2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

DocHQ: Towards Multi-modal Document Understanding via Hybrid Feature Queries

Jin Wang[1], Yingying Liu[2], and Yahong Han[1]

[1] Tianjin University, 92 Weijin RoadNankai District, Tianjin, China
[2] South China University of Technology, 381 Wushan Road, Guangzhou, China
wangjin5112@tju.edu.cn

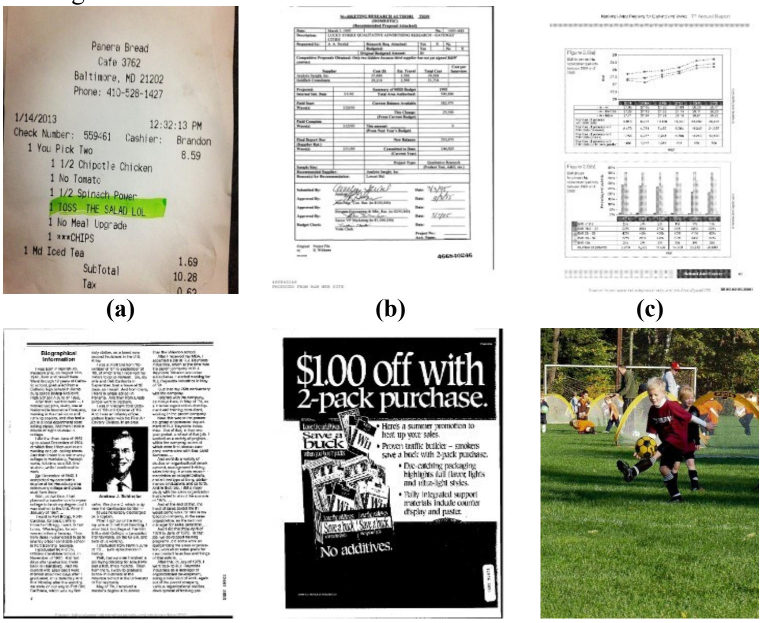
Abstract. Significant progress has been made in general multi-modal tasks leveraging pre-trained visual and language models. However, in visual document understanding tasks, enhancing performance by utilizing existing models encounters difficulties due to the fundamental differences between natural and document images. In this paper, we introduce DocHQ, a multi-modal document image understanding model with pre-trained visual and language models, employing a hybrid feature query for feature alignment between document visual information and language text. Our approach combines learnable and fixed task-oriented queries within a cross-attention visual-language alignment module to extract more fine-grained information from document images. Moreover, we utilize large-scale document images for alignment training between the pre-trained image encoder and the language model. Experimental results demonstrate that our method achieves outstanding performance across three different types of document image understanding tasks compared to existing approaches.

Keywords: Document Image Understanding, Multi-modal Feature Alignment, Document Pretrain model.

1 Introduction

In recent years, with increasing commercialization in various digital scenarios such as banking digitization, intelligent education, and smart office solutions, Document AI [1] has garnered significant attention from industrial and academic researchers. This increased interest has spurred the development of numerous document pre-training models [2,3,4,5,6,7], which typically employ self-supervised pre-training on large-scale document image datasets [8], followed by fine-tuning on various document processing tasks. These models have demonstrated remarkable performance improvements across tasks such as document information extraction [9, 10], table recognition and understanding, document image classification [11, 12], and document visual question answering [13].

To reduce the energy consumption and complexity associated with the retraining of multi-modal foundational models, [14, 15] connects visual and large language models through a multi-modal projector module, achieving significant success in most general multi-modal tasks, spanning natural images, videos, and speech tasks. [16] utilized a plug-and-play module to generate image-relevant exemplar prompts for LLM, enabling zero-shot VQA tasks without end-to-end training. However, the oversimplification of the generated image-relevant prompts poses challenges in applying the model to more complex document image tasks with richer textual information. Furthermore, mPLUG-DocOwl 1.5 [17] proposed a unified framework that emphasizes structured information processing on documents, web pages, tables, charts, and natural images. In BLIP-2 [15], a cross-attention transformer model was used to train an alignment module between pre-trained visual and language models. Although the proposed learnable queries in BLIP-2 can extract visual representations informative of the text, they may still struggle to identify detailed textual information within document images. [18] proposed a trainable bridging module with diverse instructions in a unified format to connect document images, image encoders, and large language models (LLMs). However, despite efforts to bridge document image encoders and LLM models to enhance the performance of visual document understanding tasks, these models still fall short of achieving the expected effectiveness because of the inherent difference between natural and document images, which makes it difficult to understand more fine-grained text information in document images.



