Journal of Computer Science and Cybernetics, V.41, N.3 (2025), 323–336
DOI no. 10.15625/1813-9663/20525

INTEGRATING FEATURES AND HARNESSING PRE-TRAINED VISUAL AND LANGUAGE MODELS FOR ENHANCING VQA READING COMPREHENSION

TRUONG XUAN LINH

University of Information Technology, Quarter 6, Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Viet Nam

Abstract. The Visual Question Answering (VQA) problem represents the fusion of natural language understanding (NLU) and computer vision, aiming to comprehend both textual queries and visual content. Recently, researchers have been focusing on the reading comprehension abilities of VQA models, specifically on their capacity to utilize information from scene texts to gather additional context for answering the posed questions.

In this study, we present a fundamental approach for integrating diverse information contained in both images and questions. By leveraging a Transformer model, the proposed solution effectively addresses the VQA problem.

Our approach, achieving the second position in the VLSP 2023 challenge on Visual Reading Comprehension for Vietnamese, highlights the effectiveness of our proposed method. This study contributes to the ongoing discourse on refining VQA models and emphasizes the potential for further advancements in this domain. The codes are available in the GitHub repository, i.e., https://github.com/truong-xuan-linh/FSO-implement.

Keywords. Visual reading comprehension, transformer, vision-language pre-trained model.

1. INTRODUCTION

Visual Question Answering (VQA), first introduced in [1], is a task that combines computer vision and natural language processing. A VQA task needs to comprehend a question and then use the image to provide the most accurate answer. These systems enable humans to interact directly with images through questions, leading to various applications in fields such as health-related assistance in the medical domain [2], supporting blind individuals in their daily lives [3], enhancing access to cultural heritage information [4], and enabling multi-language interaction [5].

In recent years, in addition to focusing on VQA datasets related to quantity, objects, location, and color, researchers in the field of VQA have also sought to enhance the reading comprehension abilities of VQA models. Consequently, several large datasets have emerged, emphasizing the contextual understanding capability in English, such as DocVQA [6], InfographicVQA [7], DocCVQA [8], OCR-VQA–200K [9], and textVQA [10]. However, existing

Corresponding author.
E-mail addresses: linhtx.18@grad.uit.edu.vn (T.X. Linh)

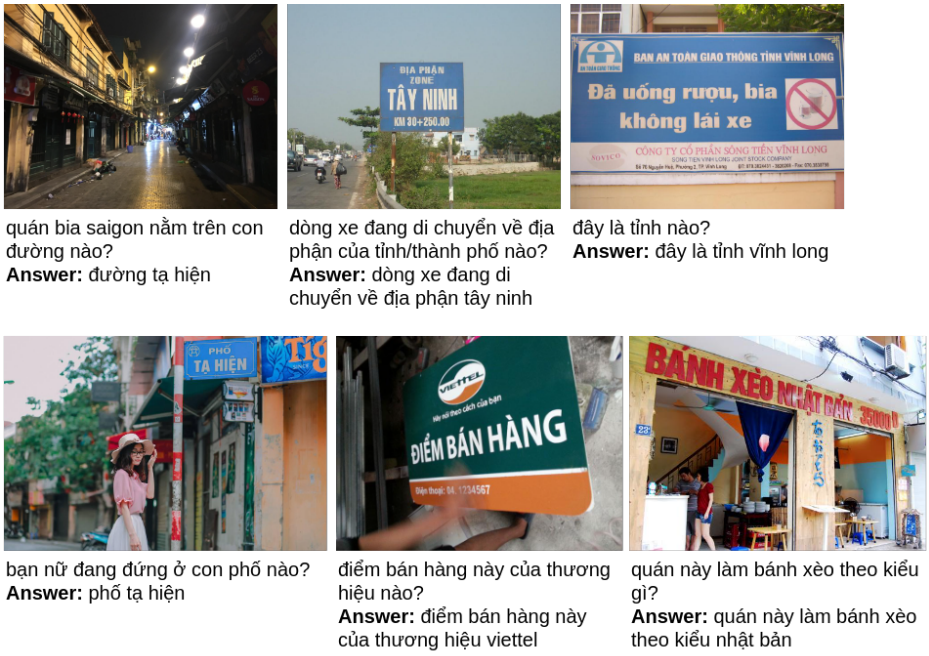

Figure 1: Examples of question-answer pairs with their corresponding images in OpenViVQA dataset

