ViConsFormer: Constituting Meaningful Phrases of Scene Texts using Transformer-based Method in Vietnamese Text-based Visual Question Answering

Nghia Hieu Nguyen[1,3], Tho Thanh Quan[1,3], Ngan Luu-Thuy Nguyen[2,3]
[1]Ho Chi Minh city University of Technology,
[2]University of Information Technology,
[3]Vietnam National University, Ho Chi minh city, Vietnam,
{nhnghia.sdh231,qttho}@hcmut.edu.vn, ngannlt@uit.edu.vn
Correspondence: Tho Thanh Quan, Ngan Luu-Thuy Nguyen

Abstract

Text-based VQA is a challenging task that requires machines to use scene texts in given images to yield the most appropriate answer for the given question. The main challenge of text-based VQA is exploiting the meaning and information from scene texts. Recent studies tackled this challenge by considering the spatial information of scene texts in images via embedding 2D coordinates of their bounding boxes. In this study, we follow the definition of meaning from linguistics to introduce a novel method that effectively exploits the information from scene texts written in Vietnamese. Experimental results show that our proposed method obtains state-of-the-art results on two large-scale Vietnamese Text-based VQA datasets. The implementation can be found at this link.

1 Introduction

Multimodal learning, particularly vision-language tasks, has recently attracted the attention of the research community. Visual Question Answering (VQA) (Antol et al., 2015) is one of the well-known tasks in vision-language studies. This task gives the machines a question and an image. The machines are required to find the evidence in the image to answer the given question.

Text-based VQA (Singh et al., 2019) is an advanced version of the VQA task in which, besides the visual information in the images, the machines are required to incorporate the information of scene texts for more accurate answers.

Various datasets were constructed for researching Text-based VQA tasks, especially in high-resource languages such as English (Antol et al., 2015; Goyal et al., 2016; Singh et al., 2019; Biten et al., 2019; Mathew et al., 2020). However, there is a limited number of hight-qualified and annotated datasets for researching this task in Vietnamese (Tran et al., 2021; Nguyen et al., 2023; Luu-Thuy Nguyen et al., 2023; Tran et al., 2023; Nguyen et al., 2024; Pham et al., 2024).

On the other hand, the main challenge of Text-based VQA is exploiting the meaning of scene texts available in the images so that deep learning methods can recognize them and depend on them to provide the most appropriate answers. They propose to tackle this challenge by introducing several modules (Biten et al., 2021; Fang et al., 2023; Kil et al., 2022). However, most of these modules explore the spatial information of scene texts in images via their bounding boxes and their meaning was obtained by using embedding layers from pre-trained language models (Hu et al., 2019; Kant et al., 2020; Gao et al., 2020; Biten et al., 2021; Fang et al., 2023).

In this study, we inspire the definition of meaning from American Distributionalsm, a field of study in linguistics, and recent works on the Vietnamese lexical system (Giáp, 2008, 2011; Xuan, 1998; Châu, 2007) to propose a novel method, **Vi**etnamese **Cons**tituent Trans**Former** (ViConsFormer), which effectively incorporate the meaning of Vietnamese scene texts to yield answers.

Our extensive experiments on the two Text-based VQA datasets in Vietnamese show that our proposed method outperforms previous baselines and proposes several research directions for future studies.

2 Related works

2.1 Datasets

Former studies in VQA defined this task as answering questions relevant to objects in images. Antol et al. (Antol et al., 2015) first introduced the VQA task by publishing the VQA dataset.

This novel way of treatment on the VQA dataset results in the language prior phenomenon as indicated by Goyal et al. (Goyal et al., 2016). This phenomenon describes that VQA methods tend to yield answers by recognizing the pattern of questions rather than based on evidence from given images.

To overcome the language prior phenomenon, (Goyal et al., 2016) introduced the VQAv2 dataset. This dataset is the rebalanced version of the VQA dataset constructed by balancing the number of answers belonging to particular patterns of questions.

Making further steps from VQAv2, (Singh et al., 2019) introduced a novel form of VQA task, which is named Text-based VQA tasks in later studies (Hu et al., 2019; Biten et al., 2021; Li et al., 2023; Fang et al., 2023; Kil et al., 2022). In particular, Text-based VQA tasks require the machines to understand scene texts beside objects in the images and use these scene texts to give the respective answers. Text-based VQA tasks become significantly challenging as the additional modality of scene texts and recently raised attention from many researchers (Hu et al., 2019; Biten et al., 2021; Li et al., 2023; Fang et al., 2023; Kil et al., 2022).