(a) (b) (c)

(d) (e) (f)

Fig. 1. The difference between document and natural image.

As shown in Fig 1, we summary the main differences between document images and natural images. (a) shows the receipt image with low resolution and image distortion in the CORD [10] dataset; (b) shows document image with abundant textual information in the RVL-CDIP [11] dataset; (c) is form and chart image with complex formatting and intricate textual details in the DocVQA [13] dataset; (d) rich text information in document image; (e) is a promotional poster image with various formats; (f) shows the natural image with coarse-grained semantic information such as people, places, and activities.

As depicted in Figure 1, there are differences between document images and natural images. Firstly, natural images typically exhibit coarser-grained semantics compared to document images. Natural images often depict broad semantic content like locations, people, activities, and landscapes, whereas document images contain finely-grained, human-readable text. Additionally, document images often have lower pixel quality than natural images, necessitating document image understanding models to focus on fine-grained pixel information within small local regions. Overall, these disparities pose significant challenges in improving document image understanding capabilities by leveraging existing visual and language models.

Considering these differences between document and natural images, we design a **Doc**ument image understanding model via **H**ybrid feature **Q**ueries (**DocHQ**) to enable efficient and effective feature alignment between pre-trained visual and language models. We fuse learnable queries with the OCR-text token embeddings as the textual token queries of document images and input them into a multi-layer cross-attention module to learn the fine-grained text information of the document images. Finally, the queried visual information is fed into the LLM for an accurate understanding of the document images.

In summary, we present our main contributions as follows: (1) We propose a visual document understanding model that avoids costly end-to-end training by leveraging existing document image encoders and LLMs. This is achieved through two training stages: multi-modal feature alignment learning and downstream task fine-tuning, which effectively bridge the pre-trained document image model and LLM. (2) A hybrid feature queries-based multi-modal alignment module comprising learnable and fixed task-oriented queries is designed to tackle the challenge of capturing textual information in document images. (3) The experimental results show the superior performance of our proposed model in five datasets covering three types of visual document understanding tasks, including document image classification, document information extraction, and document visual question answering.

2 Related Work

2.1 Document Pre-training Model

The trends in document image understanding tasks are primarily influenced by advancements in large language models (LLMs) and visual pre-training models, focusing on three key areas. First, self-supervised pretraining on large-scale document image datasets has become popular, improving model performance across various tasks, as seen in models like LayoutLMv3 [4] and UDOP [19]. Second, there is a trend towards merging multi-modal information, such as OCR-extracted text and layout, into transformer models, exemplified by works like mPLUG-DocOwl [17]. However, leveraging existing pre-trained visual and LLMs for generalization across tasks, while minimizing computational complexity, is suggested as a more efficient path. Finally, integrating multiple downstream tasks, including information extraction, table understanding, and visual question answering, into unified model architectures is gaining traction, as seen in models like UniDoc [20]. The main differences among existing models lie in inputs, architectures, and pre-training objectives. Image-only models such as Donut [6], text-layout models like LayoutLM [2], and hybrid models that integrate image, text, and layout information, such as GeoLayoutLM [21], demonstrate various approaches to improving document image understanding.

2.2 Multi-modal Models with pre-trained visual-language model

With the rapid growth of large language models (LLMs), research has increasingly focused on enhancing multi-modal tasks by fine-tuning pre-trained visual and language models [22, 23]. While models like BLIP-2 [15], and MiniGPT-4 [14] have made significant strides in natural image-language tasks, similar approaches in document image pre-training remain limited. Challenges include the diverse content of document images and the complexity of aligning them with language models. Efforts like the Q-Former alignment model [15] and approaches by MiniGPT-4 [14] aim to bridge this gap, advancing multi-modal document understanding. Furthermore, [24] introduced a multi-modal model with few shot learning capabilities across image, text, and video data, utilizing a special Perceiver-Resampler and gated-attention structure.

