Bi-directional Cross-Attention Network on Vietnamese Visual Question Answering

Duy-Minh Nguyen-Tran
Faculty of Information Technology,
University of Science, Vietnam
Vietnam National University,
Ho Chi Minh city, Vietnam
20c11041@student.hcmus.edu.vn

Tung Le
Faculty of Information Technology,
University of Science, Vietnam
Vietnam National University,
Ho Chi Minh city, Vietnam
lttung@fit.hcmus.edu.vn

Minh Le Nguyen
Japan Advanced Institute of Science
and Technology, Nomi, Ishikawa, Japan
nguyenml@jaist.ac.jp

Huy Tien Nguyen ✉
Faculty of Information Technology,
University of Science, Vietnam
Vietnam National University,
Ho Chi Minh city, Vietnam
ntienhuy@fit.hcmus.edu.vn

Corresponding author: Huy Tien Nguyen
ntienhuy@fit.hcmus.edu.vn

Abstract

Visual Question Answering (VQA) has arisen in recent public interest thanks to its applicability in many different fields. However, it requires understanding the combination of pictures and questions, which is highly challenging in both vision and language processing. Many previous works have achieved remarkable results to address this problem in many different languages. However, in the Vietnamese language, the VQA problem has not made significant progress due to the lack of data and fundamental systems. Therefore, we propose a model specifically designed and optimized for the Vietnamese Visual Question Answering problem. Our model leverages the strength of pre-trained models as well as presents Bi-directional Cross-attention architecture to learn visual and textual features more effectively. Through experimental results and ablation studies, the proposed approach obtains promising results against the existing models for Vietnamese on the ViVQA dataset.

1 Introduction

Together with the remarkable development in Computer Vision and Natural Language Processing, many problems requiring the combination of both images and languages were raised such as Image Captioning (Wang et al., 2022; Hu et al., 2021), Visual Question Answering (VQA) (Le et al., 2021a; Le et al., 2020), Visual Question Classification (Le. et al., 2022), etc. Among them, VQA is a potential area receiving a lot of attention in both research and industry. Furthermore, the recent approaches in VQA also achieved promising results. The goal of most VQA systems is to digest the content of a given image and answer the related questions. Compared to other tasks like Image Captioning and Visual Question Classification, VQA requires a deeper understanding of visual and textual inputs to give appropriate answers. In practice, VQA can be applied to various scenarios such as human-machine interaction, medical assistance, and automatic customer service or recommendation.

Undoubtedly, the VQA model plays an important role in the cutting edge of multi-modal approaches between images and languages. In recent years, many VQA approaches have been introduced and achieved promising results such as LXMERT (Tan and Bansal, 2019), SIMVL (Wang et al., 2021), OSCAR (Li et al., 2020). However, current VQA systems are almost addressed in English, Japanese, Chinese, and a few other languages. However, VQA for Vietnamese has not been strongly developed due to data limitations. Therefore, we introduce an optimized and fine-tuned VQA model specifically for the Vietnamese language. Besides, we propose bi-directional cross-attention architecture by adjusting the multi-head Attention in Transformer (Vaswani et al., 2017) to optimize learning image-question pairs with the Vietnamese language. To demonstrate the effectiveness of the method, our model is evaluated on the ViVQA dataset (Tran et al., 2021) which is built and adjusted based on the characteristics of the Vietnamese language. Through experiments, our model proves its efficiency and outperforms the competitive baselines.

The major contributions of this paper are concluded as follows: (i) We deploy to use the Vision - Text Transformer model to optimize the feature extraction of vision and language for Vietnamese. (ii) We propose a bi-directional cross-attention architecture by adjusting the attention structure of the transformer. Our proposed component is efficient to learn the combination and relationship between visual and linguistic features for Vietnamese. (iii) Through experiments and ablation studies, our proposed model achieves superior results compared to the existing approaches in the VQA dataset for the Vietnamese language.

2 Related Works

In the early stages of the Visual Question Answering problem, previous architectures are often built by two independent networks in CV and NLP to understand image and text features combined by the external components such as vector operations (Le et al., 2021b) and stacked attention (Le et al., 2021a). In particular, typical approaches in traditional VQA systems used Convolution Neural Networks (CNNs) for image embedding while a question was embedded by Recurrent Neural Networks (RNNs) (Goyal et al., 2017). Then, visual and textual features are combined by an attention mechanism.

