INTEGRATING IMAGE FEATURES WITH CONVOLUTIONAL SEQUENCE-TO-SEQUENCE NETWORK FOR MULTILINGUAL VISUAL QUESTION ANSWERING

TRIET M. THAI, SON T. LUU*

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

Abstract. Visual question answering is a task that requires computers to give correct answers for the input questions based on the images. This task can be solved by humans with ease, but it is a challenge for computers. The VLSP2022-EVJVQA shared task carries the Visual question answering task in the multilingual domain on a newly released dataset UIT-EVJVQA, in which the questions and answers are written in three different languages: English, Vietnamese, and Japanese. We approached the challenge as a sequence-to-sequence learning task, in which we integrated hints from pre-trained state-of-the-art VQA models and image features with a convolutional sequence-to-sequence network to generate the desired answers. Our results obtained up to 0.3442 by F1 score on the public test set and 0.4210 on the private test set.

Keywords. Visual question answering, sequence-to-sequence learning, multilingual, multimodal.

Abbreviations

QA	Question answering
VLSP	Association for Vietnamese language and speech processing
Seq2Seq	Sequence-to-sequence
ViT	Vision transformer
SOTA	State-of-the-art
GRU	Gated recurrent unit
GLU	Gated linear unit
LSTM	Long short-term memory
RNN	Recurrent neural network
API	Application programming interface
ConvS2S	Convolutional sequence-to-sequence network
Bi-RNN	Bi-directional recurrent neural networks
BERT	Bidirectional encoder representations from transformers

1. INTRODUCTION

Visual question answering (VQA) is a trending research topic in artificial intelligence that combines natural language processing and computer fields. This task enables computers to

---

*Corresponding author.

E-mail addresses: 19522397@gm.uit.edu.vn (T.M.Thai); sonlt@uit.edu.vn (S.T. Luu).

extract meaningful information from images and answer the question by natural language text. The VQA task has various practical applications such as chat-bot systems, intelligent assistance, and recommendation system.

The VQA can be categorized as a question-answering (QA) task. In the QA task, cross-lingual language QA has been a hot trend in recent years with the appearance of BERT [5] (trained on more than 100 languages), as well as plenty of multilingual datasets [24]. The VLSP-EVJVQA challenge [21] takes the VQA as a multilingual QA task, containing three different languages: Vietnamese, Japanese, and English. The challenge brings the first large-scale multilingual VQA dataset, UIT-EVJVQA, with approximately 5,000 images and more than 30,000 question-answer pairs. The task takes an image and a question with text form as input, and the computer must return the correct answer by text as output. The questions are written in Vietnamese, Japanese, or English, and the answers must follow the language used in the questions. To create the correct answer, the computer must understand the question content and extract the information from the corresponding image. Fig.1 illustrates several examples from the dataset provided by the organizer. According to the answer types, the VLSP-EVJVQA is a Free Form Answer QA task [7].

To solve the VLSP-EVJVQA, we propose our solution, which combines the Sequence-to-Sequence (Seq2Seq) learning with image features extraction task to generate the correct answers. We use the ViLT [14] and OFA [31] for hints extraction from the image, Vision Transformer [6] for image features extraction and the Convolutional network for Seq2Seq learning [9] to generate the answer. Our results achieve the 3rd place in the VLSP-EVJVQA. Hence, we describe our works in this paper. The paper is structured as follows. Section 1 introduces the task. Section 2 takes a brief survey about previous works for the VQA task. Section 3 overviews the VLSP-EVJVQA dataset. Section 4 describes our proposed solution. Section 5 devotes to experiments and performance analysis. Finally, Section 6 concludes our works and presents future studies.

2.   RELATED WORKS

2.1.   Existing datasets and methods for visual question answering

In computer vision, the research purpose for VQA is to make computers understand the semantic context of images. The Microsoft COCO dataset [17] is one of the large-scale datasets that impact many studies in computer vision tasks, including object detection, image classification, image captioning, and visual question answering. Several VQA datasets are built on the MS-COCO in different languages, such as the VQA [1], VQAv2 [10] in English, FM-IQA [8] for Chinese, the Japanese VQA [26] for Japanese, and the ViVQA [29] for Vietnamese. There are also other two benchmark datasets for training and fine-tuning VQA methods, including Visual Genome (VG-QA) [15] and GQA [12]. VG-QA is a VQA dataset that contains real-world photographs. It is designed and constructed to emphasize the interactions and relationship between natural questions and particular regions on the images. The creation of VG-QA lays the groundwork for building GQA, another large VQA collection that make use of Visual Genome scene graph structures to feature compositional question answering and real world reasoning. Besides, in the natural language processing field, the SQuAD dataset [22] has boosted many studies in question-answering and natural language understanding. Based on SQuAD, many corpora are created in different languages
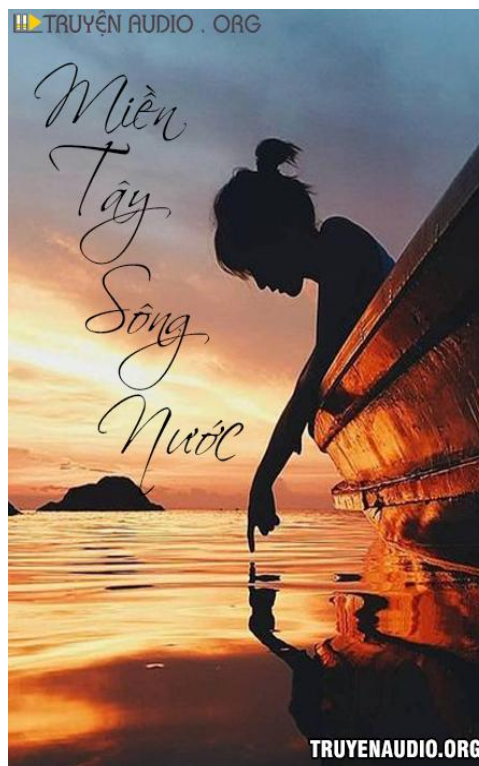
Q: how many people are using their phones to take pictures on the boat?
A: just one

