VNU Journal of Science:
Computer Science and Communication Engineering

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

Original Article

# VLSP 2021 - VieCap4H Challenge: Automatic Image Caption Generation for Healthcare Domain in Vietnamese

Le Minh Thao[1,*], Dang Hoang Long[1], Nguyen Thanh Son[2], Nguyen Thi Minh Huyen[3], Vu Xuan Son[4]

[1]*Deakin University, 75 Pigdons Rd, Waurn Ponds, Victoria, Australia,*
[2]*Institute of High Performance Computing, A*STAR, Singapore, 5 Soon Lee, Pioneer Point, Singapore*
[3]*Hanoi University of Science, 334 Nguyen Trai, Thanh Xuan, Hanoi, Vietnam*
[4]*Umeå University,4 Universitetstorget, Umeå, Sweden*

Received 1 January 2022
Revised 31 March 2022; Accepted 5 May 2022

**Abstract:** This paper presents VieCap4H, a grand data challenge on automatic image caption generation for the healthcare domain in Vietnamese. VieCap4H is held as part of the eighth annual workshop on Vietnamese Language and Speech Processing (VLSP 2021). The task is considered as an image captioning task. Given a static image, mostly about healthcare-related scenarios, participants are asked to design machine learning methods to generate natural language captions in Vietnamese to describe the visual content of the image. We introduce VieCap4H, a novel human-annotated image captioning dataset in Vietnamese that contains over 10,000 image-caption pairs collected from real-world scenarios in the healthcare domain. All the models proposed by the challenge participants are evaluated using BLEU scores against ground truths. The challenge was run on AIHUB.VN platform. Within less than two months, the challenge has attracted over 90 individual participants and recorded more than 900 valid submissions.

*Keywords:* Image Captioning in Vietnamese, VieCap4H.

## 1. Introduction

Humans are unique in their capability to interpret and describe their visual perception in natural language. In the last decade, we have witnessed ground-breaking success of modern AI powered by deep learning in different tasks in visual understanding [1-4] and natural language processing [5-8].

However, building a machine that learns to talk about what it sees remains very challenging. In this playground, image captioning, a machine learning task that automatically generates natural language descriptions of a given image, has emerged and attracted enormous attention in the AI research community [9-17]. The task is fascinating and yet challenging at the same time as it sits on the bridge between computer vision

* Corresponding author.
*E-mail address:* thao.le@deakin.edu.au

https://doi.org/10.25073/2588-1086/vnucsce.341

and natural language processing, the two most important fields of modern AI.

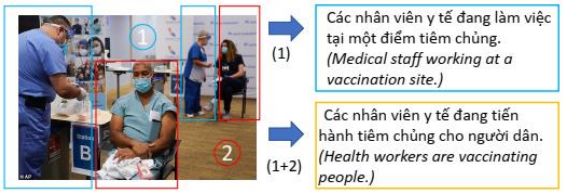


Figure 1. An example of the image captioning task in VieCap4H data challenge. This image can be described in words in two different ways: (1) describing partial information of the image; and (1+2) describing overall information of the image.

The COVID-19 pandemic has exacerbated the ongoing shortage of healthcare workers globally, posing an urgent need for smart assistants that can effectively cooperate with humans to fill the gap. More specifically, these agents require a capability to understand the healthcare domain and summarize the visual information captured by a digital camera to produce accurate descriptions and analyses to humans. While there have been multiple image captioning datasets available in prominent languages such as in English [18, 19], Chinese [20], datasets with Vietnamese captions are scarce. In this work, we introduce a so-called image captioning dataset VieCap4H to challenge machine learning models in their ability to generate descriptions in Vietnamese for visual content, mainly in the healthcare domain. Figure 1 shows an example of the image captioning task in VieCap4H data challenge - i.e., an image is associated with at least one caption in Vietnamese. In some cases, the same image can be described in multiple ways to assess the generalizability of different solutions.

Along with the introduction of the dataset, we also run a challenge held as part of the eighth annual workshop on Vietnamese Language and Speech Processing (VLSP 2021). The challenge was hosted on aihub.vn which consisted of two phases, including public test and private test during September 20 to October 25, 2021. The challenge recorded over 900 valid submitted entries in total by 14 teams of nearly 90 participants.

