning systems on the computer vision side. More complex sequential models, such attention-based recurrent neural networks, have generated more detailed caption synthesis in the context of NLP.

Inspired by neural machine translation, most conventional image captioning systems utilize an encoder-decoder framework, in which an input image is encoded into an intermediate representation of the information contained within the image, and subsequently decoded into a descriptive text sequence. This encoding can consist of a single feature vector output of a CNN (as in [1]), or multiple visual features obtained from different regions within the image. In the latter case, the regions can be uniformly sampled (e.g., [2]), or guided by an object detector (e.g., [3]) which has been shown to yield improved performance.

Although the state-of-the-art, these detection-based encoders do not currently make use of information on the spatial relationships between the detected items, such as relative position and size. However, since humans utilize this information to make inferences about the actual world, it can frequently be crucial for interpreting the information contained in a picture and corporating spatial relationships has been shown to improve the performance of object detection itself, as demonstrated in [4] Furthermore, in machine translation encoders, positional relationships are often encoded, in particular in the case of the Transformer [5], an attention-based encoder architecture. Consequently, visual encoders for picture captions should also benefit from using the relative locations and sizes of observed items.
Our contributions can be summarized as follows:

- We extract region and grid features from UIT-ViIC dataset[6] using Bottom-up mechanism.
- As a complementary contribution, we conduct experiments to compare different fully-attentive architectures on image captioning and validate the perfomances of models like Object-Relation Transformer (ORT model)[7], Meshed-Memory Transformer (M^2 model)[8], Object-Attention-on-Attention (ObjectAOA, [9]) Relative-Sensitive Transformer Network (RSTNet)[10], using the UIT-ViIC[6] corpus.

2 Related works

2.1 Image Captioning

The main development of image captioning can be divided into two stages: traditional method stage and deep learning method stage. In traditional method stage, retrieval-based and template-based methods are two common types of implementation for image captioning. Given an image, retrieval-based methods retrieve one or a set of most similar sentence from a pre-specified sentence pool, while template-based methods generate slotted sentence templates and use detected visual concepts to fill in the slots. With great progress made in deep learning, the encoder-decoder paradigm derived from neural machine translation was exploited in captioning models where CNN was used as the encoder to extract visual features from an image and RNN as the decoder to generate the corresponding output sequence. After that, the main focus of image captioning is to model the interaction between visual and lingual cues via attention mechanism to get more faithful and rich captions. For example, [2] introduced soft and hard attention into LSTM-based decoder, [11] proposed an adaptive attention mechanism to dynamically decide whether to attend visual signals when generating each word, Anderson et al.[8] proposed bottom-up and top-down attention mechanism that makes the visual features in attention upgrade from grid-level to object and salient region level.

2.2 Bottom-up Attention Model

The definition of spatial image features V is generic. However, in this work we define spatial regions in terms of bounding boxes and implement bottom-up attention using Faster R-CNN. Faster R-CNN is an object detection model designed to identify instances of objects belonging to certain classes and localize them with bounding boxes. Other region proposal networks could also be trained as an attentive mechanism. Faster R-CNN detects objects in two stages. The firststage, described as a Region Proposal Network (RPN), predicts object proposals. A small network is slid over features at an intermediate level of a CNN. At each spatial location the network predicts a class-agnostic objectness score and a bounding box refinement for anchor boxes of multiple scales and aspect ratios. Using greedy non-maximum suppression with an intersection-over-union (IoU) threshold, the top box proposals are selected as input to the second stage. In the second stage, region of interest (RoI) pooling is used to extract a small feature map (e.g. 14 × 14) for each box proposal. These feature maps are then batched together as input to the final layers of the CNN. The final output of the model consists of a softmax distribution over class labels and class-specific bounding box refinements for each box proposal. The original Faster R-CNN multi-task loss function contains four components, defined over the classification and bounding box regression outputs for both the RPN and the final object class proposals respectively. We retain these components and add an additional multi-class loss component to train the attribute predictor.