Although there are numerous VQA datasets in English, there are few VQA datasets in Vietnamese. Tran et al. (Tran et al., 2021) first introduced the ViVQA dataset, the first dataset for researching VQA in Vietnamese. Later on, various studies released many datasets, particularly UIT-EVJVQA (Luu-Thuy Nguyen et al., 2023), OpenViVQA (Nguyen et al., 2023) and ViClever (Tran et al., 2023) for VQA task as well as ViTextVQA (Nguyen et al., 2024) and ViOCRVQA (Pham et al., 2024) for Text-based VQA task.

2.2 Methods

Former methods share the same architecture that uses pre-trained CNN-based models to extract image features and RNN-based methods to learn the questions with integrated image features for producing answers (Lu et al., 2016; Yang et al., 2015).

In addition, with the introduction of the attention mechanism (Vaswani et al., 2017), former VQA methods attempted to provide this technique with the assumption of learning the attention relation between images and questions. Typical VQA methods for this approach can be categorized as Co-Attention (Lu et al., 2016; Yu et al., 2019), or Stack Attention (Yang et al., 2015).

The development of BERT (Devlin et al., 2019) provides another architecture that lots of studies inspired as well as introduced numerous novel methods such as (Lu et al., 2019; Li et al., 2019b; Tan and Bansal, 2019; Su et al., 2019; Zhou et al., 2019; Li et al., 2019a; Cho et al., 2020).

Another way of learning the correlation between images and questions is to define multilinear functions that accept features of questions and images as inputs. Various studies followed this approach and introduced deep learning methods using multilinear functions instead of Transformer (Kim et al., 2018; Do et al., 2019).

However, most of the VQA methods in English were defined as answer-selection models. Recently, text-based VQA datasets were introduced, and these answer-selection methods can not model effectively Text-based VQA datasets because of the diverse forms of answers. In particular, (Nguyen et al., 2023) defined the open-ended VQA task with the publication of the OpenViVQA dataset in Vietnamese. They showed that former VQA methods are challenging to model in this novel form of VQA task. VQA methods using language models are then developed to tackle the challenging of Text-based VQA tasks and open-ended VQA tasks (Nguyen et al., 2023).

The main challenge of the Text-based VQA task is how to model scene texts in images to yield a good answer. Many T5-based VQA methods were introduced with particular modules that try to learn the meaning and spatial relations among scene texts (Biten et al., 2021; Kil et al., 2022; Fang et al., 2023)

3 ViConsFormer - Vietnamese Constituent Transformer

3.1 Preliminaries

3.1.1 Vietnamese Scene Texts

Scene texts in images taken in Vietnam have the following rule in general: scene texts on the same line are in the same meaningful constituents. For instance, in Figure 1, there are three lines of scene texts on the truck: *VinShop x VinID* indicates the cooperation of the two companies, *Tạp hóa* (grocery) indicates the kind of business of the two companies, and *Thời công nghệ* (technological times) points out the characteristic of the grocery. There does not exist the situation where all scene texts are in the same line, but they are meaningless.

However, current Text-based VQA methods receive scene texts as the line of tokens ordered by the spatial information (the 2D coordinates of bounding boxes) (Biten et al., 2021; Fang et al., 2023; Kant et al., 2020; Gao et al., 2020; Hu et al., 2019). There is no signal to determine which constituents

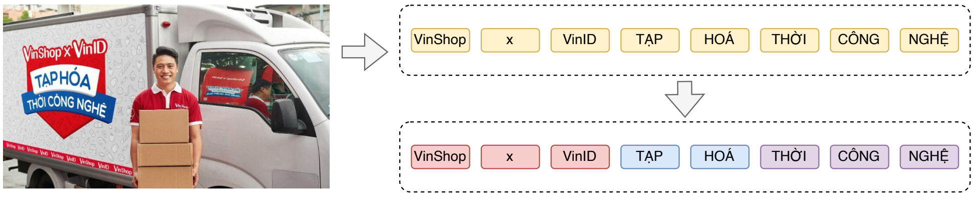


Figure 1: Forming meaningful constituents in the sequence of OCR tokens.

each scene text token belongs to. SaL (Fang et al., 2023) tackled this challenge by providing an additional special token <*context*> between scene text tokens. These tokens learn how to represent the start and end positions of every meaningful constituent.

We approach this challenge in different ways. From our observations, we find that meaningful constituents include complete lexical units, which we call Vietnamese words. To this end, we introduce a novel method that describes the Vietnamese words and hence describes the meaningful constituents of scene text of images taken in Vietnam (Figure 1). We continue the in-depth discussion of how to describe words from the line of scene text tokens in the following Section.

