ViOpenDocVQA: Novel Dataset for Vietnamese Visual Question Answering on Open-domain Structured Document Images

Dung Hoang Dao[1,2], **Minh Huu-Tuan Nguyen**[1,2], **Ngan Thi-Kim Huynh**[1,2]
[1] *University of Information Technology, Ho Chi Minh City, Vietnam*
[2] *Vietnam National University, Ho Chi Minh City, Vietnam*

Abstract

We present a new dataset for Visual Question Answering (VQA) on Open-domain Structured Document Images called ViOpenDocVQA. The dataset consists of 5,471 pairs of questions and answers defined on 104 structured document images. We report several baseline results by adopting existing VQA and reading comprehension models. Despite of the small data size, our empirical results suggest that increasing the amount of data can increase the performance of models. Hence, this opens up a new promising direction in the VQA works and expands the scope of its application in the Vietnamese language.

1 Introduction

Recently, Visual Question Answering (VQA) task has garnered widespread attention, driven by a substantial amount of data encompassing general document images and, more specifically, structured document images. Documents such as scientific researches, financial reports and invoices all contain valuable information presented in tabular form. Extracting meaningful insights from these documents is of utmost importance in the process of making well-informed decisions.

However, recent research has predominantly focused on issues related to question-answering about images which mostly feature objects and contain minimal amount of text. This trend poses a challenge for QA based on structured document images due to their intricate and diverse structures, as well as the numerical computations that extend beyond simple queries.

In this paper, we introduce the ViOpenDocVQA dataset, an open-domain Vietnamese dataset designed for the Visual Question Answering task based on Structured Document Images. Figure 1 illustrates an example from the dataset, showcasing an image which contains a table along with a question about the table and its corresponding answer

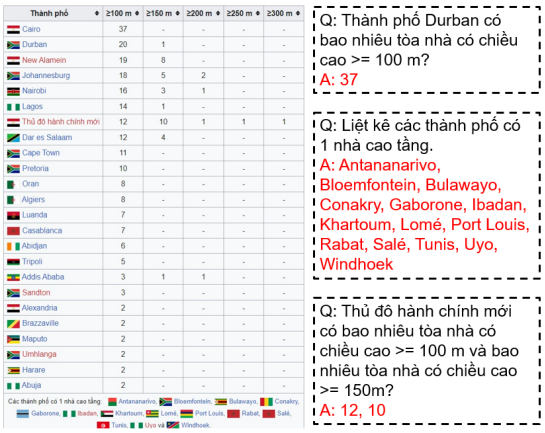


Figure 1: The example illustrates a data sample of ViOpenDocVQA. Q1: How many buildings does the city of Durban have a height of >= 100 m?, Q2: List cities with 1 high-rise building., Q3: How many buildings does the new administrative capital have a height of >= 100 m and how many buildings have a height of >= 150 m?

derived from the information provided in the image. This figure presents three examples from our dataset.

The first Question and Answer (QA) pair in figure 1 necessitate information extraction. The second pair involves the ability to count, while the remaining QA demand enumeration skills in identifying the answers.

This study makes the following contributions:

- Introducing ViOpenDocVQA, a novel dataset tailored for Visual Question Answering (VQA) specifically focused on Open-domain Structured Document Images. With approximately 104 images featuring diverse table formats, organized in various column and row structures, we have meticulously curated and defined over 5471 pairs of questions and answers.

- Our dataset presents 2 challenges for the image processing community and the VQA com-

munity. First, there needs to be a system capable of handling tables with a variety of structures. Second, my dataset includes questions that are structurally diverse because Vietnamese is an unstructured language, without a certain linguistic structure in asking questions.

2 Related Work

Open-domain Question Answering: is a task of finding answers to the question from a large collection of textual documents. Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) consisting of 100,000+ questions posed by crowdworkers on a set of Wikipedia articles. TriviaQA (Joshi et al., 2017) has over 650,000 question-answer-evidence triples, with questions originating from trivia enthusiasts independent of the evidence documents. Microsoft's Machine Reading Comprehension (MS MARCO) dataset (Bajaj et al., 2018) comprises of 1,010,916 anonymized questions - sampled from Bing's search query logs - each with a human generated answer and 182,669 completely human rewritten generated answers.