2.3 Transfomer Models

RNN-based models are limited by their sequence nature and suffer from dependencies between distant positions [5]. In order to address this problem, [12] [13] proposed to replace recurrence and convolutions with attention mechanisms and excitedly refreshed almost all the metrics of neural language processing (NLP). Subsequently, great efforts have been made to transfer this idea into image captioning. [13] explored the convolutional language model in image captioning model. [7] incorporated geometry relationships between region features into transformer architecture for captioning. [14] proposed a GLU like structure on attention mechanism to determine the relevance between attention results and queries. Li et al. [15] extended the attention module linking transformer encoder and decoder to exploit visual information and semantic knowledge extracted by a external attribute detector. [8] introduced a learnable priori information to augment the attention module in transformer encoder and a mesh structure to build full connections between each encoder layer and each decoder layer.[16] introduced Bi-linear Pooling into transformer model to exploit both spatial and channel-wise bi-linear attention distributions. Although the aforementioned transformer-based captioning models have achieved quite promising results, a serious problem still exists: all word sequences are coupled into high dimensional tensor, where visual and non-visual words are treated equally. In this paper, we explore an adaptive attention based on a transformer backbone so that the model can adaptively measure the contributions of visual signals and the current language context when predicting the word sequence for captioning.

3 Approach

3.1 Features Extraction

Due to different inputs of different models, in this work, we propose a method Fig. [1] for feature extraction. Using a detector Faster R-CNN in conjunction with the ResNeXt152++ (X-152++) https://github.com/facebookresearch/grid-feats-vqa based on 1×1 RoIPool while keeping the output architecture fixed for grid features, then apply RoI alignment to extract features for each box based on grid features, convert grid features into region features.

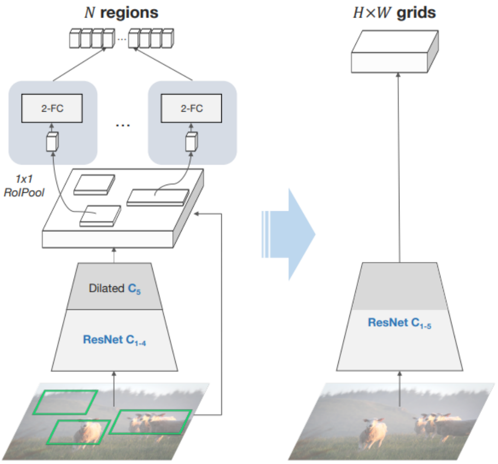


Fig. 1 Features extraction Architecture

3.2 Captioning models

3.2.1 ORT - Object Relation Transformer Model

To improve the visual information from images, inherited the benefit of Bottom-up attention recommended using additional box coordinates obtained from the RPN layer of FasterRCNN [17]. In Fig. [2], given $Q \in R^{seq \times d_q}$, $K \in R^{seq \times d_k}$, $V \in R^{seq \times d_v}$ 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 Ω_W is the matrix formed by $w_{m,n}$.

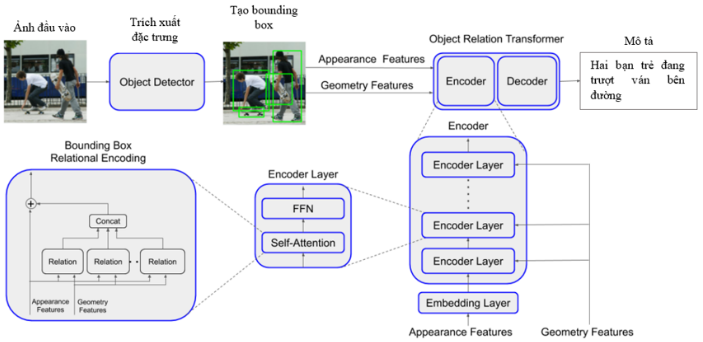


Fig. 2 Object Relation Transfomers architecture

3.2.2 AoA - Enhancing ORT model with Attention on Attention mechanism

Combining the advantages of several SOTA strategies, [9] expand the use of geometric characteristics with the concept of AoA. In their proposal, the output of the attention layer (Fig. [2] is recast as follows using the attention weights Ω of the geometric attention layer in the Object Relation Transformer:

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

where

$$G(Q, K, V) = W_q^g Q + W_v^g f_{att}(Q, K, V) + b^g$$
$$I(Q, K, V) = W_q^i Q + W_v^i f_{att}(Q, K, V) + b^i$$
$$f_{att}(Q, K, V) = Softmax(\Omega)V$$

As previously mentioned, their attention formulation now has an additional attention stage. We assume that, in addition to internal relationships discovered from visual feature vectors of these objects, geometric attention also takes advantage of geometric relationships between objects in images. However, the geometric link is not present in every visual context. Therefore, when the model learns how to employ attention weights from the encoder layers in order to generate appropriate captions, extra information regarding the locations of objects in images may slightly confuse the model. The Attention on Attention techniques are useful in this situation. According to research by [14], Attention on Attention can improve the connection between input data and output vectors from attention layers to produce information that is relevantly attended. By combining geometric attention with Attention on Attention, we can prevent uncorrelated output vectors of attention layers from the query vectors. This makes the model more understandable when learning how to combine the data from visual and linguistic features, which in turn helps the model produce better captions. Furthermore, we use transformer as its decoder to improve the model's use of linguistic information. To allow the model to focus on the things it needs to describe, we inherit the benefit of region-based visual features produced by Bottom-up Attention methods[3].

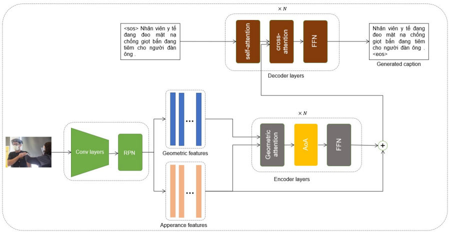


Fig. 3 ObjectAoA Architecture (the yellow AoA module is the addition to the original ORT architecture)

3.2.3 Meshed-Memory Transformer Model

Although the process of determining the attention weight matrix is improved when Bounding Box Relational Encoding [7] is used, defining the relationship between objects is not merely straightforward based on their relative location relationship. Marcella and colleagues studied [8] Transformer's self-attention technique and came to the following conclusion: at the Encoder layer of Transformer, image feature (or sequence of features of locations in image) $X \in R^{d_in \cdot l}$ is used as the matrix keys, queries and values when projected into latent spaces (through the mappings W_k, W_q, W_v), this time the formula calculates the matrix. The attention weights (Ω_A) are concerned with assessing the similarity of the objects identified in the image, and these values reflect information about the relationship between them, which is suitable for employing as a feature for the Decoder to use to provide a description of the objects, but doing so will be difficult for the model to capture the context (a-priori knowledge) to decide specifically and properly. How can one tell if a baseball player is "catching" or "throwing" the ball when both the player and the ball are present near the player's position?
Marcella et al. [8] proposed to improve the self-attention technique of the Encoder layer in the Transformer architecture by augmenting the K and V matrices with Augmented Memory matrices based on the observations made above. In particular, rather than we define:

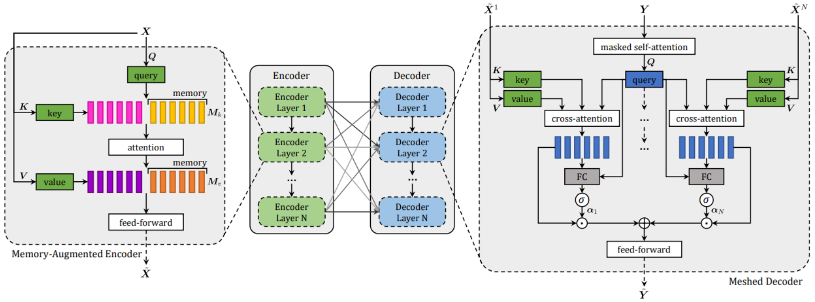


Fig. 4 Architecture of the Meshed-Memory Transformer

$$K = W_k X; \; V = W_v X$$

Then, as shown below, Augmented Memory augments K and V:

$$K = [W_k X; \; M_k]; \; V = [W_V X; \; M_V]$$

Where, $M_k \in R^{d_in \cdot (l+l_mem)}$ and $M_v \in R^{d_in \cdot (l+l_mem)}$ are Augmented Memories, used to store the representation values for a-priori knowledge and [;] is the matrix join operation. The attention weight matrix Ω_A is then still determined through the expression:

