DOI: 10.17148/IARJSET.2024.11515

DocQA: Document Driven Question Answering

Dr Vijayalaxmi Mekali[1], Monish M[2], Likhith V[3], Abhishek A[4], Chandan VK[5]

Professor and Head, Dept of Artificial Intelligence and Machine Learning, K S Institute of Technology, Bengaluru, Karnataka, India[1]

Student, Dept of Artificial Intelligence and Machine Learning, K S Institute of Technology, Bengaluru, Karnataka, India[2-5]

Abstract: The project focuses on enhancing Document Question Answering (DocQA) systems in the financial sector through the integration of the fine-tuned DoNUT model. This model enables swift and accurate extraction of crucial information from various financial documents, facilitating faster decision-making and regulatory compliance. Through training and fine-tuning on diverse synthetic financial datasets, the proposed system aims for high precision and recall in extracting key financial information from the forms. Empirical evaluations and case studies within financial institutions aim to quantify the time savings and efficiency gains achieved by AI-driven DocQA systems, highlighting the tangible benefits of the DoNUT-enhanced model. Ultimately, this research underscores the transformative potential of AI-driven document understanding in the financial sector, emphasizing the importance of sophisticated AI models for improving operational efficiency, regulatory compliance.

Keywords: Document Question Answering (DocQA), DoNUT, artificial intelligence (AI), key-value pair extraction.

I. INTRODUCTION

In recent years, document question answering (DocQA) has emerged as a vital area of research and development, with applications spanning from information retrieval to virtual assistants and automated document processing systems. The ability to accurately extract information and provide relevant answers from documents has significant implications for various industries, including finance, legal, healthcare, and government sectors.

Traditional approaches to DocQA often rely on manually annotated datasets, which can be time-consuming and expensive to create, especially for tasks involving complex document structures and diverse content types. Additionally, the availability of labeled data may be limited, leading to challenges in training robust and generalized models.

To address these challenges, we present a novel approach that leverages synthetic data generation and transfer learning techniques to enhance the performance of DocQA systems. Our methodology revolves around fine-tuning the DoNUT (Document Neural Understanding Toolkit) model on a synthetically generated dataset, enabling the model to effectively answer questions based on various types of documents, including tax forms (e.g., 1099-DIV, 1099-INT, W-2, W-3).

II. LITERATURE SURVEY

TABLE 1 LITERATURE SURVEY

Sl. No.	Year	Title	Description
1	2023	BloombergGPT: A Large Language Model for Finance	BloombergGPT, tailored for financial applications, exhibits remarkable performance in this domain. However, its applicability outside finance is constrained due to a lack of domain specificity, potentially limiting its versatility. Furthermore, the model's immense size, with 50 billion parameters, contributes to significant computational costs and extensive training times, posing practical challenges for broader deployment.

