Video Question Answering on Vietnamese

Trung Le Chi Phan [1,2†], Man Nguyen Tran [1,2†], Tuan Dat Nguyen [1,2†], Tai Huu Le [1,2†], Trong Hop Do [1,2*]

[1]Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam.
[2]Vietnam National University, Ho Chi Minh City, Vietnam.

*Corresponding author(s). E-mail(s): hopdt@uit.edu.vn;
Contributing authors: 21522725@gm.uit.edu.vn; 21522325@gm.uit.edu.vn; 21522754@gm.uit.edu.vn; 21522562@gm.uit.edu.vn;
†These authors contributed equally to this work.

Abstract

Video Question Answering (VideoQA) is an interdisciplinary research field, that presents numerous challenges by integrating computer vision, natural language processing, and multimedia comprehension. This is a rigorously designed video Question answering (VideoQA) benchmark for video enhancement understanding from description to explanation of temporal actions.VideoQA aims to develop intelligent systems capable of understanding visual content and textual queries, enabling them to answer questions about the information presented in videos.
This research specifically focuses on Video Question Answering in the context of the Vietnamese language. We set up open-ended QA tasks targeting causal action reasoning, temporal action reasoning, and a general understanding of the scene. As the demand for video content continues to rise, particularly within the Vietnamese-speaking community, the need for effective methods to understand and interact with video data becomes increasingly significant.
In this paper, the team proposes a straightforward model that utilizes fundamental computer vision techniques to analyze videos and simple natural language processing methods to comprehend questions and generate answers. However, models still have difficulty generalizing the answers.

Keywords: Video Question Answering, Computer Vision, Natural Language Processing, Vietnamese Language, Multimodal

1 Introduction

Extracting information from videos is highly crucial and holds significant value for various applications, especially considering the large volume of videos generated daily. Video Question Answering (VideoQA) can help us quickly obtain essential information from videos without the need to manually review them, benefiting diverse real-world applications.

Answering natural questions about a video is a robust expression of cognitive abilities. To assess the understanding capabilities of models on videos, various intermediate tasks are proposed, such as Video Classification, Action Recognition, and Video Captioning. Recently, VideoQA has been proposed based on Image Question Answering (ImageQA), where the input consists of a short video and a question, and the task requires the model to provide answers. Due to the arbitrary nature of videos and questions, VideoQA models necessitate the ability to analyze diverse types of videos and questions.

Currently, there is a limited availability of Vietnamese VideoQA datasets. Therefore, the research team has utilized certain tools to translate the NExT-QA question set into Vietnamese and conducted experimental tests on the model using them. However, the translation capabilities of these tools are not yet optimal, particularly for shorter questions and answers, and the tools may not comprehend the context, leading to translation errors for some synonymous terms.

About the NExT-QA dataset, actions in videos are often not independent but rather related to causal and temporal relationships. For example, in the video shown below, a toddler cries because he falls, and a lady runs to the toddler to pick him up. Recognizing the objects “toddler”, and “lady” and describing the independent action contents like “a toddler is crying” and “a lady picks the toddler up” in a video is now possible with advanced neural network models.

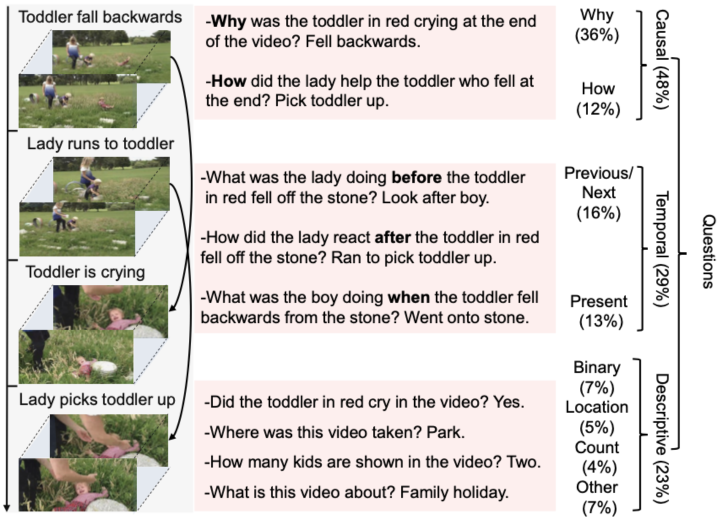
In this research, the team proposes a simple neural network model that employs basic computer vision techniques to analyze videos and straightforward natural language processing methods to understand questions and generate answers.