datasets primarily cater to English and other major languages, leaving a gap for underrepresented languages like Vietnamese. Moreover, many current models struggle with complex multi-modal reasoning tasks that require deep integration of text and visual features, especially when working with noisy or low-quality text extracted from images. More recently, a group of Vietnamese authors has introduced the OpenViVQA dataset [11], the first large-scale dataset for VQA with open-ended answers in Vietnamese. This dataset serves as a catalyst for research and development in the VQA domain for the Vietnamese language. Figure 1 shows examples of question-answer pairs with their corresponding images in the OpenViVQA dataset. Due to the relatively late release of VQA datasets for Vietnamese, existing models, such as the MLPAG model [11], have not yet achieved competitive results. These models exhibit significant limitations, including insufficient utilization of high-quality pre-trained models tailored for Vietnamese and inadequate integration of visual and linguistic features.

In this study, we present a method for Visual Reading Comprehension for Vietnamese tasks. We develop a model based on the Transformer architecture, extracting features from both the image and the question using pre-trained Vision-Language models. The main components of our model are as follows:

- For image feature extraction, we utilize a pre-trained Vision Transformer (ViT) [12] model to extract patch features from the image. Additionally, we use the pre-trained PP-OCR [13] model to detect regions containing text within the image, and the pre-trained VietOCR [14] model to recognize the text within these regions.
- For extracting features from questions and text within the image, we utilize a pre-trained ViT5 [15] model to extract features.

- Finally, we aggregate all these features and perform transfer learning on the pre-trained ViT model to generate the answer.

The remaining part of our study is as follows. In Section 2, we present our survey on some datasets and related studies in the VQA problem. Section 3 provides a detailed description of our proposed method. The experiences and achieved results of our approach are presented in Section 4, and a detailed discussion is presented in Section 5. Section 6 concludes our study and outlines our future development directions.

2. BACKGROUND AND RELATED WORK

In this section, we review of various datasets and recent research achievements in the field of Visual Question Answering (VQA). This provides an overview of how other researchers address this problem, especially in terms of the model's ability to understand text within images.

VQA Datasets: In recent years, along with the growing interest of researchers in the Visual Question Answering (VQA) problem, datasets related to this task have also been developed vigorously. The most prominent example is the VQA v2.0 dataset [16], which was one of the pioneering large-scale datasets in this field. Subsequently, other datasets were released, such as the OK-VQA dataset [17], DAQUAR dataset [18], Flickr30k dataset [19], and so on. In addition to datasets requiring reading comprehension in images, datasets like DocVQA, textVQA, and CR-VQA–200K have also been developed significantly. Apart from English datasets, Vietnamese VQA datasets have also seen strong development in recent years, such as ViVQA dataset [20], UIT-EVJVQA dataset [21], and notably the OpenViVQA dataset which was recently published, focusing heavily on the model's ability to comprehend text within images. The OpenViVQA dataset comprises over 37,000 pairs of questions and answers, with images, questions, and answers labeled in Vietnamese. The tables 1 and 2 illustrate our survey on some VQA datasets containing Vietnamese questions and answers.

Table 1: The table comparing the number of questions, answers, and images in different VQA datasets that contain Vietnamese questions and answers.

Name	Images	QAs
ViVQA	10,328	15,000
UIT-EVJVQA	4,321	33,790
OpenViVQA	11,199	37,914

Vision Models: There are numerous pre-trained models applied to the image encoding process in VQA models, such as the ResNet150 [22], EfficientNet [23], and so on. Particularly, in recent times, feature extraction models based on Transformer [24] structures have seen strong development, with notable examples like ViT, BEiT [25], Swin Transformer [26], and others. Additionally, to enhance the text comprehension capability within images for VQA models, researchers often augment the features extracted from OCR models such as EasyOCR [27], PP-OCR [13] and Tesseract OCR [28]. In the OpenViVQA paper, OpenViVQA proposed the MLPAG method. It uses FasterRCNN [29] with ResNeXt152++ [30]

Table 2: Linguistic comparison on questions and answers among VQA datasets (in words). Note that these results were obtained on train-dev and test sets.