Unlike existing visual document understanding model architectures, we integrate OCR-extracted tokens embedding with learnable queries as hybrid feature queries into a cross-attention multi-modal projector. Through two-stage training, our approach exhibits superior model capabilities across multiple tasks.

3 Method

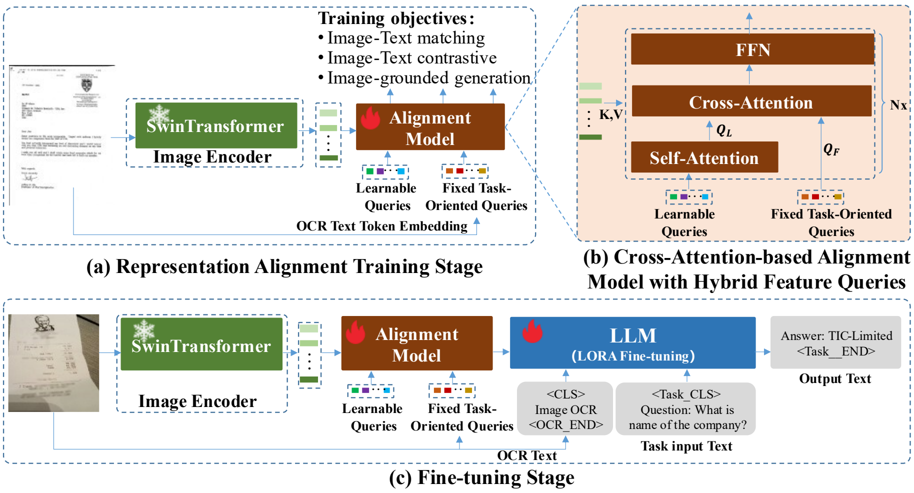
3.1 Model Architecture

We propose DocHQ, a multi-modal document understanding via hybrid feature queries, illustrated in figure 2, adopts a two-stage training approach. It consists of three main components: the image encoder, a cross-attention-based alignment model with hybrid feature queries, and the LLM generation decoder. In the multi-modal feature alignment training stage, the alignment model is trained with three similar objectives designed in [15]: image-text matching, image-text contrastive, and image-ground generation, while keeping the image encoder frozen. Subsequently, during the stage of fine-tuning of visual document understanding tasks, we use LORA [25] to finetune the LLM decoder.

3.2 Image Encoder

In our approach, we leverage the Nougat model [26] as the document image feature extractor, which is specifically designed for neural optical understanding of academic documents and pre-trained on millions of Arxiv PDF images. This pre-training process ensures that Nougat possesses powerful document image representation capabilities.

Let I denote the document image input, represented as an RGB image with three channels. The output of the document image encoder, denoted as F_{Img}, has a size of 512x1024. To align with the input requirements of the alignment model, we utilize a simple linear layer to map the image features to 768 dimensions.



(a) Representation Alignment Training Stage

(b) Cross-Attention-based Alignment Model with Hybrid Feature Queries

(c) Fine-tuning Stage

Fig. 2. The two-stage training architecture of our proposed DocHQ model. (a) Document image and language alignment pretraining with three objectives. (b) The detail of cross-attention alignment model with hybrid feature queries. (c) The fine-tuning stage on visual document understanding tasks.

3.3 Feature Alignment Module with Hybrid Queries

In Figure 2(b), we introduce a novel document image alignment model that lever-ages cross-attention mechanisms, incorporating both learnable and task-oriented textual token queries to enhance the model's performance in document image pro-cessing. Traditional models in the field often rely exclusively on learnable queries, which are typically initialized as fixed vectors and then refined through training. In contrast, our approach integrates textual token queries derived from optical charac-ter recognition (OCR) outputs, specifically from the OCR-text token embeddings, which allows the model to incorporate fine-grained textual information from the document images. This innovation enhances the model's ability to capture detailed relationships between text and visual features within the document, a critical aspect for tasks such as document retrieval, text understanding, and alignment in multi-modal settings.