2	2023	DocTr: Document Transformer for Structured Information Extraction in Documents	DocTr presents an innovative solution for structured information extraction. Despite its novel approach, the model's domain-agnostic nature may result in suboptimal performance in specialized fields. The reliance on a pre-training strategy adds a layer of complexity, especially in scenarios where annotated data for fine-tuning is limited.
3	2023	DocFormerv2: Local Features for Document Understanding	DocFormerv2 showcases state-of-the-art performance in Visual Document Understanding (VDU). However, its efficacy is contingent upon the quality and diversity of pre-training tasks, potentially limiting its adaptability to languages or domains not well-covered in pre-training. This poses challenges for understanding languages beyond the model's training scope.
4	2022	BROS: A Pre-trained Language Model Focusing on Text and Layout for Better Key Information Extraction from Documents	BROS takes a unique approach to key information extraction, encoding 2D spatial information. While it effectively addresses information extraction challenges, the model may encounter difficulties with intricate layouts or non-standard document structures. Additionally, its efficiency in handling diverse languages and domains requires further exploration.
5	2022	FormNet: Structural Encoding beyond Sequential Modeling in Form Document Information Extraction	FormNet excels in correcting token serialization issues, enhancing information extraction from forms. However, its performance might be compromised when dealing with forms that significantly deviate from the structures seen in its training data. This limitation raises concerns about the model's adaptability to unconventional layouts.
6	2022	OCR-free Document Understanding Transformer	Donut's OCR-free approach overcomes challenges associated with OCR, yet its effectiveness relies on the quality of the synthetic data generator. The Transformer architecture, while advantageous, may face difficulties with highly complex document structures. These limitations suggest considerations for improving synthetic data quality and handling intricate layouts.
7	2021	Cost-effective End-to-end Information Extraction for Semi-structured Document Images	The proposed end-to-end IE model simplifies development but may encounter performance degradation with highly specialized or complex document types. Its reliance on a sequence generation task might limit efficacy in scenarios with varied document structures, prompting further exploration of tailored solutions for diverse document types.
8	2021	Spatial Dependency Parsing for Semi-Structured Document Information Extraction	SPADEs efficiently tackle spatial complexities in information extraction. However, it may face challenges with diverse document types, particularly those with unconventional spatial relationships. The model's performance could be impacted when applied to documents with irregular layouts, suggesting a need for robustness in handling varied spatial structures.
9	2020	Representation Learning for Information Extraction from Form-like Documents	The system excels in extracting structured information but may lack domain specificity, affecting precision for specialized document types. Additionally, the computational cost associated with representation learning poses

			limitations, emphasizing the need for efficient solutions to handle large-scale datasets.
10	2020	LayoutLM: Pre-training of Text and Layout for Document Image Understanding	LayoutLM innovates document-level pre-training, incorporating layout and style information. However, challenges arise in scenarios where such information plays a less crucial role. The model's reliance on visual details may lead to computational costs, especially when dealing with large-scale document datasets.
11	2019	Table Detection in Invoice Documents by Graph Neural Networks	The graph-based approach excels in table detection, leveraging structure perception. However, its effectiveness may vary with irregular or non-tabular document structures. The model's reliance on structure perception may limit its performance on documents with unconventional layouts, necessitating improvements for enhanced adaptability.
12	2019	RoBERTa: A Robustly Optimized BERT Pretraining Approach	While shedding light on BERT's undertraining, the study's limitations include a lack of openness, hindering community collaboration. Additionally, BERT's general language understanding might not be universally strong, impacting performance across diverse linguistic contexts. Improved transparency and language adaptability could enhance the study's contributions.

III. METHODOLOGY

Steps followed for the project is as given below:

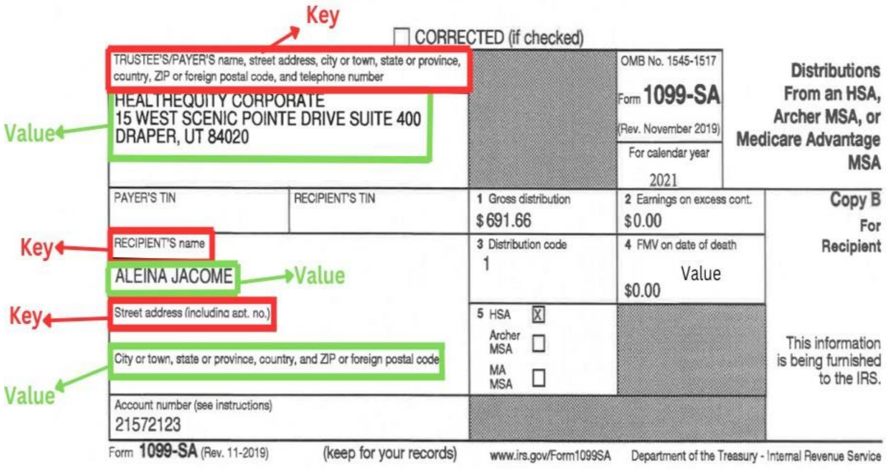
1. Dataset Preparation
2. Fine-Tuning DoNUT Model
3. Dataset Preparation for Classification Model
4. Fine-Tuning ResNet 152 model
5. Integration of RAG
6. User Interface and Interaction

