VNU Journal of Science:
Computer Science and Communication Engineering

Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

VieCap4H - VLSP 2021: A Transformer-Based Method for Healthcare Image Captioning in Vietnamese

Bui Cao Doanh*, Trinh Thi Thanh Truc, Nguyen Trong Thuan, Nguyen Duc Vu, Nguyen Duy Vo

University of Information Technology, Vietnam National University,
Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City

Received 2 April 2022
Revised 28 April 2022; Accepted 5 May 2022

Abstract: The automatic image caption generation is attractive to both Computer Vision and Natural Language Processing research community because it lies in the gap between these two fields. Within the VieCap4H contest organized by VLSP 2021, we participate and present a Transformer-based solution for image captioning in the healthcare domain. In detail, we use grid features as visual presentation and pre-training a BERT-based language model from PhoBERT-base pre-trained model to obtain language presentation used in the Adaptive Decoder module in the RSTNet model. Besides, we indicate a suitable schedule with the self-critical training sequence (SCST) technique to achieve the best results. Through experiments, we achieve an average of 30.3% BLEU score on the public-test round and 28.9% on the private-test round, which ranks 3rd and 4th, respectively. Source code is available at https://github.com/caodoanh2001/uit-vlsp-viecap4h-solution.

Keywords: Healthcare, Image Captioning, Transformer, Rstnet, Bert, Vietnamese.

1. Introduction

Automatically generating captions for images is an exciting subject in computer vision and the natural language processing field [1]. Despite the precision that recent research has achieved, training an AI model for imitating this unique human ability still has many challenges. In recent years, the standard approach for the image captioning problem is based on encoder-decoder architecture like the machine-translation problem [2]: the encoder is a CNN architecture used for extracting visual signals; the decoder is an RNN architecture to predict the possible captions for the corresponding images based on output from the encoder. Inside, the feature extraction problem in image captioning has two main approaches: grid features and region features; which one is more effective are now still being discussed. On the other hand, the VieCap4H [3] dataset is a Vietnamese dataset, so mining the aspect of the Vietnamese language by using a BERT-like model with Vietnamese pre-

* Corresponding author.
E-mail address: 19521366@gm.uit.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.371

trained models such as PhoBERT, BARTPho may achieve good results.

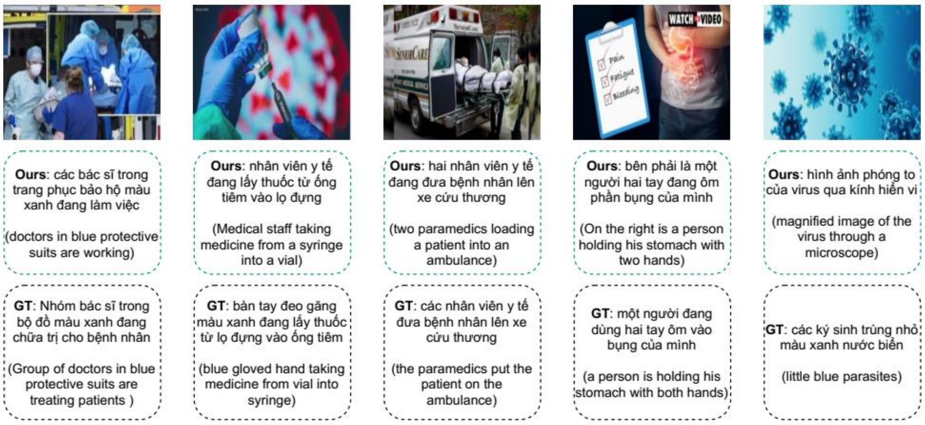
To the best of our knowledge, the main problem of the Image Captioning problem is how images and captions are presented to fit into the RNNs models for training [2]. Typically, region features, which are a set of embedding vectors that present regions with high objectness scores, are often used to present an image. [4]. Extracting these features has high complexity of computation. Random embedding vectors usually present the input captions to adapt to multiple languages. Because the VieCap4H dataset is annotated with Vietnamese captions, we suppose that the mining aspect of Vietnamese will achieve a better performance instead of just using random embedding vectors.

