ViVQA: Vietnamese Visual Question Answering

Khanh Quoc Tran, An Trong Nguyen, An Tran-Hoai Le, Kiet Van Nguyen
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{18520908, 18520434, 18520426}@gm.uit.edu.vn, kietnv@uit.edu.vn

Abstract

Visual question answering (VQA) is a hot topic that has recently drew the attention of researchers from domains as diverse as natural language processing and computer vision. In spite of the fact that numerous VQA datasets and models have been fundamentally considered in English, there are no related works in Vietnamese VQA. In this article, we have created ViVQA, a new dataset for **Vi**etnamese **V**isual **Q**uestion **A**nswering consisting of 10,328 images and 15,000 question-answer pairs in Vietnamese for evaluating Vietnamese VQA models. On the ViQA dataset, we also propose a system that uses the Hierarchical Co-Attention Model and achieves Accuracy, WUPS 0.9, and WUPS 0.0 of 0.3496, 0.4513, and 0.7786, respectively. Our system beats the two baseline models (LSTM and BiLSTM) on the ViQA dataset. These findings are optimistic for the development of visual question-answering systems in Vietnam. Finally, we discuss future prospects for ViVQA models in order to improve performance. Our dataset[1] is available freely for research purposes.

1 Introduction

Visual question answering (VQA) is a new field that has gradually gained traction and made substantial progress in recent years. VQA is also one of the potential research areas with a combination of natural language processing and computer vision. A VQA system can extract a proper answer to a question based on an image and a question related with it. Although the task is simple for humans, it is a challenge for computers.

VQA has practical applications in our life. For example, to answer inquiries and discover information, we may incorporate automated visual question answering systems into the Chatbot platform. Visual question answering systems are needed for many real-world situations, including customer support, recommendations, question answering, dialogue, and customer systems management. In addition, it has incredible capabilities for situations such as making displayers aware of essential and valuable information from their surroundings.

Visual question answering systems play an essential role in AI applications for human life. Although current research works are available in English, Japanese, and a few other languages. There are no studies on visual question answering in Vietnamese because of data limitations for research. For that reason, we decided to carry out this study to build a new dataset for evaluating Vietnamese visual question answering systems. This dataset is built based on the image data source from MS COCO. Along with implementing the VQA model, we evaluated the performance of the models on the dataset. We described the refinements in the models we implemented to find the model that gives the best results with this dataset using different state-

[1] https://github.com/khanhtran0412/ViVQA

of-the-art methods such as Deeper Long Short Term Memory, Bidirectional-Long Short Term Memory, and Hierarchical Co-Attention.

In this paper, we focus on introducing our new ViVQA as following orders. Section 2 is Related Words, where we present research relevant to this task. Then, Section 3, how we build our dataset, is carefully described. Section 4 includes methods and models for this dataset, and our evaluation is in Section 5. Finally, Section 6 has our conclusion and future works for this task.

2 Related works

Building a system that can automatically answer questions from random images is considered an ambitious goal. Recently, along with the development of modern machine learning methods and the application of a series of related studies, great strides have been made in solving the image-based question and answer problem. In this field, you could specify a few typical study, such as the VQA dataset inception. (Antol et al., 2015) with 614,163 questions and 7,984,199 answers for 204,721 images from the Microsoft COCO set (Lin et al., 2014) is the cornerstone for VQA system development. Later on, several research based on the COCO dataset, such as the Chinese Baidu FM-IQA dataset (Gao et al., 2015) using 123,287 images from the COCO set, and the Japanese VQA dataset (Shimizu et al., 2018) used 99,208 images emerged. In addition to a wide range of high-quality datasets, new models are presented with the goal of improving the efficiency of autonomous question and answer on images such as Deeper Long Short Term Memory (Antol et al., 2015), and Hierarchical Co-Attention (Lu et al., 2016).

Despite the development of AI in the world, the question and answer problem on automatic images is no longer a strange field, but this is still a new field for Vietnam. To the best of our knowledge, this is the first work in the Vietnamese VQA system. In Table 1 an incomplete list of published Visual Question Answering datasets in English and other languages.

