Journal of Computer Science and Cybernetics, V.39, N.4 (2023), 381–391
DOI no 10.15625/1813-9663/38/3/18122

OHYEAH AT VLSP2022-EVJVQA CHALLENGE: A JOINTLY LANGUAGE-IMAGE MODEL FOR MULTILINGUAL VISUAL QUESTION ANSWERING

LUAN NGO DINH[1,3,*], HIEU LE NGOC[2,3], LONG PHAN QUOC[2,3]

[1] *University of Information Technology, Ho Chi Minh City, Vietnam*
[2] *University of Technology, Ho Chi Minh City, Vietnam*
[3] *Vietnam National University, Ho Chi Minh City, Vietnam*

Abstract. Multilingual Visual Question Answering (mVQA) is an extremely challenging task which needs to answer a question given in different languages and take the context in an image. This problem can only be addressed by the combination of Natural Language Processing and Computer Vision. In this paper, we propose applying a jointly developed model to the task of multilingual visual question answering. Specifically, we conduct experiments on a multimodal sequence-to-sequence transformer model derived from the T5 encoder-decoder architecture. Text tokens and Vision Transformer (ViT) dense image embeddings are inputs to an encoder then we used a decoder to automatically anticipate discrete text tokens. We achieved the F1-score of 0.4349 on the private test set and ranked 2nd in the EVJVQA task at the VLSP shared task 2022. For reproducing the result, the code can be found at github*.

Keywords. Machine reading comprehension, Question answering

1. INTRODUCTION

Visual Question Answering (VQA) is a challenging task that has received increasing attention from both the computer vision and the natural language processing communities [1, 2], both areas of artificial intelligence use techniques based on machine learning. Visual question answering is a significantly complex problem because in VQA the form that the question taken is unknown, as is the set of operations required to answer it. In this sense, it more closely reflects the challenge of general image understanding. It becomes more difficult in the VLSP2022-EVJVQA task when applying not only one language but for multilingual including English, Japanese, and Vietnamese.

Multilingual Visual Question Answering (mVQA) shared task was proposed in the VLSP 2022 evaluation campaign to promote the development of Natural Language Processing and Computer Vision in multilingual tasks. The shared task published a training corpus of approximately 30,000 question-answer pairs with images. The task is formalized so that based on an image and a question about it, an mVQA system can predict correct answers

*Corresponding author.
E-mail addresses: 18521064@uit.edu.vn (L.N. Dinh); hieu.le6102@hcmut.edu.vn (H.L. Ngoc); long.phan2810@hcmut.edu.vn (L.P. Quoc).
*https://github.com/DinhLuan14/VLSP2022-VQA-OhYeah

Figure 1: EVJVQA Dataset Example

in several languages. Example questions are shown in Fig. 1. In this paper, we present our approach to the problem. First, we concatenate the embedding at the last hidden state text-encoder model (mBERT [3], XML-R [4], and mT5 [5]) and vision feature extraction model (ViT [6], BeiT [7], and SwinT [8]). Subsequently, the text-decoder model generates the answer. Further elaboration on this process can be found in Section 3, where additional details are provided.

The paper is organized as follows: Section 2 provides a review of related works about the topic. Section 3 presents our proposed method and workflow, which encompasses data cleaning and preprocessing, feature extraction from the utilized images, and the development of language-image models leading to our outcome. Section 4 focuses on the experiments conducted and subsequent analysis. Section 5 concludes the paper by summarizing the findings and suggesting potential avenues for future research.

2. RELATED WORKS

2.1. Multilingual models

Recent works in sequence-to-sequence models focused on the large-scale and multilingual model, which have brought immense impacts on various downstream tasks, including Question Answering, Summarization, Translation, etc. Several model architectures have been trained on other languages not solely on English language text, for example, mT5 [5], mBART [9], mBERT [3], XLM-R [4] and BLOOM [10] for its variants T5 [11], BART [12], BERT [13], RoBERTa [14], and GPT3 [15], respectively.