In an attempt to overcome these challenges, we present our experimental process based on our survey and contribute a Transformer-based solution: i) Parsing a sentence into word level by VnCoreNLP [5]; ii) Pre-training a BERT language model from pre-trained PhoBERT-base and applying Masked Language Model (MLM) to combine visual signals and hidden states to predict the next stage; iii) Using X101, X152 and X152++ grid features for presenting images;iv) Training an RSTNet model [4] with a suitable schedule; v) Inference with suitable hyperparameters. Some of our predictions compared with labelled descriptions are shown in Figure 1.

The rest of this report: Section 2 presents our quick survey on the image captioning problem; Section 3 describes more profound our proposed solution; Section 4.1 presents experimental and final results in public test and private test round; Section 5 offers valuable things we learned from the competition; Section 6 summarize the report and present some directions for future research.

2. Related Work

Since 2015, many studies have conducted experiments and proposed methods that solve the Image Captioning problem. One of the earliest studies that made a remarkable milestone in this problem was "Show and Tell", proposed by Xu et al. [6]. In this study, the authors used an LSTM model that encoded the variable length input into a fixed dimensional vector and used these embedding spaces to decode it to the desired output sentence. In the same year, the "Show, Attend and Tell" method [7] was born as an improvement of "Show and Tell". Instead of using global features of images, Vinyals et al. [7] used a CNN backbone to extract grid features and used it as embedding spaces in the LSTM model. Moreover, they proposed two visual mechanisms: Stochastic "Hard" Attention and Deterministic "Soft" Attention, to learn the parts that the LSTM model should focus on predicting the hypothesis caption. This study motivated further research on mining grid features and attention mechanism aspects. In 2018, Anderson et al. [8] proposed Bottom-up and Top-down architecture, which was highly inspired by the Faster R-CNN model [9], opening the new era of presentation of images in the Image Captioning problem. In general, they pre-train a Faster R-CNN model on the Visual Genome dataset. They used this model and fit the input image into it to obtain proposal boxes from the Regional Proposal Network. These proposal boxes were then used as the region features to fit into the RNN-based model for training the VQA and Image Captioning tasks. The year 2019 witnessed many methods of mining the self-attention mechanism to improve the performance of the VQA and Image Captioning problems. Huang et al. [11] proposed "Attention on Attention (AoA)", which extends the conventional attention mechanisms to detect the relative information between attention results and queries; This module explores the relevance among objects in region features in the encoder and filters out the irrelevant/misleading attention result to keep only the useful ones in the decoder. Cornia et al. [11] proposed the M2-Transformer model, which includes a multi-layer encoder for region features and a multi-layer decoder that generates output sentence; A mesh-like structure was also proposed to connect encoding and decoding layers to exploit both low-level and high-level contributions. The exploration of the self-attention mechanism in the Image Captioning problem is still trendy up to now; many studies have improved the performance on this problem via this direction [4], [12], [13]. Some studies have recently improved performance based on BERT-based models [14-16] which are promising.



Ours: các bác sĩ trong trang phục bảo hộ màu xanh đang làm việc

(doctors in blue protective suits are working)

GT: Nhóm bác sĩ trong bộ đồ màu xanh đang chữa trị cho bệnh nhân

(Group of doctors in blue protective suits are treating patients)

Ours: nhân viên y tế đang lấy thuốc từ ống tiêm vào lọ đựng

(Medical staff taking medicine from a syringe into a vial)

GT: bàn tay đeo găng màu xanh đang lấy thuốc từ lọ đựng vào ống tiêm

(blue gloved hand taking medicine from vial into syringe)

Ours: hai nhân viên y tế đang đưa bệnh nhân lên xe cứu thương

(two paramedics loading a patient into an ambulance)

GT: các nhân viên y tế đưa bệnh nhân lên xe cứu thương

(the paramedics put the patient on the ambulance)

Ours: bên phải là một người hai tay đang ôm phần bụng của mình

(On the right is a person holding his stomach with two hands)

GT: một người đang dùng hai tay ôm vào bụng của mình

(a person is holding his stomach with both hands)

Ours: hình ảnh phóng to của virus qua kính hiển vi

(magnified image of the virus through a microscope)