2 Related Works

In this section, we will review about the datasets of Video QA around the world and datasets that included question-answering tasks and videos in Vietnam.

2.1 ActivityNet-QA

Yu et al. provided this benchmark for testing the performance of VideoQA models on long-term spatio-temporal reasoning. The dataset consists of 58,000 QA pairs on 5,800 complex web videos derived from the popular. What makes this dataset different from other VideoQA datasets is the length of the videos which is much longer than other datasets. The second thing is that the annotations are fully annotated by humans while other datasets like MovieQA(Tapaswi et al. 2016), TGIF-QA (Jang et al. 2017),



-**Why** was the toddler in red crying at the end of the video? Fell backwards.

-**How** did the lady help the toddler who fell at the end? Pick toddler up.

-What was the lady doing **before** the toddler in red fell off the stone? Look after boy.

-How did the lady react **after** the toddler in red fell off the stone? Ran to pick toddler up.

-What was the boy doing **when** the toddler fell backwards from the stone? Went onto stone.

-Did the toddler in red cry in the video? Yes.

-Where was this video taken? Park.

-How many kids are shown in the video? Two.

-What is this video about? Family holiday.

Fig. 1: NExT-QA is a question-answering benchmark targetting the explanation of video contents. It challenges QA models to reason about causal and temporal actions and understand the rich object interactions in daily activities.

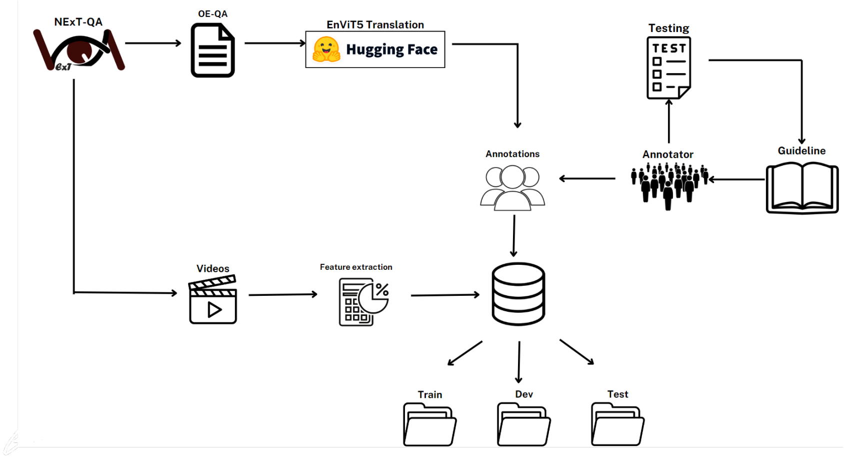


Fig. 2: Data creation Proccess

MSVDQA, MSRVTT-QA (Xu et al. 2017), and Video-QA (Zeng) are annotated by algorithms, which makes the models to be easily overfitted.

2.2 NeXT-QA

Xiao et al. introduced this VideoQA benchmark in CVPR2021, which targets achieving advanced video understanding from describing to explaining the temporal actions. The benchmark includes two different tasks: multiple choice QA and open-ended QA targeting causal action reasoning, temporal action reasoning, and common scene comprehension. NExTQA contains 5,440 videos and about 52K manually annotated question-answer pairs grouped into causal, temporal, and descriptive questions.

2.3 UIT-Anomaly

Vo et al. published the UIT-Anomaly dataset in 2022 which is a video dataset captured in Vietnam. UIT-Anonmay included 224 videos with six different types of anomalies with a total duration of 200 minutes. However, this dataset is mainly focused on the anomaly detection task and the amount of videos in this dataset is much less than the two presented datasets above. So we can see that the VideoQA dataset is still not so popular in VietNam.

2.4 OpenViVQA

Continuing with the visual question-answering task in Vietnamese. We can see that it is significantly developing with many different datasets focusing on this task. In 2023, Nguyen et al.introduced OpenViVQA, a dataset which haves both open-ended questions and answers. This dataset includes 11,199 images together with 37,914 question-answer pairs (QA) annotated manually. However the questions in this dataset only depend on image information, these questions will not contain the causal and temporal features when bringing into videos.