Name	Question		Answer	
	Avg.	Max.	Avg.	Max.
ViVQA	9.5	26	1.8	4
UIT-EVJVQA	10.2	45	6	32
OpenViVQA	10.1	32	6.9	56

to obtain region features of the images. Scene texts in images are recognized using Swin-TextSpotter [31].

Language Models: Introduced in 2017, language models based on the Transformer architecture have been increasingly advancing and achieving significant milestones in the field of natural language processing. Some prominent models include BERT [32], BART [33], T5 [34], and others. Alongside this development, pre-trained language models on large Vietnamese datasets have emerged, such as PhoBERT [35], BARTpho [36], and ViT5, thereby making significant contributions to the language encoding process for VQA tasks in general, and specifically for VQA tasks in Vietnamese. In the OpenViVQA paper, OpenViVQA performed embedding both scene texts and questions using FastText [37].

3. METHODOLOGY

3.1. Proposed architecture

Our proposed method named Fusing by Spread Out (FSO) (Figure 2) is based on the ViT5 model introduced in [15]. In contrast to using it solely for feature extraction as discussed in Subsection 3.3, we perform transfer learning on both the encoder and decoder parts of this ViT5 model to leverage the full potential of the features it has previously learned from large Vietnamese datasets. The implementation details are described below:

Grid Features: We utilize the pre-trained ViT model as mentioned in Subsection 3.2 to extract grid features from the images. Specifically, the trained model we used is "vit-base-patch16-224-in21k". Finally, the output will be a vector with a dimensionality of $R^{1\times197\times768}$.

OCR box features: After the bounding boxes are detected using the PP-OCR model as discussed in Subsection 3.2, we continue to use ViT to extract features from these text objects. Instead of using the entire vector with dimensions of $R^{1\times197\times768}$, we simply take the first element of this vector with dimensions of $R^{1\times1\times768}$.

OCR Content Features: We employ VietOCR as mentioned in Subsection 3.2 to recognize the text present in the image. Afterward, for each recognized text object, we add prefixes and suffixes according to the following pattern: "<idx>: " + text + " image:" where the <idx> represents the sequential number of the text container from left to right, top to bottom. Finally, we use ViT5 to extract features from this information.

Note that, to filter noise from the feature extraction process of the bounding box and content, we employ post-processing methods as follows: (1) remove regions with a unique