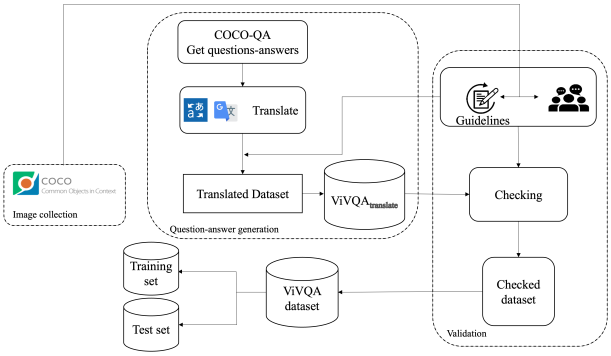


Figure 2: The overview process of creating our dataset ViVQA.

3 Dataset

3.1 Task Definition

In this paper, we aim to build a dataset for a VQA system in Vietnamese. Firstly, we need to define our task clearly as below.

Input: Given an image and a question pertaining to the image content.

Output: A correct answer to the question.

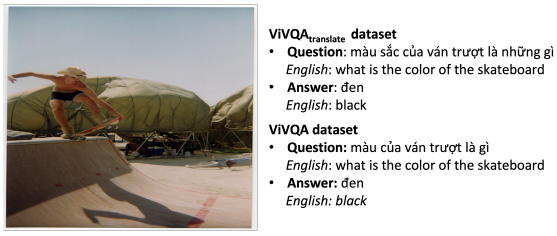


Figure 1: Several instances of the visual question answering task in Vietnamese.

3.2 Dataset creation

We examine numerous dataset creation strategies for this task (see Table 1) in order to determine the best strategy for the Vietnamese language. In the first step, we examine and estimate the range of this task to have the most profound understanding. Based on that, we search for an appropriate dataset, and our choice is MS COCO (Lin et al., 2014) because this dataset is one of the most prestigious, massive, and diverse datasets of images up to this point. As a result, we believe that the MS COCO dataset is substantial and qualified for our task.

Our dataset creation process goes through

Table 1: Background information about our novel dataset and previously published datasets for visual question answering evaluation.

Dataset	Image Source	Language	Images	Q&A	Annotation
FM-IQA (2015)	COCO	Chinese	120,360	250,569	Manual
Japanese VQA (2018)	COCO, YFCC	Japanese	99,280	793,664	
DAQUAR (2014)	NYUDv2	English	1,449	12,468	Semi-auto
COCO-QA (2015)	COCO	English	123,287	117,684	Auto
COCO-VQA (2015)	COCO	English	204,721	614,163	Manual
Visual7W (2016)	COCO	English	47,300	327,939	Manual
Visual genome (2017)	COCO, YFCC	English	108,000	1,773,258	Manual
ViVQA (Ours)	**COCO**	**Vietnamese**	**10,328**	**15,000**	**Semi-auto**

three phases: Collecting images, generating question-answer, validating the dataset. This procedure we refer to from the COCO-QA dataset (Ren et al., 2015) creation process. These phases are described in detail as follows.

3.2.1 Image collection

We extract 10,328 images randomly from the MS COCO dataset (Lin et al., 2014). After that, we preliminarily process the retrieved data to get good quality images and diverse contexts, preparing for the next stage.

3.2.2 Question-answer generation

Based on the COCO-QA dataset in English, we automatically translate question-answer pairs from English into Vietnamese. Initially, we took a simple approach to this task using two famous machine translation tools, Google Translate[2] and Microsoft Translator[3]. We found that several translated questions are unnatural and challenging to understand. Figure 1 shows a translated sample from the ViVQA$_{translate}$ dataset versus a corresponding from the ViVQA dataset. Although this is just a simple question in English, "what is the color of the skateboard". However, when translated into Vietnamese, it causes ambiguity and difficulty to understand. As a result, in order to ensure the dataset's quality and efficiently train the models, we must review and rectify the mistake translated data in the next section.

[2]https://translate.google.com/

[3]https://translator.microsoft.com/

3.2.3 Validation

Before checking and correcting mistakes in translated data, we build annotation guidelines to enhance the quality of the dataset. We take out several random images to pose question-answer pairs. Then we proceed to build guidelines for the dataset construction. In particular, we build annotation rules to make questions and their answers more qualified and natural. Annotators are guided and must strictly follow the guidelines. Question-answer pairs must comply with the rules described according to Table 2. Questions or answers that do not follow the rules are removed and replaced with other questions and answers.