1. Dataset Preparation

The initial stage of our project, dataset preparation, is foundational for the success of the entire system. It begins with the procurement of blank tax forms from the IRS website, specifically 1099-DIV, 1099-INT, W-2, and W-3 forms, which cover a broad range of financial reporting requirements. To populate these forms, we employ the Faker library, which is adept at generating realistic-looking but entirely synthetic personal and financial data.

This step is critical as it ensures that our model encounters a variety of data points similar to what would be seen in real-world scenarios. Following data generation, each form undergoes a meticulous annotation process. This process involves the manual labeling of each field with text entries and bounding boxes that are essential for the training of our models. This ensures that our models learn not only the text but also the spatial layout of the forms. Finally, the dataset is split into training, validation, and testing subsets at proportions of 75%, 10%, and 15%, respectively. This division is strategically chosen to maximize learning while ensuring robust model evaluation.

DOI: 10.17148/IARJSET.2024.11515


Key
Value
Key
Value
Key
Value
Value

□ CORRECTED (if checked)

TRUSTEE'S/PAYER'S name, street address, city or town, state or province, country, ZIP or foreign postal code, and telephone number HEALTHEQUITY CORPORATE 15 WEST SCENIC POINTE DRIVE SUITE 400 DRAPER, UT 84020			OMB No. 1545-1517 Form 1099-SA (Rev. November 2019) For calendar year 2021	Distributions From an HSA, Archer MSA, or Medicare Advantage MSA
PAYER'S TIN	RECIPIENT'S TIN	1 Gross distribution $ 691.66	2 Earnings on excess cont. $ 0.00	Copy B For Recipient
RECIPIENT'S name ALEINA JACOME		3 Distribution code 1	4 FMV on date of death $0.00	
Street address (including apt. no.) City or town, state or province, country, and ZIP or foreign postal code		5 HSA ☒ Archer MSA □ MA MSA □		This information is being furnished to the IRS.
Account number (see instructions) 21572123				

Form 1099-SA (Rev. 11-2019) (keep for your records) www.irs.gov/Form1099SA Department of the Treasury - Internal Revenue Service

Fig 1. Sample key-value pair form

The final form of the ground truth data after preprocessed will be in the form of a json for the corresponding images; an example is shown below.

```
{"items":
  {"1 Wages, tips, other compensation": 78224.41,
    "2 Federal income tax withheld": 31954.74,
    "3 Social security wages": 20808.86,
    "4 Social security tax withheld": 10203.49,
    "5 Medicare wages and tips": 77064.62,
    "6 Medicare tax withheld": 65581.02,
    "7 Social security tips": 83313.92,
    "8 Allocated tips": 54571.24,
    "10 Dependent care benefits": 8416.72,
    "11 Nonqualified plans": "",
    "12a": 71399.49,
    "12b": 6740.31,
    "12c": 68743.94,
    "12d": 25656.95,
    "14 Other": "",
    "15 State": "TN",
    "16 State wages, tips, etc. ": 48677.08,
    "17 State income tax": 11799.75,
    "18 Local wages, tips, etc.": 805.22,
    "19 Local income tax": 85816.48,
    "20 Locality name": "Austin",
    "Form": "W-2 Wage and Tax Statement",
    "OMB No.": "1545-0008",
    "For calendar year": 2024
  }
}
```

2. *Fine-Tuning DoNUT Model*

Fine-tuning the DoNUT model is a critical process tailored to enhance its performance on our specific dataset. The DoNUT model, originally designed for general document understanding tasks, is adapted through training on our annotated tax forms. Over the course of three epochs, the model learns detailed nuances of the dataset, including variations in text placement and form layouts unique to tax documentation. This training process is carefully monitored to balance learning and overfitting, with a significant achievement of 97% accuracy and an impressively low validation edit distance of 0.0434. These metrics not only signify the model's ability to accurately extract and interpret text but also its efficiency in dealing with the complex structures of tax forms.

