ViTextCaps: A Dataset for Image Captioning with Reading Comprehension

Nhi Ngoc-Yen Nguyen[1,2], Nguyen Duc Anh[1,2], Liu The Hien[1,2], Tu Le Huy[1,2], and Do Trong Hop[1,2](✉)

[1] University of Information Technology, Ho Chi Minh City, Vietnam
{21521231,21520140,21522062,21522173}@gm.uit.edu.vn
[2] Vietnam National University, Ho Chi Minh City, Vietnam
hopdt@uit.edu.vn

Abstract. The description of images with reading comprehension is becoming increasingly crucial in the fields of computer vision, natural language processing, and machine learning, particularly as text plays a vital role in understanding the human environment. This research contributes to this domain by extending the dataset to a new language - Vietnamese. This marks a significant step in developing image captioning with reading comprehension in the Vietnamese language, as, until now, there has been no dataset serving this purpose. The initiative begins with the construction of the ViTextCaps dataset, containing manually written descriptions for images sourced from the OpenVQA dataset and supplemented with self-collected data in Ho Chi Minh City. ViTextCaps comprises 23,000 Vietnamese descriptions for 5,289 images. This dataset poses numerous challenges for models, especially in text recognition, linking with the visual context, and determining which part of the text to replicate or paraphrase. Our analysis through both automated and human studies reveals that ViTextCaps introduces several new technical challenges. When employing the M4C-Captioner model, we achieve the highest results, surpassing traditional models in the task of image captioning.

Keywords: Deep Learning · Natural Language Processing

1 Introduction

In recent years, the development of media and technology has resulted in a significant impact of visual information. Images and videos dominate our online experiences, making it crucial to have the ability to infer and comprehend visual content. The lack of reading comprehension and inference capability can pose challenges for some individuals in accurately generating and describing visual information. Additionally, research on the VizWiz dataset [2] indicates that 21% of questions asked by visually impaired individuals about an image are related to the text within it. Image captioning plays a vital role in initiating a visual dialogue with the blind, enabling them to request additional information as needed. Therefore, image captioning has played a significant role in representing visual content through generating descriptive and informative textual captions for images. Furthermore, this technology has gained considerable attention and interest from researchers in various fields, including computer vision, natural language processing, and machine learning. With the emergence of large labeled datasets, progress in image captioning has seen continuous enhancement in performance and quality [11], and optical character recognition (OCR) has become mature [4]. However, while OCR focuses solely on written text, many available image captioning datasets only have a limited focus on text within the scenes of images, where substantial information is stored. They often overlook the significance of scene text, which can appear in various contexts such as street signs, product labels, or book titles. By not incorporating textual information, these datasets may fail and struggle to comprehensively capture the complete context and miss out on valuable information provided within images.

Moreover, the landscape of scene text datasets is particularly lacking when it comes to Vietnamese language-specific datasets. Despite the importance of scene text in image understanding and captioning, there has not been a dedicated Vietnamese dataset that focuses on this crucial aspect of visual content. The absence of specific Vietnamese datasets exposes the gap in this field, while Vietnamese, as a complex and rich language, are considered an enormous resource of valuable information in terms of reading comprehension in images.

In this research, we aim to leverage the strengths of widely-used datasets for image captioning by introducing the ViTextCaps dataset, a novel resource specifically designed for text-based image captioning in the Vietnamese language. ViTextCaps emphasizes enabling logical inference based on scene text, facilitating the learning and comprehension of complex captions by image captioning models. By incorporating ViTextCaps into the existing array of resources, this research aims to augment the capability of models to generate precise and contextually rich captions in Vietnamese. Additionally, it endeavors to analyze and present the unique challenges encountered within this specific field, thereby contributing to the advancement of text-based image captioning in the Vietnamese language. The research comprises the following sections: We announce a dataset for text-based image captioning in the Vietnamese language. Simultaneously, we provide each step of the entire process to create the dataset and analyze its metrics in Section 3. The dataset will be evaluated on standard captioning models and the M4C-Captioner model to assess performance on the dataset in Section 4. Finally, the conclusion and future development directions are discussed in Section 5.