3 Dataset

According to the previous datasets mentioned above, we can see that there is still not a dataset on Vietnamese that supports the video question-answering task in Vietnamese. Although the significant development in the visual question task in Vietnamese is clear, the questions and answers in this task are different from those in the video about the causality and temporality in features. This is the reason we created this dataset.

3.1 Dataset Collection

According to the large amounts of videos in video question answering datasets on English [Figure. 2], we decided to use their data to create our datasets on Vietnamese. We chose Next-QA [] as the main data source to use to create the dataset. The NeXT-QA dataset is publicly published on GitHub and we will use this dataset just for educational and research goals only.

3.2 Data Creation

3.2.1 Videos

Firstly, we downloaded the videos from Git Hub and then removed the sound in the videos. We removed the sound so it would not affect the questions if they contained information related to human sound as it was spoken in languages that are not Vietnamese. Also, the dataset does not have any questions related to using the sound from the video to answer the questions. Therefore, removing the sound from the video does not affect the model.

Besides that, we also provide the features of the video, the videos are first reduced the video fps to the constant number 2. Despite reducing the video fps leads the decreased video frames, we believe that the model got two frames in a second is still fine, this also can reduce the challenge of reading the video features because this task needs powerful GPU requirements.

Next to another video processing task, we will use the algorithm to crop the center part of the video frames and then resize them to the fixed size of (128x128) per frame. Finally, we apply an architecture in computer vision that is InceptionV3 [1] to extract features of the videos and then save them under pickle files (.pkl). These features are extracted due to our need for the proposed method in 8.

3.2.2 Question - Answer

Different from the OpenViVQA, we decided to follow the annotations' creation which was used in the ViVQA(Tran et al.,) dataset. In the NeXT-QA datasets, there are two different tasks in question answering (multiple choice, open-ended). We will only focus on the open-ended task of the original dataset. The annotations will be collected from the public repository of NeXT-QA on GitHub and go through a language translation model. As using the translation models, the annotations will not be natural for readers as well as computers to learn. That is the reason why we decided to use human knowledge to recheck the translated annotations and then improve the translation in a way that is more natural but less consuming of time.

3.3 Data Annotation

As we can see in [Figure. 2], we will first create a brief guideline based on our observation of the translated QAs. After that, we trained our annotators with the guidelines and let them do the test to evaluate the annotators' ability and similarity. Besides that, we will rely on the testing results to improve our guidelines. For the test, we randomly selected 1000 translated QAs together with their corresponding original QA and the video ID. When annotators do the test, they will have to watch the videos and then annotate if the QAs are translated correctly or not with two labels (1-correct; 0-not correct). The result will then be collected and we will calculate the F1-score between members of the annotators based on the collected result. If the annotators failed to meet the requirements, we would force them to do the test again until the requirement that needs the satisfaction of the F1-score (we need the F1-score to reach

0.8). Besides that, we observe the QAs that make the annotators conflict with each other and then update the guidelines.

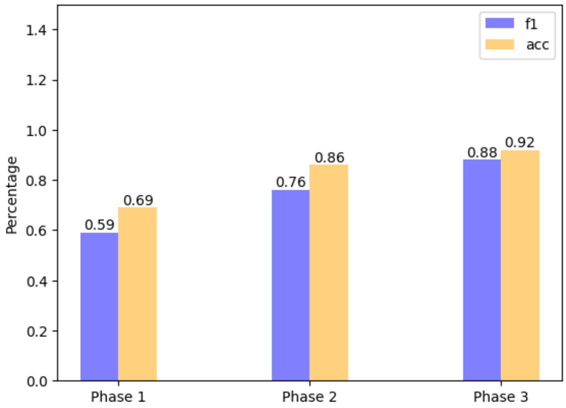


Fig. 3: Scores of three phases annotators testing

As presented in [Figure. 3]. Our annotators went through three phases of testing, the first phase seems to show the high difference between members when the calculated f1-score is not very high. However, with the updates in the guideline, the next phases have a significant improvement in results and meet our requirements. After that, we update the final version of the guideline and finish the annotators testing phase.

After finishing the final guideline, we decided to conduct experiments on the performance of different language translation models to choose the best model for the translating steps. To calculate the performance of the models, we randomly selected 100 English QAs and then translated them into 3 files with three models. Annotators will once again watch the video and label if the question is correctly translated. The QAs will be correct if 3/4 annotators label it with value 1. The performance of the model will be calculated based on the accuracy score based on 100 chosen QAs. The result of the translation models is presented in the [Figure 4] and we chose the best performance models.