3.1.2 Meaning representation

Recent studies (Yang et al., 2015; Biten et al., 2021; Gao et al., 2020; Kant et al., 2020) addressed the Text-based VQA task by proposing a module that incorporates the position of scene texts in images via the coordinates of their bounding boxes. However, the meaning of scene texts is not reflected by their spatial positions. One attempt to explore the meaning of scene texts is to use the embedding layer of pre-trained language models such as BERT (Devlin et al., 2019) and T5 (Raffel et al., 2019) as in (Biten et al., 2021; Hu et al., 2019; Fang et al., 2023; Gao et al., 2020). This approach has a limitation in that not all scene texts in images are available in the fixed vocab of the pre-trained language models. When tokens are not in the vocab, pre-trained language models usually segment them into subwords (Wang et al., 2019). This way of representation tends to introduce the ambiguity of scene texts to Text-based VQA methods.

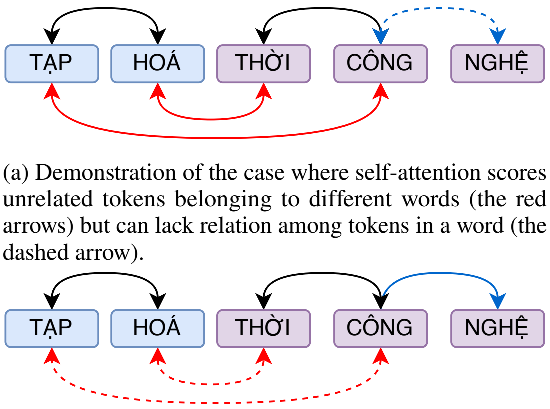
Another approach is constructing a pre-trained model, particularly for scene text representation as in (Kil et al., 2022). However, training a pre-trained model requires high-cost computational facilities.

In our study, we approach the meaning representation of scene texts following the study of American Distributionalism (Bloomfield, 1933; Harris, 1951) in linguistics. This field of study in linguistics researches language by observing how it appears. They think research in linguistics must be done via observable and measurable units. Linguisticians should avoid falling into unobservable things such as semantics or meanings. They describe languages as the distributions of their constituents, and meaning is defined as the consequence of how words are distributed and how they appear together in sentences. For instance, given the sentence *Everyone in the room knows at least two languages* and the sentence *At least two languages are known by everyone in the room*, these two sentences are different in terms of meaning although they are formed from the same set of words (Chomsky, 2014).

In addition, Vietnamese lexical structure differs from English. Words in English include one or more than two syllables. While in Vietnamese, words contain one or more syllables, and spaces separate these syllables. For instance, *đại lý* in Vietnamese is a word containing two tokens and two syllables, while in English, it means *agency*, has three syllables and one token. In other words, an English token can be translated into more than one token in Vietnamese.

We follow the recent advantages of the attention mechanism in English. From various studies (Vaswani et al., 2017; Bahdanau et al., 2014; Luong et al., 2015), the attention mechanism can describe how words are relevant to each other. However, as analyzed previously, Vietnamese words may include more than one syllable, encoded as tokens in sentences. The attention mechanism has no constraint in how it can form attentive connections. Therefore, in Vietnamese sentences, one token in this word will attend to one token in another, which does not yield any meaning (Figure 2a).

To this end, we introduce the Constituent Module. This module constructs two components: the attention score matrix $\mathcal{A}$ and the constituent score matrix $\mathcal{C}$. The constituent score matrix $\mathcal{C}$ describes



(a) Demonstration of the case where self-attention scores unrelated tokens belonging to different words (the red arrows) but can lack relation among tokens in a word (the dashed arrow).

(b) Our Constituent Module was proposed to re-correct the attention scores via the constituent scores (the blue line) while removing the unnecessary relations (dashed red arrows).

Figure 2: An example of a noun phrase in Vietnamese. The translated phrase in English: grocery in technological times.

the words of scene texts in images by highlighting which tokens belong to a word.

Moreover, as there are no technical constraints in the attention mechanism, we might have two tokens in different words, but they can be scored to attend to each other. The constituent score matrix $\mathcal{C}$ plays the role of re-correcting the attention score matrix $\mathcal{A}$ so that there are no unnecessary connections among tokens belonging to different words (Figure 2b). The description of how we construct $\mathcal{C}$ was detailed from equation 1 to equation 5.

3.2 Overall architecture

The main contribution of ViConsFormer is the Constituent module. This module is proposed to describe the meaning of scene texts, as discussed in the previous section. In general, our method has five components: Image Embedding module, Question Embedding module, Scene Text Embedding module, Constituent module, and Multimodal backbone (Figure 3).

3.2.1 Constituent module

The Constituent module includes two components: the multi-head attention (Vaswani et al., 2017) determining attention score $\mathcal{A}$ and Constituent formation determining constituent score $\mathcal{C}$.