For these achievements, many massively multilingual pre-trained models have brought the NLP field beyond its borders and attained great applications in many real-world problems. mBERT [3], which has followed BERT [13] architecture strictly, has trained on the Wikipedia corpus for 104 languages. XLM-R [4] is a multilingual version of RoBERTa [14] and is an improved variant of XLM [16]. mT5 [5] is a massively multilingual pre-trained text-to-text transformer, which leverages T5 [11] to train on a new Common Crawl-based dataset with up to 101 languages. It has been proven to outperform many multilingual models in a variety of tasks such as Question Answering and Translation.

T5 [11] is an encoder-decoder pre-trained model, which has trained on both unsupervised and supervised tasks converted into a text-to-text format. The special thing about the t5

architecture is that it adds a prefix to a language model that corresponds to allowing fully-visible masking over the input. They achieve state-of-the-art results on many benchmarks covering summarization, question answering, text classification, etc.

2.2. Vision transformer

With the advancement of Transformer architecture in the Natural Language Processing field, this architecture has gradually widespread in other domains. In the computer vision field, the transformer architecture has been applied directly to sequences of patches and attained very well on image classification tasks. Vision Transformer - ViT [6] is the first pre-trained model that has successfully applied Transformer architecture and trained on the ImageNet [17] dataset. There are many attempts to combine self-attention mechanisms with CNN architectures, but none have yet been proven effective for large-scale image recognition. Therefore, classic ResNet [18] architectures were still state-of-the-art back then.

Inspired by the Transformer architecture success, ViT applied this architecture with the fewest modifications, which considered image patches as word tokens like in NLP and trained on image classification in a supervised way. When trained on mid-sized datasets, the ViT model performs discouraging outcome due to a lack of inductive biases inherent to CNN and lead to not generalizing well. However, it attains outstanding performance when pre-trained at sufficient amounts of data and transferred to tasks with fewer data points. ViT approaches have outperformed state-of-the-art approaches on multiple image recognition benchmarks.

2.3. Multilingual visual question answering

As technology progresses, more and more real-world problems are not capable of being addressed by using solely one modality. Instead, they need collaboration on multimodal solutions. For instance, the visual question-answering task, image captioning task, and optical character recognition task (OCR) need vision-language (VL) models to be addressed. Many works for generating datasets for Visual Question Answering (VQA) have been processed but solely in one language like COCO [19], ViVQA [20]. MS COCO VQA dataset [19] contains 200 thousand, and more than 600 thousand English question-answer pairs which are manually labeled. ViVQA [20] is collected from a COCO source and uses translation tools to translate to Vietnamese, then goes through a validation process for checking the dataset. Another research on leveraging visual rather than textual information [21] is counter to many language-prior approaches and makes the training process more robust.

There are many approaches to the problem of combining visual and textual modalities in VQA. In the previous survey study on VQA, the proposals [22] have classified these approaches based on their mechanisms. The first one is the Joint embedding approach, in which image representation is extracted using a CNN-like pre-trained model, and text representations are obtained with word embeddings feeding to RNN. Then, the joint embedding goes through a classifier to predict a short answer or a recurrent network just like a decoder to generate variable-length a sentence. This method is limited by using global image features to represent visual input, leading to producing irrelevant and noisy information. The second approach tends to address this problem by applying the Attention mechanism to local image features from different regions. This approach is reported to have a better improvement over many models using global image features. The third approach is by using compositional

models, which connect distinct modules designed for specific capabilities such as memory or reasoning. Another approach leverages external knowledge bases.

Recently, the authors in [23] introduced PaLI, a unified language-image model trained to perform many tasks including VQA in more than 100 languages. It outperforms many previous models on visual question-answering benchmarks. Its approach is significantly useful and scalable for building a Multilingual Visual Question Answering system. The PaLI model architecture consists of a Transformer encoder for processing input text and a Transformer decoder for generating output text. It also uses a Vision Transformer model for processing the image and feeding them to the Transformer decoder as "visual words". This architecture reuses weights from a uni-model pre-trained model in image and text, which not only leverages the transfer learning process but also saves computational costs.

3. METHODOLOGY

Our overall system integrates three essential components: Data Preprocessing, Data Augmentation, and Modeling Architecture. Data Preprocessing ensures the data is cleaned, and standardized, and language-specific variations are handled. Data Augmentation enhances the training data by introducing variations tailored to different question types. The Modeling Architecture combines text and vision processing, leveraging pre-trained models to accurately answer questions based on contextual information. By seamlessly connecting these components, our system optimizes performance and provides robust solutions to the task at hand.