DOI: 10.17148/IARJSET.2024.11515

3. *Dataset Preparation for Classification Model*

In addition to text extraction, we extend our system's capabilities to classify entire documents. A new dataset, comprising 2,500 samples, is prepared to train a document classification model. This dataset includes various types of tax forms and non-form documents to challenge the model's ability to distinguish between different document types effectively. Utilizing the ResNet-152 architecture, a deep convolutional neural network pre-trained on the ImageNet dataset, we apply transfer learning techniques. This approach leverages learned features from generic large-scale image recognition tasks, adapting them to our specific classification task. The model is fine-tuned with adjustments in hyperparameters to optimize performance, ensuring it can accurately classify not only different types of forms but also identify non-form documents.

4. *Fine-tuning ResNet 152 Model*

In this step, the ResNet 152 model is fine-tuned to classify documents into five distinct classes: 1099-div, 1099-int, w2, w3, and non-forms. Unlike traditional fine-tuning approaches, where models are typically adjusted for specific tasks, here, the ResNet 152 architecture is adapted to accommodate document classification. By fine-tuning the model, it learns to recognize the unique features and patterns associated with each document type, enabling accurate classification. The training process involves feeding the model with labeled data from the prepared dataset, which includes examples of each document type. During training, the model adjusts its internal parameters through backpropagation, optimizing its performance for the classification task. Fine-tuning ResNet 152 for document classification enhances its ability to differentiate between various forms and non-form documents, providing a solid foundation for the subsequent stages of the pipeline.

5. *Integration of RAG*

The integration of the Retrieval-Augmented Generation (RAG) system represents a pivotal advancement in augmenting the interactivity and intelligence of our document processing system. By seamlessly incorporating LLaMAParse for content extraction from non-form PDFs and integrating it with the powerful LLaMA-3 model, our system achieves a remarkable capability to interpret and respond to natural language queries with finesse. LLaMA-3's exceptional reasoning abilities elevate the quality of responses, ensuring not only accuracy but also contextual relevance, thereby significantly enriching the user experience. This comprehensive setup empowers the system to adeptly handle intricate queries, extract pertinent information from documents, and generate coherent responses that resonate with the user's intent. By bridging the gap between mere document retrieval and interactive document comprehension, our integrated RAG pipeline establishes a new standard of sophistication in document processing systems, offering users a seamless and intuitive interface for accessing and interacting with their documents effectively.