In Vietnamese morphology, syllables in words have two kinds of semantic relations (Giáp, 2008, 2011):

- Syllables in a word contribute their meanings equally to the overall meaning of that word. For instance, *quần áo* means clothes in general, compounding *quần* (paints in general) and *áo* (shirts in general).
- One syllable defines the core meaning of the word, and other syllables play the role of modifiers. These modifiers narrow down the meaning of the main syllable so that the meaning of the whole word is more particular. For instance, *nhà ăn* (cafeteria) includes syllable *nhà* (houses in general) and *ăn* (dining in general).

Given the sequence of scene texts $f_{ocr} = (f_1^{ocr}, f_2^{ocr}, ..., f_n^{ocr})$ obtained from the embedding layer of ViT5 (Phan et al., 2022) as input, we model these kinds of semantic relations by defining a bilinear function:

$$r_{k,k+1} = f(f_k^{ocr}, f_{k+1}^{ocr}) = f_k^{ocr} W (f_{k+1}^{ocr})^T \tag{1}$$

where W is the learnable parameter. Then with every token ith, we describe the semantic relations with its neighbor tokens $(i-1)th$ and $(i+1)th$ as:

$$pr_{i-1,i}, pr_{i,i+1} = softmax(r_{i-1,i}, r_{i,i+1}) \tag{2}$$

If token ith has semantic relations with token $(i-1)th$, and token $(i-1)th$ is in another word, then we expect $pr_{i-1,i} > pr_{i,i+1}$ (and vice versa). In the case token ith has semantic relations with both token $(i-1)th$ and $(i+1)th$, the mass of $pr_{i-1,i}$ and $pr_{i,i+1}$ determine how much relevant these tokens share (contributing equally to the overall meaning, or main-secondary meaning contribution, or there is no connection among these tokens).

In addition, as the consequence of the asymmetry of matrix multiplication, we have $pr_{k,k+1} \neq pr_{k+1,k}$ while they describe the same idea. To this end, we define the probability P_k to measure the semantic relations of token k with its neighbor tokens. P_k is obtained by averaging over $pr_{k,k+1}$ and $pr_{k+1,k}$:

$$P_k = \sqrt{pr_{k,k+1} \times pr_{k+1,k}} \tag{3}$$

Defining $\mathcal{C}_{ij}$ the probability of "tokens from the position ith to the position jth are in the same constituent", together with the definition of P_k, we have:

$$\mathcal{C}_{ij} = \prod_{k=i}^{j-1} P_k \tag{4}$$

It is worth noting that $P_k \in [0, 1]$, hence $\mathcal{C}_{ij}$ rapidly converges to 0 when $k \to \infty$, which results

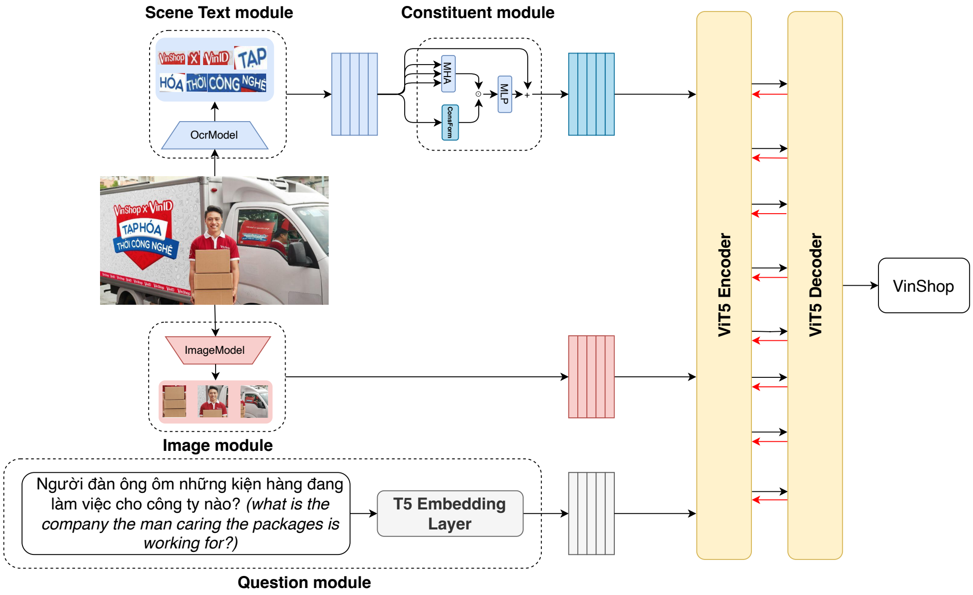


Figure 3: The overall architecture of the ViConsFormer.

in the gradient vanishing. To avoid this problem, we re-formulate $\mathcal{C}_{ij}$ as:

$$\begin{aligned}\mathcal{C}_{ij} &= exp(log(\mathcal{C}_{ij})) \\ &= exp\left(log\left(\prod_{k=i}^{j-1} P_k\right)\right) \\ &= exp\left(\sum_{k=i}^{j-1} log(P_k)\right)\end{aligned} \tag{5}$$

To construct the attention score matrix $\mathcal{A}$, we perform the self-attention by defining the query Q, key K, and value V as:

$$Q = W_q f_{ocr}$$

$$K = W_k f_{ocr}$$

$$V = W_v f_{ocr}$$

where $W_q, W_k, W_v \in \mathbb{R}^{d_{model} \times d_{model}}$ are learnable parameters.