The contribution of this work is two-fold:

i) we introduce a reliable dataset for image captioning in healthcare specifically for the Vietnamese research community as a result of a conscientious annotation process, providing a fair benchmark for interested researchers to train and evaluate their models.

ii) we run a challenge with an automatic evaluation framework to encourage participants to make use of the provided dataset and contribute their knowledge to advance the field, making potential applications of the task in either healthcare settings and general settings (e.g. virtual assistants for blind and visually impaired people, or visual content indexing and searching) accessible for the local community.

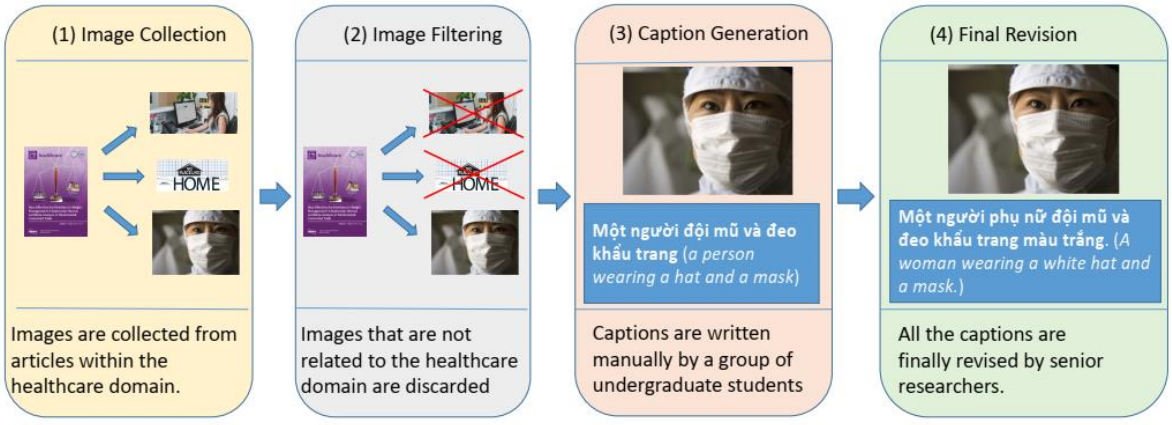


Figure 2. Overall pipeline for data collection and data annotation process. It consists of four steps: (1) ImageCollection, (2) Image Filtering, (3) Caption Generation, and (4) Final Revision.

The remaining sections of this paper are organized as follows. The next section will detail the data collection and annotation process. Section 3 will discuss different approaches in solving the task during the challenge. We will discuss the results in detail in Section 4. Finally, we conclude the paper with some discussions on future directions and challenges of the topic.

## 2. The VieCap4H Dataset

### *2.1. Data Collection*

As we focus on images reported in news, we employed the following steps for constructing the dataset. First, we selected a set of articles collected in the GDEL project. An article was selected if its url contained the substring */health/*. This was to ensure that the articles are within the healthcare domain. We then downloaded all the selected articles and extracted their images using the services provided by Diffbot. Lastly, the images were downloaded and manually examined by two undergraduate students. The images that were not truly related to the healthcare domain were discarded.

A group of undergraduate students were then recruited to write the caption for each of the remaining images. The students were given with detailed instructions. Finally, all their captions were revised by senior researchers.

This last revision helps to correct all the linguistic annotation mistakes and also to improve the consistency among the captions generated by different students. Figure 2 summarizes our data collection and data annotation process.


a) Cluster of captions related to hospital beds

b) Cluster of captions related to people wearing masks

c) Cluster of captions related to babies

d) Cluster of captions related to people wearing protective clothing

e) Cluster of captions related to cells

f) Cluster of captions related to glass bottles

Figure 3. Word clouds showing the most frequent words in representative clusters of the ground-truth captions.

### *2.2. Dataset Splits*

We carefully design a splitting method to form different sets (i.e., training, public test, and private test sets) from the collected data to provide a reliable evaluation of model performance. A good data splitting method should divide the dataset into splits in which each of them covers all usual and unusual patterns in the challenge domain. In addition, as one image can inherently associate with multiple captions and vice versa, the splitting method needs to avoid identical images and captions appear across the splits.

In VieCap4H, the K-mean clustering technique is applied to distribute the collected

data into three different splits including training, public testing, and private testing splits. To be specific, we firstly use Fasttext-Sent2Vec [21] with the pre-trained model provided in ETNLP [22] to extract the features of ground truth captions. Since an image can include multiple captions, we take mean of the extracted features of all the captions associated with the same image to return a single feature vector per image. These textual features are then served as the input to the K-mean algorithm with K = 50 to cluster the dataset into 50 distinguished non-overlapping groups. Figure 3 presents word clouds showing the most frequent words in representative clusters of the ground-truth captions.