Q: người đàn ông mặc áo xanh lá đang làm gì?
A: đang quét dọn

Q: 女の子は水に何の手を入れていますか?
A: 少女は左手を水の中に入れます

Figure 1: Multilingual samples from UIT-EVJVQA dataset. From top to bottom, left to right: English (en), Vietnamese (vi) and Japanese (ja). The dataset contains a wide variety of questions; in some cases, the image contains noises that makes it difficult for a computer to distinguish the indicated object or action, for instance, "phones" in the English example or the action of "the man in green shirt" in the Vietnamese case. Besides, the Japanese example provides a tough scenario, "which hand is the girl putting onto the water?" in English, that even humans find it challenging to deliver the proper response.

like DuReader [11] for Chinese, JaQuAD [28] for Japanese, KorQuAD [16] for Korean, and ViQuAD [13, 19] for Vietnamese.

Apart from creating high-quality datasets, the architecture also plays a vital role in constructing intelligence systems. Taking advantage of natural language processing, we have several robust models for sequence-to-sequence learning tasks such as Long-short Term Memory [3], Convolutional Neural Networks for Sequence-to-sequence [9], Transformer [30] and BERTology [25]. In computer vision, state-of-the-art models for extracting useful information from images includes YOLO [23], VGG [27], Vision Transformer (ViT) [6]. With the increasing diversity of data and the need to solve multi-modal tasks that involve both visual and textual features, recent research trends focus on developing models that combine both the vision and language modalities such as: Vision-and-Language Transformer (ViLT) [14], or OFA [31].

2.2. Vision-language models

2.2.1. Vision-and-language transformer (ViLT)

Introduced at ICML 2021, Vision-and-Language Transformer or ViLT [14] is one of the first and considerably, the simplest architectures that unifies visual and textual modalities. The model takes advantage of transformer module to extract and process visual features without using any region features or convolutional visual embedding components, making it inherently efficient in terms of runtime and parameters. The architecture of the ViLT model is originally set up to approach the VQA problem in the direction of a classification task. Thus, the output of the model contains various keyword answers with respect to probabilities.

2.2.2. The "one for all" architecture that unifies modalities (OFA)

OFA [31] is a unified sequence-to-sequence pre-trained model that can generate natural answers for visual question answering task. The architecture uses Transformer [30] as the backbone with the Encoder-Decoder framework. The model is pre-trained on the publicly available datasets of 20M image-text pairs and achieves state-of-the-art performances in a series of vision and language downstream tasks, including image captioning, visual question answering, visual entailment, and referring expression comprehension, making it a promising component in our approach toward the VQA challenge.

2.3. Convolutional sequence-to-sequence network

Sequence-to-sequence learning (Seq2Seq) is a process of training models to map sequences from one domain to sequences in another domain. Some common Seq2Seq applications include machine translation, text summarization and free-form question answering, in which the system can generate a natural language answer given a natural language question. A trivial case of Seq2Seq, where the input and output sequences are in the same length, can be solved using a single Long Short Term Memory (LSTM) or Gated Recurrent Unit (GRU) layer. In a canonical Seq2Seq problem, however, the input and output sequences are of different lengths, and the entire input sequence is required to begin predicting the target. This requires a more advanced setup, in which RNN-based encoder-decoder architectures commonly used to address the problem. The typical and generic architecture of these type of models include two components: an encoder that processes input sequence $X = (x_1, x_2, ..., x_m)$ and return state representation $z = (z_1, z_2, ..., z_m)$, also known as context vector, and a decoder that decodes the context vector and outputs the target sequence $y = (y_1, y_2, ..., y_n)$ by generating it left-to-right consecutively, one word at a time.

Different from other Seq2Seq models such as bi-directional recurrent neural networks (Bi-RNN) with soft-attention mechanism [2, 18] or the mighty Transformer [30] with self-attention, the convolutional sequence-to-sequence network (ConvS2S) has the Encoder - Decoder architecture based entirely on convolutional neural networks and originally set up for machine translation task. The model employs many convolutional layers, which are commonly used in image processing, to enable parallelization over each element in a sequence during training and thus better utilize GPU hardware and optimization compared to recurrent networks. ConvS2S applies a special activation function called the Gate Linear Unit (GLU) [4] as non-linearity based gating mechanism over the output of the convolution layer,

which has been shown to perform better in the context of language modeling. Multi-step attention is also the key component of the architecture that allows the model to make multiple glimpses across the sequence to produce better output.

3. THE DATASET

The dataset released for the VLSP-EVJVQA challenge, UIT-EVJVQA [21], is the first multilingual Visual Question Answering dataset with three languages: English (en), Vietnamese (vi), and Japanese (ja). It comprises over 33,000 question-answer pairs manually annotated on approximately 5,000 images taken in Vietnam, with the answer created from the input question and the corresponding image. Besides various types of questions, the answers are constructed in a free-form structure, making it a challenge for VQA systems. To perform effectively and achieve good results on UIT-EVJVQA, the typical VQA systems must identify and predict correct answers in free-form format for multilingual questions, due to the dataset characteristics.

