Multimodal Encoder with Gated Cross-attention for Text-VQA Tasks

Wei Yang, Arisa Ueda, and Komei Sugiura
Keio University
{wei.yang,arinko31,komei.sugiura}@keio.jp

Abstract

Visual scene understanding, such as visual question answering (VQA), is expected to improve as it benefits people with disabilities in daily life. The Text-based VQA task as an extension of VQA is more challenging to tackle, in which the questions' answers must relate to the text information with reading and reasoning like humans. In this work, we propose an integrated self- and gated cross-attention encoder module to fuse multi-modalities captured in an image effectively. We evaluated our method on the TextVQA dataset, and the results demonstrated that our model outperformed baseline models on the accuracy evaluation in the text-based VQA task.

1 Introduction

In recent decades, visual and natural language understanding has grown into crucial domains for innovation in Artificial Intelligence (AI) [1, 2], with more and more applications reshaping lifestyles [3]. Such as automatic navigation for guiding vehicles [4], dialogue systems [5–7], etc. In daily life, many visual scenes and questions contain text-related information. Thus, it should be helpful for humans to obtain an accurate answer when they ask a question about the visual scene related to the text, especially for visually impaired people.

The target task of this work refers to constructing a visual question answering (VQA) model that can handle the question-answering problem while requiring reading and reasoning the text in images (TextVQA). And this makes it more complicated and challenging to tackle. For example, in Figure 1, several words exist in the image in different colors (green, brown, and white) to introduce a small bean around the Lake Trasimeno area of Italy. When asked about '*what word is written in white text?*', it is required to generate an answer '*trasimeno*', which is written in white color and difficult to see. And this requires reading and reasoning from all text information in the image.

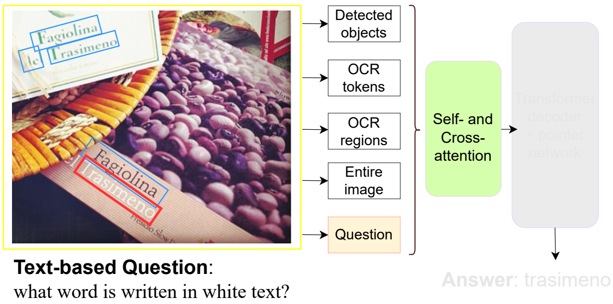

Fagiolina
le
Trasimeno
Fagiolina
Trasimeno
Detected objects
OCR tokens
OCR regions
Entire image
Question
Self- and Cross-attention
Transformer decoder + pointer network
Text-based Question: what word is written in white text?
Answer: trasimeno

Figure 1 Overview of our method for TextVQA: We integrate self- and gated cross-attention mechanisms into a V&L encoder.

In [8], they combine different modalities with a multimodal transformer over a large joint embedding space. Thus it lacks specific modality pair computation, e.g., a pair between the question and the OCR text information. On the other hand, the image-related cross-attention computation needs to be improved in [9]. And the utilization of the global image modality is absent in both works.

We propose a multimodal encoder, which maximizes using multiple modalities in an image and models the relationship with both self- and gated cross-attention mechanisms. Therefore, our model can handle the text-based VQA problem with a stronger visual-language encoder, especially to obtain visually informed question (language) features.

The main contributions of our work are as follows:

- We introduce an additional vanilla attention block for the entire image to complete the utilization of the visual information.
- To obtain richer features for text and visual modalities, we introduce using a pre-trained CLIP [1] model for OCR tokens and the entire image.
- We introduce a Flamingo's [10] gated cross-attention mechanism to further model the relationship between the entire image (visual) and the question (language).