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

At this time, the $\Omega_A^{m,n}$ of Ω_A will use more a-priori knowledge stored in augmented memory matrices, so the information collected by the Encoder will be more complete for the Decoder to use to generate Descriptive content for objects. Using this technique, Transformer can pay attention to and describe the image content more deeply and rationally, particularly abstract relationships (preparing to throw, withdrawing to catch the ball, etc.) between the image's objects.
Marcella et al. [8] proposed a new attention technique for Encoder as well as Meshed Cross-Attention for Decoder of Transformer architecture (Fig 4). Unlike the original Transformer, the Meshed Cross-Attention uses all of the features of all Encoder layers rather than just the last Encoder layer's features. This ensures that the Decoder can fully utilize all of the information provided by the Encoder, thereby increasing the efficiency of the description process.

3.2.4 RSTNet

A typical transformer-based image captioning model follows the classic encoder-decoder framework, where the encoder takes the visual features extracted from the image as input and further processes them to strengthen their relationships. Given the encoded features from an encoder, the decoder then generates the output sequence word by word. The core component of transformer is Scaled Dot-Product Attention [5] whose input consists of matrix Q, K and V , where Q is the combination of n_q query vectors, K and V are the combining results of n_k key vectors and n_k value vectors, respectively. Considering only being able to access the partially generated sentence information at testing phase, Zhang et al. [10] added a masked attention module similar to transformer decoder layer on top of the BERT model. But in this work, we train PhoBERT-based language model [18] with cross-entropy loss. All parameters are frozen and the output of masked attention module is used as the representation of language features in RSTNet.

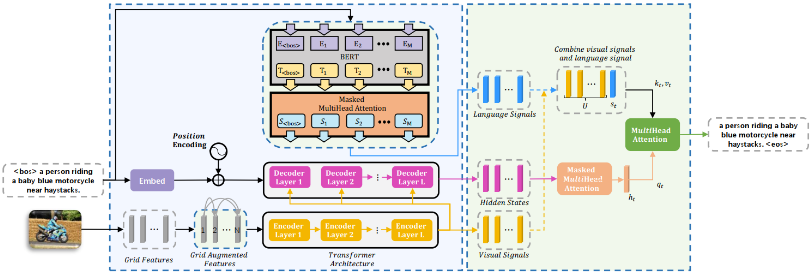


Fig. 5 Architecture of Relationship-Sensitive Transformer (RSTNet

Encoder The raw image feature is first flattened, and then embedded by a fully-connected layer followed by a *ReLU* and a dropout layer to project its dimension to $d_{model} = 512$. The embedded features are send to the first encoder layer of the transformer model. The Scaled Dot-Product Attention in the encoder layer is formulated as:

$$Q = UW_q, K = UW_k, V = UW_v,$$

$$Z = softmax(\frac{QK^T}{\sqrt{d_k}})V$$
$$U \leftarrow U + Z$$

where $U \in R^{N \times d_{model}}$ is the packed visual feature vectors passing in transformer encoder layer, W_q , W_q , W_q are matrices of learnable weights, and d_k is a scaling factor. Grid Augmented (AA) Module In order to compensate for the spatial information loss of the grid features caused by the flattening operation, we propose a grid-augmented Scaled Dot-Product Attention to enhance the encoder layer. The grid-augmented Scaled Dot-Product Attention is formally define as follows:

$$Z_{aug} = softmax(\frac{QK^T}{\sqrt{d_k}} + \lambda^g)V$$
$$U \leftarrow U + Z_{aug}$$

where λ_g is the relative geometry feature of the grid features and Z_{aug} is the result of our augmented attention.