With construction criteria to contribute, so ensuring a clean and correct dataset is our top priority. Annotators check and correct the automatic translated question-answer pairs to minimize mistakes in translated questions and their answers. Question-answer pairs that do not translate well have low similarity between translation tools. Therefore, we use the Cosine Similarity (Rahutomo et al., 2012) to measure the semantic similarity between samples translated by Google Translate and Microsoft Translator. The following is a description of formula 1.

$$\cos(\theta) = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

where A_i and B_i are components of the vectors A and B respectively, representing

Table 2: ViViQA annotation rules.

No.	Descriptions
1	Each image must contain 1 - 3 questions.
2	Each question must have one corresponding and unique answer.
3	Each answer must only contain one word.
4	Q&A only about the activities and objects visible in the image.
5	Familiar English words like laptop, TV, ok, etc. are allowed.
6	Each question must be a single sentence.
7	While annotating, personal opinion and emotion must be avoided.
8	Questions can include a variety of activities and objectives from various perspectives.

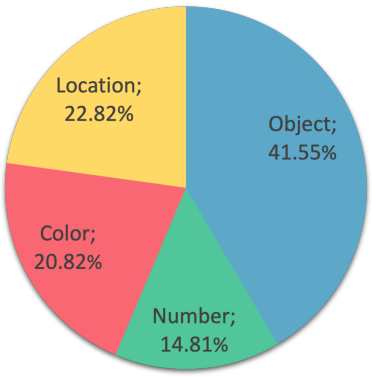
question-answer pairs that Google Translate and Microsoft Translator automatically translate.

Following the guidelines, questions or their answers where the similarity score between question-answer translations is less than 0.8 are checked and corrected by our English qualified crowdsourcing team, including five annotators, with an average IELTS band of 6.5.

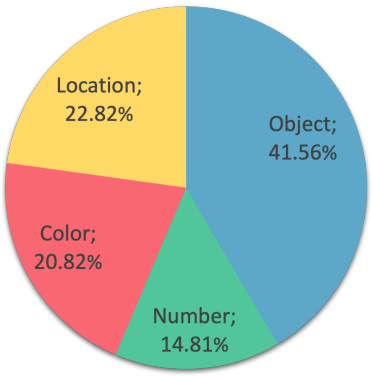
3.3 Dataset Analysis

The ViVQA dataset consists of 10,328 images and 15,000 pairs of questions and answers corresponding to the content of the images. We divide the dataset randomly into training and test sets with a ratio of 8:2. Table 3 summarizes the statistics of ViVQA dataset.

Besides, we conduct a distribution analysis of the different question types in the ViVQA dataset to have a deeper understanding of the dataset. Because the question and answer pairs in our dataset are generated based on question and answer pairs from the COCO-QA dataset, the question types in the ViVQA follow the definitions in the study of Ren et al. (Ren et al., 2015). So we also proceeded to divide the questions into four types: Object, Number, Color, Location. The distribution of question types in our dataset is seen in Figure 3. The Object questions make up a significant proportion of the ViVQA dataset, with rates of 41.55% and 41.56% in the two sets of training and testing, respectively.



Training set.



Test set.

Figure 3: The distribution of the question types on the ViVQA dataset.

4 Our Proposed System

This paper proposes a system using the main core of Hierarchical Co-Attention for Vietnamese visual question answering. The sys-

Table 3: Overview statistics of the ViVQA dataset.

	Training		Test		Total
	Frequency	**Percentage**	**Frequency**	**Percentage**	
Number of image	-	80.0%	-	20.0%	10,328
QA pairs	11,999	80.0%	3,001	20.0%	15,000
Longest question	26 words	-	24 words	-	26 words
Longest answer	4.0 words	-	4.0 words	-	4.0 words
Average question length	9.49 words	-	9.59 words	-	9.51 words
Average answer length	1.78 words	-	1.78 words	-	1.78 words

tem includes two main components: data preprocessing (see Section 4.1) and the Hierarchical Co-Attention model (see Section 4.2). Figure 4 depicts an overview of the system we presented.