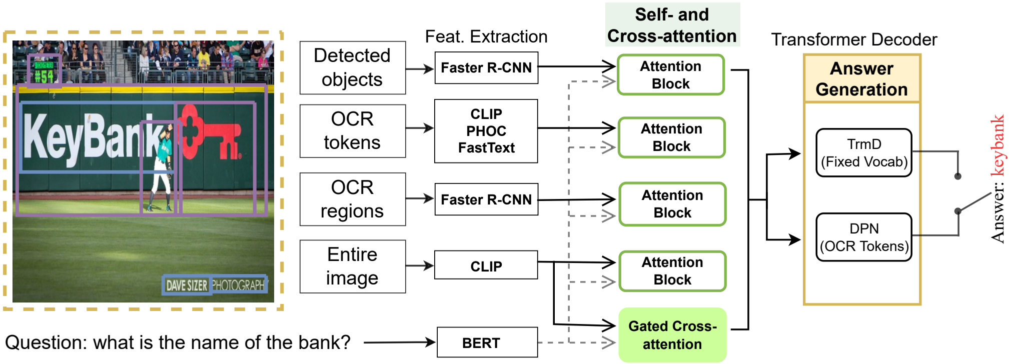


Figure 2 Overview of our proposed encoder-decoder architecture. TrmD and DPN denote the general Transformer decoder and the dynamic pointer network, respectively. The introduced integrated self- and gated cross-attention encoder are given in green.

2 Problem Statement

The focus of this paper is the TextVQA task. The model is expected to predict (output or generate) an answer to a question which should be a deduction based on the text and the visual information in the image. The input and output in our experiments are defined as follows,

- **Input:** An image that contains text and one image-related question. In detail, an image, including the entire image, detected object regions and detected OCR tokens/regions.
- **Output:** An answer that answers the question by reading and reasoning over the text in the image.

3 Proposed Method

Our method is inspired by the M4C [8] and SSbaseline model [11], multimodal Transformer networks that have been successfully applied to the TextVQA task. Figure 2 shows the overview of our proposed self- and gated corss-attention encoder-decoder architecture. Our proposed method is applicable to more than just this work and can be applied to other visual-and-language (V&L) tasks that consider language- and image-related modalities as inputs. Because efficiently fusing and modeling the relationship between visual and textual modalities is considered the basic strategy for solving any V&L problems.

Our model comprises two grand divisions: encoder and decoder modules. The encoder module consists of a sequence of stacked attention blocks with self- and gated cross-attention mechanisms. The Transformer decoder module with a dynamic pointer network (DPN) is built to predict the answers from a fixed vocabulary or the OCR tokens in an image, i.e., the answer generation module.

3.1 Input

The input $\boldsymbol{X}$ to our model is defined as follows:

$$\boldsymbol{X} = \{\boldsymbol{X}_{\mathrm{Q}}, \boldsymbol{X}_{\mathrm{img}}, \boldsymbol{X}_{\mathrm{obj}}, \boldsymbol{X}_{\mathrm{ocr,v}}, \boldsymbol{X}_{\mathrm{ocr,s}}\}$$

where $\boldsymbol{X}_{\mathrm{Q}}$, $\boldsymbol{X}_{\mathrm{img}}$, $\boldsymbol{X}_{\mathrm{obj}}$, $\boldsymbol{X}_{\mathrm{ocr,v}}$, and $\boldsymbol{X}_{\mathrm{ocr,s}}$ denote the question, entire image, detected objects, recognized OCR regions (visual) and OCR tokens (semantic), respectively.

We use the CLIP (RN50x4) model to obtain a 640-dimensional feature vector ($\boldsymbol{x}_{\mathrm{glob}}$) for each entire image.

Moreover, we encode the detected objects in an image through the fc6 layer of Faster R-CNN [12] and fine-tuning the last layer on the TextVQA dataset, $\{\boldsymbol{x}^{(m)}_{\mathrm{obj,fr}} \mid m = 1, \cdots, M\}$, where M denotes the number of the objects considered in the image. The same Faster R-CNN model and fine-tuning are also applied to the OCR region feature extraction, $\{\boldsymbol{x}^{(n)}_{\mathrm{ocr,fr}} \mid n = 1, \cdots, N\}$, where N denotes the number of the OCR tokens considered in the image. We obtain a 2048-dimensional feature vector for each region. The spatial features (4-dimensional bounding box features, e.g., $\boldsymbol{x}^{(m)}_{\mathrm{obj,bx}}$) are also used in our experiments.