The attention scores over detected scene texts are determined as follows:

$$\mathcal{A} = softmax(\frac{QK^T}{\sqrt{d_{model}}})$$

The final attention score matrix is determined by providing $\mathcal{A}$ with the constituent score matrix $\mathcal{C}$:

$$\mathcal{S} = \mathcal{A} \odot \mathcal{C} \tag{6}$$

As both $\mathcal{C}$ and $\mathcal{A}$ have the form of exponential functions, it is worth noting that the constituent scores are not added to the attention scores intuitively by summation but by element-wise multiplication.

Finally, the output of scene text features is determined as follows:

$$f^{ocr}_{out} = \mathcal{S}V \tag{7}$$

3.3 Image Embedding module

We follow the module for object features representation of SaL to represent the features of images. In particular, the features of objects $f_{obj} = (f^{obj}_1, f^{obj}_2, ..., f^{obj}_n)$ available in images are determined as follows:

$$\begin{aligned}f^{obj}_i &= ViT5LN(W^{obj}_{fr} x^{obj,fr}_i) \\ &+ ViT5LN(W^{obj}_{bx} x^{obj,bx}_i)\end{aligned} \tag{8}$$

where W^{obj}_{fr} and W^{obj}_{bx} are trainable parameters, $ViT5LN$ is the normalization layer of ViT5, and $x^{obj,fr}_i$ is the features of object ith in image.

Unlike SaL, we do not consider the features of object tags, as their appearance in our proposed methods does not yield any significant improvement in scores. We will provide more numerical information about this statement in Section 4.6.

3.4 Scene Text Embedding module

To obtain features of scene texts in images, we follow SaL to determine these features:

$$\begin{aligned} f_i^{ocr} &= ViT5LN(W_{fr}^{ocr} x_i^{ocr,fr}) \\ &+ ViT5LN(W_{bx}^{ocr} x_i^{ocr,bx}) \\ &+ W_{ViT5}^{ocr} x_i^{ViT5} \end{aligned} \tag{9}$$

where W_{fr}^{ocr}, W_{bx}^{ocr}, and W_{ViT5}^{ocr} are trainable parameters, $ViT5LN$ is the normalization layer of ViT5, $x_i^{obj,fr}$ is the features of object ith in image, and x_i^{ViT5} is the text embedding of scene text token ith produced by the embedding layer of ViT5.

3.5 Question Embedding module

Following LaTr (Biten et al., 2021), the questions are embedded into $f_Q = (q_1, q_2, ..., q_L)$ using the embedding layer of ViT5 where each position $q_i \in \mathbb{R}^d$.

3.6 Multimodal backbone

Following previous works (Biten et al., 2021; Fang et al., 2023; Kil et al., 2022), we use the T5-based pre-trained model as the multimodal backbone. However, as our experiments were conducted on the Vietnamese Text-based VQA dataset, we provide ViConsFormer with ViT5 (Phan et al., 2022) pre-trained language models.

The input to the ViT5 backbone is defined as the fused features f_f constructed by concatenating f_{obj}, f_{ocr}, and f_Q:

$$f_f = [f_{obj}; f_{ocr}; f_Q]$$

4 Experiments

4.1 Datasets

In this work, we evaluate our proposed methods on the two datasets ViTextVQA (Nguyen et al., 2024) and ViOCRVQA (Pham et al., 2024).

The ViTextVQA dataset was constructed by asking questions and answering answers relevant to scenario images. These images were street views in Viet Nam (Nguyen et al., 2024). Scene texts in this dataset are diverse in positions, colors, light conditions, transformations, shapes, and meaning. As indicated in (Nguyen et al., 2024), the smaller size of scene texts in the images lead to more challenges in producing answers.

The ViOCRVQA dataset was constructed semi-automatically by collecting book covers from websites (Pham et al., 2024). The authors built question templates, then filled in these templates and extracted answers via the corresponding metadata of the books.

4.2 Metrics

We follow the experiments on the two datasets ViTextVQA (Nguyen et al., 2024) and ViOCRVQA (Pham et al., 2024) to use the Exact Match (EM) and F1-token as main metrics in our evaluation.