2 Related works

2.1 Image Captioning

TextCaps [26]: TextCaps dataset can be considered as the contrast due to the requirements of focusing on text as an additional modality where models have to read, comprehend and generate a sentence. TextCaps dataset contains 145,329 captions for 28,408 images.

Conceptual Captions (CC) [25]: Conceptual Captions dataset is a massive dataset with over 3 million images and their following raw description which are harvested from webpages with their Alt-text HTML attribute, therefore offering a wide variety of style descriptions and range of context. Their raw description will be applied to an automatic pipeline that extracts and filters candidate image/captions pairs to achieve the balance of information and logical learnability of the processed captions.

vieCap24H-VLSP 2021 [22]: This dataset has been released through a challenge held by the Vietnamese Language and Speech Processing (VLSP). The dataset has provided 10066 images with 11,563 captions about the healthcare domain crawled by scientific articles.

UIT-VilC [12]: UIT-VilC is one of the pioneer Vietnamese Image Captioning datasets. The dataset consists of 3,850 images which are about sports played with balls from the MS COCO dataset, followed by 5 captions for each image, summing up to 19,250 captions in total.

2.2 Optical Character Recognition (OCR)

OCR contains two steps, detection - locating the text region; and extraction - based on the boundaries that were found in the previous step, extracting the texts as characters. OCR can be taken as a subtask of image captioning with reading comprehension task since the model needs to know the texts or characters present in the image to generate meaningful captions. For that reason, OCR plays an important role in this task and similar ones, in which we need to understand their semantic meaning as well as their relationship with the context of the images. In recent years, OCR models have improved in performance and reliability [18], [6], [29]. The Decoder-only Transformer for Optical Character Recognition (DTrOCR) [10], which uses a decoder-only Transformer to take advantage of a generative language model that is pre-trained on a large corpus, has outperformed current state-of-the-art methods by a large margin in the recognition of printed, handwritten, and scene text in both English and Chinese. However, in our research, we found out that OCR is just a stepping-stone for solving the problem we stated in this research.

2.3 Visual Question Answering with Text Reading Ability

In recent years, numerous datasets for the Visual Question Answering task have been published. The TextVQA dataset [27] consists of 45,336 questions, of which 37,912 (83.6%) are unique, on 28,408 images collected from selected categories of the Open Images v3 dataset. ST-VQA [3] (Scene Text VQA) dataset has a similar size of 23,038 images, which were collected from a combination of public datasets. The DocVQA [19], a dataset on document images, contains over 12,000 document images with 50,000 questions. The ViVQA dataset [16], a dataset for VQA task in Vietnamese, consists of 10,328 images collected randomly from the MS COCO dataset and 15,000 question - answer pairs. Reading and reasoning about the text are required tasks in these datasets while considering the context for answering a question, which is similar to the spirit of the dataset we use in this research.

3 The ViTextCaps Dataset

3.1 Data Collection, Preprocessing and Annotation

Overview: The process of creating ViTextCaps dataset show in Figure 1.

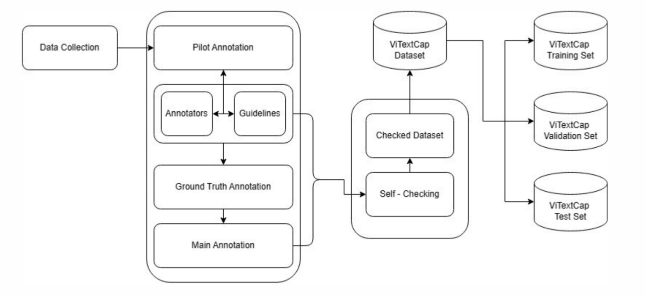


Fig. 1. The process of creating ViTextCaps

Collection: We collected images from diverse sources, not only utilizing Selenium to fetch images from Google Image but also inheriting some from the OpenViQA dataset [21]. Particularly, manually captured images of Saigon were included. The imagery spans a variety of themes, ranging from Vietnamese markets, street food, courts, and restaurants to cultural heritage sites, famous tourist destinations, traditional festivals, and more. The ViTextCaps dataset was curated to establish a tight connection between images and natural language text, providing a diverse context for various natural language processing tasks.

Preprocessing: During the filtering process, we will eliminate gif images, those with excessively small dimensions, and images that are either unreadable or lack text content.