The recognized OCR token features $\boldsymbol{x}^{(n)}_{\mathrm{ocr,tok}}$ are made up of (1) character-level Pyramidal Histogram of Characters (PHOC) [13] feature for each OCR token (604-dimensional $\boldsymbol{x}^{(n)}_{\mathrm{ocr,phoc}}$), (2) FastText features $\boldsymbol{x}^{(n)}_{\mathrm{ocr,ft}}$ for the OCR tokens in subword-level (300-dimensional) and (3) 640-dimensional CLIP-based OCR token features $\boldsymbol{x}^{(n)}_{\mathrm{ocr,clip}}$. Note that we use the same pre-trained CLIP model for image and OCR token's feature extraction. We use a three-layer BERT [14] model to obtain a 768-dimensional embedding for each token. This BERT model is fine-tuned during training.

$$\boldsymbol{h}_{\mathrm{glob}} = f_{\mathrm{LN}}(\boldsymbol{W}_{\mathrm{glob}}\boldsymbol{x}_{\mathrm{glob}}) \quad (1)$$

$$\boldsymbol{h}^{(m)}_{\mathrm{obj}} = f_{\mathrm{LN}}(\boldsymbol{W}_{\mathrm{obj,fr}}\boldsymbol{x}^{(m)}_{\mathrm{obj,fr}}) + f_{\mathrm{LN}}(\boldsymbol{W}_{\mathrm{obj,bx}}\boldsymbol{x}^{(m)}_{\mathrm{obj,bx}}) \quad (2)$$

$$\boldsymbol{h}_{\text{ocr,v}}^{(n)} = f_{\text{LN}}(\boldsymbol{W}_{\text{ocr,fr}}\boldsymbol{x}_{\text{ocr,fr}}^{(n)}) + f_{\text{LN}}(\boldsymbol{W}_{\text{ocr,bx}}\boldsymbol{x}_{\text{ocr,bx}}^{(n)}) \tag{3}$$

$$\boldsymbol{h}_{\text{ocr,s}}^{(n)} = f_{\text{LN}}(\boldsymbol{W}_{\text{ft}}\boldsymbol{x}_{\text{ocr,ft}}^{(n)} + \boldsymbol{W}_{\text{ph}}\boldsymbol{x}_{\text{ocr,phoc}}^{(n)} + \boldsymbol{W}_{\text{c}}\boldsymbol{x}_{\text{ocr,clip}}^{(n)}) \tag{4}$$

We apply Layer Normalization $f_{\text{LN}}(.)$ to various extracted features. $\boldsymbol{W}_{.}$ denotes weight matrix.

3.2 Visual-and-language encoders

The first encoder module consists of a stack of attention blocks to model the relationship between the question and other text or visual modalities using self-attention. The inputs of each attention block are the embeddings ($\boldsymbol{X}_{\text{Q}} = \{\boldsymbol{x}_{\text{q}}^{(1)}, \cdots, \boldsymbol{x}_{\text{q}}^{(L)}\}$ of the question sequence, $Q = \{q_i\}_{i=1}^{L}$, and the encoded features of another modality (e.g., $\boldsymbol{h}_{\text{ocr,v}}^{(n)}$ and $\boldsymbol{h}_{\text{glob}}$). The outputs of an attention block refer to the weighted sum of subword-based features ($\boldsymbol{X}'_{Q}$) and the summarizing feature of another modality (e.g., $\boldsymbol{X}'_{\text{ocr,v}}$).