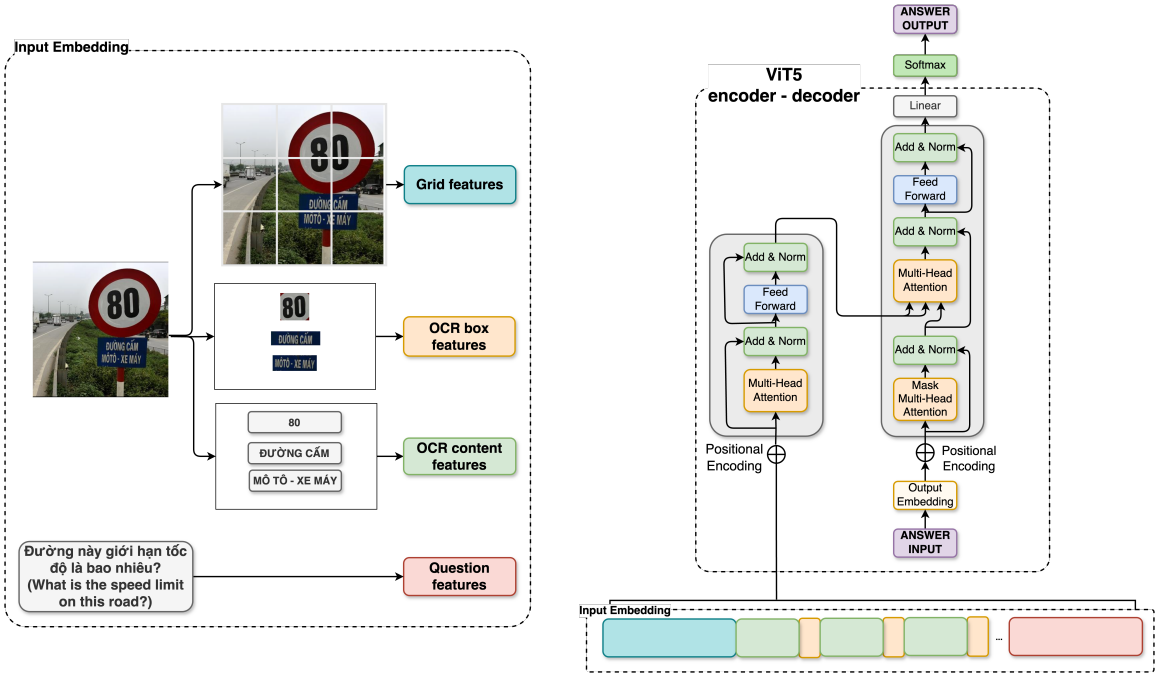

Figure 2: An overview of Fusing by Spread Out (FSO) method based on the Transformer model.

character count less than or equal to 2, (2) eliminate regions with the width smaller than the height, (3) discard regions with an area smaller than 1000 pixels.

Question Features: First, we preprocess the input data by converting all text to lowercase, removing any punctuation marks in the questions so that the input data is consistent. After that, we use the pre-trained ViT5 model as mentioned in Subsection 3.3 to extract features from the input questions. Specifically, the trained model we used is "vit5-base".

Input Embedding: Finally, we concatenate the features together following the pattern: Grid feature - [OCR content feature - OCR box feature]$_1$ -...- [OCR content feature - OCR box feature]$_n$ - Question feature. In Specially, the Grid feature is a vector with the dimension of $R^{1\times197\times768}$, Question feature is a vector has a dimension of $R^{1\times32\times768}$, OCR box feature is a vector has a dimension of $R^{1\times1\times768}$, and OCR content feature is a vector has a dimension of $R^{1\times k\times768}$, where k is the length of the corresponding tokenizer for the OCR-detected text in the image. Additionally, when concatenating OCR content features and OCR box features together, we limit the maximum length of this concatenated vector to 200. For vectors with a length less than 200, we create a padding matrix of zeros and create mask attention to force the model to ignore it. Conversely, for vectors with a length exceeding 200, we perform truncation on them. In this way, we ultimately obtain an input vector with a dimensionality of $R^{1\times429\times768}$.

3.2. Visual representation

In this section, we describe three different methods that we employ in the image representation process, including ViT, PaddleOCR, and VietOCR. Specifically, we use ViT to

extract patch features from the image, PaddleOCR is used for the text detection phase, and VietOCR is employed for the text recognition phase.

Table 3: Comparison table of different visual encoding methods on the public test, where TE stands for textual embedding. More specifically, for ease of comparison regarding the impact of different image feature extraction modules, we kept the textual embedding process constant using ViT5.

Encoder	CIDEr	BLEU
TE+ViT	2.1242	0.2465
TE+ViT+[paddle-viet]OCR	**3.3159**	**0.4413**

Visual Transformer (ViT): The first pre-trained model we use to extract visual features is ViT. The ViT model was proposed in [12]. As the original paper, the ViT pre-trained model encodes an input image $x \in R^{H \times W \times C}$ to a vector representation $x_{ViT} \in R^{1 \times N \times (P^2 \times C+1)}$ where H and W are height and width of an image, C is the number of channels, P is the resolution of each patch be split from the input image and $N = H \times W/P^2$. Besides, the shape of x_{ViT} is added by one because of $1D$ position embeddings $Epos$. In more detail, we use ViT to extract an image input $x \in R^{224 \times 224 \times 3}$ into a feature $x_{ViT} \in R^{1 \times 197 \times 768}$ and use it for our solution.