Table 1: Statistical information about the UIT-EVJVQA dataset

	Training set			Public test set		
	English	Vietnamese	fJapanese	English	Vietnamese	Japanese
Number of samples	7,193	8,320	8,261	1,686	1,678	1,651
Questions						
Vocabulary size	2,089	1,860	3,035	1,080	919	1,226
Average Length	8.52	8.70	13.03	8.76	8.87	13.27
Max Length	24	21	45	26	22	33
Min Length	3	3	4	3	4	4
Answers						
Vocabulary size	2,307	2,067	3,534	1,029	877	1,176
Average Length	5.09	6.04	7.27	3.89	4.54	5.05
Max Length	23	23	30	19	18	21
Min Length	1	1	1	1	1	1

The training set and public test set have total samples of 23,774 and 5,015, respectively. Table 1 describes the statistical information about the UIT-EVJVQA dataset on the training and public test sets. The length of sentences is computed at the word level. We use Underthesea[*] and Trankit [20] library for word segmentation. Generally, the distribution of the training and public test sets is quite similar. English has fewer training samples compared with Vietnamese and Japanese, which may affect the question answering performance on this language. The length of questions is longer than the length of answers in three languages. The questions in Japanese are significantly longer than those in the two remaining languages. For the answers, those in Japanese are still also longer than those in English and Vietnamese. However, the difference in the length is not as much as the questions. Particularly, the shortest answers in three languages have only one word. In contrast, the questions and answers in Vietnamese have fewer words than those in English and Japanese.

---

[*]https://github.com/undertheseanlp/underthesea

4.   THE PROPOSED METHOD

Figure 2 depicts an overview of the proposed approach in this study. In general, we transform the VQA problem into a sequence-to-sequence learning task, in which we take advantage of State-of-the-art (SOTA) vision-language models to offer richer information about the question-image dependencies in the input sequence. The method consists of two main phases that are carried out sequentially. At the first stage, numerous hints are extracted from question-image pairs using pre-trained vision-language models. The extracted hints are then concatenated with the question and visual features to form a sequence representation as input to the proposed Seq2Seq model to generate the corresponding answers in free-form natural language. Our source code and demo for the proposed methodology are available at this link: https://huggingface.co/spaces/daeron/CONVS2S-EVJVQA-DEMO
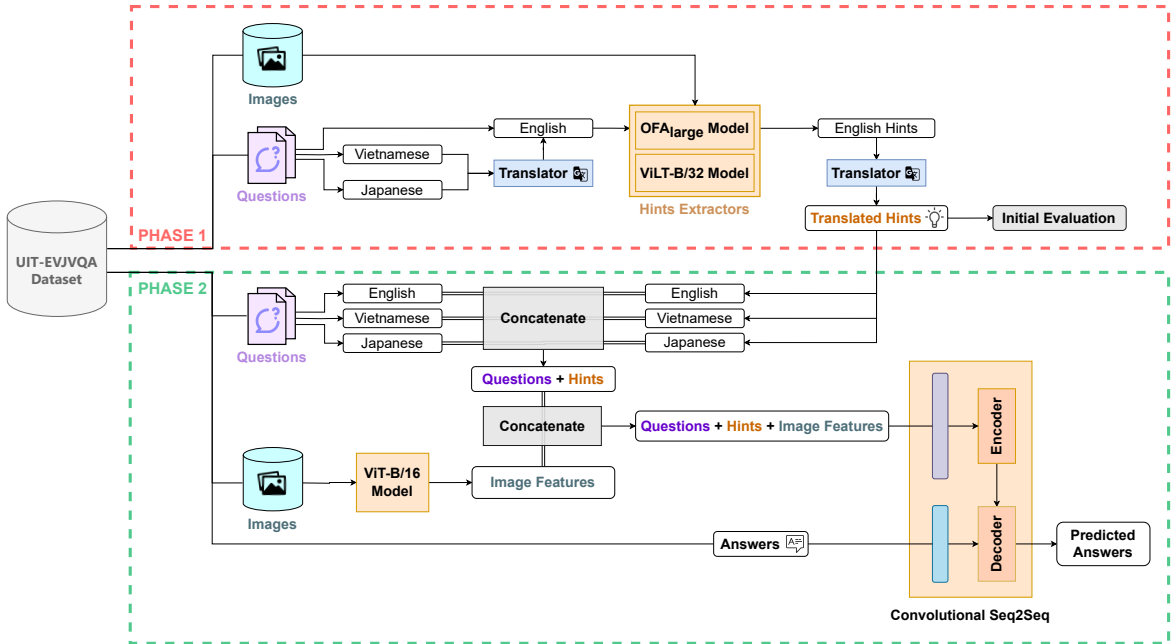
Figure 2: An overview of the proposed method for visual question answering on UIT-EVJVQA dataset.

4.1.   Hints extraction with pre-trained vision-language models

This phase concentrates on implementing SOTA vision-language models, including OFA [31] and ViLT [14] to predict the possible answers given a question and its corresponding image. Due to the diverse nature of the questions and the multilingual aspect of UIT-EVJVQA, these models are only set up to provide answers directly through zero-shot prediction, with no training or fine-tuning step on the dataset. These SOTA models, which were pre-trained and fine-tuned on various datasets (VQAv2, VG-QA, and GQA), mainly support English but do not yet support Vietnamese or Japanese. To achieve desired results, we translate the Vietnamese and Japanese questions from UIT-EVJVQA into English using Google Trans-

Table 2: Performance of SOTA vision-language models on public test set.

Model	# hints	F1
ViLT	1	0.1303
	2	0.1317
	3	0.1315
	4	0.1290
	5	0.1252
OFA	-	0.1902

late API[†] before feeding them into the models to get inferences. Once the output answers are generated, they are translated back into the original languages for evaluation and experiments in the second phase. For ViLT, we choose up to five candidate answers with the highest probability for further experiments. Using more hints is feasible, but it will put more pressure on computational resources as we approach creating long sequences based on hints probability in the next phase. We concatenate each output answer from ViLT along the sequence, respectively with decreasing relevance, to assess their quality on the new dataset. The inference performance of pre-trained ViLT and OFA models on the public test set are shown in Table 2.