Figure 4: Our proposed system for Vietnamese visual question answering.

4.1 Data Preprocessing

Data preprocessing is an essential step in most current machine learning projects. Cleaning a dataset allows more information to be extracted for model training, which enhances experimental results. In particular, we preprocess images, questions, and answers according to the steps as shown in Figure 5.

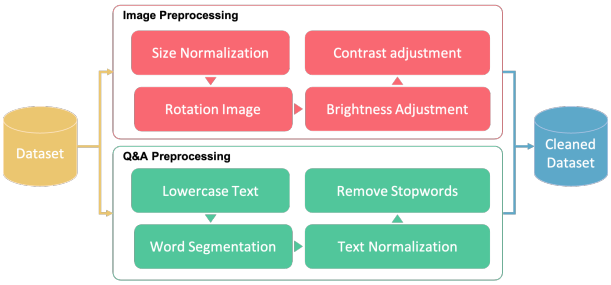


Figure 5: The overview of data preprocessing process.

We deal with resizing all the images in the dataset. Training with many non-uniform dimensions results in low model accuracy. In this paper, we normalize the images with size 64x64x3 for training models to achieve better performances. Then, we rotate the images and adjust the brightness and contrast to diversify the context of the dataset. On that basis, we improved the learning ability and increased the stability of deep learning network architectures.

After pre-processing the images, our dataset is synchronously normalized in size and has more contexts of brightness, contrast, preparing for training deep learning models: LSTM, Bi-LSTM, and our approach.

Based on the questions and their answers described in Section 3, we apply the pre-processing techniques described below to create the cleaned dataset before training the models.

- Converting into lowercase texts;
- Using nltk (Bird, 2006) for word segmentation;
- Removing special characters and spaces;
- Removing stopwords based on the study (Le, 2017).

After data cleaning, we apply two algorithms for visual question answering to evaluate and analyze the models performance on the ViVQA dataset.

4.2 Hierarchical Co-Attention Model

4.2.1 Co-attention

Co-attention is an operation that employs feature information extracted from an image and a question logically. It senses that the image feature is used to supportively attend the question feature and vice versa.

4.2.2 Question hierarchy

The hierarchical architecture concentrates on three kinds of hierarchies: word-image, phrase-image, and question-image.

4.2.3 Model architecture

Each question having T words is represented as $Q = \{q_1, q_2, ..., q_T\}$ where q_T is a feature vector of the word t. Whilst, representations of word, phrase, and question at position t are q_w^t, q_p^t, q_s^t respectively.

Words from the original question are depicted by a one-hot encoding vector, $Q^w = \{q_1^w, q_2^w, ..., q_T^w\}$. Then, the word embedding is applied to those vectors in order to have the correlation between words. This causes the models to learn effectively.

In order to have the feature from the phrase hierarchy, a 1-D convolution is applied on the word embedding vectors, and at each word t the inner product is computed with sizes s.

$$\widehat{q}_{s,t}^p = \tanh(W_s^c q_{t:t+s-1}^w), s \in \{1, 2, 3\} \quad (2)$$

The word-level features Q^w are padded with 0 before being fed into the convolution in order to maintain the length of sequence. The results of the convolution layers are then fed into max-pooling layers for different n-gram at each word to gain phrase-level features.

$$q_t^p = \max(\widehat{q}_{1,t}^p, \widehat{q}_{2,t}^p, \widehat{q}_{3,t}^p), t \in \{1, 2, ..., T\} \quad (3)$$

This pool technique takes the feature at different locations and keeps the sequence's length as well as word order. The phrase-level q_t^p given from max-pooling layers is encoded by using LSTM. Consequently, the result gained from LSTM layers is question-level q_t^s at time t.

Alternating Co-attention, as its name describes, the mechanism alternatively attends to question (or image) features guided by image (or question) features.