Accordingly, let $P = \{p_1, ..., p_m\}$ is the predicted answers, and $G = \{g_1, ..., g_n\}$ is the truth answers. The M of each predicted-truth answer is determined as follows:

$$EM = \delta_{P,G}$$

where $\delta_{x,y}$ is the Kronecker symbols which $\delta_{x,y} = 1$ when $x = y$ and 0 otherwise.

The F1-Token metric is defined as the harmonic mean of the Precision and Recall (in token level) as:

$$Pr = \frac{P \cap G}{P}$$

$$Re = \frac{P \cap G}{G}$$

$$\text{F1-Token} = \frac{2PrRe}{Pr + Re}$$

The overall EM and F1-Token are averaged over all predicted-truth answers of the whole dataset.

4.3 Configuration

In our experiment, we trained the ViConsFormer following the previous studies on ViTextVQA and ViOCRVQA datasets (Nguyen et al., 2024; Pham et al., 2024) that used ViT5 (Phan et al., 2022) as the multimodal backbone. For the $ImageModel$ we deployed the VinVL (Zhang et al., 2021) pre-trained image models. We used SwinTextSpotter (Huang et al., 2022) to obtain Vietnamese scene texts from images to extract their detection features and recognition features. The ViConsFormer was trained in a single run, using Adam (Kingma and Ba, 2014) as optimizer on an NVIDIA A100 80GB GPU. The batch size was set to 32 and the learning was set to $1e^{-4}$. We applied the early stopping technique to train ViConsFormer.

4.4 Baselines

To evaluate the effectiveness of our proposed ViConsFormer on the Vietnamese Text-based VQA dataset, we compared this method with the following baselines:

- M4C (Hu et al., 2019): M4C is the first vision-language learning task that was constructed based on the Transformer architecture (Vaswani et al., 2017). Its multimodal backbone is BERT (Devlin et al., 2019). M4C approaches the text-based VQA task by sequentially generating tokens to form the answers. Tokens of answers can be obtained from the vocabulary or copied from the scene texts available in the images using the Pointer Network (Hu et al., 2019) module.

- LaTr (Biten et al., 2021): This is the first method that integrated spatial information of scene texts into the multimodal backbone. They encoded the coordinates of the bounding boxes into 4-dimensional vector space, then projected them directly to the latent space of the multimodal backbone and added them to the features of scene texts. Unlike M4C, LaTr proposed using T5 (Raffel et al., 2019) as its multimodal backbone and using a subword tokenizer to encode scene texts. Scene texts in the images that are not available in the vocabulary of the T5 pre-trained model are subsegmented into sequences of chunks. Hence, instead of copying scene texts from images via a particular module, it learns how to form out-of-vocabulary scene texts from respective subwords.

- PreSTU (Kil et al., 2022): Instead of modeling the relation among scene texts via their spatial relations, PreSTU pre-trained the T5 backbone to approximate the distribution of the scene texts. The particular technique of PreSTU differs from other methods in that they sort the scene texts in left-right top-bottom orders.

- SaL (Fang et al., 2023): SaL proposed integrating the labels of objects and tokens of scene texts to their respective visual features. These labels and tokens are embedded by the embedding layer of the T5 backbone to yield the textual meaning of the objects and scene texts. Moreover, instead of encoding the coordinates of bounding boxes, they introduce another way, which is to measure the relative position among scene texts in the images.

- BLIP-2 (Li et al., 2023): BLIP-2 proposed the Q-Former module, which is fine-tuned to connect the latent space between two frozen pre-trained models: pre-trained language model and pre-trained image model. This method was pre-trained using three objective functions: Image-Text matching, Image-Text Contrastive learning, and Image-grounded Text generation. The adaption of BLIP-2 is to fine-tune the Q-Former on the downstream tasks while keeping the pre-trained image and language models frozen.

4.5 Results

Table 1: Main results of the ViConsFormer and the baselines on the ViTextVQA and ViOCRVQA datasets. The scores of baselines are obtained from previous studies (Nguyen et al., 2024; Pham et al., 2024).