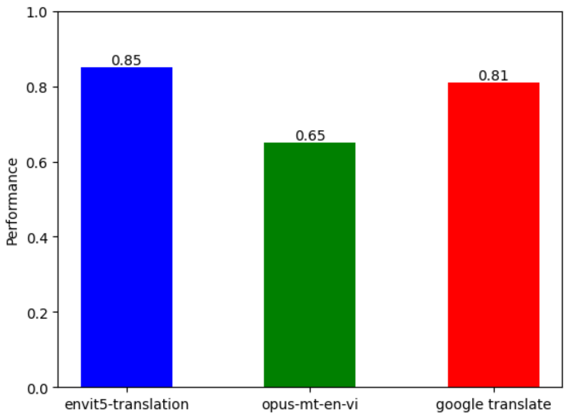


Fig. 4: Performance between three translation models

3.4 Dataset Analysis

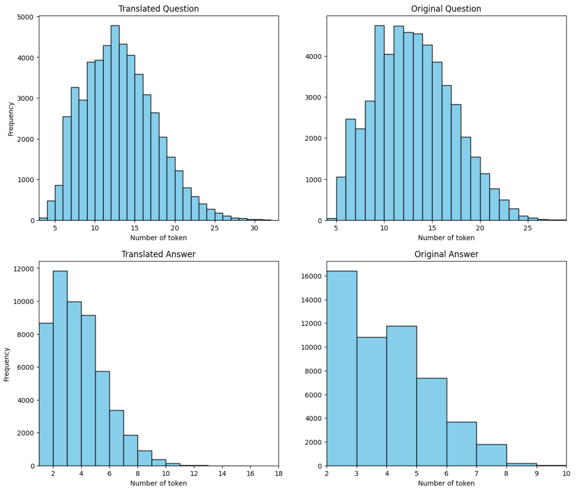


Fig. 5: Answers and questions length between ours and NeXT-QA

3.4.1 Comparison with original questions

As shown in [Figure. 5], our dataset has a diversity in the length of the questions with a maximum length of over 30 tokens which is a little bit longer than the original dataset.

Based on the comparison between the original question and the translated question, we could also see that the question with lengths from 10 to 12 tokens in our dataset seems to decrease while the number of questions that are longer than 13 tokens increases.

We could explain this situation due to the reason for the word segmentation in Vietnamese. A word in English will be presented with only one token while in Vietnamese words could be presented with more than one token. This would probably make the difference between our dataset and the original one.

3.4.2 Comparison with original answers

The same trend happens in answers when the length of ours tends to increase and decrease the number of short answers. Besides that with our annotators checking, we could explain this situation in the case of synonyms in English. While questions will have the concepts and relations when translated, the original dataset's answers are mostly from 2 to 3 tokens and are not long enough to bring in the concepts or the relations. This caused the problem for the model to translate synonyms in English wrongly because those synonyms are two different words in Vietnamese.

Example 1. *Org: "Remove the cap"*
Vietnamese: "Cởi nón"
Translated: "Bỏ cái nắp" (remove the jar)

As shown in the example above, we can see that "cap" has two meanings in English while it has two meanings in Vietnamese. In addition, the answer that is not related to the question's context makes the translation model give the answer incorrectly. Therefore, this can also be a challenge for old models that were used for this task on NeXT-QA to approach our dataset.

3.4.3 Wordcloud of questions and answer

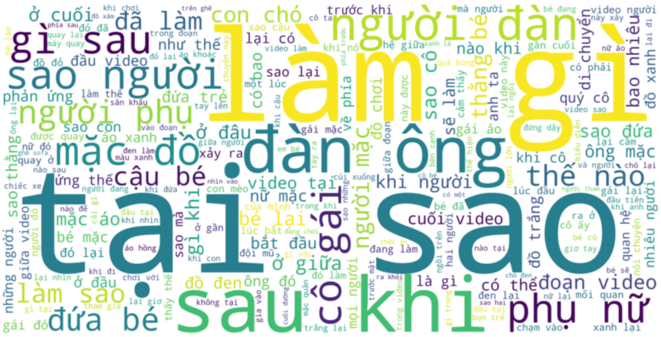


Fig. 6: Wordcloud of our questions

As we can see in [Figure. 6], the words that mostly appear are "tại sao" - "Why" and "làm gì" - "what is someone doing". This shows that the types of questions which is mostly asked are causal questions (why) and Temporal questions (Questions about actions).