We used the widely adopted and open-source Tesseract-OCR tool for OCR text extraction from documents. We initialize the model with the Q-Former from BLIP-2 and generate task-oriented queries Q_F by embedding the top 256 tokens from the OCR-text. The task-oriented textual queries Q_F are constructed by embedding the top 256 tokens from the OCR-extracted text. These queries are represented as a matrix of size 256×2048, capturing detailed token-level information. However, to ensure proper interaction with the cross-attention mechanism, these queries are projected down to a lower-dimensional space of size 256×768 via a linear transformation layer.

In the self-attention module, we define the output as Q_L, which represents the learnable queries. These queries are updated through the self-attention mechanism, where the query matrix Q_L is iteratively refined as Q_L=self-attention(Q_L). The self-attention module allows the model to capture intricate dependencies within the learnable queries, which can be used for various downstream tasks in document processing.

The cross-attention module plays a critical role in aligning the textual token queries with image features. In this module, the image features are treated as keys (K) and values (V), which are used to compute the cross-attention between the queries and the image features. Specifically, the cross-attention for the learnable queries is defined as C_L=Attention(Q_L,K,V), where C_F denotes the cross-attention output corresponding to the learnable queries. Similarly, the cross-attention for the task-oriented textual token queries is defined as C_F=Attention(Q_F,K,V), where C_F is the output for the task-specific queries. It is important to note that the cross-attention layers for both Q_L and Q_F operate independently with distinct parameters. This ensures that each type of query, the learnable and the textual token queries, can interact with the image features in a specialized manner.

Furthermore, the feed-forward layers following the cross-attention operations do not share parameters between the learnable and task-oriented queries, preserving the independence of the two query types. After the cross-attention and feed-forward operations, the outputs of the final layer are projected through a linear layer, ensuring that the dimensionality matches the required input size for large language models (LLMs), facilitating further multimodal processing.

3.4 Visual Document Understanding Decoder with LLM

In this paper, we use Tinyllama [22] as the text generation module for downstream tasks. Tinyllama is a decoder-only pre-trained language model with 1.1 billion parameters, comprising 22 decoder layers and a hidden size of 2048. It was trained on a text dataset containing 3 trillion tokens. Additionally, for our ablation experiments, we incorporate GPT-Neo-1.3B, a transformer model with 1.3 billion parameters. This model was designed using EleutherAI's replication of the GPT-3 [27] architecture, featuring a hidden size of 2048 and 24 transformer layers.

During the fine-tuning stage, the decoder model (LLM) receives inputs from the feature align module, the OCR results of the input image, and the instruction text related to the visual document understanding task. We employ the image-grounded next token prediction objective using the LORA method. Moreover, we introduce special tokens associated with the task into the LLM tokenizer to denote the beginning and end of the downstream task.

4 Experiments

4.1 *Visual-Language Feature Alignment Training*

Training Dataset: We conducted pre-training using the IIT CDIP collection [8] dataset. To ensure data quality, we filtered out short texts and non-Latin languages, rotated images, and applied the Nougat model to exclude samples with more than 30% line breaks or significant duplicate text in the extracted content.

Training Objectives: We adopted objectives similar to Q-former's pre-training in [15], including image-text matching, image-text contrastive, and image-grounded generation. The image-text contrastive objective maximizes the mutual information between the document image and the OCR text, while image-text matching involves fine-grained binary classification for pair alignment. The image-grounded text generation trains the model to generate text based on document image features, with information transfer via queries and self-attention layers.

Training Details: About three million high-quality image-text samples were obtained after pre-processing. The pretraining batch size was set to 512, with a gradient accumulation interval of 32. The model was trained for 10 epochs with an initial and stop learning rate of 3e-5 and 1e-5 respectively. The warming step was set to 10% of the total steps. Adam optimizer with $\beta1$= 0.9, and $\beta2$ = 0.98, was utilized, along with AMP mixed precision training. The pre-training phase was conducted on 16 V100 GPUs for 12 hours.