Decoder The word sequence features is first processed by word embedding and incorporated with word sequence position encoding before used as the input of the first decoder layer of the transformer model. The decoder of transformer can be formulated as:

$$h_t = Decoder(U, W_{<t})$$

where $U \in \mathrm{R}^{N \times d_{model}}$ is the output of the last layer of transformer encoder, $W_{<t} = (w_0, w_2, ..., w_{t1})^T$, $W_{<t} \in \mathrm{R}^{t \times d_{model}}$ is word sequence feature of the partially generated sentence, and h_t is the hidden state output by transformer decoder to predict the current word w_t.

Adaptive Attention (AA) Module The adaptive attention module was built on top of the classic transformer decoder. Instead of predicting word using the hidden state h_t directly, language signals, visual signals output by encoder and the hidden states are combined together to measure the contribution of visual signals and language signal for each word prediction.

$$q_{i,t} = h_t {W_i}^Q, k_{i,t} = [U;\ s_t]{W_i}^K, v_{i,t} = [U;\ s_t]{W_i}^V,$$
$$head_{i,t} = softmax(q_{i,t}k_{i,t}^T)v_{i,t},$$
$$head_i = Concat(head_{i,1}, ..., head_{i,M}),$$
$$att = Concat(head_1, ..., head_h)W^O,$$

where $q_{i,t}$ is the query vector for the t-th word word in head i of multi-head attention, $k_{i,t}$, $v_{i,t}$ are the key matrix and value matrix for the t time step word in head i of multi-head attention respectively, head$_{i,t}$ is the attention result for the t-th word in head i, head$_i$ is the attention result for the word sequence in head i, *att* is the attention result of multi-head attention for sequence generation.

4 Experiments

4.1 Configuration

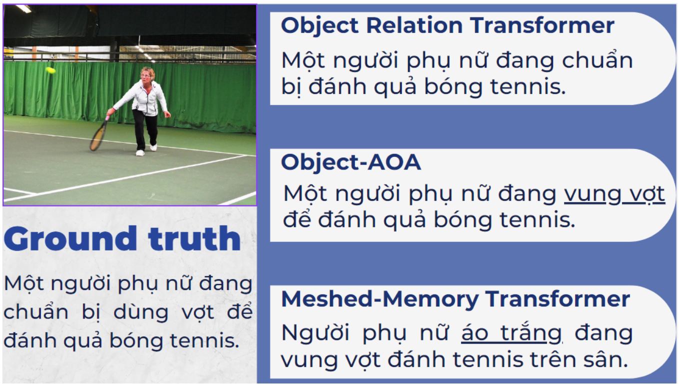
We have evaluated all the mentioned models on the UIT-ViIC [6] dataset. BLEU [19], METEOR [20] ROUGE [21], CIDEr [22] are popular metrics for evaluating image captioning methods. We trained the methods on the UIT-ViIC dataset using the training set, the development set to select the best models, and the test set to evaluate them. All methods were trained using Adam [23] optimizer with the learning rate adjusted to follow the learning rate scheduler from [5]. The best models on dev set were chosen using the CIDEr metric. We set the batch size to 32 and trained the models using the Early Stopping technique (with patience equals to 5).

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
Baseline	0.710	0.575	0.476	0.394		0.626	1.005
RSTNet	0.743	0.639	0.548	0.477	0.357	0.635	1.387
ORT	0.759	0.646	0.550	0.475	0.378	0.639	1.480
ObjectAOA	**0.765**	**0.659**	**0.567**	**0.495**	**0.381**	**0.657**	1.488
M^2	0.752	0.636	0.546	0.476	0.375	0.650	**1.512**

4.2 Experiment results

Table 1 lists the findings of four models that we examined using various metrics. The ObjectAOA model outperforms all other models except the M^2 model in the CIDEr metric. Figure 6, 7 shows some captions generated by the three best model in Table 1. From these two examples, it can be seen that Object-AOA describes things more thoroughly than ORT. It recognizes the woman swinging the racket ("vung vot), not just about to hit the ball ("chuan bi"). But the Meshed-Memory model is more descriptive than the other models. It recognizes the woman is in the white shirt ("ao trang") or even the tennis player is standing on the edge ("dung ngoai bien").