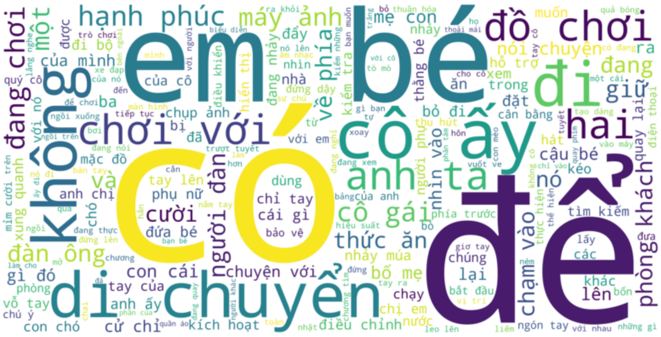


Fig. 7: Wordcloud of our answers

Continue with the word cloud of answer. The most appeared object is "em bé" - "baby" and the most answered keyword is "có" - "Yes". From that, we can see that babies-related video seems to play a leadership role in the dataset and the "yes" answer explained for the high frequency of one token answers presented in [Figure. 7] on both the English and Vietnamese datasets

4 Proposed Model

The overview of the model architecture is shown in [Figure. 8], we have two things that need to be pre-processed before it is given as the Input of the model, there are the open-ended QAs and the videos.

The QAs will be pre-processed by word segmentation and word tokenizer from Pyvi, then will be embedded by word embedding which we choose to experiment with: FastText and PhoBert. The question has a padding with a maximum length of 33 tokens.

The videos will also be pre-processed by reducing FPS task based on ffmpeg-python module [2], and will be center cropped and resized through OpenCV module [3], next to the task that is video features extracting, we choose the InceptionV3 model [1] for this task. The video also has a padding with a maximum length of 400 frames x 2048 features. This means there are 2048 features per frame and any videos will be padded to have 400 frames.

We provide a simple self-building neural network with two inputs as presented above: video features and embedded questions. The embedded question and the video features extracted will be given to the LSTM layer. We provide a language model layer at the last that is needed to process 11 units, the output is expected to have one dimension that has 11 elements equal to 11 elements from the embedded answer.

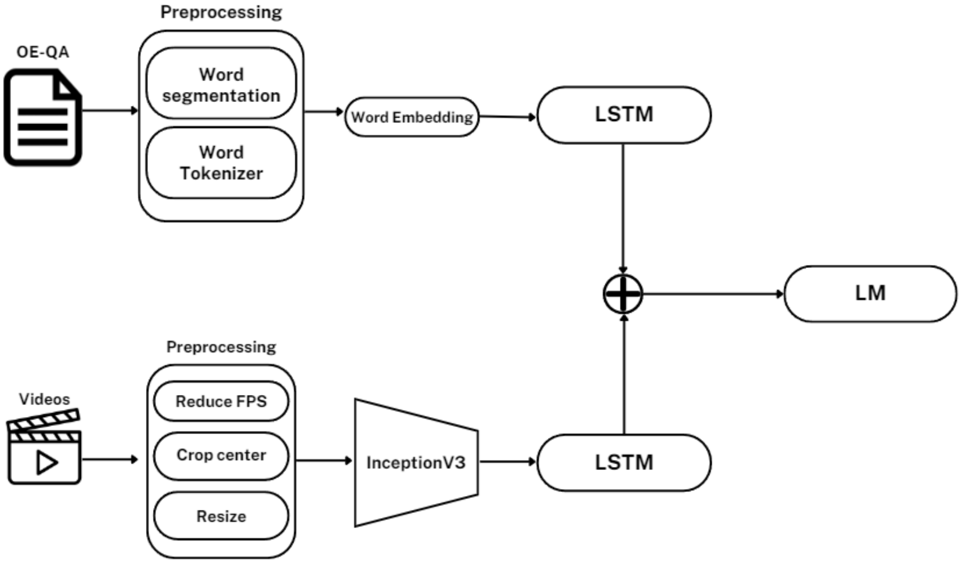


Fig. 8: Our proposed model Architecture

5 Experiment and Results

To test the performance of our proposed model on different word embeddings over our dataset, we decided to compare the performance of different language models at the final layer of our proposed model. Each language model will be combined with a word embedding in Vietnamese for the evaluation of our models with different settings.