OpenBookQA (Mihaylov et al., 2018) introduces about 6000 questions which probe an understanding of a set of 1326 elementary level science facts. Quasar-T (Dhingra et al., 2017) consists of 43000 open-domain trivia questions and their answers obtained from various internet sources. HotpotQA (Yang et al., 2018) is a dataset with around 113,000 Wikipedia-based question-answer pairs requiring multi-hop reasoning, which means that the models have to integrate information from multiple documents to formulate accurate responses.

Natural Questions (Kwiatkowski et al., 2019) provides more than 300,000 QAs with real anonymized, aggregated queries issued to the Google search engine. An annotator in this dataset is presented with a question along with a Wikipedia page from the top 5 search results, and annotates a long answer (typically a paragraph) and a short answer (one or more entities) if present on the page, or marks null if no long/short answer is present.

There are also open-domain QA datasets in Vietnamese, a typical example is OpenViVQA dataset (Nguyen et al., 2023a) which consists of 11,000+ images associated with 37,000+ questions and their corresponding open-ended answers.

Structured Document Images Datasets: many datasets on this topic have been created. They are commonly used to address problems related to structured document recognition and information extraction. TabLeX (Desai et al., 2021), Table2Latex (Deng et al., 2019) are some datasets that contain table images generated from scientific articles, with their corresponding ground-truth LaTeX. TableBank (Li et al., 2020) and PubTabNet (Zhong et al., 2020) are also notable datasets which include table images and their corresponding HTML representation.

Scene-text VQA: is a branch of Visual Question Answering (VQA) which aims to solve the cases where scene-text is required to answer the given question. Some notable datasets in this task include TextVQA (Singh et al., 2019) with 45,000+ questions over 28,000+ images, ST-VQA (Biten et al., 2019) with 31,000+ questions over 23,000+ images. In recent years, this problem has attracted a lot of attention, and many more specific problems, with more specialized datasets, have been created.

Document Images VQA: this is a specific task of Scene-text VQA, where the datasets comprise only document images. Many datasets have been created to make challenges in this task. DocVQA (Mathew et al., 2021b) has 50,000 question-answer pairs over more than 12,000 document images, which consist of many different document types like graphs, tables, figures, paragraphs, ect. VisualMRC (Tanaka et al., 2021) is a visual machine reading comprehension dataset which contains 30,000+ pairs of a question and an abstractive answer for 10,000+ document images sourced from multiple domains of webpages. VQAonBD 2023 dataset (Raja et al., 2023) has more than 1,5 million question-answer pais over approximately 50,000 business document images. Another remarkable dataset is OCR-VQA (Zhou et al., 2021), which comprises more than 1 million question-answer pairs over 207K+ images of book covers. The questions in this dataset are domain-specific, generated based on template questions and answers extracted from available metadata.

For more specific type of images, DVQA (Kafle et al., 2018) consists of more than three million question-answer pairs over 300,000 images. In this dataset, the images are automatically generated from bar charts and the questions are also defined from available templates. FigureQA (Kahou et al., 2018) is a dataset using similar generation method, but in this case, 5 types of charts are used: line plots, dot-line plots, vertical and horizontal bar graphs, and pie charts. It has over one million

question-answer pairs grounded in over 100,000 images, and 15 template questions. InfographicVQA (Mathew et al., 2021a) comprises 5,485 infographics along with 30,035 question-answer annotations.

Methodologies: Given the introduction of numerous datasets, various models and methods have emerged to address the challenge of Visual Question Answering (VQA) specifically on document images.

TabIQA (Nguyen et al., 2023b) is a system designed for question-answering using table images in business documents. It utilizes a table recognition algorithm to convert the table images into HTML format, and then process to transform each HTML table into a dataframe. Finally, it uses an encoder-decoder architecture with the processed dataframe and the question as the input to generate the final answer. This model achieved an accuracy of 0.8997 on VQAonBD 2023 dataset (Raja et al., 2023).