Annotation: The annotation process was facilitated through the inheritance of labeling tools from the UIT-OpenViIC dataset [5]. The tool was developed using the Qt framework with C++ on the Windows platform, as illustrated in Figure 1. This labeling tool is designed to manage annotations at the folder level. Upon completing labeling for a folder, labels are saved in a *.json file with a structure including Annotations, Delete, Filename, and Filepath. In addition to the labeling tool, annotators were provided with a labeling guideline. These simple and memorable guidelines were implemented to ensure the authenticity and quality of the dataset. Crucially, images were always displayed to support annotators during the translation process

3.2 Data Validation

Stage 1 - Pilot Labeling We instructed annotators to perform initial labeling on a pre-segmented dataset consisting of 100 images. The output was then evaluated using concise combined guidelines: text in the image must be readable; correspond accurately to the provided image; include a complete sentence; have correct grammar; and not contain subjective language. The quality of annotators' work was controlled using known good/bad quality gold annotations. For *.json files that did not meet the requirements, feedback was provided to annotators, and relabeling was carried out until the specified criteria were met.

Stage 2 - Main Labeling This stage involved simultaneously labeling, self-evaluating, and conducting cross-checks on the labeled cases to ensure the quality of labeling. In summary, this clear verification step aims to ensure that the final segmented results are of high quality and reliability.

3.3 Data Analysis

Our dataset consists of 5289 images. The topics range from street food courts, restaurants, and markets to cultural monument sites, tourist attraction locations, traditional festivals, and so on, which are all taken in Vietnam. Images in our dataset have a height starting from 400 pixels and up to 4624 pixels. Similarly, the width has a length between 500 and 4624 pixels. The ViTextCaps dataset has a total of 23,000 captions, and the average length of the captions is 23 words (shown in Figure 2), which is longer than a normal sentence in Vietnamese. This can be explained by the fact that the captions in ViTextCaps include scene description and texts from it in one sentence. More than 90% of the dataset's images have 5 captions (shown in Figure 2), which will make the model be able to learn more about how to describe an images in different ways.

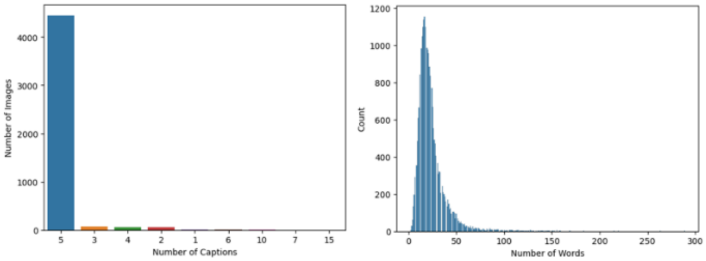


Fig. 2. # Captions per Image (left) and # Words per Caption (right)

In addition, the OCR tokens or texts in the images make the caption much clearer and easier to understand. Most of the captions in the ViTextCaps dataset contain around 3 OCR tokens. Some of the images contain more than three OCR tokens in their captions since our dataset does contain infographic and document-type images. There are 30,383 unique OCR tokens detected in the ViTextCaps dataset, some of which only appear less than 10 times in all captions. We also find out that there are 2,177 out of 30,383 unique OCR tokens in the test set, have neither appeared in the train set nor the validation set, which makes it necessary for the model to be able to read new texts in the images.

4 Experiments and Results

4.1 Baselines

Bottom-Up Top-Down Attention model (BUTD): The Bottom-Up Top-Down Attention model (BUTD) [1] is a prevalent image captioning model that utilizes Faster R-CNN [24] object detection features (Bottom-Up) combined with attention-weighted LSTM layers (Top-Down).

Attention on Attention model (AoANet): The AoA [14] Net for Image Captioning is a module that implements the Attention on Attention mechanism in both the encoder and decoder. This mechanism enables the model to effectively capture the relationships and interactions between ground-truth labels and image features, as well as among the features themselves. We have extracted features by using Faster R-CNN, passed it through the Refiner Layer and obtained LSTM + AoA mechanism to generate the candidate outputs.

Meshed-Memory Transformers: [7] is a transformer-based architecture that can improve both the image encoding and the language generation steps: it learns a multi-level representation of the relationships between image regions, integrating learned a priori knowledge, and uses mesh-like connectivity at the decoding stage to exploit low- and high-level features.