#	Method	ViTextVQA		ViOCRVQA	
		F1-token	**EM**	**F1-token**	**EM**
1	M4C	30.04	11.60	-	-
2	BLIP2	37.78	15.01	55.23	21.45
3	LaTr	43.13	20.42	60.97	30.80
4	PreSTU	43.81	20.85	66.25	33.86
5	SAL	44.89	20.97	67.25	**39.08**
6	ViConsFormer (ours)	**45.58**	**22.72**	**70.92**	37.65

In general, the evaluation scores on the ViTextVQA dataset are lower than those on the ViOCRVQA dataset. This can be explained by the questions in the ViTextVQA dataset being annotated manually by Vietnamese native speakers, while questions in the ViOCRVQA were constructed semi-automatically using constructed templates. Therefore, the patterns of questions in the ViOCRVQA dataset are easier to explore. In addition, images from the ViTextVQA dataset are scenery views in Viet Nam, including street signs, signboards, addresses, banners, places, etc. Scene texts available in images from ViTextVQA are complicated under various transformations, colors, light conditions, and sizes and are relevant to diverse objects. In the ViOCRVQA dataset, scene texts are more clarified and belong to particular categories such as titles, names of authors, publishers, and translators (Pham et al., 2024).

On the ViTextVQA dataset, our proposed methods decisively outperformed all the given baselines. In particular, M4C using BERT as its multimodal backbone yielded the lowest scores. Text-based VQA methods using T5 as their multimodal VQA backbone significantly achieved higher scores.

On the ViOCRVQA, our method significantly outperforms all the baselines on the F1-Token metric. However, on EM, our method drops down its score compared to SaL.

4.6 Ablation study

4.6.1 Ablation study on $\mathcal{A}$ and $\mathcal{C}$

Table 2: Ablation study on attention score matrix $\mathcal{A}$ and constituent score matrix $\mathcal{C}$

Dataset	$\mathcal{A}$	$\mathcal{C}$	F1-token	EM
ViTextVQA	✓	✗	0.4309	0.2038
	✗	✓	0.4025	0.1740
	✓	✓	**0.4558**	**0.2272**
ViOCRVQA	✓	✗	0.6503	0.3204
	✗	✓	0.6172	0.3132
	✓	✓	**0.7092**	**0.3765**

The Constituent module determines two matrices: the attention score matrix $\mathcal{A}$ and the constituent score matrix $\mathcal{C}$. We expect the constituent score matrix will re-correct the unnecessary relations scored by the attention score matrix to represent the meaning of scene texts in images appropriately. To show how these two matrices interact with each other, we conduct experiments in case only the attention score matrix is calculated, and only the constituent score matrix is calculated.

According to Table 2, the Constituent module with only attention score matrix $\mathcal{A}$ performed better than when being replaced by the constituent score matrix $\mathcal{C}$. However, having the constituent score matrix $\mathcal{C}$ to re-correct the attention score matrix $\mathcal{A}$ leads to significant improvement in both metrics.

4.6.2 The necessity of object labels

Table 3: Ablation study of the ViConsFormer on the Scene Text module and Image module. *Labels* indicate the labels of detected objects and *Tokens* indicate the tokens of detected scene texts.

Dataset	Metrics	Labels		Tokens	
		✔	✘	✔	✘
ViTextVQA	F1-Token	45.58	↓0.25	45.58	↓1.22
	EM	22.72	↓0.40	22.72	↓1.19
ViOCRVQA	F1-Token	70.92	↑0.12	70.92	↓2.55
	EM	37.65	↓0.52	37.65	↓1.60

As mentioned in Section 3, in the Image module of ViConsFormer, the labels of detected objects are not required but the tokens of scene texts. To show this claim, we conducted an ablation study for ViConsFormer on the ViTextVQA and ViOCRVQA datasets.

As indicated in Table 3, there is no significant performance degradation in both F1-Token and EM scores if we do not provide ViConsFormer with object labels. However, ViConsFormer obtained significantly lower scores when it did not see scene text tokens. These results indicate that not the labels of detected objects but tokens of detected scene texts influence the overall performance.

5 Conclusion

In this study, we introduced the Constituent module. Hence, the ViConsFormer, inspired by the SaL method, approaches the main challenge of Text-based VQA in Vietnamese in a novel way. Experimental results indicate that our proposed method is effective on both Text-based VQA datasets.

6 Limitations

Although our ViConsFormer addresses the challenge of Text-based VQA task by proposing the Constituent module, this method has some limitations that need to be improved and studied in the following studies.

The first limitation is our assumption of linearity while modeling the semantic relationship between two continuous scene text tokens. This assumption is proposed for the simplicity in our novel method. It is necessary to explore the form of this semantic relationship and find the appropriate ways of modeling it in subsequent studies.

The second limitation is that we give a naive treatment for the fused futures f_f when passing them forward to the multimodal backbone. There are various ways of obtaining these fused features, such as using the Co-Attention mechanism (Yu et al., 2019; Lu et al., 2016; Yang et al., 2015), or multilinear functions (Do et al., 2019; Kim et al., 2018). We will leave these directions in our future work.

7 Acknowledgement

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under the grant number DS2024-26-01.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R. Manmatha. 2021. Latr: Layout-aware transformer for scene-text vqa. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16527–16537.