PaddleOCR (PP-OCR): For text detection, we utilize a pre-trained model called PP-OCR. This is an ultra-lightweight model. In PP-OCR, the authors use Differentiable Binarization (DB) [38] as a text detector which is based on a simple segmentation network. Additionally, to boost effectiveness, during the training process, the authors use six training strategies: light backbone, light head, remove SE module, cosine learning rate decay, learning rate warm-up, and FPGM pruner. In summary, despite being an ultr -lightweight model, the PP-OCR model proves to be a very powerful tool in the OCR problem. However, due to its limitations in recognizing Vietnamese characters, we only employ the text detection component of this model in our solution. More specifically, during deployment, to eliminate common noise, we implement rules such as: skipping text regions with height greater than width, ignoring regions with a $height \times width$ smaller than 1000 pixels, and only extracting a maximum of 32 text regions from left to right, from top to bottom from each image.

Vietnamese OCR (VietOCR): For text recognition, we use a pre-trained model called VietOCR. This model is trained specifically for the Vietnamese language, utilizing VGG19 [39] for feature extraction from images and a Transformer architecture for the encoder-decoder phase. The results achieved by this model have been remarkable and it is widely used in various OCR solutions for Vietnamese.

As shown in Table 3, employing a combination of image feature extraction methods has resulted in higher performance for us.

3.3. Textual representation

In this section, we describe the use of various pre-trained models for extracting features from questions. Specifically, we employ and compare the BARTPho, PhoBERT, and ViT5 pre-trained models.

Table 4: Comparison table of different textual encoding methods on the public test, where VE stands for visual embedding. More specifically, for ease of comparison regarding the impact of different textual encoding methods, we kept the visual embedding process constant using ViT, PP-OCR, and VietOCR.

Encoder	CIDEr	BLEU
VE+BARTPho	3.0223	0.4202
VE+PhoBERT	2.7448	0.4131
VE+ViT5	**3.3159**	**0.4413**

BARTPho: BARTPho is a sequence-to-sequence pre-trained model specifically designed for Vietnamese, developed by VinAI Research [36]. It is based on the BART architecture and trained on the PhoBERT pre-training corpus, which contains 20GB of uncompressed Vietnamese text. The pre-training of BARTPho involves two main stages: (i) introducing noise to the input text through a noising function (e.g., token masking, deletion, and shuffling), and (ii) training the model to reconstruct the original text by minimizing the cross-entropy loss between the decoder's output and the original text.

PhoBERT: PhoBERT is a model based on BERT. It was pre-trained on approximately 20GB of textual data from the Vietnamese News and Wikipedia corpus. Since these models are trained on the Vietnamese dataset, they achieve outstanding results on several NLP tasks in Vietnamese.

ViT5: Through T5-style self-supervised pretraining, ViT5 is trained on a large corpus of high-quality and diverse Vietnamese texts. These texts include a wide range of genres such as Vietnamese news articles, Wikipedia entries, literary works, and other publicly available Vietnamese datasets, ensuring comprehensive coverage of the language's structure and nuances. ViT5 is evaluated based on two downstream text generation tasks: Abstractive Text Summarization and Named Entity Recognition. In the case of the Abstractive Text Summarization task, ViT5 demonstrates outstanding results compared to current models. As for the Named Entity Recognition task, ViT5 produces competitive results with existing models.

Specifically, we use a common feature extraction flow for the above language models as follows: we use only the encoder part of each model to extract an input into a feature $\in R^{max-len \times 768}$, where max-len is the maximum length of the input tokenizer. For $max-len$, we use a value of 32 for the question and do not apply a maximum length for the text identified from the OCR model. In case the question length is less than 32, we simply pad it by zeros and create mask attention to force the model to ignore it.