In recent years, transfer learning has arisen the community's interest in many works to take advantage of huge datasets via self-supervised learning. There are more and more pre-trained models, which are the cornerstone for many areas such as BERT (Devlin et al., 2018), Vision Transformer (Dosovitskiy et al., 2021), and SpeechBERT (Chuang et al., 2020), etc. In recent VQA approaches, pre-trained models are also applied in many feature extraction modules to inherit the development of the CV and NLP domain. By using pre-trained models that have been trained on huge datasets, the feature extraction components are more efficient thereby improving the efficiency of the model.

Together with the necessity of feature extraction, the combination of RNNs and CNNs is considered one of the most important components in VQA systems. In understanding signals from images and texts, in recent years, with the development of deep learning, the Transformer model has been introduced and achieved many admirable results in both Natural Language Processing and Computer Vision fields. Transformer's Attention Architecture was applied to many VQA systems and achieved promising results. LXMERT (Tan and Bansal, 2019), one of the best models for the VQA problem, uses the self-attention and multi-head attachment architecture of Transformer to propose a cross-modality encoder architecture in combining visual and textual features. In addition, another famous model, SIMVLM (Wang et al., 2021), applied Transformer encoder architecture in feature extraction and achieved superior results against previous models in the VQA problem.

However, the Visual Question Answering problem for the Vietnamese language (ViVQA) developed at a modest rate due to the lack of resources. Among many previous works in ViVQA, Tran et al. (Tran et al., 2021) use PhoW2Vec (Nguyen and Nguyen, 2020) to address Vietnamese questions and Hierarchical Co-Attention (Lu et al., 2016) to combine visual and textual features. This model currently achieves the best performance in the ViVQA dataset. However, the performance of the model is still very limited against VQA systems in other languages. In addition, due to specific language characteristics, recent approaches in English also meet the difficulties as being applied to the Vietnamese dataset. This issue inspired us to build a VQA model specifically for Vietnamese. By using pre-trained models and the proposed bi-directional cross-attention architecture, our model improves the efficiency compared to previous models by deploying a feature understanding module compatible with the Vietnamese language.

3 Model Architecture

3.1 Visual and Textual Feature Extraction

Our model leverages the power of pre-trained models in the feature extraction of both images and texts. Using pre-trained models makes feature extraction more effective with prior knowledge from huge datasets in both CV and NLP domains. Details of our feature extraction modules are shown in Figure 1.

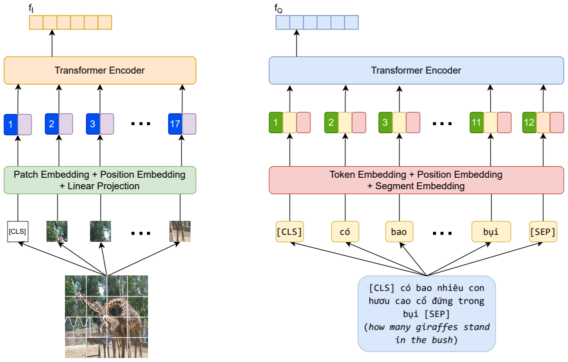
f_I
Transformer Encoder
1
2
3
...
17
Patch Embedding + Position Embedding + Linear Projection
[CLS]
...
f_Q
Transformer Encoder
1
2
3
...
11
12
Token Embedding + Position Embedding + Segment Embedding
[CLS]
có
bao
...
bụi
[SEP]
[CLS] có bao nhiêu con hươu cao cổ đứng trong bụi [SEP]
(*how many giraffes stand in the bush*)

Figure 1: Detail of Feature Extraction Modules: Images are embedded by the pre-trained Vision Transformer model (left) and Questions are embedded by the pre-trained PhoBERT model (right)

In particular, images are embedded by the pre-trained Vision Transformer model (Dosovitskiy et al., 2021). With the success of the Transformer model in the field of Natural Language Processing (NLP), Alexey Dosovitskiy et al. (Dosovitskiy et al., 2021) applied this architecture to Computer Vision and achieved better results than the state-of-the-art models in image classification. The Vision Transformer model divides an image into a sequence of regions called patches and treats them as words in a sentence. In our image embedding, a special patch $[CLS]$ is added to the first position of the visual patches to represent the aggregated information of an image. After being processed by Transformer architecture, visual features f_I are extracted from the $[CLS]$ representation vector.

For question embedding, a pre-trained PhoBERT model (Nguyen and Nguyen, 2020) is used to extract question features in our model. Although the pre-trained model BERT (Devlin et al., 2018) is considered the best performing model in many NLP tasks, it has not achieved good results when applied to Vietnamese datasets. Therefore, based on BERT, PhoBert is built and trained on the Vietnamese dataset to address the typical characteristics of this language. In particular, for extracting textual features, we put two special characters $[CLS]$ and $[SEP]$ in the first and last positions respectively. After going through pre-trained PhoBERT, we utilize the outputs of $[CLS]$ vector as question features f_Q.