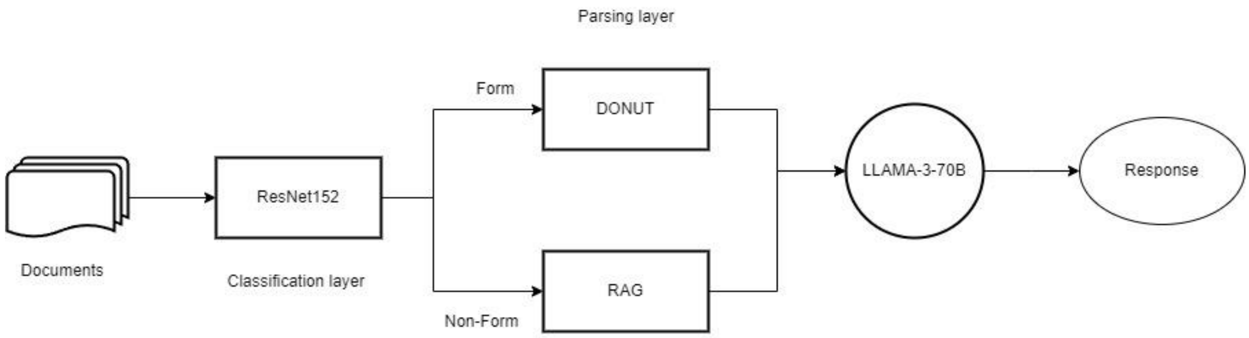


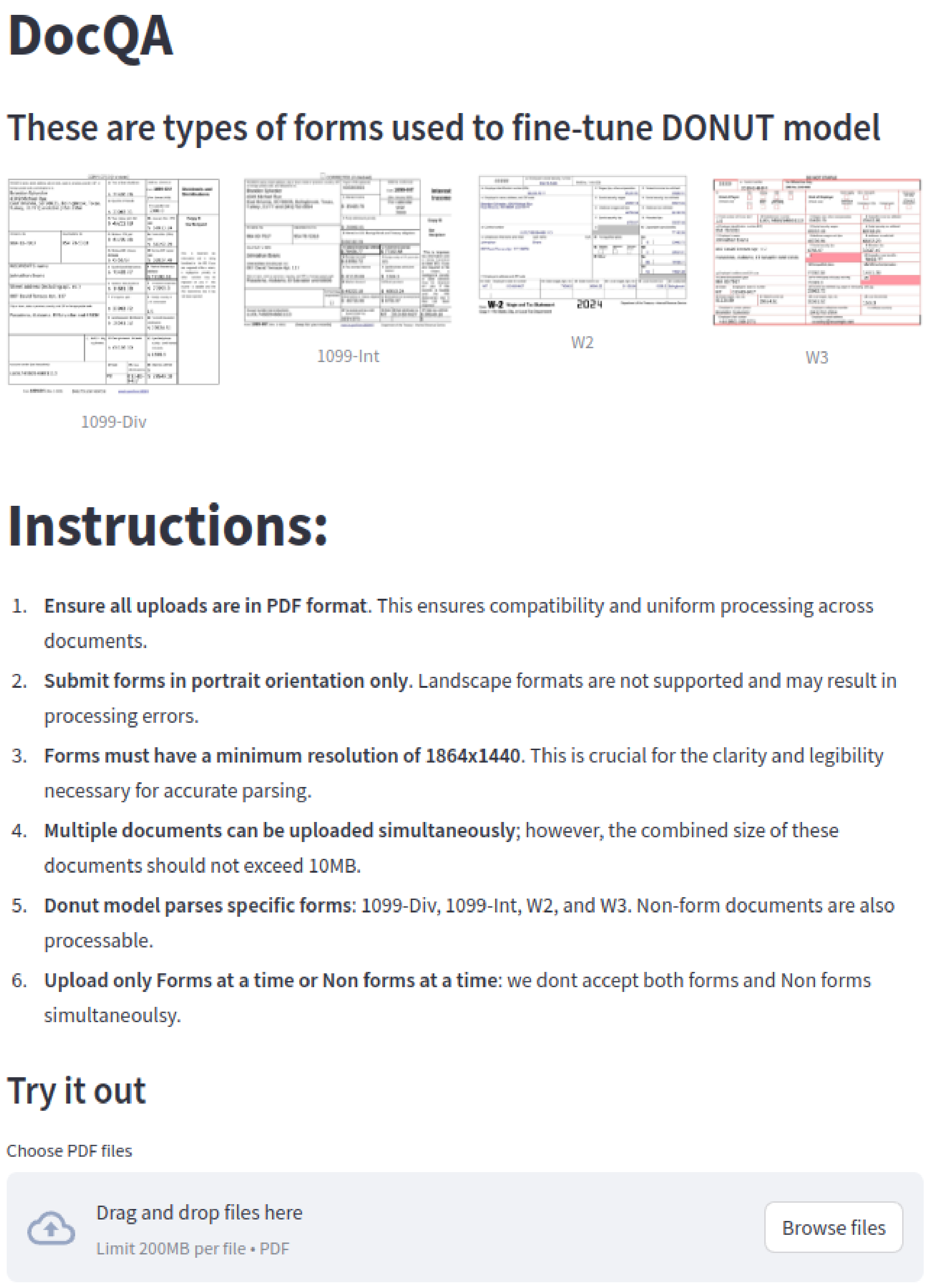
Fig 2. Pipeline of the project

6. *User Interface and Interaction*

The user interface is designed with simplicity and usability in mind, aiming to accommodate users from varied technical backgrounds. Through a minimalist and intuitive design, users are able to upload PDF files to the system effortlessly. Once a document is uploaded, the system classifies and processes the document, preparing it for interaction. Users can then pose questions directly related to the document content, engaging in a chat-like interaction. This interface is not just about providing answers but also about facilitating an interactive dialogue between the user and the document, making the data contained within accessible and easily navigable.

Following dataset generation, we manually annotated each sample in the synthetic dataset to provide ground truth labels for training our DocQA model. This annotation process involved generating questions based on the content of the tax forms and providing accurate answers.

Clear guidelines were established to maintain consistency and accuracy in the annotation process, and quality control measures were implemented to resolve any discrepancies or ambiguities. Through an iterative refinement process based on feedback and quality assessments, we continually improved the accuracy and reliability of the ground truth labels.



DocQA

These are types of forms used to fine-tune DONUT model

Instructions:

1. **Ensure all uploads are in PDF format**. This ensures compatibility and uniform processing across documents.
2. **Submit forms in portrait orientation only**. Landscape formats are not supported and may result in processing errors.