GT: các ký sinh trùng nhỏ màu xanh nước biển

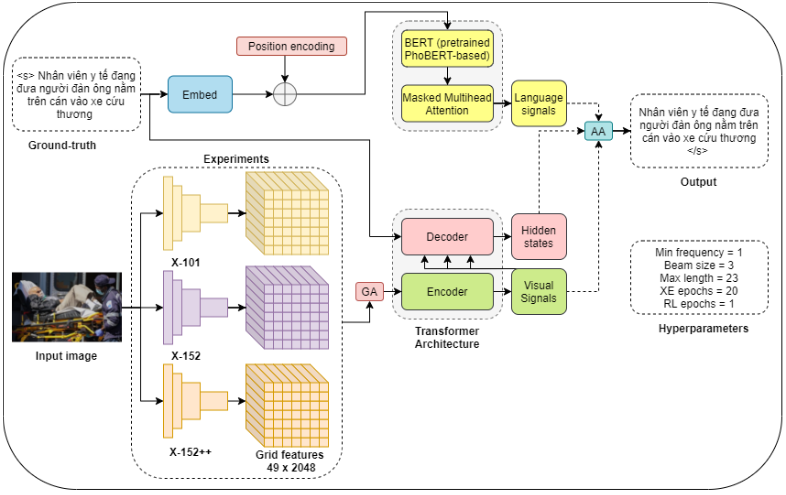
(little blue parasites)

Figure 1. Some predictions by our approach and compare with the labelled description.

3. Methodology

3.1. Image Embedding

We follow [17] to extract grid features; in detail, Jiang et al. [17] use bottom-up, top-down architecture [8] to compute feature maps from lower blocks of ResNet to block C4. But instead of using 14 14 RoIPooling to compute C4 output features, then feeding to C5 block and applying AveragePooling to compute per-region features, they convert the detector in [8] back to the ResNet classifier and compute grid features at the same C5 block. By experiments, they observe that using converted C5 block directly helps reduce computational time but achieves surprising results. We use their X101, X152 and X152++ pre-trained models for grid features extraction. The raw grid features have the shape of (H, W, 2048); we apply an AdaptiveAvgPool2D (7,7) to reshape from (H, W, 2048) to (7, 7, 2048). A single grid is flattened, and then the final output has the shape of (49, 2048).



<s> Nhân viên y tế đang đưa người đàn ông nằm trên cán vào xe cứu thương

Nhân viên y tế đang đưa người đàn ông nằm trên cán vào xe cứu thương </s>

Min frequency = 1
Beam size = 3
Max length = 23
XE epochs = 20
RL epochs = 1

Figure 2. Overview of our experimental process. GA and AA are two modules proposed in [4]. To present an image, we conduct experiments with 3 backbones: X-101, X-152, X-152++. To present an input caption, we train the BERT-based language model from PhoBERT-base.

3.2. Language Embedding

To get language presentations, we train a BERT-based language model (BBLM) that can be expressed by the Equation 1, 2, 3 below this:

$$lf = BERT(W) \quad (1)$$

$$S = MaskedAttention(FF1(lf) + pos) \quad (2)$$

$$\widehat{W} = \log\left(\text{sotfmax}\left(FF2(S)\right)\right) \quad (3)$$

Where $W = (< bos >, W_1, W_2, \ldots, W_M)$ is input sequences; $< bos >$ is an abbreviation of "begin of sentence", this token is fit to the decoder first to begin to predict the captions; $pos \in \mathbb{R}^{d_{bert}}$ is the position encoding of word sequences (Position encoding is an embedding vector that includes positions of each component in the input sequence); FF1 and FF2 are the point-wise feed-forward networks containing two linear layers with ReLU activation. These feed-forward networks are familiar with the Transformer-based model, which process the attention output from the previous Multi Self-Attention head and give its richer presentation version; $lf \in \mathbb{R}^{d_{bert}}$ is output of BERT model; $S \in \mathbb{R}^{d_{bert}}$ is output of masked attention module; $\widehat{W}$ is the log softmax probability distribution of the predicted words.