The authors of DocVQA (Mathew et al., 2021b) proposed a system that reaches an accuracy of 0.5577 on the same dataset. It first serialize the OCR tokens recognized on the document images to a single string, separated by space, in top-left to bottom-right order. Then it utilizes a pretrained BERT (Devlin et al., 2019) question answering model which is also pre-finetuned on the SQuAD (Rajpurkar et al., 2016) dataset. BERT is a pretrained model which uses a deep bidirectional transformer (Vaswani et al., 2023) to understand languages.

The authors of OCR-VQA dataset (Zhou et al., 2021) also introduced a novel baseline for this dataset, called OCR-VQA model. It combine 4 feature vectors for each input. There is a 300-dim BiLSTM representation for the question. For the image it's a 4096-dim feature vector using VGG-16. For the two other vectors, one is the representation of the OCR-ed text blocks which includes indices, coordinate positions and Named Entity Recognition (NER) tags; while the other is a 300-dim Word2Vec representation of the OCR-ed text itself. Finally, they are concatenated to form a 4731-dimensional composite vector. This vector is then fed to a fully connected feed forward network (2 layers of size 1024) followed by a softmax layer.

Vietnamese VQA: in contrast to the abundance of English language datasets, Vietnamese lacks extensive datasets for Visual Question Answering (VQA). ViVQA (Tran et al., 2021) dataset was created using MS-COCO images source with semi-automatical annotation method, resulting in 15,000 QAs over 10,000+ images. Nguyen et al. introduced the OpenViVQA dataset (Nguyen et al., 2023a) which consists of 11,000+ images associated with 37,000+ questions and their corresponding open-ended answers. ViCLEVR (Tran et al., 2023) is a visual reasoning dataset which comprises over 26,000 images and 30,000 QAs.

Another notable dataset is EVJVQA (Luu-Thuy Nguyen et al., 2023), which is used as a benchmark dataset at the 9th Workshop on Vietnamese Language and Speech Processing (VLSP 2022). It includes 33,000+ pairs of question-answer over three languages: Vietnamese, English, and Japanese, on approximately 5,000 images taken from Vietnam and is used to evaluate multilingual VQA systems or models.

With our understanding, there has not been any official open-domain dataset in Vietnamese related to the Document Image VQA task. Inspired by this, we create a new dataset, named ViOpenDocQA, which consists of 5471 pairs of questions and answers defined on 100+ structured document images which include the table. Our project aims to augment the resource of datasets and make contributions to VQA research community in Vietnam.

To make comparisons, we focus on constructing our dataset with diverse topics, long document images, and contributions of answers from annotators with linguistic diversity and open mindset. The datasets we use in this research for detailed comparisons are described in Table 1

3 Dataset

In this section, we elucidate the data collection process, expound upon the data manipulation procedures, and present the statistical outcomes and data analysis within ViOpenDocQA.

3.1 Data Collection

3.1.1 Images collection:

Images in the dataset are screenshots of tables sourced from Wikipedia. We collect the images by taking screenshots of the tables and save them in JPG format. The images are collected according to a set of standards, ensuring that each table meets the criteria of having at least three columns and rows, along with the number of text attributes for each table being greater than three.

Dataset	Task	Query	Images	Source of Query	Source of Images	Answer Type	Answer span Type
TextVQA	VQA + DocVQA	45k	28k	Crowdsourced	Open Images v3 Dataset	Extractive, Abstract	Single, Multi
DocVQA	DocVQA	30k	12k	Crowdsourced	Industry Documents	Extractive	Single
InfographicVQA	DocVQA	30k	5.4k	Crowdsourced	Infographic	Row 4, Col 7	Single, Multi
ViOpenDocVQA	Open-domain DocVQA	5471	104	Crowdsourced	Wikipedia	Extractive	Single

Table 1: Comparison between ViOpenDocVQA and existing DocVQA datasets. # denotes "the number of".

