Transformer-Based Visual Feature and Textual Feature Fusion Model for Document Key Information Extraction

Wooseok Kim , **Juhyeong Kim** , **Sangyeon Yu** , **Gyunyeop Kim** , **Sangwoo Kang**
School of Computing, Gachon University
{kws9208, stephano12, teryas, gyop0817, swkang}@gachon.ac.kr

Abstract

The VRDIU-Track A competition requires finding the value corresponding to a natural language query in a document image. To address this challenge, we designed a model that fuses both visual and textual features of the ROI object. This approach has substantially improvement scores compared to baseline, with an increase of approximately 0.46 points, and demonstrated a notable advantage on the private leaderboard, showing an improvement of appoximately 0.13 points over a vision-language pretrained model that relies solely on visual features.

1 Introduction

The VRDIU competition tackles the challenge of document understanding, with Track A focusing on the task of extracting key information by identifying the corresponding ROI region based on a user's query. These tasks are closely related to various aspects of document understanding, including document classification, question answering, and form analysis. When handling document images, capturing both visual and textual aspects is crucial, as the location and textual information are just as important as the visual features. Therefore, we propose a transformer-based model that learns both visual and textual features from identifiable text regions within a document.

2 Related Works

Various studies have been investigated to improve performance in vision language tasks. For vision language tasks, some pre-trained models, such as VisualBERT [Li *et al.*, 2019] and LXMERT [Tan and Bansal, 2019] learned both visual and textual information about ROI object by analyzing how ROI objects are related to text captions in images. In the text VQA task, [Wang *et al.*, 2022] leveraged textual information from the scene to generate and learn training data, resulting in enhanced performance. [Hegde *et al.*, 2023] demonstrated improved performance in a Visual Question Answering task by training with a blend of text-based VQA and VQA datasets. [Shen *et al.*, 2023] enhanced its performance in the Video Question Answering task by incorporating the textual embeddings of nouns found in the questions into the input.

3 Method

Motivation of our methodology is that both visual and textual features evenly serves crucial role to discern the ground truth of a question. For instance, titles or company names incline to correspond to nouns, while the ACN/ARSN numbers or dates correspond to the numerical values. These relationships are essential for encoding object features in the model. Therefore, unlike other existing vision-language models that focus solely on the visual features of the ROI object, we designed a novel vision object & text encoder that learns visual features, location information, and textual features.

3.1 ROI Feature Extraction

We extracted visual and textual features from each ROI object based on the bounding boxes and text provided in the dataset. The visual feature was obtained by cropping the bounding box of each ROI in the images and fed into a Resnet model [He *et al.*, 2016], using avgerage 2D pooling to obtain a 2048-dimensional visual feature V_i. For textual features, we input the ROI-specific text from the image into the transformer encoder-based pre-trained model and used the 1024-dimensional output embedding of the [CLS] token as the textual feature T_i.

3.2 Model Architecture

Text Encoder Initially we split the text of the given question into tokens and fed them into a transformer-based text encoder. The encoded vector, which represents contextual information between question tokens, is obtained.

Vision Object & Text Encoder Given a set of ROI features in a document image, the transformer-based vision object & text encoder pair each visual feature V_i with its corresponding textual feature T_i and the location information, resulting in a contextual relationship vector. We also performed a linear projection to align text features with visual features before inputting them into the encoder.

Cross Encoder The cross encoder generates a vector through cross-attention, capturing the contextual interactions between the textual features of the questions and the visual and textual features of each Region of Interest (ROI) object. This vector integrates the outputs from both text and vision object & text encoders.

Classifier Text embedding E_i^T and vision embedding E_i^V are concatenated for each ROI object from the vector using