Although BARTPho and ViT5 share an encoder-decoder architecture, we specifically utilize only their encoder parts to extract features, ignoring their decoder components. Conversely, PhoBERT is designed as an encoder-only architecture, aligning seamlessly with our feature extraction approach. For this study, we utilize the "vit5-base" model available at the following link: https://huggingface.co/VinAI/vit5-base.

As indicated in Table 4, utilizing ViT5 for question feature extraction has led to superior results for us.

3.4. Training strategy

All of our models and experiments were trained on a Tesla P100 GPU with 16GB of RAM. The input image size is 224×224. The batch size is 8. We utilize the Adam optimizer with an initial learning rate of $lr = 1e^{-4}$, and set the exponential decay rate for the 1st moment estimates to $beta_1 = 0.9$, and for the 2nd moment estimates to $beta_2 = 0.999$, with epsilon $eps = 1e^{-7}$. Furthermore, during the training process, if the dev CIDEr score does not improve over two epochs, we promptly adjust $lr = lr \times 0.1$ and continue training.

4. CHALLENGE RESULT AND ANALYSIS

4.1. Dataset

Table 5: Data statistics of the OpenViVQA dataset. Text and Non-Text columns represent the number of question-image (Q-I) pairs focused on text and non-text aspects of the images, respectively.

	Images	Text (Q-I pairs)	Non-Text (Q-I pairs)
Train	9,129	13,104	17,729
Dev	1,070	1,733	1,772
Test	1,000	1,766	1,770
Total	11,199	16,643	21,271

The OpenViVQA [11] dataset is the first large-scale dataset for VQA with open-ended answers in Vietnamese. It is divided into three subsets: the training set, the development set, and the test set. The dataset consists of 11,199 images and 37,914 pairs of different questions and answers. The detailed dataset statistics are presented in table 5.

4.2. Metrics

To evaluate the proposed methods, the VLSP2023 - ViVRC competition employs two metrics: the BLEU score and the CIDEr score. In particular, the BLEU score is the average score of BLEU-1, BLEU-2, BLEU-3, and BLEU-4 as the evaluation metric for VQA. The final ranking is evaluated on the test set according to the CIDEr score (BLUE as a secondary metric when there is a tie).

4.3. Results

The results of our proposed method in the VLSP2023 - ViVRC Challenge, along with other participating teams, are presented in Table 6. On the private test, we got a CIDEr score of 3.4293 and an avg BLUE of 0.4609. In the end, with the achieved results, we secured the 2^{nd} position in the final competition rankings.

Through the competition, we have gained a wealth of new techniques and knowledge related to image and text encoding methods, as well as techniques to address VQA problems in general, and VQA with Reading Comprehension in particular.

Table 6: Performance of the FSO method on the private test set of the OpenViVQA dataset. Which ICNLP, DS@ViVRC, DS@UIT - Multimodal Team, and NTQ Solution are the names of the competing teams in the VLSP 2023 competition. MLPAG [11], M4C [40], and MCAN [41] are the names of the SOTA models tested on the OpenViVQA dataset.

Team name	CIDEr	BLEU
ICNLP	**3.6384**	0.4663
Our approach (linh)	3.4293	0.4609
DS@ViVRC	3.4121	0.4457
DS@UIT - Multimodal Team	3.3172	0.4742
NTQ Solution	3.2926	**0.4876**
MLPAG	1.6104	0.2739
M4C	1.5073	0.2542
MCAN	1.0613	0.1699

4.4. Analysis

Our Fusing by Spread Out (FSO) method leveraged the power of pre-trained models on large datasets. Specifically, by utilizing Vision Transformer (ViT) and PaddleOCR (PP-OCR), our model comprehends images from various perspectives, particularly those containing text. Additionally, leveraging the ViT5 model for question feature extraction and fine-tuning the transformer block further exploits the capabilities of this model, enabling a deeper understanding of questions and accurate answer generation. Moreover, combining features by spreading them into a long sequence simplifies the training process while ensuring the model retains all necessary information.

In summary, by employing the proposed methodology, we have enhanced the model's ability to understand contextual text within images and generate accurate answers. Some results of the FSO method are described in Figure 3. The proposed model demonstrates strong performance on OCR-related questions. However, it struggles with questions that require understanding object locations and counting objects in an image. Future models can address these limitations to achieve more competitive results.