Under our expectations, the OFA model with unified Seq2Seq structure outperforms ViLT with F1 0.1902, while ViLT achieves the best performance with F1 0.1317 using 2 keyword answers. The evaluation results are not quite good compared to the ground truth because of the special characteristic of the dataset with long answers and, since no training has been done, the predicted answers lack sufficient vocabulary. The predicted answer may match or not match the ground truth completely but gives a similar and proper response to the question. Regardless of accuracy, these simple keyword answers provide valuable insights about question-image interactions. Due to this, we consider these answers as hints or suggestions for each question-image pair and apply them to the training of the main model in the following phase.

4.2. Experiment with convolutional sequence-to-sequence network

The second phase of the approach concentrates on developing and training the main model for this challenge: the convolutional sequence-to-sequence network (ConvS2S) [9] with different combinations of textual and image features for visual question answering task. ConvS2S has significant capabilities to accelerate training progress and reduce our computational resource limitations due to its efficiency in terms of GPU hardware optimization and parallel computation. This is why the architecture is preferred over other Seq2Seq models for the competition.

In this study, each convolutional layer of ConvS2S uses many filters with a width of 3. Each filter will slide across the sequence, from beginning to end, looking at all 3 consecutive

---

[†]https://cloud.google.com/translate

elements at a time to learn to extract a different feature from the questions, hints, visual factors and answers. With these special settings, the model has a significant capacity to extract meaningful features from the input sequence and generate free-form content. Due to its proven performance in other Seq2Seq learning tasks, such as machine translation, we anticipate the model to perform well on question-image features combination and produce good results on the visual question-answering task.

4.3. Textual and visual features combination

In the early stage, a set of various useful hints is achieved using pre-trained ViLT and OFA. In order to train the proposed Seq2Seq model with the existing materials, the textual features, including questions and hints, and image features have to be combined in the form of sequence representations as input to the Seq2Seq model.

As shown in Table 2, adding more ViLT hints to the sequence tend to reduce the F1 score performance. These simple answers, on the other hand, may passively contribute to the overall understanding of the scenario of the corresponding images. Therefore, our approach focuses on using hint probability to generate sequences with repeated keywords while avoiding noise from outliers. This method allows hints that have a higher probability to appear more frequently in the sequence. For efficiency and reducing cost, the number of times a ViLT hint occurs in the sequence is the integral part of its half probability. For experiments involving the output of two models, the hint from the OFA model is set to appear 10 times in the sequence. The newly created sequence is concatenated with the question to form the final sequence for question and hint. We then remove special characters, lowercasing, and tokenize the text contents before passing them into the encoder. For English content, we tokenize the text simply by splitting them word-by-word. For Vietnamese and Japanese content, Underthesea toolkit and Trankit [20] libraries are applied for word segmentation, respectively. Figure 3 illustrates an example of a question and hints combination in our approach.

Question: where is this girl putting her hands on?
OFA: a tree branch
ViLT: {'tree': 0.2237, 'nothing': 0.1213, 'trees': 0.0920, 'face': 0.0292, 'hat': 0.0277}

Combined sequence: <sos>where is this girl putting her hands on<sep>a tree branch a tree branch a tree branch a tree branch a tree branch a tree branch a tree branch a tree branch a tree branch a tree branch a tree branch<sep>tree tree tree tree tree tree tree tree tree tree tree nothing nothing nothing nothing nothing nothing trees trees trees trees face hat<eos>
Groundtruth answer: on a tree branch

Figure 3: An example of question and hints combination. The hint 'tree' occurs 11 times in the sequence since the half of its probality is 11.19 (%).

Besides the hints from question-image pairs, we also apply the Vision Transformer (ViT) [6] to extract visual features from the image. The input image is passed into ViT model to obtain a sequence of patches called the patch embeddings, which then pass through a Transformer encoder with multi-head attention to output the image features with the size of

196 x 768. Once the image features are achieved, we remove the vector at [CLS] token position and concatenate these visual features with text embeddings along the sequence dimension to have the final representative embedding matrices for questions, hints, and images.

5. EXPERIMENTS AND ANALYSIS

5.1. Experiment settings

The ConvS2S model has 512 hidden units for both encoders and decoders. All embeddings, including the output produced by the decoder before the final linear layer, have a dimensionality of 768. This setup allows the encoders to concatenate with patch embeddings from ViT model. To avoid overfitting, dropout is applied on the embeddings, decoder output, and the input of the convolutional blocks with a retaining probability of 0.5. Teacher forcing with probability of 0.5 in the is also applied in the architecture to accelerate the training progress.

Many experiments are carried out in order to evaluate the proposed approach toward the VLSP-EVJVQA challenge. Typically, the training and evaluation of ConvS2S model is conducted using four types of input sequence: question only, question-image, question-hint and question-hint-image. First, we initialize the baseline result of ConvS2S with only question as input sequence and no image information. This scenario is similar to the Knowledge-based question answering (KBQA) task in that the generated answers are entirely based on the question-answer dependencies learned during the training phase. The second experiment involved image features combined with question as typical VQA approach. We then add visual hints to the input sequences used in the two prior experiments and investigate their effect on overall performance.

Because of the limitation in computational resources as well as the strict timeline of the competition, we only deploy the fine-tuned ViLT-B/32 with 200K pretraining steps and pre-trained OFA_{large} with 472M parameters for hints inference given the question and image. For feature extraction from image, we deploy pre-trained ViT-B/16 with base-sized version. To have the comparative result, we set up the same hyperparameters for all experiments with ConvS2S. The model is trained in 30 epochs with batch size of 128 using Adam optimizer with a fixed learning rate of 2.50e-4. After each epoch, the performance loss on the train and development sets is calculated using the Cross-Entropy Loss function.