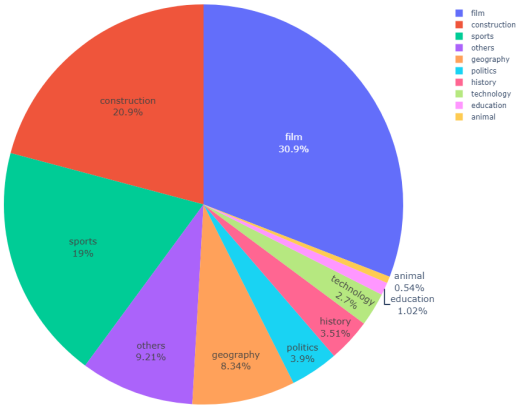


Figure 2: ViOpenDocVQA's domains pie chart.

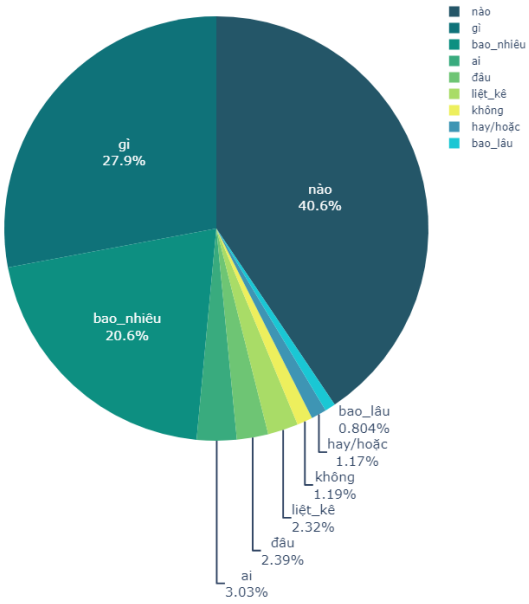


Figure 3: ViOpenDocVQA's keywords pie chart.

The final dataset includes a diverse set of images, comprising 104 images from abandoned topics and periods on Wikipedia to ensure the open-domain property of the dataset. We have cataloged the domains and their occurrences in the dataset as shown in Figure 2

3.1.2 Questions and Answers:

Our dataset focuses on using questions to retrieve information from document images, and the answers will be the information contained in those document images. Questions and answers on the selected document images are collected with the help of remote workers, using a web-based annotation tool. In total, five annotators with enough cultural and academic standard participated in posing questions and providing answers to ensure the diversity of linguistic representation and the expansion of vocabulary from various regions. The annotation process was organized in four stages.

Stage 1: We, as authors, personally served as annotators for 20 sample table images with the goal of creating as many question-answer pairs as possible. Subsequently, we convened, compared, cross-referenced, discussed, and reached an agreement on a comprehensive set of guidelines with accompanying hints (details in figure 3). We ensured certainty in covering the spatial diversity of question-answer pairs within the document images, guaranteeing the diverse querying nature of the questions and meeting certain requirements regarding the common format of the answers.

Stage 2: We trained two annotators with the content provided in the guidelines and instructed them to use our tools to create question-answer pairs for document images. To ensure accuracy and diversity in the dataset, we required annotators to generate at least 10 question-answer pairs for each document image, using only the text content to formulate questions and refraining from using any visual cues beyond the text content in images for answering.

Simultaneously, during this stage, we closely supervised the annotators' results. In case of reported discoveries or new insights in question formulation, we conducted reviews to update the guidelines. However, if there were instances of annotators that do not meet the specified requirements, we assessed the situation and either provided reminders or initiated re-training, depending on the severity of the deviation.