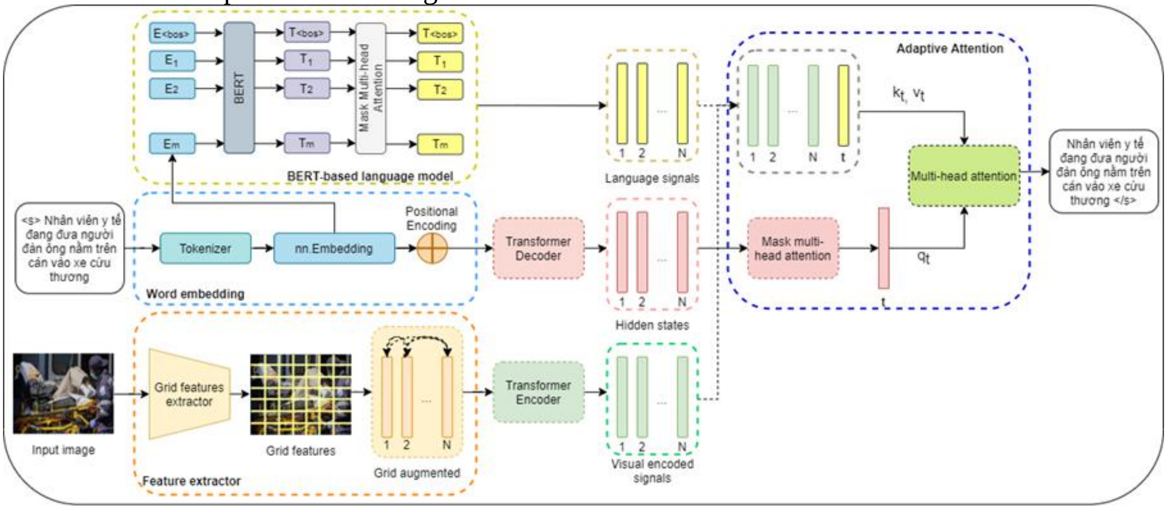


Figure 3. The visualization of the training process of the RSTNet model. The inference is similar, but the input tokens fit the decoder are previous words.

Since VieCap4H is a Vietnamese dataset, we use a pre-trained vinai/phobert-base model [18] which is available on HuggingFace for pre-training a BERT-based language model. Furthermore, the phobert-base model is the small architecture that is adapted to such a small dataset as the VieCap4H dataset, leading to a quick training time, which helps us conduct more experiments. We also try PhoBERT-large, BARTPho-syllable and BARTPho-word [19] pre-trained models, but it does not seem to operate well. The reason may be that the large architectures are not suitable for the small dataset as VieCap4H (contains 8032 samples).

Two tokenizers that we use for experiments is VnCoreNLP [5] and Underthesea. Following [4], we apply the Masked Language Model technique. In detail, the only output of Masked Attention is used as language presentation of a single reference sequence.

3.3. The RSTNet Model

For training, we use the RSTNet model, a Transformer-based architecture proposed by Zhang et al. [4], in which the authors contribute two significant modules for enhancing the performance: Grid Augmented (GA) and Adaptive Attention (AA). We chose this method to conduct the experiments on the VieCap4H dataset as it is a new method whose two proposed modules are novel. The architecture is adapted with grid features, which reduces computation complexity compared with region features [17]. Moreover, besides using the grid features and random embedding vectors to train the Transformer-based model, RSTNet pre-trains a BERT-based model to get language signals. Then they combine three modals: visual encoded features (output from the encoder), hidden states (output from the decoder) and language signals by using the Adaptive Attention module to predict the next word. Because they train the BERT-based model from bert-base-uncased pre-trained model, we replace it with PhoBERT-base to adapt with Vietnamese.

3.3.1. Grid Augmented (GA)

Zhang et al. [4] draw inspiration from two works [20], [21] to calculate the relative geometry matrix between grids $\lambda^g \in \mathbb{R}^{N\times N}$. The authors then incorporated this information into the Transformer's attention mechanism by adding to the attention matrix.