In model development, resource constraints pose a significant challenge, especially when integrating multiple smaller models. To address this, we divide the training process into separate stages. Specifically, for steps that utilize pre-trained models without retraining, such as image feature extraction and OCR extraction, we pre-process these features beforehand and then use them in model training. While this approach optimizes efficiency, it also significantly increases the time and resources required during training.

5. SYSTEMS AND RESULTS

During the problem-solving process, we applied several different methods, some of which proved to be effective in improving accuracy, but some didn't. The following section outlines the methods we employed, including those that worked and some that did not work.

Figure 3: Some examples of results of the FSO method on the public test set

5.1. Working techniques

Firstly, in the process of image representation, we utilized the Vision Transformer (ViT) for feature extraction. This method has good results and outperformed some basic CNN-based models such as ResNet150 and VGG19. Next, we employed PP-OCR and VietOCR for OCR detection in images, which also proved to be more effective than other OCR detection models like easyOCR.

Finally, for textual representation, we employed ViT5, which also demonstrated superior effectiveness compared to other methods such as PhoBERT and BARTpho.

5.2. Non-working techniques

In the problem-solving process, we initially attempted to concatenate all the recognized text in the image into a single long paragraph and then encode it using the ViT5 model. However, this method proved less effective than the approach proposed in this study.

6. CONCLUSION

In summary, we have proposed a method to address the VQA problem by utilizing ViT for image feature extraction. We enhance text comprehension in images by employing PP-OCR and VietOCR. For language feature extraction, we use ViT5. These components are then integrated through a transfer learning approach based on the ViT5 model.

In the future, we plan to fine-tune the VietOCR model to further enhance its capability to comprehend complex text in images. Additionally, we will continue to experiment to discover more suitable methods for image encoding and question formulation, to increase the model's accuracy.

REFERENCES

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[2] A. Ben Abacha, V. V. Datla, S. A. Hasan, D. Demner-Fushman, and H. Müller, "Overview of the VQA-Med task at imageclef 2020: Visual question answering and generation in the medical domain," in *CLEF 2020 Working Notes*, ser. CEUR Workshop Proceedings. Thessaloniki, Greece: CEUR-WS.org, September 22-25 2020.

[3] D. Gurari, Q. Li, A. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:3831582

[4] P. Bongini, F. Becattini, A. D. Bagdanov, and A. D. Bimbo, "Visual question answering for cultural heritage," *IOP Conference Series: Materials Science and Engineering*, vol. 949, no. 1, p. 012074, nov 2020. [Online]. Available: https://dx.doi.org/10.1088/1757-899X/949/1/012074

[5] L. X. Truong, V. Q. Pham, and K. Van Nguyen, "Transformer-based approaches for multilingual visual question answering," *International Journal of Asian Language Processing*, vol. 32, no. 04, p. 2350010, 2022. [Online]. Available: https://doi.org/10.1142/S2717554523500108

[6] M. Mathew, D. Karatzas, R. Manmatha, and C. V. Jawahar, "DocVQA: A dataset for VQA on document images," *CoRR*, vol. abs/2007.00398, 2020. [Online]. Available: https://arxiv.org/abs/2007.00398

[7] M. Mathew, D. Karatzas, and C. V. Jawahar, "DocVQA: A dataset for vqa on document images," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 2199–2208.

[8] R. Tito, D. Karatzas, and E. Valveny, "Document collection visual question answering," in *Document Analysis and Recognition – ICDAR 2021*, J. Lladós, D. Lopresti, and S. Uchida, Eds. Cham: Springer International Publishing, 2021, pp. 778–792.

[9] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty, "OCR-VQA: Visual question answering by reading text in images," in *ICDAR*, 2019.

[10] A. Singh, V. Natarjan, M. Shah, Y. Jiang, X. Chen, D. Parikh, and M. Rohrbach, "Towards VQA models that can read," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8317–8326.