3.2 Bi-directional Cross-Attention Network

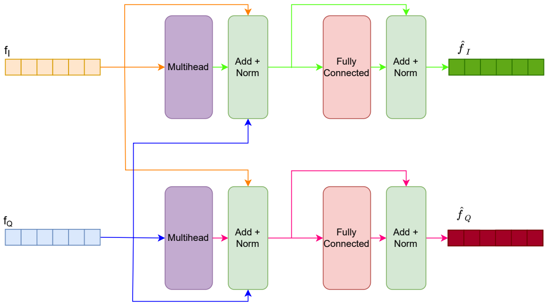

Figure 2: Bi-directional Cross-Attention

To understand visual and linguistic combinations and relationships, we propose a Bi-directional Cross-Attention architecture for simultaneous learning of textual and visual features. Bi-directional Cross-Attention is inspired by Transformer's Attention with some adjustments. While Transformer's Attention is only suitable for paying attention to individual features, our proposed architecture is capable of learning simultaneous and mutual relationships between images and languages. Our proposed component is visualized in Figure 2. Our architecture is divided into 2 sub-modules. In the first part, visual and textual features are computed through a separate multi-head layer. This phase is effective enough to mingle the visual and textual features together.

In particular, with each head h in a multi-head attention block, the attentive process is similar to self-attention(SA) calculated by Equation 1 and Equation 2.

$$f_R^h = SA(f_R, f_R, f_R) \tag{1}$$

$$SA(q, K, V) = softmax(\frac{qW^Q(KW^K)^T}{\sqrt{d}})VW^v \tag{2}$$

With a stack of many self-attention layers, the visual and textual features are activated by observing the in-context components and created more powerful representation in Equation 3 and 4.

$$f'_I = Multihead(f_I) \tag{3}$$

$$f'_Q = Multihead(f_Q) \tag{4}$$

Then, the image and language signals are combined with their original features and the other's original features by a vector operation. The final signal is normalized to produce intensified visual and textual features. Although this change seems a little small, its effect is highly valuable in performance. The whole process is calculated via Equation 5 and Equation 6:

$$f_{IQ} = Norm(G(f'_I, f_I, f_Q)) \tag{5}$$

$$f_{QI} = Norm(G(f'_Q, f_I, f_Q)) \tag{6}$$

Where $G(., ., .)$ is the feature combination operation. In our model, $G()$ function is chosen to achieve the best performance in Vietnamese VQA. Through our proposed combination, the features of each component are bi-directionally enhanced by the original information of the image and language. Therefore, our architecture can simultaneously and mutually learn textual and visual features.

In the second part, enhanced visual and textual features are handled by a fully connected and normalization layer similar to Transformer's Attention architecture. The final image and text signal is calculated by the Equation 7- 10.

$$f'_{IQ} = W^T f_{IQ} + b \tag{7}$$

$$f'_{QI} = W^T f_{QI} + b \tag{8}$$

$$\hat{f_I} = Norm(G(f'_{IQ}, f_{IQ})) \tag{9}$$

$$\hat{f_Q} = Norm(G(f'_{QI}, f_{QI})) \tag{10}$$

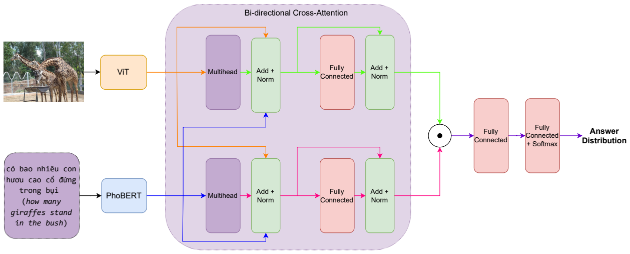
Bi-directional Cross-Attention
ViT
Multihead
Add + Norm
Fully Connected
Add + Norm
có bao nhiêu con hươu cao cổ đứng trong bụi (how many giraffes stand in the bush)
PhoBERT
Multihead
Add + Norm
Fully Connected
Add + Norm
Fully Connected
Fully Connected + Softmax
Answer Distribution

Figure 3: VQA Model System

3.3 Our model

After extracting and combining the features in the previous steps, we integrate the Bi-directional Cross-attention module into our completed system. The details of our models are presented in Figure 3. After visual and textual features are mutually augmented through the Bi-directional Cross-attention module, we only utilize simple vector operation to combine visual and textual features. Similar to most VQA approaches, we consider VQA as a classification task. So the answer distribution is calculated by the Softmax function via Equation 11.

$$y = Softmax(W_a^T.K(\hat{f_I}, \hat{f_Q}) + b_a) \tag{11}$$

Where $K(., .)$ is the feature fusion operation. In our model, we consider $K(., .)$ as the vector operations including addition, multiplication, and concatenation.