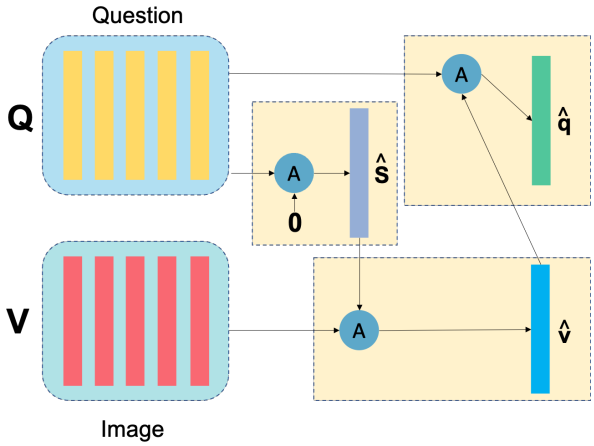


Figure 6: Our core approach using alternating co-attention mechanism inspired by Lu et al. (2016).

5 Experiments and Results

5.1 Baselines

To compare with our proposed system, we use two SOTA methods including Deeper LSTM and Bi-LSTM as baseline systems. We describe the two baselines as follows.

5.1.1 Deeper Long Short Term Memory (LSTM)

The first model we implement for the baseline is the Deeper Long Short Term Memory (LSTM) model proposed by S Antol et al. (2015). According to the authors, this model was born to make the training and inference of the model faster than previous models. The model structure consists of two parts, the multi-layer perceptron (MLP) and the LSTM model based on a softmax layer to generate the answer. The LSTM model encodes the sentence words in question using one-hot encoding, followed by a linear transformation of the image features to the size required by the LSTM encoder. Questions and pictures are coded based on matrix multiplication (element-wise multiplication).

5.1.2 Bidirectional-Long Short Term Memory (Bi-LSTM):

The LSTM could be a famous variant of RNN (Medsker and Jain, 1999). The Bidirectional Long Short Term Memory can be trained using all available input information within the past

and way forward for a selected timeframe. This method is robust in various problems, and most of its achieved high-performance results. Therefore, during this task, we plan to choose it to compare with other VQA models.

LSTM network architecture includes memory cells and ports that allow the storage or retrieval of information. We also use Bi-LSTM with Bidirectional (Schuster and Paliwal, 1997), BiLSTM can learn more contextual information extracted from two directions.

5.2 Experimental Settings

On the training set, the models are trained, and on the test set, they are tested. We analyze experiments using three models in this paper: Deeper Long Short Term Memory, Bi-Long Short Term Memory, and Hierarchical Co-Attention. Our implementation is based on PyTorch. Then, we can also compare their performance on the data set.

- Hierarchical Co-Attention: we choose to experiment with the alternating co-attention. This model is run with 30 epochs, batch size equals 64, max sequence length is 40, dropout is 0.4. This model has four conv2D layers with 32 filters at sizes 1, 2, 3, 5, respectively. Its optimizer is Adam and activation is sigmoid.

- Deeper Long short term memory: this model is run with 30 epochs, batch size equals 128, dropout is 0.1. This model has an internal state dimension of 512, dense layers with 50 in size with relu activation, and the number of hidden layer MLP is 3. Its optimizer is Adam.

- Bi-LSTM: this model is run with 30 epochs, batch size equals 128, dropout is 0.1. This model has a bidirectional layer followed by max-pooling 2D, dense layers with 50 in size with relu activation, and dense layers have 50 in size with sigmoid activation. Its optimizer is Adam.

For word embeddings, we use the Vietnamese word embedding ETNLP[4] provided by Vu et al. (2019) and PhoW2V: pre-trained Word2Vec syllable and word embeddings for Vietnamese provided by Nguyen et al. (2020), which in dimension 300 with character n-grams to perform feature extraction for questions. This pre-trained embedding model be used as an embedding layer in the neural network models training to find the best embedding set for this ViVQA dataset.

5.3 Evaluation Metrics

Before going through experimental results, we first discuss the evaluation metrics used to evaluate the model performance. Following Ren et al. (Ren et al., 2015), we evaluated the accuracy between the ground truth answer and a predicted answer with the relevant question using Accuracy and the Wu-Palmer similarity (WUPS) (Wu and Palmer, 1994) metric. WUPS determines how similar two words are based on their classification tree's longest common subsequence. If the similarity between two words is less than a specific threshold, the answer could be incorrect (0 points). According to Malinowski and Fritz (Malinowski and Fritz, 2014), we evaluate all models based on Accuracy, WUPS 0.9, and WUPS 0.0.