5.1 Word Embedding

FastText: developed by Facebook AI Research, FastText is trained on large-scale text corpora from various sources. FastText incorporates subword information by breaking words into n-grams, enabling it to capture morphological nuances and handle out-of-vocabulary words effectively.

ELMo: is a type of word embedding technique that utilizes deep contextualized representations of words, and can enhance the quality of word embeddings by considering the context in which words appear in sentences. This contextualized approach enables ELMO to better capture the intricacies of the Vietnamese language, which is characterized by tones, accents, and a rich vocabulary.

PhoBert: is a Vietnamese-language pre-trained language model based on the BERT architecture, developed by VinAI Research. PhoBERT has achieved state-of-the-art performance on various Vietnamese NLP benchmarks, showcasing its effectiveness in understanding and generating Vietnamese text. PhoBERT represents a significant advancement in natural language processing for the Vietnamese language

5.2 Language Model

Multi Layer Perceptron (MLP) or Fully Connected (FC): is a type of artificial neural network with multiple layers of interconnected nodes, each layer having weights and biases. It excels in non-linear mappings, making it effective for complex tasks like pattern recognition and classification in machine learning and deep learning applications.

LSTM: is a type of recurrent neural network (RNN) architecture designed to address the vanishing gradient problem in traditional RNNs. LSTM was introduced to overcome the vanishing gradient problem. The gating mechanisms enable LSTMs to selectively remember or forget information, allowing them to capture long-term dependencies in sentences or videos

5.3 Evaluation

5.3.1 Accuracy

One of the simplest and most widely used metrics for Natural Language Processing (NLP) is *Accuracy*, which measures the percentage of correct predictions or outputs by models.

$$\text{Accuracy} = \frac{total\_number\_of\_correct\_predictions}{total\_number\_of\_predictions}$$

Accuracy is easy to calculate and interpret, but it may not be enough to capture the nuances and complexities of natural language. In such scenarios, *Precision*, *Recall*, and *F1-score* may offer a more nuanced understanding of model performance.

5.3.2 F1 score

F1 score formal definition is the following:

$$\text{F1} - \text{Score} = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$$

Where TP stands for true positive, FP for false positive, and FN for false negative. The definition of an F1-score is not trivial in the case of NLP. In NLP span-based Question Answering task:
- TP: number of tokens that are shared between the correct answer and the prediction.
- FP: number of tokens that are in the prediction but not in the correct answer.
- FN: number of tokens that are in the correct answer but not in the prediction.
(A token is a unit of language that is used as input for our models in NLP.)

5.3.3 Bleu

Some NLP tasks, such as machine translation and summarization, involve generating natural language texts as outputs Traditional accuracy, precision, and recall may not effectively evaluate the quality and fluency of generated texts. Instead, metrics like BLEU and ROUGE are utilized, comparing generated texts with reference texts provided by human experts.

BLEU score measures the similarity between the machine-translated text and the reference translations using n-grams, which are contiguous sequences of n words. Then, the formula for the BLEU score is as follows:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

- BP (Brevity Penalty) is a penalty term that adjusts the score for translations that are shorter than the reference translations.

- p_n is the precision of n-grams, which is calculated as the number of n-grams that appear in both the machine-generated translation and the reference translations divided by the total number of n-grams in the machine-generated translation.

BLEU score heavily relies on n-grams and may not capture the overall meaning or fluency of the translated text accurately. It may also penalize translations that are longer than the reference translations, which can be unfair in some cases.

5.3.4 Rouge

The ROUGE score is a set of metrics commonly used for text summarization tasks, where the goal is to automatically generate a concise summary of a longer text.

ROUGE score calculates the recall of n-grams in the machine-generated summary by comparing them to the reference summaries.

However, similar to the BLEU score, the ROUGE score also has limitations. It may not fully capture the semantic meaning or coherence of the summary, and it relies solely on the n-gram overlap, which may not always be an accurate measure of summary quality.

5.3.5 Meteor

METEOR score is a metric that measures the quality of generated text based on the alignment between the generated text and the reference text. The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision.

It also has several features that are not found in other metrics, such as stemming and synonymy matching, along with the standard exact word matching. The metric was designed to fix some of the problems found in the more popular BLEU metric, and also produce a good correlation with human judgment at the sentence or segment level.