Corresponding to the classification approach, we use Cross-Entropy Loss as the loss function. The cross-Entropy function is calculated as follows:

$$L(y, p) = -(y \log(p) + (1 - y) \log(1 - p)) \tag{12}$$

$$p = Probability(y = 1) \tag{13}$$

4 Experiments

4.1 Dataset

To solve the problem of Visual Question Answering for Vietnamese, our model is evaluated on the ViVQA dataset. The ViVQA dataset is extracted and translated from the MS COCO dataset. Particularly, to avoid ambiguity when translating directly from other languages, the question-answer pairs are selected and adjusted based on the unique characteristics of the Vietnamese language. Therefore, the dataset is more accurate and realistic than using the translated data sets from English. Thus, this dataset is considered a benchmark in the VQA task for Vietnamese. The detail of the dataset is shown in Table 1.

	Train	Test
No. Samples	11999	3001
Longest Question Length	26	24
Longest Answer Length	4.0	4.0
Average Question Length	9.49	9.59
Average Answer Length	1.78	1.78

Table 1: Detail of ViVQA dataset

Component	Value
Vision Transformer	google/vit-base-patch16-224-in21k
PhoBERT	vinai/phobert-base
No. Cross-Attention Layer	1
No. Head	12
Dropout	0.1
Fully-connected layers	768 - 512 - 353
Optimizer	AdamW(lr = 3e-5, eps = 1e-8)

Table 2: Detail of Component Setting

4.2 Experimental Setting

In our model, we leverage the power of pre-trained models in extracting visual and textual features. Besides, we also propose a Bi-directional Cross-Attention architecture to effectively improve signal learning from images and languages. Details of the configurations of each module are presented in Table 2 to facilitate the future reproduction process. Due to limited infrastructures in the experimental environment, we use some default parameters for configuration. However, through experimental results, these parameters are effective enough to implement a compact and high-performance VQA system.

4.3 Evaluation

Similar to previous VQA works, we use an accuracy score for model evaluation. With the ViVQA dataset, each question only corresponds to one answer, so the accuracy score is calculated based on the number of questions correctly answered on the entire number of questions. Let N be the number of samples, the accuracy score is calculated via Equation 14.

$$Acc = \frac{1}{N}\sum_{i=1}^{N} 1\{\hat{y}_i == y_i\} \tag{14}$$

4.4 Result

Because the model built for the Visual Question Answering problem is specifically for Vietnamese, there are no published works to solve this problem. Therefore, we evaluate and compare our model with the 3 best models proposed by the author in the ViVQA dataset paper. All three models use Resnet (He et al., 2016) for visual representations and PhoW2Vec for textual features. First, we compare the model using Long Short Memory(LSTM) (Antol et al., 2015) for associative attribute learning between images and questions. Besides, we also compare the model with the model using Bidirectional Long Short Memory(Bi-LSTM) (Schuster and Paliwal, 1997) architecture for learning textual and visual features. Finally, the results of our model are compared with the model using the Hierarchical Co-Attention (Lu et al., 2016) . Details of the result are shown in Table 4.4. Obviously, our model obtains remarkable results against the existing approaches. First, our model outperforms others by using pre-trained transformers for both visual and textual representation. Second, the performance of bi-directional cross-attention architecture in concurrent learning between images and questions is more effective than LSTM, Bi-LSTM, or Hierarchical Co-Attention. This demonstrates the strength of the bi-directional cross-attention compared to previous models.

5 Discussion

To demonstrate the effectiveness of each component in our proposed model, we have analyzed them through some experiments.

First, we evaluate the strength of the Bi-directional Cross Attention module compared to Transformer Attention. Besides, we also clarify the

	Accuracy
Hierarchical Co-Attention + PhoW2Vec	34.96
LTSM + PhoW2Vec	33.85
Bi-LTSM + PhoW2Vec	33.97
Our model	**51.3**

Table 3: Detail our model results against other competitive baselines

contribution of our proposed attention and its modified variants. Details of the results are shown in Table 4. Obviously, our module outperforms the original Transformer Attention through its simultaneous and mutual learning between visual and textual features. Among the different configurations, our model achieves the most promising performs when $G()$ function in Equation 5, 6, 9, and 10 is equal to the addition operation. In most variants, our proposed attention also outperforms the system without Bi-directional Cross-Attention.