4.2 Fine-tuning on VDU Tasks

Downstream Tasks and Metrics. We fine-tuned and evaluated our model on three document understanding tasks. For document Image classification, we finetuned on the RVL-CDIP dataset (16 classes, 320K train-set) and Tobacco-3482 (10 classes, 2782 training images). For document image content understanding, we used FUNSD and CORD datasets, with CORD containing 1,000 receipts and FUNSD consisting of 199 forms with over 9,700 semantic entities. For Document Visual Question Answering, we evaluated our model on the DocVQA dataset, which includes 50k questions on 12k+ images. We used accuracy for document image classification, F1 score for content understanding, and the commonly-used edit distance-based metric ANLS (also known as Average Normalized Levenshtein Similarity) for DocVQA as evaluation metrics.

Table 1. Results of the document image classification task on RVL-CDIP and Tobacco3482 dataset.

Method	**Visual Encoder**	**RVL-CDIP (ACC)**	**Tobacco3482 (ACC)**
StructalLM [28]	-	96.08	-
LayoutLM [2]	-	91.90	-
TransferDoc [29]	ViT-B/16	93.18	-
EmmDocClassifier [30]	EfficientNet-B0	95.70	90.30
VLCDoC [31]	ViT-B/16	-	90.94
SelfDoc [32]	FasterR-CNN	92.81	-
DocFormer [33]	ResNet-50	96.17	-
LayoutLMv2 [3]	ResNeXt-FPN	95.64	-
LayoutLMv3 [4]	Linear	95.93	-
UDOP [19]	-	96.00	92.10
DocHQ (ours)	**SwinTransformer**	**96.47**	**94.30**

Baseline models. We compare our DocHQ with the following baseline methods. The LayoutLM series (v1, v2, and v3) [2, 3, 4] are pre-trained models designed for document understanding tasks that combine textual and layout information. OCR-free models analyze and interpret document content directly from images, including Donut [6], UniDoc [20], and StrucTexTv2 [35]. Structure-centric models, like GeoLayoutLM [21] and StructalLM [28], focus on learning and leveraging geometric and relational structures within documents to improve content comprehension. Other influential models, such as DocFormerv2 [7], and Formnet [36].

4.3 Results and Analysis

Document Image Classification. As demonstrated in Table 1, our model has achieved remarkable state-of-the-art (SOTA) performance on two widely recognized document image classification datasets. Notably, our model outperformed the second-best model by an impressive margin of 0.3% in accuracy on the first dataset and 0.8% on the second dataset. These results underscore the effectiveness of our approach in accurately classifying document images.

Table 2. Experiment Results on FUNSD and CORD.

Method	Visual Encoder	FUNSD (F1)	CORD (F1)
StructalLM [28]	-	85.14	-
LayoutLM [2]	-	77.89	-
BROS [5]	-	84.52	97.28
DocFormerv2 [7]	ViT	88.89	97.70
SelfDoc [32]	FasterR-CNN	83.36	-
DocFormer [33]	ResNet-50	84.55	96.69
LayoutLMv2 [3]	ResNeXt-FPN	84.20	96.01
LayoutLMv3 [4]	Linear	92.08	97.46
UDOP [19]	-	91.62	97.58
FormNetV2 [34]	3-layerConvNet	86.35	97.37
GeoLayoutLM [21]	ConvNeXt	92.86	**97.97**
DocHQ (ours)	**SwinTransformer**	**96.47**	94.30

Document Image Content Understanding. In terms of document image content understanding, we present the comparative results of our proposed model on the benchmark datasets FUNSD and CORD. The findings highlighted in Table 2 reveal that our model achieves F1 scores that rank first and second for these tasks, respectively. This signifies a superior ability to understand and interpret the content embedded within document images.

Document Visual Question Answering. Regarding the document visual question answering task, we rigorously compared our model with the latest architectures, including UDOP and LayoutLMv3, on the challenging DocVQA dataset. As shown in Table 3, our model demonstrated similarly excellent performance. Additionally, the cases illustrated in Fig 3 clearly indicate that our model effectively captures intricately detailed content information from document images.

4.4 Ablation Study

Hybrid Feature Queries. As shown in Table 4, our model outperforms the methods of using only learnable queries or only textual token queries on all three types of document image tasks. This is because by adding textual token queries, the model can retrieve features from the document image feature space through the cross-attention module that are more relevant to the input vector space of the LLM decoder model compared with the learnable queries. At the same time, the learnable queries make the model focus on some diverse and dynamic features within the document image.