6 Result Analysis

In this section, we provide a detailed result analysis of the best performance from the embedding and the method of building a neural network that is PhoBert embedding and LSTM by reporting through three cases based on the comparison between the original tokens and the predicted tokens from the model:

Model	Accuracy	Absolute Accuracy	F1	Bleu1	Bleu2	Rouge1	Rogue2	RougeL	Meteor
FastText + FC	0.427	0.0255	0.0747	0.0489	0.0400	0.0894	0.0316	0.0763	0.0411
FastText + LSTM	0.505	0.0279	0.0883	**0.0593**	**0.0465**	0.0935	0.0389	0.0873	0.0518
Elmo + FC	0.420	0.0269	0.0798	0.0453	0.0395	0.0890	0.0332	0.0802	0.0383
Elmo + LSTM	0.489	0.0281	0.0889	0.0580	0.0461	0.0943	0.0396	0.0879	0.0537
PhoBert + FC	0.492	0.0272	0.0822	0.0513	0.0455	0.0902	0.0358	0.0837	0.0445
PhoBert + LSTM	**0.536**	**0.0293**	**0.1008**	0.0588	0.0426	**0.1096**	**0.0421**	**0.0895**	**0.0541**

Table 1: Results from the experiments

6.1 Correct answer

Almost all the correct answers have the same common form which only has one meaningful element, other elements are the padding. Some samples are shown at [Table. 2].

index	Answer embedded	Meaning
1	[458 0 0 0 0 0 0 0 0 0 0 0 0 0]	có
2	[2103 0 0 0 0 0 0 0 0 0 0 0 0 0]	Hai
3	[1789 0 0 0 0 0 0 0 0 0 0 0 0 0]	nói

Table 2: Samples of correct answers

In the table, we show that the word 'có' or 'Hai' is the correct answer from the model, on the other predictions, also have the same type as this sample, almost correct answers that the model gives are answered for the counting question as 'Có bao nhiêu...' or yes/no question as 'Họ có... không?'. The final model predicts answers for these types of questions mostly correctly. Also, some exception predictions that have one token are predicted exactly.

6.2 A part of the answer that is correct

Different from the correct answer that the model gives, this section shows the ability to give a part of the correct answers, there are two types shown in [Table. 3].

index	Type	Answer embedded	Meaning
1	True answer	[1554 3777 0 0 0 0 0 0 0 0 0 0 0 0]	đang đọc
	Predicted answer	[3777 2606 0 0 0 0 0 0 0 0 0 0 0 0]	đọc truyện
2	True answer	[915 2057 2136 0 0 0 0 0 0 0 0 0 0 0]	cha_mẹ và con_cái
	Predicted answer	[499 2057 4773 0 0 0 0 0 0 0 0 0 0 0]	mẹ và con

Table 3: Samples of some correct parts of the answer

The table shows that the predicted answers have some tokens same to the true answers, there are two types, type 1, the model gives a right token and then provides some tokens that may relate to the given token or not, type 2, the model gives a token that links some nouns as 'và' in the presented sample, in this type 2 sample, we also observe that some nouns which are related to each other but they are not the same token, therefore, they are different during the identification process as 'mẹ' and 'cha_mẹ' or 'con_cái' and 'con' by word embedding index.

6.3 Completely incorrect answer

There is likely the same in the last section, instead of giving some tokens and there is a correct part, there are no similarities between the original tokens and the predicted tokens in this section [Table. 4]

index	Type	Answer embedded	Meaning
1	True answer	[2300 1656 0 0 0 0 0 0 0 0 0 0 0 0]	giữ chặt
	Predicted answer	[1510 472 0 0 0 0 0 0 0 0 0 0 0 0]	người khác
2	True answer	[1900 2294 4751 0 0 0 0 0 0 0 0 0 0 0]	nhặt bóng lên
	Predicted answer	[1395 4043 2602 1752 1149 0 0 0 0 0 0 0 0 0]	bảo_vệ về đồ_chơi búp_bê đi

Table 4: Samples of incorrect answers

7 Conclusions and Future Work

In this report, we had some experiences with text and video preprocessing and did some experiments that applied some word embeddings and language models for the VideoQA task. Specifically, we used FFmpeg [2] and OpenCV [3] to preprocess the videos and get the video extractions through InceptionV3 [1]. Word embeddings used for this task are FastText, Elmo, and PhoBert, we experiment with two methods of building neural networks: Fully Connected (FC) and LSTM. The experiment ends with the best performance from PhoBert word embedding and LSTM neural network building.