3. **Forms must have a minimum resolution of 1864x1440**. This is crucial for the clarity and legibility necessary for accurate parsing.
4. **Multiple documents can be uploaded simultaneously**; however, the combined size of these documents should not exceed 10MB.
5. **Donut model parses specific forms**: 1099-Div, 1099-Int, W2, and W3. Non-form documents are also processable.
6. **Upload only Forms at a time or Non forms at a time**: we dont accept both forms and Non forms simultaneoulsy.

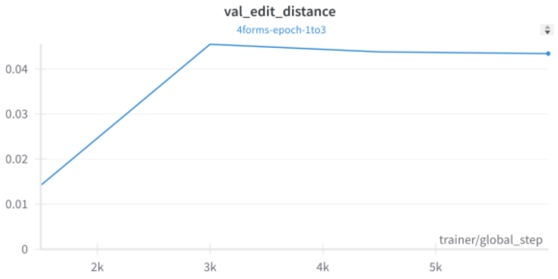
Try it out

Choose PDF files

Fig 3. UI screenshot

IV. RESULT

The culmination of our project represents a significant advancement in the field of document analysis and interaction, focusing on the robust integration of machine learning technologies to process and understand diverse document types. By synthesizing and annotating a specialized dataset containing types such as 1099-DIV, 1099-INT, W-2, and W-3 forms, and fine-tuning the DoNUT model on this data, we achieved a remarkable accuracy of 97%, with a validation edit distance of 0.0434 and train loss of 0.008. This demonstrates the model's exceptional ability to accurately extract and interpret complex textual data from structured documents.



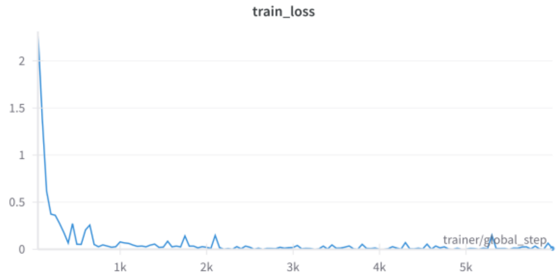


Fig 4. Val edit distance of our model

Fig 5. Train loss of our model

We extended our system's capabilities by developing a classification model capable of distinguishing between different form types as well as identifying non-form documents. Utilizing a transfer learning approach with the ResNet-152 model, our system was adept at handling a dataset of 2,500 samples, ensuring robust document categorization.