The proposed architecture and SOTA vision and language models are implemented in PyTorch and trained on the Kaggle platform with hardware specifications: Intel(R) Xeon(R) CPU @ 2.00GHz; GPU Tesla P100 16 GB with CUDA 11.4.

5.2. Experimental results

The two metrics: F1 and BLEU, are used in the challenge to evaluate the results. The BLEU score is the average of BLEU-1, BLEU-2, BLEU-3, and BLEU-4. F1 is used for ranking the final results. Table 3 presents the performance of the proposed ConvS2S model with different combinations of pre-trained models on the UIT-EVJVQA public test set.

According to Table 3, the original ConvS2S model using only question obtained 0.3005 by F1 and 0.1932 by BLEU. Using question-image pairs, ConvS2S achieves a marginally better performance on both metric. When visual hints are integrated into questions, the

Table 3: Performance of ConvS2S with different combinations of pre-trained models on the public test set.

Model	F1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU
ConvS2S (Question only)	0.3005	0.2592	0.2034	0.1677	0.1425	0.1932
ConvS2S + ViT	0.3109	0.2683	0.2119	0.1747	0.1480	0.2007
ConvS2S + ViLT	0.3294	0.2692	0.2109	0.1723	0.1446	0.1993
ConvS2S + OFA	0.3331	0.2858	0.2269	0.1876	0.1598	0.2150
ConvS2S + ViLT + OFA	**0.3442**	0.2797	0.2205	0.1808	0.1529	0.2085
ConvS2S + ViT + ViLT	0.3361	0.2833	0.2243	0.1845	0.1564	0.2122
ConvS2S + ViT + OFA	0.3390	0.2877	0.2276	0.1877	0.1593	**0.2156**
ConvS2S + ViT + ViLT + OFA	**0.3442**	0.2747	0.2148	0.1747	0.1465	0.2027
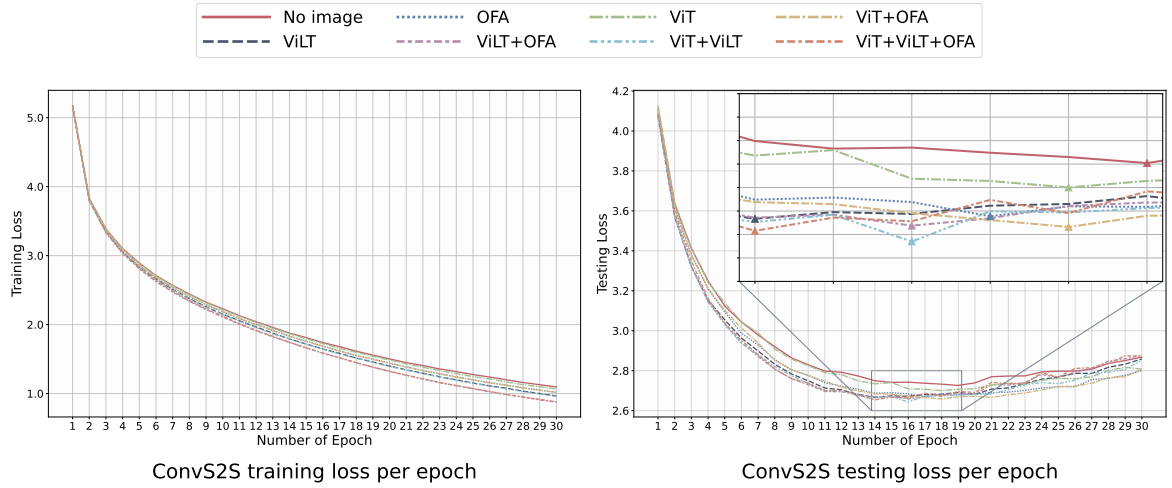
Figure 4: Training loss and public testing loss comparison of ConvS2S model with different combinations of hint and image features.

F1 score improves by at least 2.89%, and the model achieves the best performance with 0.3442 by F1 and 0.2085 by BLEU when both ViLT and OFA hints are used. At final stage, adding image feature from ViT to question-hint sequences help improve the performance of previous models. Based on F1, these two combinations ConvS2S + ViLT + OFA and ConvS2S + ViT + ViLT + OFA are considered as our best methods on the public test set. Figure 4 depicts the gradual improvement in both training loss and testing loss as more image features and hints are added to the ConvS2S model. Knowledge-based ConvS2S (red line) does not catch the image context and thus have the highest loss. Though ConvS2S with ViT+VILT features does not obtained a competitive result on evaluation metrics, it gives the best loss among methods in the public test phase. In general, the optimal testing loss of methods is achieved between 14th and 20th epoch, then the models tend to be overfitting.

We manage to deploy two ensembles of ConvS2S using features from ViT combined with hints from ViLT and OFA, respectively, for the final evaluation on private test set. As shown

in Table 4, the ConvS2S + ViT + OFA model obtained the better result, which is 0.4210 by F1 and 0.3482 by BLEU, and ranked 3rd in the challenge. Table 5 shows the final standing at the EVLSP-EVJVQA competition, in which our best model perform poorer 1.82% and 1.39% by F1 compared with the first and second place solutions. In terms of methodology, our approach comes in second place after the ViT + mT5 method, which has a large amount of pre-trained data. Overall, there is a gap between F1 and BLEU scores.

Table 4: Performance on the private test set.

Model	F1	BLEU
ConvS2S + ViT + ViLT	0.4053	0.3228
ConvS2S + ViT + OFA	**0.4210**	**0.3482**

Table 5: Our performance compared with other teams at VLSP2022-EVJVQA [21].