[11] N. H. Nguyen, D. T. Vo, K. Van Nguyen, and N. L.-T. Nguyen, "OpenViVQA: Task, dataset, and multimodal fusion models for visual question answering in Vietnamese," *Information Fusion*, vol. 100, p. 101868, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253523001847

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[13] Y. Du, C. Li, R. Guo, X. Yin, W. Liu, J. Zhou, Y. Bai, Z. Yu, Y. Yang, Q. Dang, and H. Wang, "PP-OCR: A practical ultra lightweight ocr system," *ArXiv*, vol. abs/2009.09941, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221819010

[14] pbcquoc. [Online]. Available: https://github.com/pbcquoc/vietocr

[15] L. Phan, H. Tran, H. Nguyen, and T. H. Trinh, "ViT5: Pretrained text-to-text transformer for Vietnamese language generation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, Jul. 2022, pp. 136–142. [Online]. Available: https://aclanthology.org/2022.naacl-srw.18

[16] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] D. Schwenk, A. Khandelwal, C. Clark, K. Marino, and R. Mottaghi, "A-OKVQA: A benchmark for visual question answering using world knowledge," 2022. [Online]. Available: https://arxiv.org/abs/2206.01718

[18] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," *CoRR*, vol. abs/1410.0210, 2014. [Online]. Available: http://arxiv.org/abs/1410.0210

[19] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *CoRR*, vol. abs/1505.04870, 2015. [Online]. Available: http://arxiv.org/abs/1505.04870

[20] K. Q. Tran, A. T. Nguyen, A. T.-H. Le, and K. V. Nguyen, "ViVQA: Vietnamese visual question answering," in *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*. Shanghai, China: Association for Computational Lingustics, 11 2021, pp. 683–691. [Online]. Available: https://aclanthology.org/2021.paclic-1.72

[21] N. Luu-Thuy Nguyen, N. H. Nguyen, D. T.D. Vo, K. Q. Tran, and K. V. Nguyen, "VLSP 2022 - EVJVQA challenge: Multilingual visual question answering," *Journal of Computer Science and Cybernetics*, p. 237–258, Sep. 2023. [Online]. Available: http://dx.doi.org/10.15625/1813-9663/18157

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385

[23] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6000–6010.

[25] H. Bao, L. Dong, S. Piao, and F. Wei, "BEit: BERT pre-training of image transformers," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=p-BhZSz59o4

[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10 002, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232352874

[27] JaidedAI. [Online]. Available: https://github.com/JaidedAI/EasyOCR

[28] A. Kay, "Tesseract: An open-source optical character recognition engine," *Linux J.*, vol. 2007, no. 159, p. 2, jul 2007.

[29] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[31] M. Huang, Y. Liu, Z. Peng, C. Liu, D. Lin, S. Zhu, N. Yuan, K. Ding, and L. Jin, "Swintextspotter: Scene text spotting via better synergy between text detection and text recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 4593–4603.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[33] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural

language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

[34] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[35] D. Q. Nguyen and A. Tuan Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 1037–1042. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.92

[36] N. L. Tran, D. M. Le, and D. Q. Nguyen, "Bartpho: Pre-trained sequence-to-sequence models for vietnamese," *ArXiv*, vol. abs/2109.09701, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:237571389

[37] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov, "Fasttext.zip: Compressing text classification models," *arXiv preprint arXiv:1612.03651*, 2016.

[38] M. Liao, Z. Zou, Z. Wan, C. Yao, and X. Bai, "Real-time scene text detection with differentiable binarization and adaptive scale fusion," *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, vol. 45, no. 01, pp. 919–931, jan 2023.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[40] R. Hu, A. Singh, T. Darrell, and M. Rohrbach, "Iterative answer prediction with pointer-augmented multimodal transformers for textvqa," *CoRR*, vol. abs/1911.06258, 2019. [Online]. Available: http://arxiv.org/abs/1911.06258

[41] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," *CoRR*, vol. abs/1906.10770, 2019. [Online]. Available: http://arxiv.org/abs/1906.10770

Received on April 07, 2024
Accepted on April 07, 2025