For each cluster, we take 80%, 10% and 10% of the total images and their associated captions for the train, public test and private to obtain the train, public test and private test test split, respectively. Finally, we join the respective subsets over all the clusters together. The VieCap4H splits are made available for public use. Interested researchers are required to sign up for an aihub.vn account to access the dataset.

During the two phases of the challenge, only the visual part of the test splits is accessible to participants while ground-truth captions are used at inference time for evaluation purposes only.

*2.2.1. Data Statistics*

Table 1. Data statistics of the VieCap4H dataset

| | Images | Captions | Avg. captions per image | Avg. caption's length (tokens) |
|---|---|---|---|---|
| Train | 8032 | 9429 | 1.17 | 11.88 |
| Public test | 1002 | 1039 | 1.04 | 11.86 |
| Private test | 1034 | 1095 | 1.05 | 11.97 |
| All | 10068 | 11563 | 1.15 | 11.89 |

We provide in Table 1 the statistics of the VieCap4H dataset. The train split contains over 8,000 images with over 9,000 associated captions, while the public test split and private test split have approximately 1,000 images and ground-truth captions each. All splits have captions with similar average length of approximately 12 words per caption.

## 3. The VieCap4H Challenge

*3.1. Methods*

During the challenge, participants have utilized different network architectures to learn to generate textual captions for visual content. These methods can be broadly classified into two categories (See Table 2): Encoder-decoder framework-based methods and Unified vision-language pretraining based methods. As deep neural models severely suffer from data hungry, all the participants leveraged transfer knowledge by adapting different pre-trained image captioning models and fine-tuning them on the provided data to deal with the problem of small training data provided by the challenge.

To be fair between all the participants, we asked all the teams to register with the organizers beforehand to use these existing pre-trained models. Table 3 summarizes all the pre-trained models that are used by the challenge's participants.

Table 2. Methods used by top teams in the final roundof the VieCap4H challenge

| Method / Team | Encoder-decoder | Unified V-L Pretraining |
|---|---|---|
| tiendv | | ✓ |
| gpt-team | | ✓ |
| coder_phuho | ✓ | |
| caodoanhuit | ✓ | |
| vingovan | ✓ | |
| NguyenNghia | ✓ | |

*3.1.1. Methods Using Encoder-Decoder Framework*

A large body of works proposed for image captioning task utilizes an encoder-decoder framework. Specifically, the encoder is often a CNN based model in early works [23] or a transformer-based model in more recent works [24], which takes as input a given image and returns a vector representation of the image. The visual encoding is then served as input of the

decoder, which is often a recurrent neural network, to generate the output caption. These models predict one word at a time using the visual encoding and a ground-truth caption during training, while inputs are the visual encoding and words that are generated in previous time steps during inference time. Figure 4 illustrates a typical encoder-decoder framework for image captioning. Four out of six teams in the final round of the VieCap4H challenge made use of the encoder-decoder framework for their solutions with different choices for network models for the encoder and decoder. These teams have tried different settings with different network architectures for visual encoding, ranging from RoI pooling region features extracted by Faster R-CNN [2] to more recent Transformer-based models such as Swin-Transformer [25] and Vision Transformer [26]. Regarding the decoder, all the teams either used LSTM coupled with attention mechanism or Transformer [7].

#### *3.1.2. Fine-tuning Unified Vision-language Pre-training Models*

A number of unified vision and language pre-training models [33-35] have been proposed in the last few years to solve multiple vision and language tasks with a single model by a simple fine-tuning process. These models are usually trained with a large amount of image-text data using self-supervised learning techniques. They achieved superior performance on various tasks including Visual Question Answering, Image-Text Retrieval and Image Captioning.

In the scope of our challenge, the UIT AI team, who achieved the highest performance on both public and private leaderboards, made use of the VLP model proposed by Zhou et. al. [33]. Meanwhile, the GPT-team, who is the runner-up of the competition, fine-tuned a variant of the GPT model [30] to solve both image captioning and machine translation at once. Thanks to the superb generalization capabilities of these pre-trained models and the access to larger training data via data augmentation and external data, these teams outperformed other teams in the challenge utilizing more traditional approaches by a large margin.

### *3.2. Training Strategies*

As the performance of deep neural networks is largely reliant on the amount of training data, multiple teams of the challenge have applied different augmentation techniques to diversify the training data in order to better facilitate the learning of their models.