No.	Team name	Models	Public Test		Private Test	
			F1	BLEU	F1	BLEU
1	CIST AI	ViT + mT5	0.3491	0.2508	0.4392	0.4009
2	OhYeah	ViT + mT5	0.5755	0.4866	0.4349	0.3868
3	**DS_STBFL**	**ConvS2S+ViT+OFA**	**0.3390**	**0.2156**	**0.4210**	**0.3482**
4	FCoin	ViT + mBERT	0.3355	0.2437	0.4103	0.3549
5	VL-UIT	BEiT + CLIP + Detectron-2 + mBERT + BM25 + FastText	0.3053	0.1878	0.3663	0.2743
6	BDboi	ViT + BEiT + SwinTransformer + CLIP + OFA + BLIP	0.3023	0.2183	0.3164	0.2649
7	UIT_squad	VinVL+mBERT	0.3224	0.2238	0.3024	0.1667
8	VC_Internship	ResNet-152 + OFA	0.3017	0.1639	0.3007	0.1337
9	Baseline	ViT + mBERT	0.2924	0.2183	0.3346	0.2275

5.3. Performance analysis

According to the final result in the private test phase, the generated output from ConvS2S +ViT+OFA model are chosen for further analysis. Generally, the model manages to generate answers with correct language with the input question.

5.3.1. Quantitative analysis

We randomly choose 100 samples from the generated result to perform quantitative analysis. The average length, vocabulary size, and the number of POS tags in the ground truth and generated answers are calculated for each language. Table 6 shows the statistics of the ground truth answer compared with the predicted answer by the model.

Table 6: The quantitative statistic of 100 generated samples compared with the ground truth

Language	Stats.	Ground Truth	Predicted
English	Avg.length	3.74	6.18
	Vocab. size	78	72
	# POS tag	12	9
Vietnamese	Avg.length	4.42	5.97
	Vocab. size	97	101
	# POS tag	10	9
Japanese	Avg.length	4.67	8.43
	Vocab. size	77	83
	# POS tag	10	11
All	Avg.length	4.26	6.78
	Vocab. size	252	256
	# POS tag	14	14

From Table 6, it can be seen that although the model gave the answers longer than the ground truth answers, the semantics is not as much as the ground truth. It can be seen from Table 6 that the predicted answers in English have an average length higher than the ground truth answers. Also, the vocabulary in the generated answers is more than the original. In contrast, the number of POS tag components in the predicted answers is lower than the ground truth. This is similar to the answers in Vietnamese. For the Japanese, the characteristics of the predicted answers in average length and vocabulary size are the same as the two remaining languages. However, the number of POS tags in the predicted answers is more than in the ground truth answers. To make it clear, we propose three types of error on our model in Subsection 5.3.2.

In addition, Figure 5 illustrates the distributions of F1 and BLEU scores for each language. Generally, the histograms skewed to the right and the model performs inconsistently across languages. The proportion of samples with F1 and BLEU scores less than 0.2 dominates the overall result across all three languages. In Vietnamese, the number of generated samples with F1 and BLEU scores greater than 0.4 is significantly higher than in other languages. Meanwhile, English and Japanese responses rarely score greater than 0.6 on both metrics, furthermore, no Japanese samples scoring greater than 0.8 in BLEU. This illustrates that our model faces numerous challenges in producing the desired responses, with specific limitations on each language.

5.3.2. Qualitative analysis

Attention visualization. Figure 6 shows several samples of attention weights between each element from the generated answer with those in the input sequence that contains no image features, OFA hints, and ViT+OFA combined features, respectively. The visualization provided an intuitive way to discover which positions in the input sequence were considered more important when generating the target answer word. The brighter a pixel's color, the
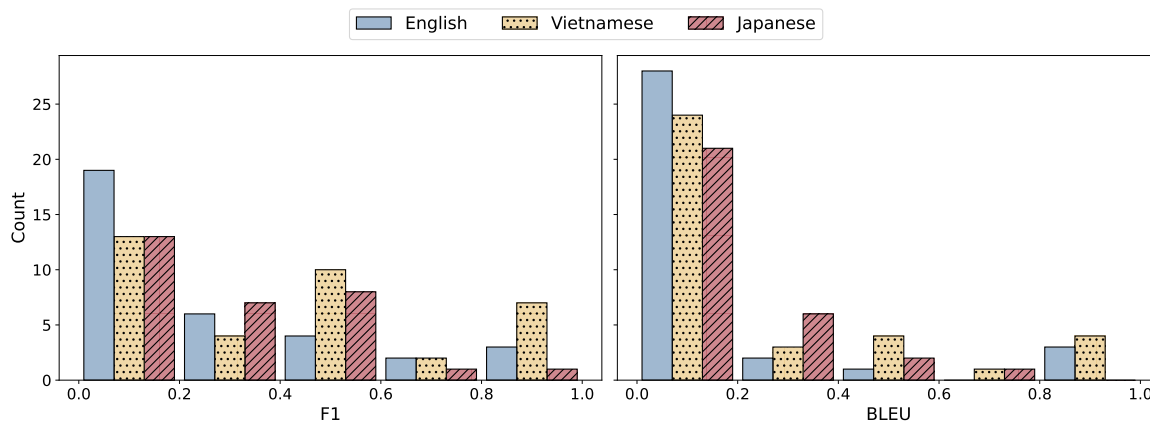
Figure 5: Distributions of F1 and BLEU scores for each language from 100 generated samples.