5.4 Experimental Results

Our experiments achieve novelty results, checking translations that enhance our model's performance 10%. Our proposed system achieves the best performance for the Vietnamese VQA system with Accuracy, WUPS 0.9, and WUPS 0.0 scores of 0.3496, 0.4513, and 0.7786, respectively. In general, PhoW2Vec word-level embeddings achieve optimistic results on this task, and our proposed system obtains better results than other models. Table 4 shows our results through experiments performed.

However, there are still many wrongly predicted answers. This problem could be explained by the fact that although we have checked and corrected the bad translations, a significant number of COCO-QA questions have grammatical errors. That makes it difficult to

[4]Vietnamese Embedding ETNLP - `https://github.com/vietnlp/etnlp`

Table 4: The experimental results of different Vietnamese visual question answering systems.

System	$\textbf{ViVQA}_{translate}$			ViVQA		
	Acc.	WUPS 0.9	WUPS 0.0	Acc.	WUPS 0.9	WUPS 0.0
LSTM + W2V	0.2521	0.3701	0.6325	0.3228	0.4132	0.7389
LSTM + FastText	0.2585	0.3896	0.6237	0.3299	0.4182	0.7464
LSTM + ELMO	0.2462	0.3534	0.6120	0.3154	0.4114	0.7313
LTSM + PhoW2Vec	0.2924	0.3617	0.6423	0.3385	0.4318	0.7526
Bi-LSTM + W2V	0.2761	0.3309	0.6241	0.3125	0.4252	0.7563
Bi-LSTM + FastText	0.2796	0.3805	0.6419	0.3348	0.4368	0.7542
Bi-LSTM + ELMO	0.2758	0.3669	0.6331	0.3203	0.4247	0.7596
Bi-LTSM + PhoW2Vec	0.2887	0.3757	0.6373	0.3397	0.4215	0.7616
Our proposed system	**0.2919**	**0.3811**	**0.6570**	**0.3496**	**0.4513**	**0.7786**

check the similarity between translations that do not really achieve the expected effect when the data source has no guaranteed quality.

6 Conclusion and Future Works

We introduced the new ViVQA dataset for evaluating Vietnamese VQA models in this study. The ViVQA dataset included 15,000 question-answer pairs on 10,328 images. In addition, we conducted experiments on SOTA models, including LSTM and Bi-LSTM with pre-trained word embeddings as first baseline models. We also proposed a system-based Hierarchical Co-Attention Model for Vietnamese VQA. As a result, our system outperformed the two baselines (LSTM and BiLSTM) with 0.3385% and 0.3397% (in accuracy), respectively. Our system obtained 0.3496 (in accuracy), 0.4513 (in WUPS 0.9), and 0.7786 (in WUPS 0.0).

We plan to enhance the image-question-answer triples' quality in the future while expanding the dataset's size and diversity of image-question-answer triples. We undertake studies on BERTology (Rogers et al., 2020) and transfer learning (Ruder et al., 2019) to improve the performance of visual question answering systems in Vietnamese, inspired by the success of ViReader (Nguyen et al., 2021).

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.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question answering. *arXiv preprint arXiv:1505.05612*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Van-Duyet Le. 2017. stopwords: Vietnamese. https://github.com/stopwords/vietnamese-stopwords.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *arXiv preprint arXiv:1606.00061*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *arXiv preprint arXiv:1410.0210*.

Larry Medsker and Lakhmi C Jain. 1999. *Recurrent neural networks: design and applications*. CRC press.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020. A Pilot Study of Text-to-SQL Semantic Parsing for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4079–4085.

Kiet Nguyen, Nhat Duy Nguyen, Phong Nguyen-Thuan Do, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Vireader: A wikipedia-based vietnamese reading comprehension system using transfer learning. *Journal of Intelligent & Fuzzy Systems*, pages 1–19.

Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. In *The 7th International Student Conference on Advanced Science and Technology ICAST*, volume 4, page 1.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. *arXiv preprint arXiv:1505.02074*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18.

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

Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. 2018. 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*, pages 1918–1928.

Xuan-Son Vu, Thanh Vu, Son N Tran, and Lili Jiang. 2019. Etnlp: A visual-aided systematic approach to select pre-trained embeddings for a downstream task. *arXiv preprint arXiv:1903.04433*.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.