Stage 3: We selected a group of annotators and

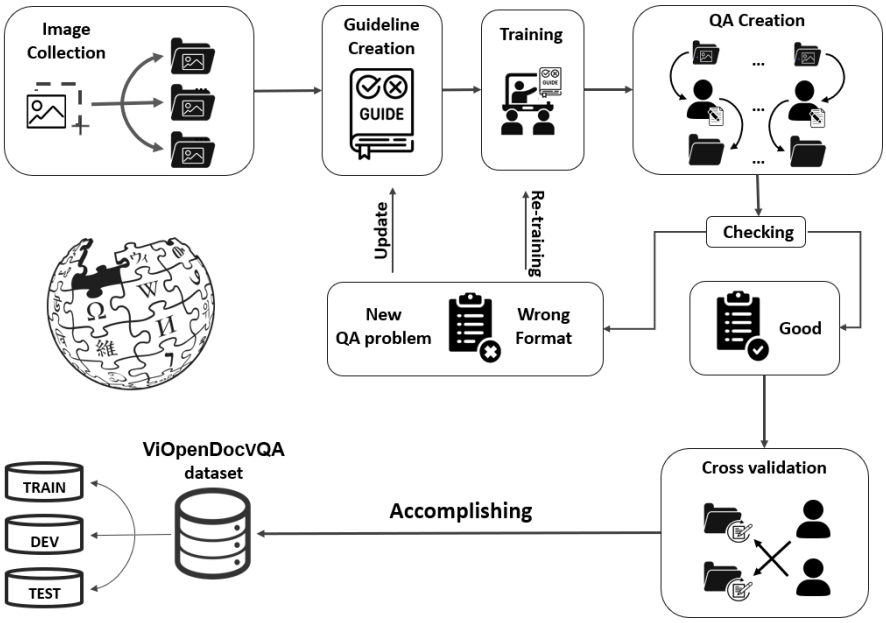


Figure 4: ViOpenDocVQA's data collection process.

a randomly chosen subset representing 10 percent of the dataset for cross-validation. The annotators reviewed document images and their corresponding question-answer pairs to assess and identify any errors in spelling or inaccuracies in the QA pairs. When they detected discrepancies, they marked the respective data points. The subset of data that underwent this stage became part of the dataset.

Stage 4: The authors will review the marked data points to decide whether to make corrections, remove, or retain them.

Finally, we obtained 5471 QA pairs with the corresponding set of images. The overall pipeline of ViOpenDocVQA's data collection process is visualized in figure 4.

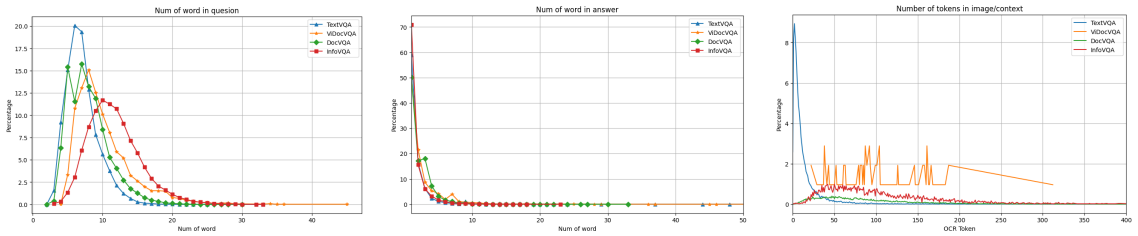
3.2 Statistics and Analysis

ViOpenDocVQA comprises 5471 question and answers pairs framed on 104 images. The data is split randomly in an 8:1:1 ratio to train, validation, and test splits.

We have a total of 9 question keywords covering the entire dataset. Among them, the keyword 'nào' has the highest coverage in the posed questions (40.6%) because it is a common and versatile keyword. For example: 'Khi nào' (when), 'Nơi nào' (where), 'Người nào' (who), 'Cái nào' (which), 'Như thế nào' (how), ... This keyword can be placed at the beginning or end of a sentence, usually combined with a noun immediately before to ask a specific question or stand alone to support a previously mentioned issue. The word 'nào' is a particularly special relative word due to its unpredictable functional changes in various combinations, making it challenging for the dataset to determine the correct answer to the object that keywords refers to.