more important the word in the input sequence is in producing the respect answer word. The first heatmaps illustrate the case where no image information is used during training but only question. This is similar to Knowledged-based QA task where the model gives the answer solely based on the context of question. As a result, the generated answer is just simply a guess from ConvS2S model and has poor evaluation results on both metrics. Through attention visualization, we study that OFA hint is importance feature to model's attention as it provides the near-correct insight for the question and reduces the reliance on question words when generating the answer. This reduction in attention is not completely common for all question tokens, and it still depends on the importance of other elements in the whole sequence. However, in some cases, the model focuses too much on a specific element of the hint, which may lead to bias. ViT features has shown to control the affection of OFA hint, neutralizing it with other elements from question if hint appears to be off-topic. It may enhance the attention, making the model focus stronger on specific parts of the provided hint, for instance, the hint token "nhà hàng" (*restaurant*) in Figure 6c is given more attention when adding ViT image features. These features can also reduce the attention in one element and distributes concentration on other parts of the sequence. Figures 6a and 6b depict the reduction in hint attention into question elements, while Figures 6d and 6e show the changes in attention weight distribution among hint tokens.

Error analysis. For better understand the generation performance on the VQA task, we examine the generated answers of our best ensemble, ConvS2S +ViT+OFA, to identify the limitations and analyze factors that may cause the model to perform poorly. Through the error analysis process, various errors and mistakes have been pointed out in the outputs of the model. The typical examples of various types of errors are illustrated in Figure 7. In summary, we divide these errors into three groups:

- The generated answer does not match the question and has no correct tokens compared with the ground truth answer, as shown in Figure 7a. This error case sometimes accompanied by text degeneration.

- The output response gives the wrong answer to the question but shares some insignificant tokens or has a similar structure with the ground truth answer, as shown in
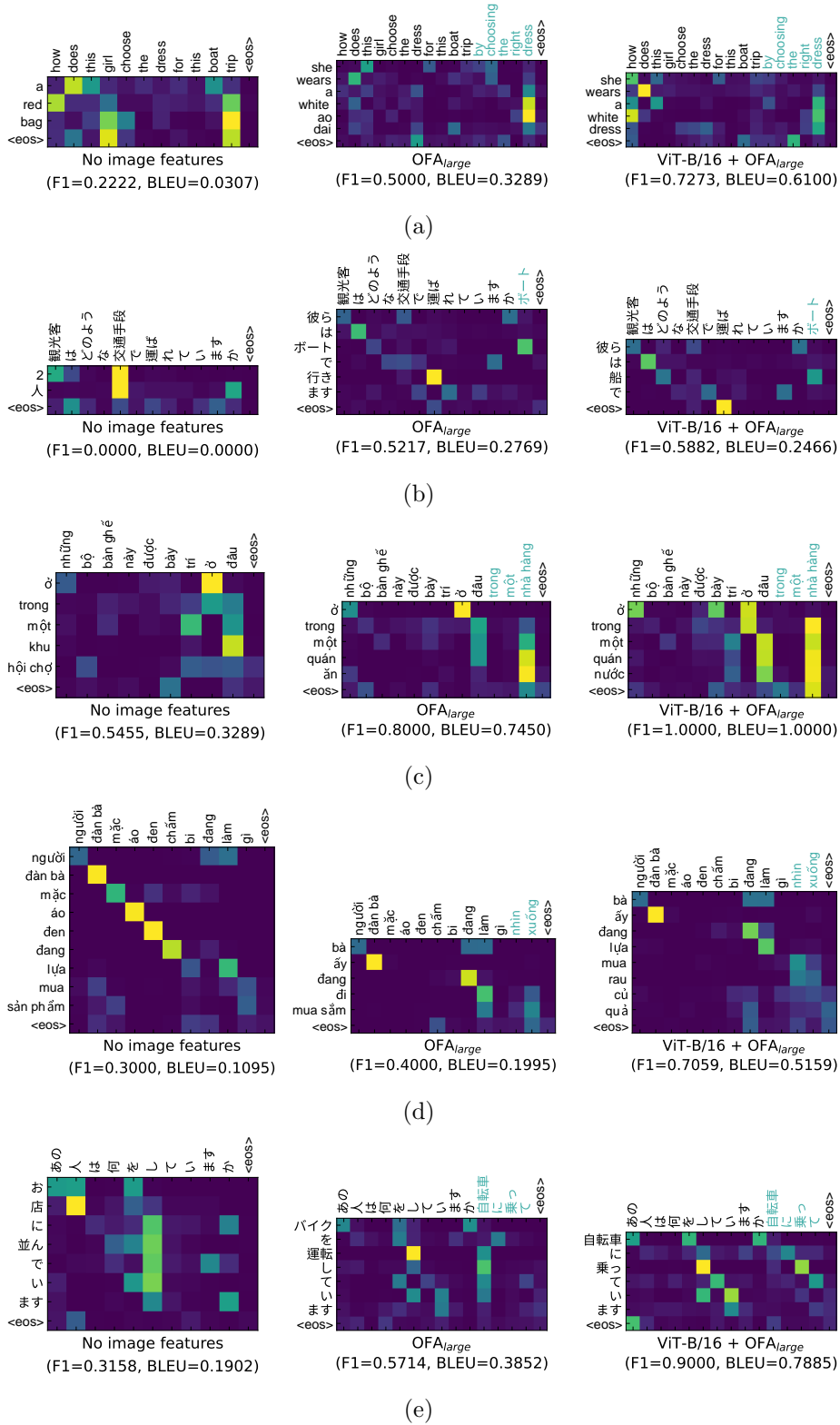
Figure 6: Numerous samples of attention alignment from ConvS2S and the changes in attention when adding features from ViT and OFA models. The x-axis and y-axis of each plot correspond to the words in the question and the generated answer, respectively, while each pixel illustrates the weight w_{ij} of the assignment of the j-th question word for the i-th answer word.