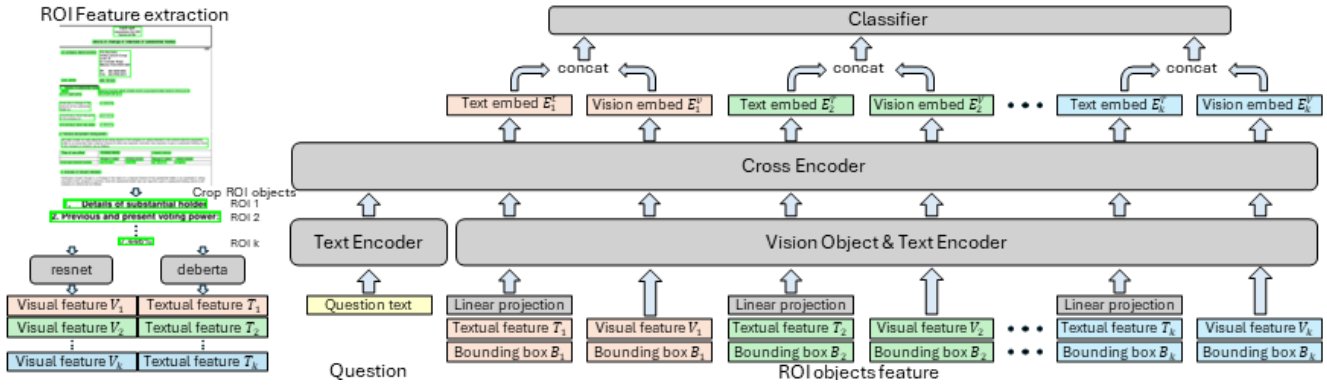

Figure 1: Model Architecture

a Cross Encoder. This concatenated vector is then fed into the classifier, which uses the binary cross-entropy (BCE) loss to determine whether each ROI object represents the ground truth of a question.

The overall architecture is shown in Fig 1. Our model utilizes different vision objects and text encoders compared with the LXMERT model; however, it also uses three encoders, similar to LXMERT. Therefore, we initialize each encoder in our model using the same weights as those used in LXMERT.

3.3 Image Augmentation

We modulated the training and validation datasets to steer our model to address the printed images better. To noise digital images such as printed images, Gaussian noise was added with a mean of 0 and a standard deviation of 5, rotating images between $-2°$ and $2°$, and applying blur effects.

4 Experiments

4.1 Dataset and Experimental Settings

We used the Form-NLU dataset [Ding *et al.*, 2023], which includes form images along with details, such as bounding boxes and textual content, for identifiable ROI objects in each form image. The training dataset consisted solely of digital images, whereas the evaluation dataset cosisted of both printed and digital images.

We fine-tuned the model with binary cross-entropy loss using the AdamW optimizer with a learning rate of 1e-4 for 50 epochs. We used the Resnet-101 model and Deberta-v3-large model [He *et al.*, 2021] to extract the visual and textual features of ROI objects.

4.2 Results

We assessed the performance using the F1 score, and our proposed model showed a significant improvement in scores over the competition's baseline and highlighted a score difference on the private leaderboard compared to VisualBERT or LXMERT, which are vision-language pre-trained models. Our model achieved a superior score of 1.0000 on the public leaderboard and 0.9695 on the private leaderboard, ranking third overall.

Our ablation study showed a significant score difference when using both visual and textual features versus visual features alone. The visual feature only model, similar to LXMERT, is trained using only visual features as input from our model's visual object and text encoder. The early feature sum model is trained using only one feature as input, which is the sum of the textual and visual features before input to the model's visual objects and text encoder. Ours(w/ aug) model is trained using image augmentation.

Rank	Team	Test_{public}	$\text{Test}_{private}$
1	Team1	1.0000	1.0000
2	Team2	0.9778	0.9793
3	**Team3(ours)**	1.0000	0.9695
	Baseline(Bert)	0.5401	0.5040
	VisualBERT	0.9662	0.5809
	LXMERT	0.9930	0.8316

Table 1: Performance comparison for main results

Model	Val	Test_{public}	$\text{Test}_{private}$
Visual feature only	0.9922	0.9930	0.8316
Early feature sum	1.0000	0.9979	0.9424
Ours	1.0000	0.9969	0.9718
Ours(w/ aug)	1.0000	1.0000	0.9660

Table 2: Ablation Study on feature utilization

5 CONCLUSIONS

This study includes experiments conducted on the Key Information Extraction task of the VRDIU-Track A challenge. When handling ROI objects in documents, our model was designed to better identify the ground truth for questions by considering both visual and textual features. Unlike other models that only focus on visual features, our model demonstrates a significant performance advantage on the private leaderboard achieving an improvement of approximately 0.13 points.

6 Acknowledgements

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A2C1005316). And this research was supported by funding from HMC/KIA.

References

[Ding *et al.*, 2023] Yihao Ding, Siqu Long, Jiabin Huang, Kaixuan Ren, Xingxiang Luo, Hyunsuk Chung, and Soyeon Caren Han. Form-nlu: Dataset for the form natural language understanding, 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[He *et al.*, 2021] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Wei Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *2021 International Conference on Learning Representations*, May 2021. Under review.

[Hegde *et al.*, 2023] S. Hegde, S. Jahagirdar, and S. Gangisetty. Making the v in text-vqa matter. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5580–5588, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society.

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

[Shen *et al.*, 2023] Ruoyue Shen, Nakamasa Inoue, and Koichi Shinoda. Text-guided object detector for multimodal video question answering. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1032–1042, 2023.

[Tan and Bansal, 2019] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, November 2019. Association for Computational Linguistics.

[Wang *et al.*, 2022] Jun Wang, Mingfei Gao, Yuqian Hu, Ramprasaath R. Selvaraju, Chetan Ramaiah, Ran Xu, Joseph JaJa, and Larry Davis. Tag: Boosting text-vqa via text-aware visual question-answer generation. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022.