Ali Furkan Biten, Rubèn Pérez Tito, Andrés Mafla, Lluís Gómez, Marçal Rusiñol, Ernest Valveny, C. V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4290–4300.

Leonard Bloomfield. 1933. *Language*. Henry Holt.

Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. 2020. X-lxmert: Paint, caption and answer questions with multi-modal transformers. *ArXiv*, abs/2009.11278.

N. Chomsky. 2014. *Aspects of the Theory of Syntax, 50th Anniversary Edition*. Aspects of the Theory of Syntax. MIT Press.

Dỗ Hữu Châu. 2007. *Từ vựng ngữ nghĩa tiếng Việt*. Nhà xuất bản Giáo dục Việt Nam.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Tuong Khanh Long Do, Thanh-Toan Do, Huy Tran, Erman Tjiputra, and Quang D. Tran. 2019. Compact trilinear interaction for visual question answering. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 392–401.

Chengyang Fang, Jiangnan Li, Liang Li, Can Ma, and Dayong Hu. 2023. Separate and locate: Rethink the text in text-based visual question answering. *Proceedings of the 31st ACM International Conference on Multimedia*.

Chenyu Gao, Qi Zhu, Peng Wang, Hui Li, Yuliang Liu, Anton van den Hengel, and Qi Wu. 2020. Structured multimodal attentions for textvqa. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:9603–9614.

Nguyễn Thiện Giáp. 2008. *Từ vựng học tiếng Việt*. Nhà xuất bản Giáo dục Việt Nam.

Nguyễn Thiện Giáp. 2011. *Vấn đề "từ" trong tiếng Việt*. Nhà xuất bản Giáo dục Việt Nam.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127:398 – 414.

Z.S. Harris. 1951. *Methods in Structural Linguistics*. Methods in Structural Linguistics. University of Chicago Press.

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2019. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9989–9999.

Mingxin Huang, Yuliang Liu, Zhenghao Peng, Chongyu Liu, Dahua Lin, Shenggao Zhu, Nicholas Jing Yuan, Kai Ding, and Lianwen Jin. 2022. Swintextspotter: Scene text spotting via better synergy between text detection and text recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4583–4593.

Yash Kant, Dhruv Batra, Peter Anderson, Alexander G. Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. 2020. Spatially aware multimodal transformers for textvqa. In *European Conference on Computer Vision*.

Jihyung Kil, Soravit Changpinyo, Xi Chen, Hexiang Hu, Sebastian Goodman, Wei-Lun Chao, and Radu Soricut. 2022. Prestu: Pre-training for scene-text understanding. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15224–15234.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Neural Information Processing Systems*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *ArXiv*, abs/1908.06066.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019b. Visualbert: A simple and performant baseline for vision and language. *ArXiv*, abs/1908.03557.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Neural Information Processing Systems*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *ArXiv*, abs/1606.00061.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. *ArXiv*, abs/1508.04025.

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, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2020. Docvqa: A dataset for vqa on document images. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208.

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

Quan Van Nguyen, Dan Quang Tran, Huy Quang Pham, Thang Kien-Bao Nguyen, Nghia Hieu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2024. Vitextvqa: A large-scale visual question answering dataset for evaluating vietnamese text comprehension in images. *ArXiv*, abs/2404.10652.

Huy Quang Pham, Thang Kien-Bao Nguyen, Quan Van Nguyen, Dan Quang Tran, Nghia Hieu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2024. Viocrvqa: Novel benchmark dataset and vision reader for visual question answering by understanding vietnamese text in images. *ArXiv*, abs/2404.18397.

Long Phan, Hieu Trung Tran, Hieu Chi Nguyen, and Trieu H. Trinh. 2022. Vit5: Pretrained text-to-text transformer for vietnamese language generation. *ArXiv*, abs/2205.06457.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

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.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *ArXiv*, abs/1908.08530.

Hao Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Conference on Empirical Methods in Natural Language Processing*.

Khanh Quoc Tran, An Trong Nguyen, An Tran-Hoai Le, and Kiet Van Nguyen. 2021. Vivqa: Vietnamese visual question answering. In *Pacific Asia Conference on Language, Information and Computation*.

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. *ArXiv*, abs/2310.18046.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2019. Neural machine translation with byte-level subwords. *ArXiv*, abs/1909.03341.

Hao Cao Xuan. 1998. The problem of phoneme in vietnamese. *Vietnamese studies*.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2015. Stacked attention networks for image question answering. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–29.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6274–6283.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Making visual representations matter in vision-language models. *ArXiv*, abs/2101.00529.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2019. Unified vision-language pre-training for image captioning and vqa. *ArXiv*, abs/1909.11059.