#### *3.2.1. Data Augmentation*

The UIT AI team used NLP tools such as Underthesea and PyVi library to partly change the linguistic structure of the annotated captions while keeping the semantic meaning of the sentences unchanged.

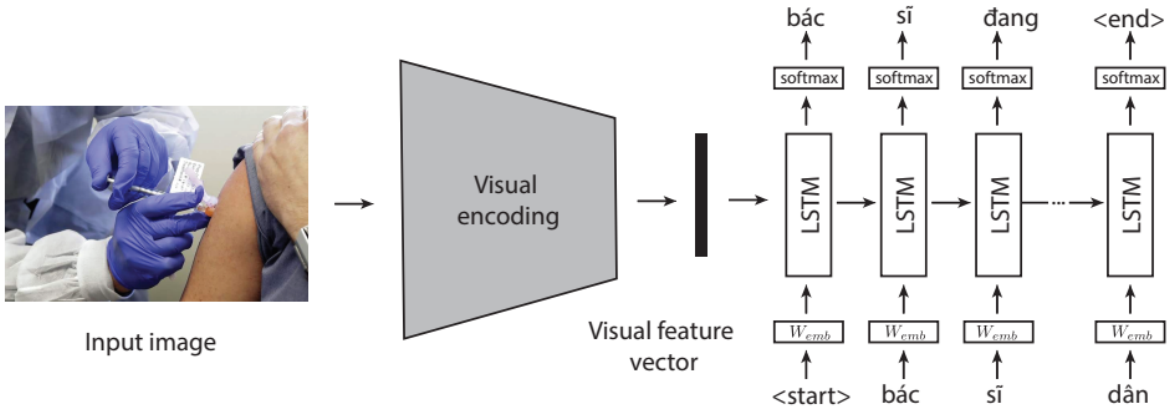


Figure 4. Illustration of a typical encoder-decoder framework for image captioning task. Example is taken from theVieCap4H dataset.

They used a pre-trained PhoBERT [32] to ensure the similarity of the augmented captions and the original captions. The Fruit AI team, who is ranked third in the challenge, instead injected noise to each of the annotated captions by replacing some characters in the captions with random characters and assigned the augmented caption-image pairs with fake labels.

Regarding image augmentation, most of the participants used standard image augmentation techniques such as horizontal flip, random crop to obtain more training data. Compared to text augmentation, image augmentation is more common and convenient to do. The UIT AI team additionally made use of external images crawled from external sources by matching the semantic of the images with keywords in the VieCap4H annotated captions.

Table 3. Pre-trained models used by all the participants of VieCap4H challenge

| Model | Language | Vision | Description |
|---|---|---|---|
| Faster R-CNN [2] | | ✓ | Trained on Visual Genome [27]/COCO [18] |
| ResNet-152 [1] | | ✓ | Trained on ImageNet [28] |
| Swin-Transformer [25] | | ✓ | Trained on ImageNet-1K [28] |
| ViT Transformer [26] | | ✓ | Trained on ImageNet [28] |
| EfficientNet [29] | | ✓ | Trained on ImageNet [28] |
| GPT2News [30] | ✓ | | https://huggingface.co/imthanhlv/gpt2news |
| BERT [6] | ✓ | | Trained on large scale text dataset extracted from BooksCorpus [31] and English Wikipedia |
| PhoBERT [32] | ✓ | | Trained on 20GB texts obtained from Vietnamese news and Vietnamese Wikipedia |

## 4. Results

### *4.1. Data Format*

The VieCap4H dataset includes images and their annotated captions split into three subsets, including train split, public test split and private split. The textual annotations are provided in JSON files in the following format:

[“id” : “uuid_img1”, “captions” : “caption1”,

“id” : “uuid_img2”, “captions” : “caption2”],

where “uuid_img1” refers to the unique index (ID) of a specific image in the dataset and “caption1” denotes the corresponding annotated caption by human annotators as described in Section 2.

The submission format is as follows:

[“id” : “uuid_img1”, “captions” : “caption1”,

“id” : “uuid_img2”, “captions” : “caption2”].

### *4.3. Evaluation Metrics*

Referring to [37], we use the BiLingual Evaluation Understudy (BLEU), which is a common metric for machine translation tasks, to compute the n-gram co-occurrence between ground truth captions and generated caption. Let

$$G = \left\{\{g_{i,t}\}_{t=1}^{T}\right\}_{i=1}^{D} \text{ and } P = \{p_i\}_{i=1}^{D}$$

denote all the ground truth captions $g_{i,t}$ and corresponding generated captions $p_i$ of all the images in a test set, where the index $i$ refers to the $i$-th image in the corpus, $D$ is the size of the test set and $T$ is the number of ground truth captions of the image $i$.