M4C-Captioner: M4C [13] used to be model with state-of-the-art performance on TextVQA task. The model fuse different modalities by embedding them into a common semantic space and processing them with a multimodal transformer. Apart from that, unlike conventional VQA models where a prediction is made via classification, it enables iterative answer decoding with dynamic pointer network, allowing the model to generate a multi-word answer, which is not limited to a fixed vocabulary. This feature make is also suitable for reading-based caption generation. We adapt M4C to our task by removing the question input and directly use its multi-word answer decoder to generate a caption conditioned on the detected objects and OCR tokens in the image.

4.2 Experimental Settings

AoANet and BUTD Due to resource limitations, the training of the AoANet and BUTD was conducted for 30 epochs, with each batch consisting of 8-20 samples, lr=0.001.

Meshed-Memory Transformers we used the original implementation and hyper-parameters, which consists of 3 Memory-Augmented Encoder and 3 Meshed

Decoder. We used Adam as optimizer with the learning rate is adjusted by LambdaLR scheduler, and NLLLoss as loss function.

M4C-Captioner we follow the same implementation details as used for TextCaps task. For visual objects, we detect objects with Faster R-CNN detector pretrained on the Visual Genome dataset, and keeps 100 top-scoring objects per image. Then, the fc6 feature vector is extracted from each detected object. We apply the Faster R-CNN fc7 weights on the extracted fc6 features to output 2048-dimensional fc7 appearance features and finetune fc7 weights during training. Finally, we extract text tokens on each image using the SwimTextSpotter [15] model. From each OCR token we extract FastText feature, detection feature, recognition feature, and bouding box feature. In our multimodal transformer, we use L = 4 layers of multimodal transformer with 12 attention heads. Other hyper-parameters (such as dropout ratio) follow BERTBASE. However, we note that the multimodal transformer parameters are initialized from scratch rather than from a pretrained BERT [9] model. We use PhoBERT [20] vocabulary as our answer vocabulary. During training, we use a batch size of 16. Our model is trained using the Adam optimizer, with a learning rate of 1e-4 and a staircase learning rate schedule, where we multiply the learning rate by 0.9 at 80% each epoch. The best snapshot is selected using the validation set loss. Given that an answer word can appear in both fixed answer vocabulary and OCR tokens, we apply multi-label sigmoid loss as the author implementation, but the result didn't improve so we decided to change it to cross entropy loss.

Datasets We partitioned the dataset into train,validation,test sets in the ratio of 70:20:10. We first evaluate models trained on traditional Vietnamese image captioning datasets to demonstrate how existing datasets and models lack reading comprehension. After that, we train and evaluate each baseline using TextCaps.

Metrics We evaluated the models' performance using BLEU [23], METEOR [8], ROUGE-L [17], and CIDEr [28] for the validation and test set. We gather over 450 samples from the test dataset. When comparing different methods, our emphasis is on CIDEr, which assigns greater importance to informative n-grams in the captions (such as OCR tokens) and reduces the weight of commonly occurring words through TF-IDF weighting. For notation convenience, we perform a percentage calculation on the metrics' scores.

4.3 Experimental Results

We have observed that the complexity of models correlates with improved results. Through the analysis of results (see Table 1, Table 2), it can be observed that the BUTD model, when trained on the ViTextCaps annotation dataset (line 1), achieves the lowest CIDEr score of 0.1030, significantly lower than the M4C-Captioner model. This indicates that BUTD struggles in describing text within images. Transitioning to the UIT-ViIC dataset (line 6), BUTD exhibits a CIDEr score of 1.3810, higher than expected and also surpassing the performance with ViTextCaps data. This suggests that the model is more suitable for Image

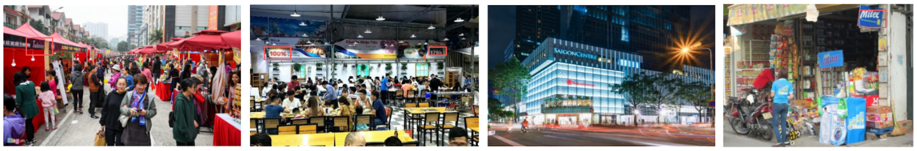
Table 1. Performance of our benchmarks on the ViTextCaps dataset. M4C-Captioner benefits significantly from OCR input and achieves the highest CIDEr score, indicating the importance of copying text from images in this task. BUTD and AoANet yield the poorest results, highlighting that the simplicity of the baseline used, without incorporating OCR input, contributes to their lower performance.