3.1. Data preprocessing

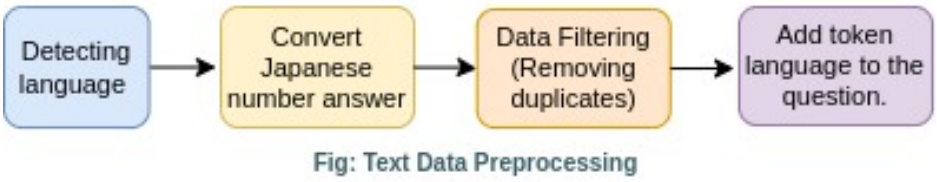

Figure 2: Text data preprocessing with 4 steps

To extract useful features, we started with some preprocessing steps before putting data into models, which are outlined below:

We found that the number questions in English and Vietnamese have the answer as a string, while in Japanese it is usually a number. We convert these types of answers to the correct format, for example, number two in Japanese into "2" or number six in Japanese into "6".

Remove duplicate data from images that have the same question and answer. Additionally, the data available for an image have one question but multiple answers. With experience, we remove duplicate questions and keep those with the longest answers.

Using polyglot to detect language and add a token <lang> for the language of the question before text-encoder: <[CLS], [lang], [question], [SEP]> .

3.2. Data augmentation

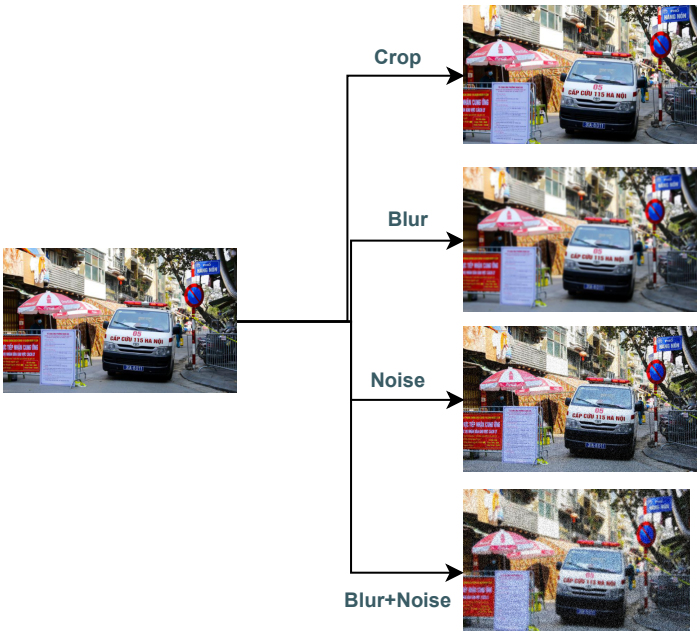

Figure 3: Image data augmentation example with an image in the training dataset

The effectiveness of image augmentation is demonstrated in [24]. This method would make the training process more robust and enhance performance. In this challenge, we suggest a strategy for dealing with any type of question.

For example, with a question about the color or the material, techniques that adjust the color is likely to cause confusion leading to wrong answer. On the other hand, we observe that there are many questions regarding the position of an object so many techniques such as flipping, and rotating at a large angle cause confusion leading to learning the wrong features. For that reason, our strategy is first randomly cropping each side by around 10 of each size, then randomly applying either Gaussian blur or additive Gaussian noise in Figure 3. Thus, we apply this augmented sequence to each image in each question-answer pair, with a probability of 50%, which resulted in different levels of augmenting an original picture for each question-answer pair. This leads to an improvement which is presented in the next section.

3.3. Modeling architecture

The proposed model architecture, as depicted in Figure 4, is an innovative extension of the PaLI architecture [23]. The design aims to generate text-based responses that align with the context provided by an image and a text string, functioning akin to the T5 model's processing of textual prompts.