With the two keywords 'gì' and 'bao nhiêu,' both are used to query information, but 'gì' is used to inquire about textual information, while 'bao nhiêu' is used to query numerical and datetime information. 'ai' functions similarly to 'who' in English, although it can be flexible in its placement at the beginning or end of a sentence when combined with certain modifiers. The keyword 'đâu' is primarily used to inquire about locations, similar to 'where' in English. However, in some cases in Vietnamese, 'đâu' is used to inquire about the subject's information. For example, 'Đâu là vị vua cuối cùng của nhà Trần' can be translated to 'Who is the last king of the Tran Dynasty' in English. Questions containing the keyword 'không' in our dataset are used to query yes/no attributes while still ensuring the extractive nature of the answers. 'hay/hoặc' is a keyword that appears in selection questions, a type of information retrieval question where the answer is present in the question itself. 'Bao lâu' serves a similar function to the English question 'how long'.

'Liệt kê' is a keyword representing a request to list the answers in the English language. However, in Vietnamese, there are various question formats where the answer needs to be listed, and this key-



(a) Question num (b) Answer num (c) OCR token

Figure 5: Question, answer and OCR tokens' statistics in ViOpenDocVQA compared to TextVQA, DocVQA and InfographicVQA

Dataset	**Questions**				**Answer**				**OCR-Token**		
	Min	Avg	Max	Unique	Min	Avg	Max	Unique	Min	Avg	Max
TextVQA	2	7.05	33	80.36%	1	1.46	**102**	51.74%	0	14.36	199
DocVQA	2	8	30	70.72%	1	2.17	33	**64.29%**	0	89.53	951
InfographicVQA	3	**11.68**	33	**99.11%**	1	1.61	23	11.54%	0	**106.12**	**1424**
ViOpenDocVQA	**4**	10.13	**45**	97.02%	1	**2.58**	54	48.47%	**22**	98.81	212

Table 2: Question, answer and OCR tokens' statistics in ViOpenDocVQA compared to TextVQA, DocVQA and InfographicVQA

word in the dataset is used to indicate such questions. For example: 'Những diễn viên nào đã tham gia vào bộ phim Hi! Emma?', which means 'Which actors/actresses joined the movie Hi! Emma?' in English. In this question format, the answer will be multi-span.

Statistics in Figure 3 show the coverage of question keywords across the entire dataset in terms of how questions are posed in an open-domain context. Through this analysis, we can see that Vietnamese questions are of a non-structured nature, which means that they lack a strict set of rules for question formulation in Vietnamese. This demonstrates the diversity and expansion of linguistic expressions in the Vietnamese language.

We compare our dataset with similar datasets, here we use TextVQA, DocVQA, and InfographicVQA, in terms of the number of words in the question, answer, and image text. To ensure a visual comparison, we employ the same EasyOCR method for all datasets.

OCR Tokens: In Figure 5c, the frequency of the number of tokens in images for the ViOpenDocVQA dataset is significantly higher than that of the other datasets, with a minimum value of 22. This is because our dataset focuses exclusively on structured and textual data, whereas the other datasets include objects as well. Regarding the maximum number of OCR tokens, it indicates that our dataset has a lower limit on the number of words in an image than the other datasets. This is explained by the fact that capturing screenshots does not guarantee quality if the image is too large and the text is too small. We only use images with a moderate font size to ensure that the models can extract complete information from the images. This prevents the images from being too small, ensuring the integrity of the information in the image. In terms of average OCR tokens, our dataset ranks second after the InfographicVQA dataset. Additionally, the frequency of the number of tokens indicates that the image quality in our dataset ensures an adequate amount of information for querying.