Question: what hat does the narrator of the historical site wear?
Ground truth: non la
Predicted answer: the boy wears a white shirt and white and white
F1: 0.0000
BLEU: 0.0000

(a) Error Case I

Question: có bao nhiêu người đứng bên phải chàng trai?
(*English: how many people on the right of the man?*)
Ground truth: có ba người đứng bên phải chàng trai
(*English: there are three people on the right of the man*)
Predicted answer: có hai người đứng bên phải chàng trai
(*English: there are two people on the right of the man*)
F1: 0.8750
BLEU: 0.7799

(b) Error Case II

Question: 小船手は何本のオールを使っていますか?
(*English: how many paddles does the boatman use?*)
Ground truth: 2
Predicted answer: 2本の船を使っています
(*English: using 2 boats*)
F1: 0.0000
BLEU: 0.0000

(c) Error Case III

Figure 7: Three typical error cases from generated results.

Figure 7b, which significantly improves the evaluation score. This incorrect scenario exemplifies the limitation of the evaluation measures.

- The model managed to generate the correct key answer while also adding unnecessary information compared to the ground truth, which may lead to the response's meaning being distorted. As shown in Figure 7c, the model correctly predicted quantity but then added unnecessary tokens afterward, resulting in a low score on both evaluation metrics.

6.   CONCLUSION

We have used the Convolutional Sequence-to-sequence network combined with the ViT and OFA model for our proposed system in the VLSP-EVJVQA task. The final results are 0.3390 on the public test set and 0.4210 on the private test set by the F1 score. From the result, we placed the 3rd rank in the competition. Through errors analysis, various errors have been found in the output answer, which are our limitations in this study. In summary,

there are factors that have significant impact on our solution for the multilingual VQA task: the diversity of each language, the translation performance, the effects of vision and language models and the generation capability of the core Seq2Seq model.

Our future research for this task is to improve the accuracy of the model in giving the correct answer by enriching the features from images and questions. Other SOTA vision-language and image models such as BEiT, DeiT and CLIP can be applied to assess the performance on UIT-EVJVQA dataset. Besides, from the proposed system, we will implement an intelligence chat-bot application for question-answering from images.

ACKNOWLEDGMENT

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] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016.

[3] I. Chowdhury, K. Nguyen, C. Fookes, and S. Sridharan, "A cascaded long short-term memory (lstm) driven generic visual question answering (vqa)," in *2017 IEEE International Conference on Image Processing (ICIP)*, 2017, pp. 1842–1846.

[4] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *International Conference on Machine Learning.* PMLR, 2017, pp. 933–941.

[5] 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

[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," *ICLR*, 2021.

[7] D. Dzendzik, J. Foster, and C. Vogel, "English machine reading comprehension datasets: A survey," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 8784–8804. [Online]. Available: https://aclanthology.org/2021.emnlp-main.693

[8] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question answering," in *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'15. Cambridge, MA, USA: MIT Press, 2015, p. 2296–2304.

[9] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 1243–1252.

[10] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, p. 6904–6913.

[11] W. He, K. Liu, J. Liu, Y. Lyu, S. Zhao, X. Xiao, Y. Liu, Y. Wang, H. Wu, Q. She, X. Liu, T. Wu, and H. Wang, "DuReader: a Chinese machine reading comprehension dataset from real-world applications," in *Proceedings of the Workshop on Machine Reading for Question Answering.* Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 37–46. [Online]. Available: https://aclanthology.org/W18-2605

[12] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] N. V. Kiet, T. Q. Son, N. T. Luan, H. V. Tin, L. T. Son, and N. L.-T. Ngan, "VLSP 2021 - ViMRC Challenge: Vietnamese Machine Reading Comprehension," *VNU Journal of Science: Computer Science and Communication Engineering*, vol. 38, no. 2, 2022.

[14] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 5583–5594. [Online]. Available: http://proceedings.mlr.press/v139/kim21k.html

[15] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vision*, vol. 123, no. 1, p. 32–73, may 2017. [Online]. Available: https://doi.org/10.1007/s11263-016-0981-7

[16] S. Lim, M. Kim, and J. Lee, "Korquad1. 0: Korean QA dataset for machine reading comprehension," *arXiv preprint arXiv:1909.07005*, 2019.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision.* Springer, 2014, pp. 740–755.

[18] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 1412–1421. [Online]. Available: https://www.aclweb.org/anthology/D15-1166

[19] K. Nguyen, V. Nguyen, A. Nguyen, and N. Nguyen, "A Vietnamese dataset for evaluating machine reading comprehension," in *Proceedings of the 28th International Conference on Computational Linguistics.* Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 2595–2605. [Online]. Available: https://aclanthology.org/2020.coling-main.233

[20] M. V. Nguyen, V. Lai, A. P. B. Veyseh, and T. H. Nguyen, "Trankit: A light-weight transformer-based toolkit for multilingual natural language processing," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 2021.

[21] 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*, vol. 39, no. 3, 2023.

[22] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2383–2392. [Online]. Available: https://aclanthology.org/D16-1264

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[24] A. Rogers, M. Gardner, and I. Augenstein, "QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension," *ACM Comput. Surv.*, sep 2022. [Online]. Available: https://doi.org/10.1145/3560260

[25] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in BERTology: What we know about how BERT works," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 842–866, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.54

[26] N. Shimizu, N. Rong, and T. Miyazaki, "Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps," in *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1918–1928. [Online]. Available: https://aclanthology.org/C18-1163

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] B. So, K. Byun, K. Kang, and S. Cho, "Jaquad: Japanese question answering dataset for machine reading comprehension," *arXiv preprint arXiv:2202.01764*, 2022.

[29] 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

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[31] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *International Conference on Machine Learning*. PMLR, 2022, pp. 23 318–23 340.