Table 3. Experiment results of the document visual question answer task on DocVQA.

Method	Visual Encoder	DOCVQA (ANLS)
Donut [6]	SwinTransformer	67.5
DocFormerv2 [7]	ViT	**87.8**
StructalLM [28]	-	83.9
LayoutLMv2 [3]	ResNeXt-FPN	78.8
LayoutLMv3 [4]	Linear	83.4
UDOP [19]	-	84.7
DocHQ (ours)	**SwinTransformer**	86.5

Table 4. Ablation study results of the effectiveness of the hybrid feature queries and the LLM decoder models on FUNSD, RVL-CDIP, and DocVQA dataset.

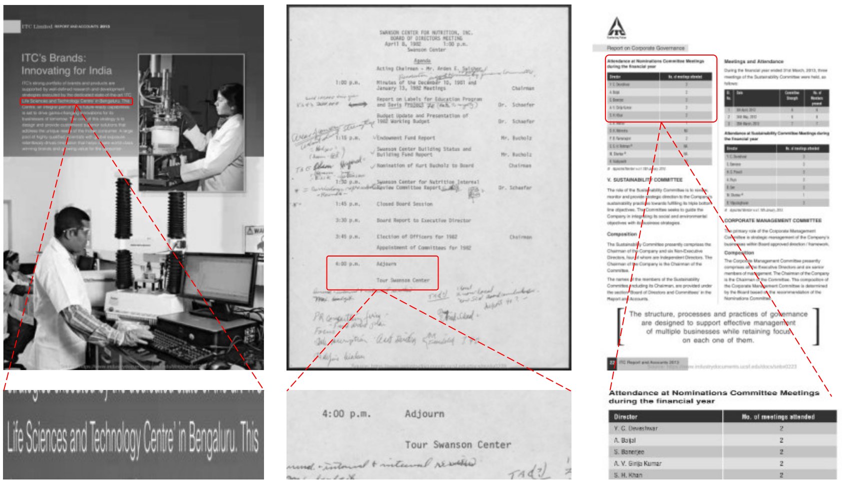
Method	RVL-CDIP (ACC)	FUNSD (F1)	DocVQA (ANLS)
Only Learnable Queries	92.5	89.19	84.9
Only textual token queries	86.13	86.50	83.67
DocHQ (with tinyllama)	**96.47**	**93.30**	**86.58**
DocHQ (with GPT-Neo-1.3B)	95.89	92.55	85.2

For the analysis of the number of learnable queries (LQ) and textual token queries (TQ), as shown in Table 4, the results indicate that changes in the number of learnable queries have almost no effect on model performance while increasing the number of textual token queries results in a slight performance improvement. We think that the learnable queries are learnable vectors, so their size only changes the vector space dimensions of the latent representation without affecting the representation capability. However, changes in the number of textual token queries to adding extra prior information can influence model performance to some extent.

The Influence of Different Image Encoders and LLM Decoders. We conducted a set of comparative experiments on the influence of different image encoders and LLM decoders. The results in Table 4 indicate that using ViT-L/14 or EfficientNet-B0 as the document image encoder performs significantly worse than using the Nougat encoder, which is pre-trained on a large-scale document image dataset, whereas ViT-L/14 or EfficientNet-B0 is pre-trained on general images. As shown in Table 5, we report the two DocHQ models with different LLM decoders, respectively, which validates the effectiveness of our approach across different LLM models. The two LLM decoders are Tinyllama and GPT-Neo-1.3B.

Table 5. Ablation study of the components of our method.