Given $K_n$ is the number of selected *n-gram*for a specific value of $n$ included in a caption,the *n-gram* precision scores at the corpus-level are calculated by:

$$\mathrm{PS}_n(P,G) = \frac{\sum_{i=1}^{D}\sum_{j=1}^{K_n} \min\left(\max_{t\in T}\left(C_j(g_{i,t})\right), C_j(p_i)\right)}{\sum_{i=1}^{D}\sum_{j=1}^{K_n} C_j(p_i)}. \quad (1)$$

Here $C_j(g_{i,t})$ and $C_j(p_i)$ are functions to count the number of $n\_gram_j$ in the ground truth and generated captions. To overcome the weakness of the *n-gram* precision measure in which it can receive a high score if the generated caption is a substring of the groundtruth caption, the $PS_n$ score calculated above is penalized by the brevity penalty defined by:

$$\mathrm{BP}(P,G) = \begin{cases} e^{1-\frac{L_P}{L_G}} & L_P \le L_G \\ 1 & L_P > L_G \end{cases}, \qquad (2)$$

where $L_P$ is the sum of the length of the generated captions and $L_G$ is the sum of the effective length of the ground truth captions. In case multiple ground-truth captions of an image are available, a representative ground-truth caption is chosen as the one that has closest length to the generated caption. As a result, the BLEU score is computed as:

$$s = \exp\left(\sum_{n=1}^{N} w_n \log \mathrm{PS}_n(P,G)\right), \qquad (3)$$

$$\mathrm{BLEU}(P,G) = s \cdot \mathrm{BP}(P,G), \qquad (4)$$

with $w_n$ served as the held constant weight for each value of $n$.

### *4.4. Participation*

During nearly two months of the competition, 90 individual participants registered for the challenge. In which eight teams have signed the terms and conditions upon the use of the provided VieCap4H dataset. Out of all the participated teams, seven teams are selected based on their results submitted on the private test leaderboard. One team has eventually withdrawn their results from the challenge, therefore there are six teams competing for the top three prizes. All the teams in the final round submitted their technical reports explaining in detail their method and findings from the challenge. Table 4 summarizes the competition participation.

### *4.5. Outcomes*

There were 927 valid submissions recorded on our challenge submission site. The best results measured by averaged BLEU scores (BLEU-0, BLEU-1, BLEU-2 and BLEU-3) on the public test split and the private test split are 0.306 and 0.329 (See Figure 5). We provide statistics of the results in Table 5.

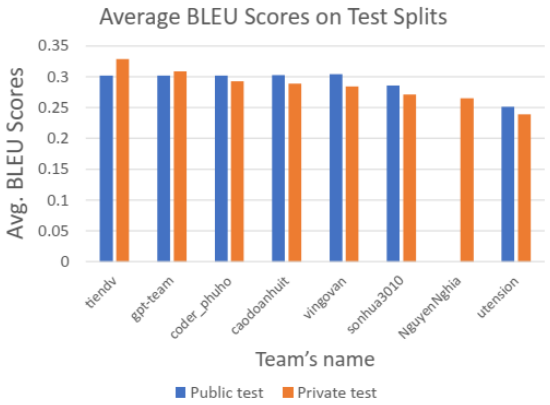


Figure 5. BLEU scores of submissions in top 8 on public and private test splits.

Table 4. The competition participation summary

| | Value |
|---|---|
| # individual participants | 90 |
| # valid teams | 8 |
| # selected teams | 7 |
| # submitted technical reports | 6 |

Table 5. Statistics of the results

| | Public Test | Private Test |
|---|---|---|
| # valid submissions | 830 | 97 |
| Best result | 0.306 | 0.329 |

## 5. Conclusion

In this paper, we introduced VieCap4H, a novel image captioning dataset in Vietnamese, which provides over 10,000 image-captions pairs. The images in the VieCap4H dataset are crawled from healthcare related news while the captions are done by human annotators. The dataset serves as a reliable benchmark to advancing research on automatic caption generation in Vietnamese, specifically in the healthcare context. We also reported details of the VieCap4H challenge which was held as part of the eighth annual workshop on Vietnamese Language and Speech Processing (VLSP 2021).