3.3.2. Adaptive Attention (AA)

The authors in [4] found cases where the prediction of the next word is based on linguistic context rather than on image features. Therefore, instead of predicting directly from hidden states of decodes, the AA module is proposed to combine all three: language presentations (output from MaskAttention module at BERT model), visual signals from encoder output and hidden states to predict the next word probabilities. In detail, the output from the decoder at timestep t is fit to another Mask Multi-head Attention to producing attention feature; it then becomes a query for further usage. All visual signals at the current time step t produced from the encoder become a key. The language signal produced by the BERT-based language model at timestep t becomes a value. Then query, key, and value fit the Multi-head Attention to predict the next word.

Figure 3 shows clearly the detail of training and inference process.

3.4. Training With Self-critical Sequence Training

Following [22] study, we apply Self Critical Sequence Training (SCST) after training 20 epochs with Cross-Entropy Loss. SCST is the training strategy that models caption generation as a Reinforcement Learning problem. Thus, the RSTNet model is considered an "agent", the image and language features are considered "environment", and the learning parameters of the network (θ) are defined as "policy" p_θ. After each loop, the "agent" will update its "state" (which are cells and hidden states, RSTNet's attention matrix). An evaluation metric computes thereward by comparing the generated sequence to corresponding annotated sequences. Then the goal of SCST training is to minimize the negative expected reward function (Equation 4):

$$L(\theta) = -\mathbb{E}_{\mathrm{w}^s \sim p_\theta}[r(w^s)] \qquad (4)$$

Where the reward r(·) in the negative expected reward function is the CIDEr-D metric and $w^s = (w_1^s, w_2^s, \dots, w_T^s)$ is the sampled sequences. To compute the negative reward function's gradients, we follow the expressions that are present in the original work:

$$\nabla_\theta L(\theta) \approx -r(w^s)\nabla_\theta \log p_\theta(w^s) \qquad (5)$$

We also use baseline b - the value is approximately equal to the expected reward to reduce the variance of gradients [22]. In this study, baseline b is the reward from the RSTNet model during inference time. The final gradient of negative expected reward is computed by the Equation 6 below:

$$\nabla_\theta L(\theta) \approx -r(w^s - b)\nabla_\theta \log p_\theta(w^s) \qquad (6)$$

In theory, using the SCST technique should improve the results because it optimizes the CIDEr metric directly via REINFORCE algorithm [22]. So we apply this technique to push the result after training 20 epochs with the cross-entropy loss function.

However, through the experimental process, we observe that the more SCST training on the VieCap4H dataset, the lower the precision. Therefore, to achieve the best acceptable precision, we apply the training strategy 20 + 1 (i.e. train 20 epochs using the normal CE loss function, and train only further one epoch RL). The teacher forcing technique is also

Figure 3. The visualization of the training process of the RSTNet model. The inference is similar, but the input tokens fit the decoder are previous words.

applied in training. The results will be presented in Section 4.1.

3.5. Other Hyperparameters

Observe that the maximum length of captions in the public train is 54, so we set $maxlength =$ 54, the minimum frequency of captions $\mathrm{minfreq} = 5$ when training. In Transformer architecture, we use $N_{encoder} = 3$, $N_{decoder} =$ 3, $d_{model} = 512$ and the number of attention heads is 8.

During inference, we use three values $maxlength \in [20, 22, 23]$, apply beam search with $beamsize \in [3, 4, 5]$.and use 50 submissions in public-test round to evaluate. For the private test, we use two values $maxlength \in$ $[22, 23, 24]$.

4. Experiments

4.1. Metric

For evaluation, we use the Bilingual Evaluation Understudy (BLEU) metric, which was first used for evaluating the performance of the captioning model in [1]. This metric is also used in the VieCap4H challenge [3]. BLEU score is commonly used in machine translation tasks. In short, this metric calculates the difference between ground-truth captions and hypothesis captions at the $\mathrm{n-gram}$ level, at which n is a specific value. The VieCap4H challenge uses the average of BLEU values at $\mathrm{n} \in \{1, 2, 3, 4\}$ as the final metric.

4.2. Main Results

In this section, we report the experimental results on both public-test and private-test sets in two Tables 1 and 2. Our machine configuration: 1) Processor: Intel(R) Core(TM) i9-10900X CPU @ 3; 2) Memory: 64GB; 3) GPU: 1× GeForce RTX 2080 Ti 11GiB; 4) OS: Ubuntu 20.04.1 LTS. The experimental results show that X152 is more effective than X101 (+1.3581%); the X152++ backbone achieve higher average of BLEU score when compared to X152 (+0.6519%).