In practice, we firstly input the question features into a fully connected feed-forward network consisting of two convolutions with kernel size 1.

$$\boldsymbol{h}_{\text{q}}^{(i)} = \text{Conv1D}\left(\text{ReLU}\left[\text{Conv1D}(\boldsymbol{x}_{\text{q}}^{(i)}, 1)\right]\right),\ i = 1, \ldots, L \tag{5}$$

Then, the output $\boldsymbol{h}_{\text{q}}^{(i)}$ goes through a self-attention process before it works with other modalities. We define the self-attention operation as follows,

$$\boldsymbol{h}' = \text{SelfAttn}(\boldsymbol{h}) = \text{softmax}\left(\frac{\boldsymbol{h}\boldsymbol{W}_{Q}\boldsymbol{W}_{K}{}^{\top}\boldsymbol{h}^{\top}}{\sqrt{d_k}}\right)\boldsymbol{h}\boldsymbol{W}_{V} \tag{6}$$

where $\boldsymbol{W}_{.}$ denotes a learnable weight, d_k is obtained as d_k = H/heads (**Appendix A.2**). Here, we parallelly perform self-attention on the question modality with independent parameters as the input for different attention blocks.

$$\boldsymbol{\alpha}_i = \text{SelfAttn}(\boldsymbol{h}_{\text{q}}^{(i)}) \tag{7}$$

For each attention block, we obtain the first output, weighted sum of subword-based features ($\boldsymbol{X}'_{Q}$) for the question sequence, as follows,

$$\boldsymbol{X}'_{Q} = \sum_{i=1}^{L} \boldsymbol{\alpha}_i \boldsymbol{x}_{\text{q}}^{(i)} \tag{8}$$

It is also used as the guidance for calculating the cross-modality (e.g., Question and OCR regions) attention weights (see formula (9) and (10)). In this example, the value of n ($n = 1, \cdots, N$) varies with the number of OCR tokens recognized in the image. It also can be the number of detected objects in the image.

$$\boldsymbol{u}_n = \text{ReLU}(\boldsymbol{W}_q\boldsymbol{X}'_{Q}) \odot \text{ReLU}(\boldsymbol{W}_h\boldsymbol{h}_{\text{ocr,v}}^{(n)}) \tag{9}$$

$$\boldsymbol{\beta}_n = \text{SelfAttn}(\boldsymbol{u}_n), n = 1, \ldots, N \tag{10}$$

We obtain the summarizing feature of the OCR visual modality as the second output of the attention block.

$$\boldsymbol{X}'_{\text{ocr,v}} = \sum_{n=1}^{N} \boldsymbol{\beta}_n \boldsymbol{h}_{\text{ocr,v}}^{(n)} \tag{11}$$

Finally, the element-wise multiplication is applied to the outputs of each attention block for final fusion.

$$z_{\text{ocr,v}} = \boldsymbol{X}'_{Q} \odot \boldsymbol{X}'_{\text{ocr,v}} \tag{12}$$

As described above, our module consists of four this kind of attention blocks for the modality pair Q-and-OCR-visual (as an example), Q-and-OCR-token, Q-and-Object, and Q-and-Image.

Moreover, to further model the relationship between an image and its corresponding question, a learnable gated cross-attention mechanism is introduced. We firstly define the cross-attention operation as follows,

$$\tilde{\boldsymbol{h}} = \text{CrossAttn}(\boldsymbol{h}_1, \boldsymbol{h}_2) = \text{softmax}\left(\frac{\boldsymbol{h}_1\boldsymbol{W}_{Q}\boldsymbol{W}_{K}{}^{\top}\boldsymbol{h}_2{}^{\top}}{\sqrt{d_k}}\right)\boldsymbol{h}_2\boldsymbol{W}_{V}$$

h_1 and h_2 can be any two modalities' features. Here they refer to the entire image and the question modalities. Then,

$$\boldsymbol{h}_{Q} = \boldsymbol{X}_{Q} + \tanh(\boldsymbol{W}_a) \odot \text{CrossAttn}(\boldsymbol{X}_{Q}, \boldsymbol{h}_{\text{glob}}), \tag{13}$$

where $\boldsymbol{W}_a$ denotes the attention-gating parameter.

$$\boldsymbol{h}_f = \boldsymbol{h}_{Q} + \tanh(\boldsymbol{W}_b) \odot \text{FFW}(\boldsymbol{h}_{Q}), \tag{14}$$

where FFW is feed-forward network, and $\boldsymbol{W}_b$ denotes the FFW-gating parameter. These layers followed by a regular self-attention and another FFW on language modality to obtain visually informed question (language) features,

$$\boldsymbol{h}_s = \boldsymbol{h}_f + \text{SelfAttn}(\boldsymbol{h}_f) \tag{15}$$

$$z_{img\rightarrow q} = \boldsymbol{h}_s + \text{FFW}(\boldsymbol{h}_s) \tag{16}$$

Finally, the obtained $z_{img\rightarrow q}$ will be concatenated with the four outputs by a sequence of stocked attention blocks to obtain a final context embedding for the decoding process.