The challenge attracted over 90 individuals within only two months, revealing an enormous interest of the local research community in the task.

**Acknowledgements**

This work is partially supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2020.DA14. We would like to thank the students from the VNU University of Science who participated to the process of data annotation.

**References**

[1] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[2] 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, Vol. 28, 2015, pp. 91–99.

[3] D. Xu, Y. Zhu, C. B. Choy, L. Fei-Fei, Scene Graph Generation by Iterative Message Passing, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5410–5419.

[4] A. Gordo, J. Almazan, J. Revaud, D. Larlus, End-to-end Learning of Deep Visual Representations for Image Retrieval, International Journal of Computer Vision, Vol. 124, No. 2, 2017, pp. 237–254.

[5] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural Architectures for Named Entity Recognition, Proceedings of the 2016 Conference of the North American Chapterof the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 260–270.

[6] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is All You Need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[8] H. Zhang, J. Xu, J. Wang, Pretraining-Based Natural Language Generation for Text Summarization, Proceedings of the 23rd Conference on Computational Natural Language Learning, 2019, pp. 789–797.

[9] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, A Comprehensive Survey of Deep Learning For Image Captioning, ACM Computing Surveys (CsUR), Vol. 51, No. 6, 2019, pp. 1–36.

[10] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image Captioning with Semantic Attention, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4651–4659.

[11] T. Yao, Y. Pan, Y. Li, Z. Qiu, T. Mei, Boosting Image Captioning with Attributes, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 4894–4902.

[12] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: Lessons learned from the 2015 mscoco image captioning challenge, IEEE transactions on pattern analysis and machine intelligence, Vol. 39, No. 4, 2016, pp. 652–663.

[13] J. Aneja, A. Deshpande, A. G. Schwing, Convolutional image captioning, in: Proceedingsof the IEEE conference on computer vision and pattern recognition, 2018, pp. 5561–5570.

[14] 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.

[15] 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.

[16] Y. Feng, L. Ma, W. Liu, J. Luo, Unsupervised image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4125–4134.

[17] X.-S. Vu, T.-S. Nguyen, D. T. Le, L. Jiang, Multimodal Review Generation with Privacy and Fairness Awareness, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 414–425.

[18] 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.

[19] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions, Transactions of the Association for Computational Linguistics 2 (2014) 67–78.

[20] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, et al., Ai challenger: A Large-Scale Dataset for Going Deeper in Image Understanding, arXiv preprint arXiv:1711.06475.

[21] M. Pagliardini, P. Gupta, M. Jaggi, Unsupervised Learning of Sentence Embeddings Using Compositional N-Gram Features, in: NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics, 2018.

[22] X.-S. Vu, T. Vu, S. N. Tran, L. Jiang, Etnlp:A Visual-Aided Systematic Approach to Select Pre-Trained Embeddings for A Downstream Task, in: Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP), 2019.

[23] 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.

[24] S. Herdade, A. Kappeler, K. Boakye, J. Soares, Image captioning: Transforming Objects into Words, Advances in neural information processing systems, 2021.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical Vision Transformer Using Shifted Windows, International Conference on Computer Vision (ICCV).

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, An Image Is Worth 16x16 Words: Transformers for Image Recognition At Scale, ICLR 2021, 2021.

[27] 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, arXiv:1602.07332.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A Large-Scale Hierarchical Image Database, In: 2009 IEEE Conference on Computer Vision And Pattern Recognition, Ieee, 2009, pp. 248–255.

[29] M. Tan, Q. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR, 2019, pp. 6105–6114.

[30] V. T. Le, Pretrained gpt-2 on vietnamese news, https://huggingface.co/imthanhlv/gpt2news. Accessed on 17th September 2021.

[31] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, S. Fidler, Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 19–27.

[32] D. Q. Nguyen, A. T. Nguyen, Phobert: Pre-Trained Language Models for Vietnamese, 2020, pp. 1037–1042.

[33] 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. 13041–13049.

[34] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Learning universal image-text representations, ECCV20.

[35] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-And-Language Tasks, Advances in Neural Information Processing Systems.

[36] R. Mokady, A. Hertz, A. H. Bermano, Clipcap: Clip Prefix for Image Captioning, arXiv preprint arXiv:2111.09734.

[37] X. Chen, H. Fang, T. Y. Lin, R. Vedantam, S. Gupta, P. Dollár, C. L. Zitnick, Microsoft Coco Captions: Data Collection and Evaluation Server, arXiv preprint arXiv:1504.00325.