Table 1. Evaluation results on public-test set

#	Visual features	Tokenizer	Training Schedule	Min freq	Beam size	Max length	Average of BLEU (%)
1	X101	Underthesea	20 + 5	5	5	20	25.1919
2		Underthesea	20 + 5	5			26.8172
3		Underthesea	20 + 1	1			27.0318
4		VnCoreNLP	20 + 1	1			27.5599
5	X152	VnCoreNLP	20 + 1	1			28.918
6	X152++	VnCoreNLP	20 + 1	1	4	22	29.5699
7		VnCoreNLP	20 + 1	1	3	23	30.3156

Table 2. Evaluation results on private-test set

#	Visual features	Beam size	Max length	Average of BLEU (%)
1	X-152++	5	22	28.0192
2		4	22	28.5096
3		3	24	28.5219
4		4	23	28.6741
5		3	23	28.858

Two hyperparameters maxlength and beamsize show significant changes when tunning. Through experiments, the best hyperparameter set is ($maxlength = 23$, beamsize = 3) which achieve an average BLEU of 30.3156% (3rd in public-test leaderboard). This hyperparameter set also gives the highest results in the private-test set among our experiments, achieving 4th on the scoreboard (28.858%). The results on two leaderboards witness our solution is effective and competitive when compared with other solutions.

Table 3. Public-test and private-test leaderboard

#	User name	Team name	Average of BLEU (%)
Public-test leaderboard			
1	namnh	AI Club - UIT	30.6
2	vingovan	vc-tus	30.4
3	***caodoanhuit***	***UIT - Together (ours)***	***30.3***
4	gpt-team	VietAI	30.2
5	khiemle	AI Club - UIT	30.2
6	coder_phuho	Fruit AI Club	30.2
7	sonhua3010	QSC	28.6
8	attempt_solution	-	26.6
Private-test leaderboard			
1	tiendv	AI Club - UIT	32.9
2	gpt-team	VietAI	30.9
3	coder_phuho	Fruit AI Club	29.3
4	***caodoanhuit***	***UIT - Together (ours)***	***28.9***
5	vingovan	vc-tus	28.4
6	sonhua3010	QSC	27.1
7	NguyenNghia	ViIC@UIT	26.5
8	utension	di thi	23.9

Table 3 show results among the participating teams. Through experiments, the RSTNet model [4] has been proven to be effective in the VieCap4H dataset. Notably, the architecture of the RSTNet model is similar to other Transformer-based architectures, but the difference is the Adaptive Attention module. When training, besides visual signals and hidden states, they also consider language signals provided by the BERT-based model and employ more Multi-head attention layers to explore more information about these language signals. It can be concluded that this contribution helped the RSTNet becomes a robust model. Because we train the BERT-based model from PhoBERT-base pre-trained model [18], the RSTNet model is adapted well in the VieCap4H dataset with competitive results. Furthermore, the SCST training [22] can push the performance after training with Cross-Entropy loss because it optimizes the CIDEr metric directly by the REINFORCE algorithm. But this training strategy has its limitation. As the baseline b in Equation 6 is the value of the model trained with Cross-Entropy loss, if the model is more accurate when we do SCST training, the performance can be improved by the next epoch. Otherwise, the accuracy can get worse quickly. In the VieCap4H dataset, we find the best schedule for the SCST training is 20 + 1, which is training 20 epochs with Cross-Entropy loss and just one more epoch with SCST training.

5. Lession Learned

During the experimental process, besides grid features, we also tried the region features obtained from the RPN network at the detector at [8] and the specific object features that are finally classified (box features) by using a pre-trained model at [16]. We found that region features and box features do not perform as well as grid features. Perhaps the detectors were trained on the Visual Genome dataset, which differs from the healthcare domain, so when using specific regions or objects will be very confusing, while using grid features to extract the image's global information will work well with multi-domain data, so we decide to experiment and report on this type of feature. On the other hand, the VieCap4H dataset is extremely sensitive; small modifications can significantly improve or decrease the results. The effects that impact the results include the type of features, tokenizer, training and inference hyperparameters. Therefore, many experiments on these effects should be conducted to find a suitable solution that achieves competitive results on this dataset.