3.3 Transformer decoder

We introduce a transformer decoder module with a dynamic pointer network to interactively generate answers from a fixed answer vocabulary ($v = 1, \cdots, V$) or copy from the OCR tokens ($n = 1, \cdots, N$) in an image alternatively. In the implementation, the probability computation for the selection from the fixed answer vocabulary $p(\hat{\boldsymbol{y}}_{t,v}^{\text{voc}})$ or OCR tokens $p(\hat{\boldsymbol{y}}_{t,n}^{\text{ocr}})$ is given as follows,

$$p(\hat{\boldsymbol{y}}_{t,v}^{\text{voc}}) = \text{softmax}\big((\boldsymbol{w}_{\text{voc}}^{(v)})^{\top}\boldsymbol{z}_{\text{dec}}^{(t)}\big)$$
$$p(\hat{\boldsymbol{y}}_{t,n}^{\text{ocr}}) = \text{softmax}\big((\boldsymbol{W}_{\text{ocr}}\boldsymbol{z}_{\text{ocr}}^{(n)})^{\top}(\boldsymbol{W}_{\text{dec}}\boldsymbol{z}_{\text{dec}}^{(t)})\big), \tag{17}$$

where $\boldsymbol{W}_{\text{ocr}}$ and $\boldsymbol{W}_{\text{dec}}$ denote $d \times d$ matrices. $z_{\text{ocr}}^{(n)}(n = 1, \ldots, N)$ denotes the d-dimensional transformer output of the N OCR tokens in the image. For both two formula, the decoder embedding (d-dimensional $z_{\text{dec}}^{(t)}$) for the current time-step t is obtained depending on the previously predicted token at time-step t-1 and its corresponding representation $\boldsymbol{x}_{\text{dec}}^{(t)}$. In other words, in the case of the prediction at time-step t-1 is a word from the fixed vocabulary, we feed its corresponding weight vector $\boldsymbol{w}_{\text{voc}}^{(v)}$ as the transformer input. On the other hand, if the prediction (at time-step t-1) from the OCR tokens, its OCR representation $\boldsymbol{h}_{\text{ocr,all}}^{(n)}$ (obtained from OCR visual and semantic modalities) is considered as the transformer input to obtain the decoder embedding for the time-step t.

For the final prediction, we take the argmax on the concatenation of both probabilities $[p(\hat{y}_{t,v}^{\text{voc}}); p(\hat{y}_{t,n}^{\text{ocr}})]$ and select the top element with the highest score as the answer from the concatenation of $V + N$ candidates.

4 Experiments

4.1 Quantitative results

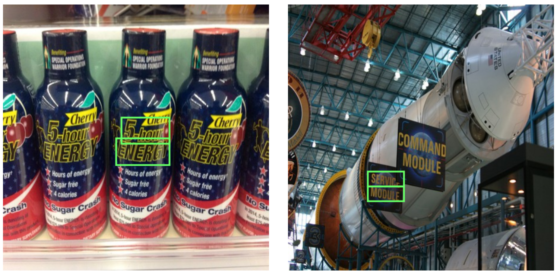
We performed the experiments using the TextVQA dataset (See **Appendix A.3**) with the experimental setting given in **Appendix A.2**. We used the standard evaluation for VQA tasks, accuracy (Acc.), which was evaluated based on the predicted answer against the ground truth (GT) answers provided by humans.