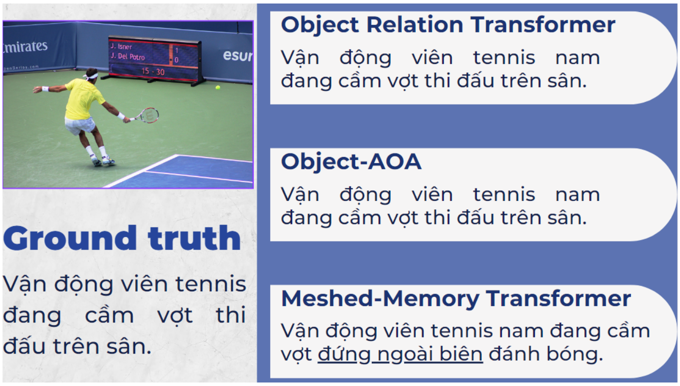
Một người phụ nữ đang chuẩn bị đánh quả bóng tennis.

Một người phụ nữ đang vung vợt để đánh quả bóng tennis.

Người phụ nữ áo trắng đang vung vợt đánh tennis trên sân.

Một người phụ nữ đang chuẩn bị dùng vợt để đánh quả bóng tennis.

Fig. 6 Example captions by ORT, ObjectAOA and M2 - 1



Vận động viên tennis nam đang cầm vợt thi đấu trên sân.

Vận động viên tennis nam đang cầm vợt thi đấu trên sân.

Vận động viên tennis nam đang cầm vợt đứng ngoài biên đánh bóng.

Vận động viên tennis đang cầm vợt thi đấu trên sân.

Fig. 7 Example captions by ORT, ObjectAOA and M2 - 2

5 Conclusion and future work

In this article, we implemented four different transformer models for image captioning task. In the future, we will install some Word Embeddings, try to improve the RSTNet [10] to get better results. Furthermore, we will try to expand the UIT-ViIC dataset into other fields and apply the models mentioned on it, rather than just sportball to conduct a cross-domain method for Vietnamese image captioning task.

References

[1] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)

[2] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation

with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015). PMLR

[3] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: 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, pp. 6077–6086 (2018)

[4] Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3588–3597 (2018)

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[6] Lam, Q.H., Le, Q.D., Nguyen, V.K., Nguyen, N.L.-T.: Uit-viic: A dataset for the first evaluation on vietnamese image captioning. In: International Conference on Computational Collective Intelligence, pp. 730–742 (2020). Springer

[7] Herdade, S., Kappeler, A., Boakye, K., Soares, J.: Image captioning: Transforming objects into words. Advances in Neural Information Processing Systems **32** (2019)

[8] Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10578–10587 (2020)

[9] Nguyen, N.H., Vo, D.T., Ha, M.-Q.: Viecap4h-vlsp 2021: Objectaoa–enhancing performance of object relation transformer with attention on attention for vietnamese image captioning. arXiv preprint arXiv:2211.05405 (2022)

[10] Zhang, X., Sun, X., Luo, Y., Ji, J., Zhou, Y., Wu, Y., Huang, F., Ji, R.: Rstnet: Captioning with adaptive attention on visual and non-visual words. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15465–15474 (2021)

[11] Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 375–383 (2017)

[12] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv

preprint arXiv:1810.04805 (2018)

[13] Aneja, J., Deshpande, A., Schwing, A.G.: Convolutional image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5561–5570 (2018)

[14] Huang, L., Wang, W., Chen, J., Wei, X.-Y.: Attention on attention for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4634–4643 (2019)

[15] Li, G., Zhu, L., Liu, P., Yang, Y.: Entangled transformer for image captioning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8928–8937 (2019)

[16] Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10971–10980 (2020)

[17] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)

[18] Nguyen, D.Q., Nguyen, A.T.: Phobert: Pre-trained language models for vietnamese. arXiv preprint arXiv:2003.00744 (2020)

[19] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)

[20] Banerjee, S., Lavie, A.: 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, pp. 65–72 (2005)

[21] ROUGE, L.C.: A package for automatic evaluation of summaries. In: Proceedings of Workshop on Text Summarization of ACL, Spain (2004)

[22] Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575 (2015)

[23] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)