6. Conclusion

In this technical report, we report our experimental process and propose a Transformer-based solution that uses grid features as visual presentation and pre-training BERT-based model from PhoBERT-base for image captioning in the healthcare domain within the VieCap4H challenge organized by VLSP2021. Our results are very competitive with other methods on the public-test leaderboard. Instead of using grid features, we will try to extract features of specific objects appearing in the image along with the embedding vector of object tags, which promises to yield better results.

Acknowledgments

This work was supported by the Multimedia Processing Lab (MMLab) and UIT-Together research group (UTRG) at the University of Information Technology, VNUHCM. We would also like to thank the contest organizers for creating an opportunity and environment for the participant teams to compete and learn new pieces of knowledge.

References

[1] X. Chen, H. Fang, T.-Y. Lin, Microsoft coco captions: Data collection and evaluation server, arXiv preprint arXiv:1504.00325, 2015.

[2] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, R. Cucchiara, From Show to Tell: A Survey on Image Captioning, arXiv preprint arXiv:2107.06912, 2021.

[3] T. M. Le, L. H. Dang, T.- S. Nguyen, T. M. H. Nguyen, and X. -S. Vu, "VLSP2021 - VieCap4H Challenge: Automatic Image Caption Generation for Healthcare Domain in Vietnamese," VNU Journal of Science: Computer Science and Communication Engineering, vol. 38, no. 2, 2022

[4] X. Zhang, X. Sun, Y. Luo, et al., "RSTNet: Captioning with Adaptive Attention on Visual and Non-Visual Words, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15 465–15 474.

[5] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, M. Johnson, VnCoreNLP: A Vietnamese Natural Language Processing Toolkit, arXiv preprint arXiv:1801.01331, 2018.

[6] K. Xu, J. Ba, R. Kiros, Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, in International conference on machine learning, PMLR, 2015, pp. 2048–2057.

[7] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and Tell: A Neural Image Caption Generator, in Proceedings of the IEEEconference on computer vision and pattern recognition, 2015, pp. 3156–3164.

[8] P. Anderson, X. He, C. Buehler, 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.

[9] S. Ren, K. He, R. Girshick, J. 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.

[10] 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.

[11] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, Meshed-memory Transformer for Image Captioning, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10 578–10 587.

[12] Y. Pan, T. Yao, Y. Li, T. Mei, X-Linear Attention Networks for Image Captioning, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10971–10980.

[13] X. Zhu, W. Wang, L. Guo, J. Liu, AutoCaption: Image Captioning with Neural Architecture Search, arXiv preprint arXiv:2012.09742, 2020.

[14] X. Li, X. Yin, C. Li, Oscar: Object-Semantics Aligned Pre-Training for Vision-Language Tasks, in European Conference on Computer Vision, Springer, 2020, pp. 121–137.

[15] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, J. Gao, Unified Vision-Language Pre-Training for Image Captioning and Vqa, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 13 041–13 049.

[16] P. Zhang, X. Li, X. Hu, Vinvl: Revisiting Visual Representations in Vision-Language Models, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5579–5588.

[17] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, X. Chen, In Defense of Grid Features for Visual Question Answering, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10 267–10 276.

[18] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in Findings of the Associationfor Computational

Linguistics: EMNLP 2020, 2020, pp. 1037–1042.

[19] N. L. Tran, D. M. Le, and D. Q. Nguyen, “BARTpho: Pre-trained Sequence-to-Sequence Models for Vietnamese,” arXiv preprint arXiv:2109.09701, 2021.

[20] L. Guo, J. Liu, X. Zhu, P. Yao, S. Lu, H. Lu, Normalized And Geometry-Aware Self-Attention Network For Image Captioning, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10327–10336.

[21] S. Herdade, A. Kappeler, K. Boakye, J. Soares, Image captioning: Transforming objects into words, 2019, arXiv preprint arXiv:1906.05963.

[22] 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.