Table 1 shows the quantitative results of the baselines and our method. The scores in Table 1 represent the evaluations reported in the papers or with average and standard deviation obtained based on five trials in our experiments. Note that the baselines and proposed method share $\boldsymbol{x}_{\text{ocr,phoc}}^{(n)}$, $\boldsymbol{x}_{\text{ocr,ft}}^{(n)}$, $\boldsymbol{x}_{\text{ocr,fr}}^{(n)}$, $\boldsymbol{x}_{\text{ocr,bx}}^{(n)}$, $\boldsymbol{x}_{\text{obj,fr}}^{(m)}$, and $\boldsymbol{x}_{\text{obj,bx}}^{(m)}$ as inputs.

Table 1 Qualitative results.

	Method	OCR system	$\boldsymbol{x}_{\text{ocr,clip}}^{(n)}$, $\boldsymbol{x}_{\text{glob}}$	Acc. on val.
(1)	M4C [8]	Rosetta-en		39.40
(2)	SSbaseline [11]	SBD-Trans		43.95
(3)	SSbaseline (reproduced)	SBD-Trans		43.90 ±0.11
(4)	Ours	SBD-Trans	✓	**44.93** ±0.16

The experimental results demonstrate that the proposed method outperformed the SSbaseline model (43.9%), with an increase of 1%. More significant improvements (5.5%) were obtained compared with the M4C method (39.4%).



Q: what is in the bottles?
A: (ours) [5-hour] [energy]
(SSbaseline) sugar energy
(GT) ['cherry 5 hour energy',
'5 hour energy']

Q: what does the sign say after command model?
A: (ours) [service] [module]
(SSbaseline) command module
(GT) ['module', 'service module',
'usa', 'sevice module']

Figure 3 Qualitative samples of the predicted answers based on our proposed method and the SSbaseline method.

4.2 Qualitative results

Figure 3 shows the qualitative samples of our model compared with the SSbaseline model. The left subfigure shows that the baseline model failed to predict the correct answer due to the wrong OCR token selection. In contrast, according to the question asked about the contents of those bottles, our model correctly selected to use the accurate OCR tokens '*5-hour*' with '*energy*'. The subfigure on the right shows that our model predicted the correct answers by considering the OCR tokens '*service*' and '*module*' and their corresponding objects (the sign after the command model). The baseline method failed in these two samples.

5 Conclusion

In this paper, we focused on the TextVQA task that generates answers according to the questions that need text understanding and reasoning. We would like to emphasize the following contribution of this wok:

- We proposed using image modality with an additional attention block to complete the utilization of the visual information in an image.
- To obtain the rich features for different modalities, we introduced using a pre-trained CLIP for OCR tokens and an entire image.
- We introduced a Flamingo's gated cross-attention mechanism to further model the relationship especially for the entire image and the question.
- Our method outperformed the baseline methods on accuracy evaluation with the TextVQA dataset.

Acknowledgements

This material is based upon work supported by JSPS KAKENHI Grant Number 20H04269, JST Moonshot, and NEDO.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **ICML**, pp. 8748–8763, 2021.

[2] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Wang Lijuan. GIT: A generative image-to-text transformer for vision and language. **arXiv preprint arXiv:2205.14100**, 2015.

[3] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied question answering in photorealistic environments with point cloud perception. In **CVPR**, pp. 6659–6668, 2019.

[4] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In **CVPR**, pp. 6629–6638, 2019.

[5] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. **ACM Transactions on Information Systems (TOIS)**, Vol. 38, No. 3, pp. 1–32, 2020.

[6] Shoya Matsumori, Kosuke Shingyouchi, Yuki Abe, Yosuke Fukuchi, Komei Sugiura, and Michita Imai. Unified questioner transformer for descriptive question generation in goal-oriented visual dialogue. In **ICCV**, pp. 1878–1887, 2021.

[7] Komei Sugiura, Naoto Iwahashi, Hideki Kashioka, and Satoshi Nakamura. Active learning for generating motion and utterances in object manipulation dialogue tasks. In **AAAI Fall Symposium on Dialog with Robots**, pp. 115–120, 2010.