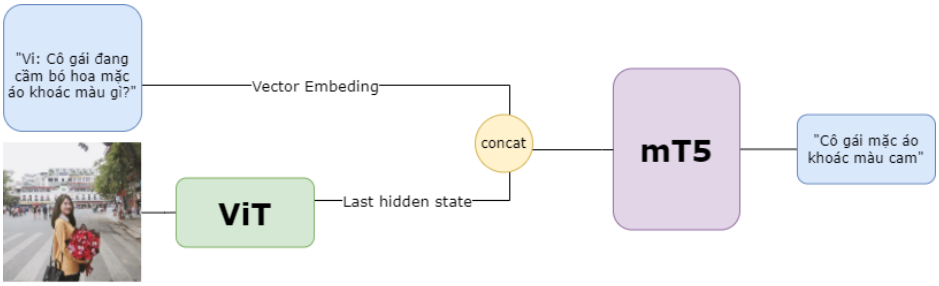

Figure 4: Architecture for Vision Question Answering Task

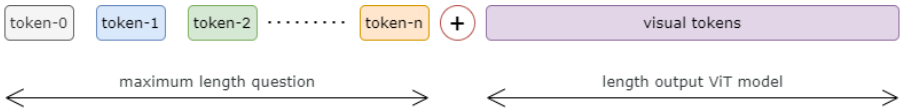

Figure 5: Present the combination of feature vectors before passing through the encoder-decoder model

Initially, the textual query undergoes a transformation in the T5 model's embedding layer, resulting in an embedded vector representation. The core of this architecture is a transformer-based text encoder-decoder [25]. The design leverages the capacity of Transformers to handle various data types, as demonstrated by the incorporation of visual information into the model.

In the proposed architecture, the text encoder is supplied with visual "tokens" derived from the output features of a Vision Transformer, which processes the input image. In contrast to common practice, the Vision Transformer's output bypasses any pooling before being forwarded to the encoder-decoder model through cross-attention, preserving the comprehensive information encapsulated in the visual tokens.

As demonstrated in Figure 5, these vector word tokens and visual tokens are concatenated, creating a unique blend of visual and textual data that is then fed into the encoder-decoder model. This integrated input encapsulates the original image and question, poised for effective processing and subsequent output generation.

Our experimentation leverages previously trained uni-model checkpoints as a robust foundation. We perform fine-tuning of the text encoder-decoder using pre-trained mT5 [5] models. Simultaneously, we are conducting trials with various pre-trained vision models, including but not limited to ViT [6], Beit [7], and Swin [8] architectures.

This innovative approach synergizes text and image data processing methodologies, offering promising prospects for advancements in image interpretation and contextual understanding, contributing to the broader quest for more sophisticated AI models.

4. EXPERIMENT AND ANALYSIS

4.1. Dataset

In our experiment, we used UIT-EVJVQA [26], the first mVQA dataset with three languages, including English and Vietnamese released by VLSP-2022 Organizers for EVJVQA

challenge (https://vlsp.org.vn/vlsp2022/eval/evjvqa). This dataset includes question-answer pairs created by humans on a set of images taken in Vietnam, with the answer created from the input question and the corresponding image. UIT-EVJVQA consists of about 30K question-answer pairs for evaluating the mQA models.

Table 1: Overview statistics of the UIT-EVJVQA dataset

	Train	**Public test**
Number of data	23775	5015
Number of vi-data	8320	1678
Number of en-data	7194	1686
Number of ja-data	8261	1651
Question length (min, mean, max)	4,11,46	
Answer length (min, mean, max)	1,6,30	

4.2. Evaluation metrics

Two metrics, including F1 and BLEU scores were used for evaluation. In particular, BLEU is the average score of BLEU-1, BLEU-2, BLEU-3, and BLEU-4 as the evaluation metric for visual question answering. F1 is calculated from the precision and recall.

4.3. Hyper-parameter setting

We based on statistics of the UIT-EVJVQA dataset on Table 1 to set the hyper-parameters and strategies for our training. The dataset was divided into a training set and a validation set using a 10-fold cross-validation method. Given the characteristics of the dataset, k-folds were grouped by image id. In this sense, data with the same image id were grouped in the same fold to prevent information leakage into the validation set. AdamW was used for optimization, with a learning rate of 3e-4 and a batch size of 16. A linear scheduler was applied with a warm-up ratio of 0.2. The weight decay was set to at 0.01 and applied to all layers except the bias and LayerNorm weights in the AdamW optimizer. The maximum gradient norm for gradient clipping was also set to 3. The number of training epochs was 15. The

Table 2: Experiment results in BLUE and F1 score on the validation set and public test set in the model development stage