Questions: Our dataset exhibits a higher frequency of word counts skewed towards the right as shown in Figure 5a, meaning that the number of words in each question tends to be higher than both TextVQA and DocVQA, but less than the InfographicVQA dataset. However, these differences are not substantial and can be understood considering that InfographicVQA and ViOpenDocVQA focus more on queries, even multi-object queries simultaneously. In terms of statistical comparison, our dataset consistently ranks second in both unique indices and the average length of questions. Although the differences are not overly large, this indicates a need for improvement in question diversity within the dataset. However, currently our dataset still surpasses the other two similar datasets. The minimum and maximum limits of our dataset are the highest, which is generally expected, as formulating a linguistically rich and clear query

in Vietnamese often requires a larger number of words compared to English.

Answers: Concerning the number of words in the answers in Figure 5b, the frequency graph shows a relatively small difference. However, the statistics in the table show that ViOpenDocVQA has the highest average number of words in answers. This aligns with our expectations as our dataset not only deals with single-object queries but also multi-object queries. In contrast, the other datasets tend to focus on providing answers to short and numerical questions.

From these observations, it can be noted that although the InfographicVQA dataset has nearly absolute uniqueness in its questions, the answers in the graph-related domain tend to be limited in diversity and mainly revolve around numerical values. On the other hand, despite DocVQA having only 70.72% unique questions, it possesses the highest uniqueness in answers. This can be explained by the dataset being designed to query information within document-type materials, encompassing both numerical values and text. As for our dataset, both uniqueness indices are high, as it employs an open-domain approach, where questions and answers are not specifically specialized in any particular field, resulting in a high level of diversity.

4 Experimental Settings

4.1 Models

To evaluate the performance of our model on our dataset, we fine-tuned the model LayoutLMv3 and LayoutLMv2 (Huang et al., 2022) implemented by Microsoft. These models requires OCR token extraction as input. While these models use the Tesseract OCR engine for this purpose, even when using a pretrained dataset in Vietnamese with Tesseract, we encountered many difficulties in extracting OCR tokens from our dataset. Instead, we opted for EasyOCR, which proved to be much more effective than the Tesseract OCR engine.

The LayoutLMv2 and LayoutLMv3 model we are using will be fine-tuned with our dataset using the following parameters. The training will be conducted on a P100 GPU, spanning 4 epochs with a batch size of 8. We will employ the AdamW optimizer with a learning rate of 2e-5.

4.2 Evaluation Metrics

One problem of normal accuracy score is that it will return 0 to a pair of answer and prediction even when they only have differences in a few characters, mostly conducted by the imperfection of OCR. We do not want the minor mistakes of OCR to be severely penailized, so we need to use another evaluation metric.

ANLS (Average Normalized Levenshtein Similarity) metric was proposed to evaluate VQA models on ST-VQA dataset (Biten et al., 2019). It is based on the Levenshtein distance between two strings, which is the minimum number of single-character edits (including insertions, deletions, or substitutions) required to transform one string into another. The Levenshtein distance is named after the Soviet mathematician and linguist Vladimir Levenshtein, who introduced the concept in 1965.

This metric utilizes Normalized Levenshtein Distance (NLD) between the ground-truth answer and the prediction of the model. The distance's formula is described as below:

$$NLD(A,B) = \frac{\text{Levenshtein Distance}(A,B)}{\max(\text{Length Of}A, \text{Length Of } B)}$$

Here, A and B are the strings being compared, and Levenshtein Distance(A,B) is the Levenshtein distance between the two strings. The denominator max(length of A,length of B) represents the maximum length of the two strings. This normalized value is between 0 and 1.

Since the smaller the distance, the greater the accuracy we obtain, we can calculate the normalized Levenshtein similairy between one prediction and its ground-truth answer by this formula:

$$s(a_i, o_i) = \begin{cases} 1 - NLD(a_i, o_i), & \text{if } NDL(a_i, o_i) < \tau \\ 0, & \text{otherwise} \end{cases}$$

Here, a threshold τ (we set it to 0.5) is defined to filter NLD values larger than this value by returning a score of 0 if the NLD is larger than τ . The intuition behind the threshold is that if a pair of prediction and answer returns a normalized distance which is more than the threshold, we reason that this is due to returning the wrong text block in the document image, rather than the imperfection of OCR.