#	Method	Trained on	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE-L	METEOR	CIDEr
1	AoANet	ViTextCaps	0.3150	0.1224	0.046	0.0152	0.3635	0.1652	0.1748
2	BUTD	ViTextCaps	0.3120	0.1920	0.1210	0.0780	0.1680	0.2240	0.1030
3	Meshed Memory	ViTextCaps	0.3643	0.2225	0.1599	0.1251	0.4438	**0.2976**	0.5954
4	M4C-Captioner	ViTextCaps	**0.4813**	**0.3295**	**0.2340**	**0.1667**	**0.5619**	0.2688	**1.9496**

Table 2. Performance of methods on the UIT-ViIC dataset

#	Method	Trained on	BLEU 1	BLEU 2	BLEU 3	BLEU 4	ROUGE-L	METEOR	CIDEr
5	AoA	UIT-ViIC	0.5794	0.4741	0.3883	0.3176	0.6098	0.5138	0.6591
6	BUTD	UIT-ViIC	0.4340	0.3150	0.2420	0.1930	0.4170	0.2330	1.3810
7	Meshed Memory	UIT-ViIC	0.5871	0.4592	0.3736	0.3094	0.6450	0.6056	1.2213
8	M4C-Captioner	UIT-ViIC	-	-	-	-	0.2386	0.0901	-

Captioning tasks unrelated to text within the scene. AoANet (lines 1, 5), while stronger than BUTD, still falls short in handling text comprehension and performs lower than M4C-Captioner (-1.8086) with a slight improvement over BUTD (+0.038). Upon transitioning to the UIT-ViIC dataset (line 7), the model's performance significantly improves. M4C-Captioner attains the highest scores across most metrics, with average scores of Bleu at 0.3028, Rouge at 0.5619, and CIDEr at 1.9496. There is a substantial gap, especially in CIDEr, between training processes and methods without OCR input (lines 1, 2, and 3). In the case of the Meshed Memory Transformer, the model exhibits the highest complexity. Although excluding OCR input has a negative impact on performance, the METEOR score remains the highest, with a marginal difference from M4C-Captioner. Results across the remaining metrics are superior compared to AoANet (+0.4544) and BUTD (+0.4924). The findings indicate that encoding OCR features and the ability to directly copy OCR tokens play crucial roles. It is also observed (lines 1-3 versus 4) that the model's simultaneous use of position, image, and semantic features of OCR tokens is important, especially in their complex combinations, where both spatial and semantic relationships play vital roles in linking words. However, on the test set, the team has not yet conducted human performance benchmarks to compare with machine performance on this task, providing a direction for future extended comparisons. In many studies, the use of actual OCR may mitigate this gap, but it has not been entirely eliminated, implying the potential for improvement in both theoretical reasoning and text recognition capabilities in the future. Figure 3 illustrates examples of quality from various methods. It can be observed that Meshed Memory Transformers rarely utilize OCR inputs to reference text in images. In contrast, the M4C-Captioner method extensively learns to read text in images and references it in the generated captions. Furthermore, M4C-Captioner learns and recognizes relationships between objects and has the capability to combine multiple OCR tokens into complex descriptions. However, in the Vietnamese model, there is still a lack of clear recognition of numbers, as seen in ID:00000004995, with the output " %" without the appearance of the number. Nevertheless, for simple sentences like ID:00000001359, the model produces fairly accurate results. We also observe the misplacement of OCR tokens on objects or in the wrong semantic context in the captions. In ID:000000234, the model makes fairly accurate predictions regarding the text in the image and the colors of the surrounding scene. However, there is an error in the output "@@ " - the position of OCR tokens is misplaced in this case. Many errors stem from a misunderstanding of the scene and object identification, which is a common issue in captioning algorithms. In ID:00000003641, OCR tokens are not placed in the output, possibly due to errors in the OCR detection algorithm. This indicates numerous potential directions for future development in this challenging generative task, requiring an understanding of both images and text, and demanding new model designs that go beyond existing captioning models.