Model	**BLEU-dev**	**F1-dev**	**BLEU-test**	**F1-test**
Vit-base + mT5-base [1]	0.4018	0.4512	0.2409	0.3369
Clip-vit-base + mT5-base	0.3783	0.4194	0.1734	0.2599
Beit-base + mT5-base	0.3924	0.4330	0.2288	0.325
Swin-base + mT5-base	0.3853	0.4281	0.2181	0.3106
[1] + Filter Data	0.3970	0.4557	0.2326	0.3353
[1] + Filter Data + Data Augmentation	**0.4082**	**0.4616**	**0.2415**	**0.3459**

maximum length of questions and answers was set to 70 and truncated if it exceeded this maximum length.

Our experiments were conducted on a machine with an A100 40GB GPU and 250 GB RAM.

4.4. Training strategies

Our model was fine-tuned based on the PALI [23] model architecture, which has achieved state-of-the-art performance in the visual question-answering task on COCO benchmarks [19]. We used pre-trained models of image processing such as ViT, BeiT, and Swin, and combined these with the text-to-text T5 model. Freezing the vision pre-trained model (ViT) during pretraining has been proven to produce better results, leading to an improvement in downstream fine-tuning [23]. Subsequently, we fine-tuned the model through incremental changes to identify methods for improvement.

The decoupled weight decay regularization method was used. Research proposed in [27] has empirically shown that their version of Adam with decoupled weight decay substantially outperforms the standard implementation of Adam with L2 regularization in terms of generalization. They also demonstrated the use of warm restarts for Adam to improve its time performance.

In addition, the combination of two pre-trained models as encoder and decoder, where the encoder used the ViLT [28] model and the decoder class utilized models like mBERT and XLM-R was experimented. However, the performance of these models was not satisfactory.

4.5. Result and analysis

Table 2 compares the effectiveness of different pre-trained models and highlights the importance of the data utilization process. Our baseline architecture was applied to various pre-trained models, and the best result in the Visual Question Answering task was achieved using the ViT model, outperforming the Swin, BeiT, and Clip models. The ViT model showed superior performance over other pre-trained image models, with a BLEU score of 0.4018 and an F1 score of 0.4512 on the validation set, as well as a BLEU score of 0.2415 and an F1 score of 0.3459 on the public test set. These results advocate for reusing pre-trained unimodal models and emphasize their capabilities, while also offsetting the considerable cost of large-scale training efforts.

Then, the data preprocessing and the model fine-tuning using the ViT model for further experiments were proceeded. In addition to the hyperparameters presented in the previous section, text data was also preprocessed and data augmentation techniques were employed for image data. The results were improved significantly on our validation set, specifically by 0.0064 points on the BLEU score and 0.0104 points on the F1 score. Our experimental results showed that data augmentation for images should primarily involve creating noise and blur while rotating or cropping the images was ineffective because these processes eliminated positional information. This reflects reality, as the number of questions about left, right, top, and bottom positions in the dataset is substantial, and rotating the images alters the position of the objects. These findings demonstrate that reasonable use of data augmentation for visual input and pre-processing for text input leads to improved performance.

Table 3: Experiment results in BLUE and F1 score on private test set with 3 submitted models for EVJVQA challenge

	Private-test	
	F1-score	BLEU
M1: ViT base	0.4273	0.3797
M2: ViT Large	0.4135	0.3543
M3: Vit base & Data Augmentation	**0.4349**	**0.3868**

Table 3 presents the results submitted for the private test set using three types of models: M1, M2, and M3. The M1 and M2 models share the same parameters, with the only difference being the hidden layer. While the M1 model employed base versions of the ViT and mT5 models over 15 epochs, the M2 model utilized their larger versions over 10 epochs. It is evident that training with the base version resulted in higher scores: an increase of 0.0138 in F1 and 0.0257 in BLEU. We found that the larger mT5 model demonstrated better performance for long text generation. However, as the answer data of the VLSP competition is relatively short, this led to poorer results on the F1 and BLEU metrics. Additionally, increasing the number of training epochs also helped the model achieve better results on the private test set. Ultimately, our best model was the M3 model, which achieved an F1-score of 0.4349, a BLEU score of 0.3868, and secured 2nd place in the VLSP 2022 competition. For this model, data augmentation and preprocessing on the training data were performed, leading to better results than the base model. Despite achieving 2nd place in the VLSP 2022 EVJVQA competition with the M3 model, the results were not outstanding, underscoring the challenging nature of this task.