The final score is then obtained by averaging the normalized Levenshtein similairies over every

pair of ground-truth answer and predicted answer. Since OCR is not perfect, we propose to use ANLS as our primary evaluation metric, so that minor answer mismatches stemming from OCR errors are not severely penalized.

4.3 Results

The model results show relatively low performance, achieving a 0.1% accuracy and an ANLS score of 0.3% for LayoutLMv2 and achieving a 0.5% accuracy and an ANLS score of 1% for LayoutLMv3 (Table 3). This can be attributed to several reasons. The first one is the issue of OCR extraction persists. Despite EasyOCR demonstrating much better extraction capabilities compared to Tesseract (illustrated in Table 4), it still encounters difficulties in capturing diacritics in Vietnamese. For instance, EasyOCR may extract "công ty" instead of the correct answer "công ty," or "quỗc gia" instead of the correct "quốc gia." Part of the problem arises from the large image size and images captured from screenshots, causing challenges in text recognition.

	Accuracy	**ANLS score**
LayoutLMv2	0.00112	0.00315
LayoutLMv3	0.00509	0.01009

Table 3: Experimental results of models on ViOpenDocVQA

Tesseract	**EasyOCR**	**Ground Truth**
Quéc gia	Quỗc gia	Quốc gia

Table 4: An example of OCR's mistake

The other issue is the problem with the small dataset size and limited resources for training the model. This limitation poses challenges to training in terms of both time and performance.

5 Conclusion

We represent a new dataset for the Document Question Answering on Structured Table Image task, presenting two main challenges: handling images containing tables with complex structures and processing non-structured questions due to the characteristics of the Vietnamese language. Experimental results indicate the need for improvements in text extraction from images, as current models exhibit suboptimal performance in this aspect, particularly in capturing diacritics in Vietnamese.

References

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering.

Yuntian Deng, David Rosenberg, and Gideon Mann. 2019. Challenges in end-to-end neural scientific table recognition. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 894–901.

Harsh Desai, Pratik Kayal, and Mayank Singh. 2021. *TabLeX: A Benchmark Dataset for Structure and Content Information Extraction from Scientific Tables*, page 554–569. Springer International Publishing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Bhuwan Dhingra, Kathryn Mazaitis, and William W. Cohen. 2017. Quasar: Datasets for question answering by search and reading.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. Figureqa: An annotated figure dataset for visual reasoning.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020. Tablebank: A benchmark dataset for table detection and recognition.

Ngan Luu-Thuy Nguyen, Nghia Hieu Nguyen, Duong T.D. Vo, Khanh Quoc Tran, and Kiet Van Nguyen. 2023. Evjvqa challenge: Multilingual visual question answering. *Journal of Computer Science and Cybernetics*, 39(3):237–258.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V Jawahar. 2021a. Infographicvqa.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021b. Docvqa: A dataset for vqa on document images.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Nghia Hieu Nguyen, Duong T.D. Vo, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023a. Openvivqa: Task, dataset, and multimodal fusion models for visual question answering in vietnamese. *Information Fusion*, 100:101868.

Phuc Nguyen, Nam Tuan Ly, Hideaki Takeda, and Atsuhiro Takasu. 2023b. Tabiqa: Table questions answering on business document images.

Sachin Raja, Ajoy Mondal, and C. V. Jawahar. 2023. Icdar 2023 competition on visual question answering on business document images. In *Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part II*, page 454–470, Berlin, Heidelberg. Springer-Verlag.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images.

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

Khiem Vinh Tran, Hao Phu Phan, Kiet Van Nguyen, and Ngan Luu Thuy Nguyen. 2023. Viclevr: A visual reasoning dataset and hybrid multimodal fusion model for visual question answering in vietnamese.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. 2020. Image-based table recognition: data, model, and evaluation.

Fang Zhou, Bei Yin, Zanxia Jin, Heran Wu, and Dongyan Zhang. 2021. Text-based visual question answering with knowledge base. In *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, MMAsia '20, New York, NY, USA. Association for Computing Machinery.