Method	**RVL-CDIP (ACC)**	**FUNSD (F1)**	**DocVQA (ANLS)**
DocHQ (ViT-L/14)	77.80	75.83	67.33
DocHQ (EfficientNet-B0)	58.35	64.16	61.57
DocHQ (SwinEncoder)	96.47	93.30	86.58
DocHQ (LQ=24, TQ=256)	95.53	93.57	86.14
DocHQ (LQ=32, TQ=256)	96.47	93.30	86.58
DocHQ (LQ=40, TQ=256)	95.46	92.40	85.91
DocHQ (LQ=48, TQ=256)	96.15	93.18	86.38
DocHQ (LQ=32, TQ=224)	93.15	91.50	84.59
DocHQ (LQ=32, TQ=256)	96.47	93.30	86.58
DocHQ (LQ=32, TQ=288)	96.39	93.41	86.62
DocHQ (LQ=32, TQ=320)	96.46	93.78	86.20



Life Sciences and Technology Centre' in Bengaluru. This
4:00 p.m. Adjourn
Tour Swanson Center
Attendance at Nominations Committee Meetings during the financial year

Question: Where is the ITC Life Sciences and Technology Centre?
Answer: bengaluru
DocHQ: bengaluru
LayoutLM: India

Question : What time the 'meeting' was adjourned?
Answer : 4:00 p.m.
DocHQ: 4:00 p.m.
LayoutLM: 3:45 p.m.

Question : How many nomination committee meetings has S. Banerjee attended?
Answer : 2
DocHQ: 2
LayoutLM: 3

Fig. 3. Three cases from DocVQA test set compared with LayoutLM

5 Conclusion

In this paper, we introduce a novel multi-modal document understanding model designed to enhance the analysis of document images. Our approach leverages hybrid feature queries in conjunction with pre-trained visual and language models to achieve this goal. Through extensive experiments conducted on three distinct types of five-document visual understanding datasets, our method consistently demonstrates superior performance compared to existing models. These remarkable results not only validate the effectiveness of our proposed model but also highlight its significant potential in enhancing various document image understanding tasks.

References

1. Cui, L., Xu, Y., Lv, T., Wei, F.: Document AI: benchmarks, models and applications. CoRR abs/2111.08609 (2021)
2. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: KDD 2020. pp. 1192–1200 (2020)
3. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D.A.F., Zhang, C., Che, W., Zhang, ˆ M., Zhou, L.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In: ACL/IJCNLP. pp. 2579–2591 (2021)
4. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document AI with unified text and image masking. In: MM 2022. pp. 4083–4091. ACM (2022)
5. Hong, T., Kim, D., Ji, M., Hwang, W., Nam, D., Park, S.: BROS: A pre-trained language model focusing on text and layout for better key information extraction from documents. In: AAAI 2022. pp. 10767–10775. AAAI Press (2022)
6. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: ECCV 2022. vol. 13688, pp. 498–517. Springer (2022)
7. Appalaraju, S., Tang, P., Dong, Q., Sankaran, N., Zhou, Y., Manmatha, R.: Docformerv2: Local features for document understanding. pp. 709–718. AAAI (2024)
8. Lewis, D.D., Agam, G., Argamon, S., Frieder, O., Grossman, D.A., Heard, J.: Building a test collection for complex document information processing. In: SIGIR 2006. pp. 665–666. ACM (2006)
9. Jaume, G., Ekenel, H.K., Thiran, J.: FUNSD: A dataset for form understanding in noisy scanned documents. In: ICDAR 2019. pp. 1–6. IEEE (2019)
10. Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., Lee, H.: Cord: A consolidated receipt dataset for postocr parsing (2019)
11. Pramanik, S., Mujumdar, S., Patel, H.: Towards a multimodal, multi-task learning based pre-training framework for document representation learning. CoRR **abs/2009.14457** (2020)
12. Harley, A.W., Ufkes, A., Derpanis, K.G.: Evaluation of deep convolutional nets for document image classification and retrieval. CoRR **abs/1502.07058** (2015)
13. Mathew, M., Karatzas, D., Jawahar, C.V.: Docvqa: A dataset for VQA on document images. In: WACV 2021. pp. 2199–2208. IEEE (2021)
14. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. ArXiv **abs/2304.10592** (2023)
15. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: TCML 2023. ICML'23, JMLR.org (2023)
16. Guo, J., Li, J., Li, D., Tiong, A.M.H., Li, B., Tao, D., Hoi, S.C.H.: From images to textual prompts: Zero-shot visual question answering with frozen large language models. In: CVPR 2023. pp. 10867–10877. IEEE (2023)