However, the model we proposed still shows a good performance over this dataset and this task. We could explain that the features of the video cause this result. We still have not extracted the video's specific features like spatial information, and temporal information. Therefore, in the future, we will continue experimenting with this task with different ways of video feature extraction. Besides that, anyone who wants to extract the video features with higher frames per second and higher quality (without resizing and center-cropping) could continue to develop the dataset as we provided raw videos.

References

[1] Chollet, F., et al.: Keras. https://keras.io (2015)

[2] Tomar, S.: Converting video formats with ffmpeg. Linux Journal **2006**(146), 10 (2006)

[3] Itseez: Open Source Computer Vision Library. https://github.com/itseez/opencv (2015)

[4] Heilbron, F.C., Escorcia, V., Ghanem, B., Niebles, J.C.: Activitynet: A large-scale video benchmark for human activity understanding. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 961–970 (2015). https://doi.org/10.1109/CVPR.2015.7298698

[5] Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. Proceedings of the AAAI Conference on Artificial Intelligence **33**(01), 9127–9134 (2019) https://doi.org/10.1609/aaai.v33i01.33019127

[6] Xiao, J., Shang, X., Yao, A., Chua, T.-S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9777–9786 (2021)

[7] Vo, D.T.T., Tran, T.M., Vo, N.D., Nguyen, K.: Uit-anomaly: A modern vietnamese video dataset for anomaly detection. In: 2021 8th NAFOSTED Conference on Information and Computer Science (NICS), pp. 352–357 (2021). https://doi.org/10.1109/NICS54270.2021.9701556

[8] Nguyen, N.H., Vo, D.T.D., Van Nguyen, K., Nguyen, N.L.-T.: Openvivqa: Task, dataset, and multimodal fusion models for visual question answering in vietnamese. Information Fusion **100**, 101868 (2023) https://doi.org/10.1016/j.inffus.2023.101868

[9] Tran, K.Q., Nguyen, A.T., Le, A.T.-H., Nguyen, K.V.: Vivqa: Vietnamese visual question answering. In: Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, pp. 546–554. Association for Computational Lingustics, Shanghai, China (2021). https://aclanthology.org/2021.paclic-1.72/

[10] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. CoRR **abs/1512.00567** (2015) arXiv:1512.00567

[11] Ngo, C., Trinh, T.H., Phan, L., Tran, H., Dang, T., Nguyen, H., Nguyen, M., Luong, M.-T.: MTet: Multi-domain Translation for English and Vietnamese.

arXiv (2022). https://doi.org/10.48550/ARXIV.2210.05610

[12] Tiedemann, J.: The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In: Proceedings of the Fifth Conference on Machine Translation, pp. 1174–1182. Association for Computational Linguistics, Online (2020). https://www.aclweb.org/anthology/2020.wmt-1.139

[13] Nguyen, D.Q., Nguyen, A.T.: PhoBERT: Pre-trained language models for Vietnamese (2020)

[14] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

[15] Dey, R., Salem, F.M.: Gate-variants of gated recurrent unit (gru) neural networks. In: 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), pp. 1597–1600 (2017). https://doi.org/10.1109/MWSCAS.2017.8053243

[16] Mouselimis, L.: fastText: Efficient Learning of Word Representations and Sentence Classification using R. (2022). R package version 1.0.3. https://CRAN.R-project.org/package=fastText

[17] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Isabelle, P., Charniak, E., Lin, D. (eds.) Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (2002). https://doi.org/10.3115/1073083.1073135 . https://aclanthology.org/P02-1040

[18] Lin, C.-Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004). https://aclanthology.org/W04-1013

[19] Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein, J., Lavie, A., Lin, C.-Y., Voss, C. (eds.) Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation And/or Summarization, pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (2005). https://aclanthology.org/W05-0909

[20] Gerz, D., Su, P., Kusztos, R., Mondal, A., Lis, M., Singhal, E., Mrksic, N., Wen, T., Vulic, I.: Multilingual and cross-lingual intent detection from spoken data. CoRR **abs/2104.08524** (2021) 2104.08524

[21] Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., Bapna, A.: Fleurs: Few-shot learning evaluation of universal representations of speech. arXiv preprint arXiv:2205.12446 (2022)