5. CONCLUSION AND FUTURE WORK

This paper presents the architecture using joint feature vectors in the language-image model for the multilingual visual question-answering task at VLSP 2022 – EVJVQA Challenge. Also, we have demonstrated experimentally using pre-trained models of the Encoder-Decoder model and ViT model. The exprimental results showed that the combination of T5 and ViT effectively outperformed other models. Additionally, the importance of processing and augmenting data in this challenge was also discussed.

In the future, the more state-of-the-art models including generative language models such as FLAN-T5, BLOOM will be experimented. In addition, the more efficient architectures will be explored to improve performance in this task.

ACKNOWLEDGMENT

This work was supported by the UIT Natural Language Processing Group, UIT-VNUHCM.

REFERENCES

[1] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, and D. Parikh, "Vqa: Visual question answering," *arXiv preprint arXiv:1505.00468*, 2015.

[2] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel, "Visual question answering: A survey of methods and datasets," *arXiv preprint arXiv:1607.05910*, 2016.

[3] J. Devlin, "Multilingual bert readme," 2018. [Online]. Available: https://github.com/google-research/bert/blob/master/multilingual.md

[4] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *CoRR*, vol. abs/1911.02116, 2019. [Online]. Available: http://arxiv.org/abs/1911.02116

[5] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *CoRR*, vol. abs/2010.11934, 2020. [Online]. Available: https://arxiv.org/abs/2010.11934

[6] 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," *CoRR*, vol. abs/2010.11929, 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[7] H. Bao, L. Dong, and F. Wei, "Beit: BERT pre-training of image transformers," *CoRR*, vol. abs/2106.08254, 2021. [Online]. Available: https://arxiv.org/abs/2106.08254

[8] 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," *CoRR*, vol. abs/2103.14030, 2021. [Online]. Available: https://arxiv.org/abs/2103.14030

[9] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, and L. Zettlemoyer, "Multilingual denoising pre-training for neural machine translation," *CoRR*, vol. abs/2001.08210, 2020. [Online]. Available: https://arxiv.org/abs/2001.08210

[10] T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.

[11] 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," *CoRR*, vol. abs/1910.10683, 2019. [Online]. Available: http://arxiv.org/abs/1910.10683

[12] 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," *CoRR*, vol. abs/1910.13461, 2019. [Online]. Available: http://arxiv.org/abs/1910.13461

[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[14] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[15] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever,

and D. Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[16] G. Lample and A. Conneau, "Cross-lingual language model pretraining," *CoRR*, vol. abs/1901.07291, 2019. [Online]. Available: http://arxiv.org/abs/1901.07291

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014. [Online]. Available: http://arxiv.org/abs/1409.0575

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

[19] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[20] K. Q. Tran, A. T. Nguyen, A. T. Le, and K. V. Nguyen, "Vivqa: Vietnamese visual question answering," in *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, PACLIC 2021, Shanghai International Studies University, Shanghai, China, 5-7 November 2021*, K. Hu, J. Kim, C. Zong, and E. Chersoni, Eds. Association for Computational Lingustics, 2021, pp. 683–691. [Online]. Available: https://aclanthology.org/2021.paclic-1.72

[21] 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," *CoRR*, vol. abs/1612.00837, 2016. [Online]. Available: http://arxiv.org/abs/1612.00837

[22] Q. Wu, D. Teney, P. Wang, C. Shen, A. R. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *CoRR*, vol. abs/1607.05910, 2016. [Online]. Available: http://arxiv.org/abs/1607.05910

[23] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer *et al.*, "Pali: A jointly-scaled multilingual language-image model," *arXiv preprint arXiv:2209.06794*, 2022.

[24] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," *CoRR*, vol. abs/1712.04621, 2017. [Online]. Available: http://arxiv.org/abs/1712.04621

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[26] N. L.-T. 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*, 2023.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017. [Online]. Available: https://arxiv.org/abs/1711.05101

[28] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5583–5594.

Received on February 26, 2023
Accepted on October 06, 2023