id: 000000001359
M2 Transformers: có ảnh con chợ có một hàng bán m
M4C-Captioner: dòng chữ jadmire có màu trắng.
Human: trong một khu chợ có cửa hàng jadmire.

id: 00000004995
M2 Transformers: Có ăn có đúc người có qua thể biển bảng màu dòng chữ màu màu gà tươi sống màu v
M4C-Captioner: nổi bật trên tấm bảng hiệu của quán ăn ăn đúc là hàng hàng chữ " % và sản phẩm và và bánh gà ăn 179k ..
Human: nổi bật trên tấm bảng hiệu của quán ăn đông đúc khách là hàng chữ 100% hải sản tươi sống và cơ trưa văn phòng 179k.

id: 00000003641
M2 Transformers: Ở phía bên phải của bức ảnh nh, có một cây cột đèn màu được sáng vào đèn màu xanh và thời ở một cái hiệu màu hình trụ được màu xanh dương đồng gắn ở phía bên đường cột đèn n.
M4C-Captioner: ở phía bên phải của bức ảnh, có một cây cây đèn đang được sáng vào đèn cây xanh được thời ở một cây hiệu màu cây chữ được màu xanh lá được treo ở bên bên trái...
Human: trong bức ảnh đang chụp trung tâm mua sắm saigon centre và ở phía bên trái của bức ảnh có một chiếc xe máy đang chạy trên đường và trên chiếc xe máy đó có một người đàn ông mặc áo màu đỏ đang ngồi phía sau xe.

id: 000000000234
M2 Transformers: Một cửa hàng có hóa có treo hàng tạp biển hiệu dòng chữ màu mart màu trên
M4C-Captioner: cửa hàng trước một cửa hàng h@@ mitex óa có biển màu màu xanh dương chữ màu trắng.
Human: biển có dòng chữ milex màu trắng trên nền màu xanh được treo bên phải của một cửa hàng tạp hóa mà phía trước có một người phụ nữ mặc áo xanh đang đứng.

Fig. 3. Illustration of positive and negative predictions from different models on TextCaps validation set. M4C-Captioner with OCR inputs results is more better.

4.4 Error analysis

Based on the observation and analysis of the experimental results on the test set, we have identified and evaluated several key challenges that our model commonly encounters. We propose the following solutions to enhance the performance of the image captioning model. Below is a detailed description of these challenges and our recommendations: **(1) Difficulty in handling complex descriptions:** Our model often struggles with the complex structure of image descriptions, particularly in cases that demand high inference capabilities. To address this challenge, we suggest expanding the dataset by collecting additional samples with more complex sentence structures. Additionally, we encourage the use of state-of-the-art models in this domain to boost the processing capabilities of our model; **(2) Labeling errors:** The current dataset still exhibits some noise-related issues related to abbreviations and spelling errors, impacting the model's overall quality. To tackle this problem, we propose conducting a thorough examination and standardization of the labeling process to minimize errors and improve data quality; **(3) Token merging capability from OCR:** Our current model has not yet achieved strong proficiency in combining tokens from OCR into complete sentences. We suggest strengthening the token merging ability by either improving the OCR model or applying text normalization techniques to ensure the quality of the tokens; **(4) Limitations in dataset size and computational resources** Moreover, the available dataset for our study is constrained in size, limiting the scope and diversity of our analyses. Additionally, the computational requirements and associated costs for our experiments are substantial, posing challenges to scalability. We acknowledge that the current OCR models still face challenges in terms of accuracy, leading to the identification of low-quality tokens by our model. Nevertheless, compared to models without text reading capabilities, our model has shown significant improvement. These recommendations lay the groundwork for our ongoing research, aiming to achieve optimal performance in generating comprehensive image descriptions.