This step was crucial for enhancing the system's applicability across various document-related tasks and environments.

Table 2 MODEL ACCURACY

TASK	MODEL	ACCURACY
Document parsing	Donut-base fine-tuned	97.43%
Document classification	ResNet transfer learning	98.06%

Integration of the Retrieval-Augmented Generation (RAG) system using LLaMAParse and the advanced LLaMA-3 model has been a game-changer. This integration allows for seamless interaction between the user and the document content, enabling the system to not only retrieve information but also generate context-aware responses to queries. This capability ensures a dynamic, interactive user experience, leveraging deep learning to provide intelligent, reasoned responses.

The user interface, designed with a focus on simplicity and accessibility, facilitates an intuitive interaction where users can easily upload documents and engage in a natural, conversational manner with the system. This interface bridges the gap between complex document processing technologies and end-user usability, making advanced document analysis accessible to a broader audience.

Overall, the project successfully combines advanced document understanding, content classification, and user interaction into a cohesive, highly functional system, setting a new benchmark for document-oriented query-answer platforms.

DOI: 10.17148/IARJSET.2024.11515

V. CONCLUSION

The project has successfully demonstrated the integration of advanced document processing technologies into a highly functional query-answer system. By fine-tuning the DoNUT model and employing sophisticated classification techniques, the system achieves high accuracy in text extraction and document categorization. The use of a Retrieval-Augmented Generation system enhances user interaction, allowing for dynamic and contextually aware responses. The intuitive user interface further ensures that the system is accessible to a broad user base. The project not only pushes the boundaries of document interaction technologies but also significantly enhances the ease and efficiency with which users can engage with document content.

REFERENCES

[1]. Wu, Shijie, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg and Gideon Mann. “BloombergGPT: A Large Language Model for Finance.” ArXiv abs/2303.17564 (2023).

[2]. Haofu Liao, Aruni RoyChowdhury, Weijian Li, Ankan Bansal, Yuting Zhang, Zhuowen Tu, Ravi Kumar Satzoda, R. Manmatha, Vijay Mahadevan. DocTr: Document Transformer for Structured Information Extraction in Documents. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19584-19594 (2023).

[3]. Appalaraju, Srikar & Tang, Peng & Dong, Qi & Sankaran, Nishant & Zhou, Yichu & Manmatha, R.. DocFormerv2: Local Features for Document Understanding.(2023).

[4]. 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. Proceedings of the AAAI Conference on Artificial Intelligence 36(10), 10767–10775 (Jun 2022).

[5]. Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, Tomas Pfister. FormNet: Structural Encoding beyond Sequential Modeling in Form Document Information Extraction (2022).

[6]. Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-Free Document Understanding Transformer. In Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII. Springer-Verlag, Berlin, Heidelberg, 498–517

[7]. Hwang, W., Lee, H., Yim, J., Kim, G., Seo, M.: Cost-effective end-to-end information extraction for semi-structured document images. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3375–3383. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021)

[8]. Hwang, W., Yim, J., Park, S., Yang, S., Seo, M.: Spatial dependency parsing for semi-structured document information extraction. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 330–343. Association for Computational Linguistics, Online (Aug 2021).

[9]. Majumder, B.P., Potti, N., Tata, S., Wendt, J.B., Zhao, Q., Najork, M.: Representation learning for information extraction from form-like documents. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 6495–6504. Association for Computational Linguistics, Online (Jul 2020).

[10]. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 1192–1200. KDD ’20, Association for Computing Machinery, New York, NY, USA (2020)

[11]. Riba, P., Dutta, A., Goldmann, L., Forn´es, A., Ramos, O., Llad´os, J.: Table detection in invoice documents by graph neural networks. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 122–127 (2019).

[12]. Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized Bert pretraining approach." 2019.