17. Hu, A., Xu, H., Ye, J., Yan, M., Zhang, L., Zhang, B., Li, C., Zhang, J., Qin, J., Huang, F., Zhou, J.: mPLUGDocOwl 1.5: Unified Structure Learning for OCR-free Document Understanding. arXiv (3 2024)
18. Tanaka, R., Iki, T., Nishida, K., Saito, K., Suzuki, J.: Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In: AAAI 2024. pp. 19071–19079. AAAI Press (2024)
19. Tang, Z., Yang, Z., Wang, G., Fang, Y., Liu, Y., Zhu, C., Zeng, M., Zhang, C., Bansal, M.: Unifying vision, text, and layout for universal document processing. In: CVPR 2023. pp. 19254–19264. IEEE (2023)
20. Gu, J., Kuen, J., Morariu, V.I., Zhao, H., Jain, R., Barmpalios, N., Nenkova, A., Sun, T.: Unidoc: Unified pretraining framework for document understanding. In: NeurIPS 2021. pp. 39–50 (2021)
21. Luo, C., Cheng, C., Zheng, Q., Yao, C.: Geolayoutlm: Geometric pre-training for visual information extraction. In: CVPR 2023. pp. 7092–7101. IEEE (2023)
22. Zhang, P., Zeng, G., Wang, T., Lu, W.: Tinyllama: An open-source small language model. CoRR **abs/2401.02385** (2024)
23. Black, S., Gao, L., Wang, P., Leahy, C., Biderman, S.: GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow (Mar 2021)
24. Alayrac, J., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Katherine Millican, e.a.: Flamingo: a visual language model for few-shot learning. In: NeurIPS 2022 (2022)
25. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR 2022. OpenReview.net (2022)
26. Blecher, L., Cucurull, G., Scialom, T., Stojnic, R.: Nougat: Neural optical understanding for academic documents. CoRR **abs/2308.13418** (2023)
27. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, e.a.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
28. Li, C., Bi, B., Yan, M., Wang, W., Huang, S., Huang, F., Si, L.: Structurallm: Structural pretraining for form understanding. In: ACL/IJCNLP 202. pp. 6309–6318. Association for Computational Linguistics (2021)
29. Bakkali, S., Biswas, S., Ming, Z., Coustaty, M., Rusinol, M., Terrades, O.R., Llad ˜ os, J.: Transferdoc: ´ A self-supervised transferable document representation learning model unifying vision and language. CoRR **abs/2309.05756** (2023)
30. Kanchi, S., Pagani, A., Mokayed, H., Liwicki, M., Stricker, D., Afzal, M.Z.: Emmdocclassifier: Efficient multimodal document image classifier for scarce data. Applied Sciences **12**(3) (2022)
31. Bakkali, S., Ming, Z., Coustaty, M., Rusinol, M., Ter- ˜ rades, O.R.: Vlcdoc: Vision-language contrastive pretraining model for cross-modal document classification. Pattern Recognit. **139**, 109419 (2023)
32. Li, P., Gu, J., Kuen, J., Morariu, V.I., Zhao, H., Jain, R., Manjunatha, V., Liu, H.: Selfdoc: Self-supervised document representation learning. In: CVPR 2021. pp. 5652–5660 (2021)
33. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: Docformer: End-to-end transformer for document understanding. In: ICCV 2021. pp. 973–983. IEEE (2021)
34. Lee, C., Li, C., Zhang, H., Dozat, T., Perot, V., Su, G., Zhang, X., Kihyuk Sohn, e.a.: Formnetv2: Multimodal graph contrastive learning for form document information extraction. In: ACL 2023. pp. 9011–9026. Association for Computational Linguistics (2023)

35. Yu, Y., Li, Y., Zhang, C., Zhang, X., Guo, Z., Xiameng Qin, e.a.: Structextv2: Masked visual-textual prediction for document image pre-training. In: ICLR 2023. OpenReview.net (2023)
36. Lee, C., Li, C., Dozat, T., Perot, V., Su, G., Hua, N., Ainslie, J., Wang, R., Fujii, Y., Pfister, T.: Formnet: Structural encoding beyond sequential modeling in form document information extraction. In: ACL 2022. pp. 3735–3754 (2022)