5 Conclusion and Further Improvements

We have created ViTextCaps, which consists of 23,000 Vietnamese descriptions for 5,289 images. Our analysis, conducted through both automated and human studies, uncovers several novel technical challenges introduced by ViTextCaps. When utilizing the M4C-Captioner model, we achieve outstanding results, surpassing conventional models in the field of image captioning, with a CIDEr score reaching 1.9496, indicating significant potential for future development. In addressing these limitations, our future work will focus on several avenues for improvement. Firstly, we plan to explore different models for feature extraction to enhance the overall efficiency of our system. Secondly, there is a need to augment the dataset, aiming to not only increase its size but also improve its quality. Lastly, we intend to implement experiments using well-established and widely recognized benchmark methods, ensuring robust evaluations and comparisons in our future research endeavors.

Acknowledgements: We would also like to express our gratitude to the academic mentors for their valuable discussions and insights. In addition, we would like to thank our workers for their cooperation.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question

answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6077–6086 (2018), cited in sections 1, 2, 9, 10, 11

2. Bigham, J.P., Jayant, C., Ji, H.Y., Little, G., Miller, A., Miller, R.C., Tatarowicz, A., White, B., White, S., et al.: Vizwiz: Nearly real-time answers to visual questions. In: Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology. pp. 333–342. ACM (2010), 1, 4
3. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4291–4301 (2019)
4. Borisyuk, F., Gordo, A., Sivakumar, V.: Rosetta: Large scale system for text detection and recognition in images. In: ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 71–79. ACM (2018), 1, 4, 5, 7, 19
5. Bui, D.C., Nguyen, N.H., Nguyen, K.: Uit-openviic: A novel benchmark for evaluating image captioning in vietnamese (2023)
6. Chen, J., Li, B., Xue, X.: Scene text telescope: Text-focused scene image super-resolution (2020)
7. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10578–10587 (2020)
8. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation. pp. 376–380 (2014), cited in section 10
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
10. Fujitake, M.: Dtrocr: Decoder-only transformer for optical character recognition (2020)
11. Goyal, P., Mahajan, D.K., Gupta, A., Misra, I.: Scaling and benchmarking self-supervised visual representation learning. In: International Conference on Computer Vision (2019), arXiv preprint arXiv:1905.012
12. Hoang Lam, Q., Duy Le, Q., Van Nguyen, K., Luu-Thuy Nguyen, N.: Uit-viic: A dataset for the first evaluation on vietnamese image captioning. arXiv e-prints pp. arXiv–2002 (2020)
13. Hu, R., Singh, A., Darrell, T., Rohrbach, M.: Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9992–10002 (2020)
14. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4634–4643 (2019)
15. Huang, M., Liu, Y., Peng, Z., Liu, C., Lin, D., Zhu, S., Yuan, N., Ding, K., Jin, L.: Swintextspotter: Scene text spotting via better synergy between text detection and text recognition. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4593–4603 (2022)
16. Khanh, T., Nguyen, A., Ân, L.T., Nguyen, K.: Vivqa: Vietnamese visual question answering (11 2021)
17. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004), cited in section 10
18. Lyu, P., Zhang, C., Liu, S., Qiao, M., Xu, Y., Wu, L., Yao, K., Han, J., Ding, E., Wang, J.: Maskocr: Text recognition with masked encoder-decoder pretraining (2020)

19. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
20. Nguyen, D.Q., Nguyen, A.T.: Phobert: Pre-trained language models for vietnamese. arXiv preprint arXiv:2003.00744 (2020)
21. Nguyen, N., Vo Tran, D., Nguyen, K., Nguyen, N.: Openvivqa: Task, dataset, and multimodal fusion models for visual question answering in vietnamese (05 2023)
22. Nguyen, T.T., Nguyen, L.H., Pham, N.T., Nguyen, L.T., Do, V.H., Nguyen, H., Nguyen, N.D.: viecap4h-vlsp 2021: Vietnamese image captioning for healthcare domain using swin transformer and attention-based lstm. arXiv preprint arXiv:2209.01304 (2022)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318 (2002), cited in section 10
24. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems. pp. 91–99 (2015), cited in sections 9, 10
25. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2556–2565 (2018)
26. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 742–758. Springer (2020)
27. Singh, A., Natarjan, V., Shah, M., Jiang, Y., Chen, X., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8317–8326 (2019)
28. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4566–4575 (2015), cited in section 10
29. Yu, D., Li, X., Zhang, C., Liu, T., Han, J., Liu, J., Ding, E.: Towards accurate scene text recognition with semantic reasoning networks (2020)