	Accuracy
No Bi-directional Cross-Attention	38.5
Transformer Attention	48.6
Bi-directional Cross-Attention **+ Equation 5, 6: G() = Add** **+ Equation 9, 10: G() = Add**	**51.3**
Bi-directional Cross-Attention **+ Equation 5, 6: G() = Multiply** **+ Equation 9, 10: G() = Add**	38.7
Bi-directional Cross-Attention **+ Equation 5, 6: G() = Add** **+ Equation 9, 10: G() = Multiply**	50.2
Bi-directional Cross-Attention **+ Equation 5, 6: G() = Multiply** **+ Equation 9, 10: G() = Multiply**	35.8

Table 4: The effect of bi-directional cross-attention in our architecture

Besides, we also compare the effect of vector operations: multiplication, concatenation, and addition in combining image and question features (Equation11). The choice of multi-model fusion function makes an important contribution to improving system efficiency. Despite the simplicity of vector operation, its effect is impressive in the whole system. The results of this comparison are shown in Table 5. Obviously, the multiplication operation gives the most promising results than other operations in our model.

Operation	**Accuracy**
Equation 11: K() = Add	46.9
Equation 11: K() = Concatenate	37.8
Equation 11: K() = Multiply	**51.3**

Table 5: The effect of fusion operation in our architecture

Finally, we conduct experiments in selecting the number of Bi-directional Cross-Attention blocks. In our system, this module plays a critical role in the model's performance. The effect of the Bi-directional Cross-Attention block is visualized in Figure 4. The change in the model's accuracy reflects the dependency of our system on the Bi-directional Cross-Attention Block. In this experiment, our model performed best with one Bi-directional Cross-Attention block. Because of the small number of samples, It is less effective to optimize the stacked structure of attention blocks.

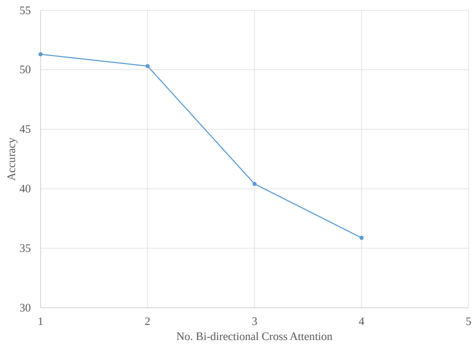

Figure 4: The effect of the number of Bi-directional Cross-Attention in our model

6 Conclusion

In this work, we propose a Bi-directional Cross-Attention Network that was fine-tuned specifically for the Vietnamese language. Our model leverages the power of pre-trained models in both Vision and Language to optimize the model's feature extraction. In particular, we adjusted the architecture of Transformer's Attention in the fusion of visual and textual features. This attention allows the image and question features to be learned simultaneously in the Transformer block. Through experiments and ablation studies, our model is proved to be more effective than other competitive baselines in Visual Question Answering for the Vietnamese language. Besides the contributions in Visual Question Answering, we also consider the extension in scientific documents where the images and figures are surrounded by many contexts in the near future works.

Acknowledgment

This work is funded by the University of Science, VNU-HCM under grant number CNTT 2022-05.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Yung-Sung Chuang, Chi-Liang Liu, Hung yi Lee, and Lin-Shan Lee. 2020. Speechbert: An audio-and-text jointly learned language model for end-to-end spoken question answering. In Helen Meng, Bo Xu 0011, and Thomas Fang Zheng, editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 4168–4172. ISCA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. Scaling up vision-language pre-training for image captioning. *arXiv preprint arXiv:2111.12233*.

Tung Le, Nguyen Tien Huy, and Nguyen Le Minh. 2020. Integrating transformer into global and residual image feature extractor in visual question answering for blind people. In *2020 12th International Conference on Knowledge and Systems Engineering (KSE)*, pages 31–36.

Tung Le, Huy Tien Nguyen, and Minh Le Nguyen. 2021a. Multi visual and textual embedding on visual question answering for blind people. *Neurocomputing*, 465:451–464.

Tung Le, Huy Tien Nguyen, and Minh Le Nguyen. 2021b. Vision and text transformer for predicting answerability on visual question answering. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 934–938.

Tung Le., Khoa Pho., Thong Bui., Huy Tien Nguyen., and Minh Le Nguyen. 2022. Object-less vision-language model on visual question classification for blind people. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART,*, pages 180–187. INSTICC, SciTePress.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Khanh Quoc Tran, An Trong Nguyen, An Tran-Hoai Le, and Kiet Van Nguyen. 2021. Vivqa: Vietnamese visual question answering. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 546–554.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*.