[8] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for TextVQA. In **CVPR**, pp. 9992–10002, 2020.

[9] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In **CVPR**, pp. 8317–8326, 2019.

[10] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. **NeurIPS**, 2022.

[11] Qi Zhu, Chenyu Gao, Peng Wang, and Qi Wu. Simple is not easy: A simple strong baseline for TextVQA and TextCaps. In **AAAI Conference on Artificial Intelligence**, pp. 3608–3615, 2020.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. **NeurIPS**, 2015.

[13] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. **IEEE Trans. PAMI**, Vol. 36, No. 12, pp. 2552–2566, 2014.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In **ACL**, pp. 4171–4186, 2019.

A Appendix

A.1 Loss function

We combine using the multi-label binary cross-entropy and a new policy gradient loss introduced in [11].

A.2 Experimental setup

The experimental setting for the hyperparameters is summarized in Table 2. L, H, and heads denote the number of hidden layers, hidden size, and the number of attention heads of a Transformer model, respectively. Our experiments were performed on four Tesla V100 GPUs with 64GB memory in total. For each performance, it required approximately 43 hours to train over 34,000 iterations. Our model had 187,456,002 (187M) trainable parameters. The prediction for one sample took approximately 70 ms.

Table 2 Hyperparameter setting of our experiments

Attention blocks	L: 12	H: 768	heads: 12
Gated cross-attention	L: 8	H: 1024	heads: 10
Optimizer	Adam		
Learning rate	1×10^{-4}		
Max iteration	34,000		
Batch size	256		
Fixed vocab. size	5k		
Decoding steps (max)	12		
# of OCR tokens (max)	50		
# of Objects (max)	100		

A.3 Dataset

In our experiments, we used the standard dataset TextVQA, released to facilitate the progress of the Text-based image captioning task in 2019. This dataset was annotated via crowd-sourcing based on the Open Images (v3) dataset. The annotators were asked to identify images containing text, then collected 1-2 questions requiring reading and reasoning about the text in the image; ten answers were collected according to different questions.

The TextVQA dataset contains 45,336 questions (samples) in English with their corresponding answers collected based on 28,408 images. The total number of tokens is 308,753, and the number of unique tokens is 9,568. The average question length is 7.44 words. Note that we did not perform any pre-processing for the questions' statistics. The average answer length is 1.58 [9]. The dataset is divided into training, validation, and test sets with sizes 34,602, 5,000, and 5,734, respectively. There is no overlap between any two splits.

We used the training set to update the parameters of our model. We evaluated our model on the validation set because the test set's ground truth (GT) answers were not provided. The experiments followed standard procedures in which the validation set was not used for training or tuning hyperparameters.

A.4 Ablation Studies

Table 3 presents **Ablation Studies** of our proposed method. We defined the following three ablation conditions:

(i) Without (w/o) image-related attention block & gated cross-attention operation for language and image (=SSbaseline).

(ii) Without (w/o) gated cross-attention operation for language and image.

(iii) Our proposed method (full).

In Table 3, compared with Condition (i), the accuracy was increased by approximately 0.4% by introducing an additional vanilla attention block with image modality and CLIP features (Condition (ii)). Thus, the image modality, absent in the baseline model introduced in our proposed method, increases the TextVQA task's performance. In comparing Condition (iii) with Condition (ii), a further improvement (0.6%) was obtained by introducing a gated cross-attention module for image and question modalities in the encoding process. In summary, these results indicate that only using the object and OCR modalities with questions is insufficient; image as one of the visual modalities is also indispensable to the text-based VQA performance. Furthermore, more robust attention computation between the image and question modality is essential in the encoder process in our TextVQA task.

Table 3 Ablation studies on TextVQA.

Condition	(i)	(ii)	(iii)	$\boldsymbol{x}^{(n)}_{\text{ocr,clip}}$, $\boldsymbol{x}_{\text{glob}}$	Acc. on val.
(1)	✓				43.90 ±0.11
(2)		✓		✓	44.31 ±0.14
(3)			✓	✓	**44.93** ±0.16