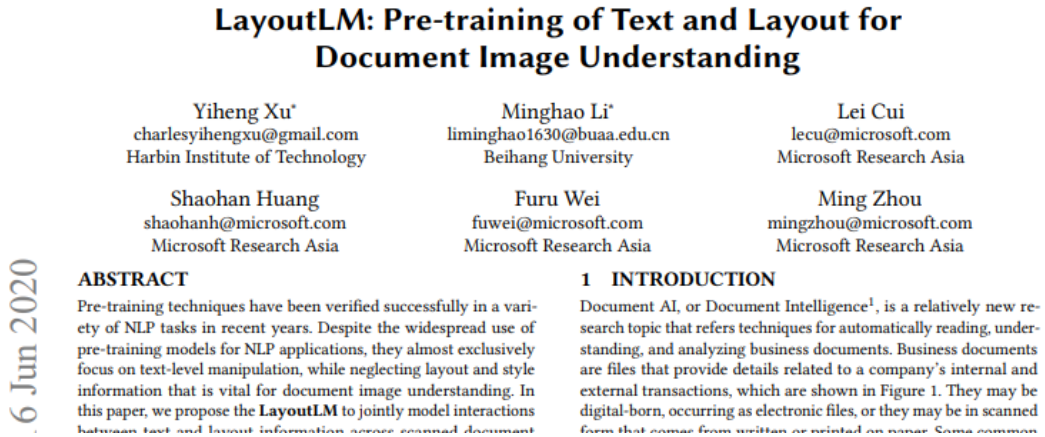
Our paper: Recent progress in multimodal large language models (LLMs) has significantly improved document interpretation by jointly modeling textual, visual, and spatial information. LayoutLM introduced multimodal pretraining that integrates text tokens with two-dimensional layout embeddings, enabling improved performance in tasks such as form understanding and invoice processing \cite{b1}.

Literature reference: [1] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of Text and Layout for Document Image Understanding," in *Proc. ACM SIGKDD*, 2020.

Link: Paper (arXiv): https://arxiv.org/abs/1912.13318

https://arxiv.org/pdf/1912.13318

Screenshot:



LayoutLM: Pre-training of Text and Layout for Document Image Understanding

Yiheng Xu*
charlesyihengxu@gmail.com
Harbin Institute of Technology

Minghao Li*
liminghao1630@buaa.edu.cn
Beihang University

Lei Cui
lecu@microsoft.com
Microsoft Research Asia

Shaohan Huang
shaohanh@microsoft.com
Microsoft Research Asia

Furu Wei
fuwei@microsoft.com
Microsoft Research Asia

Ming Zhou
mingzhou@microsoft.com
Microsoft Research Asia

ABSTRACT

Pre-training techniques have been verified successfully in a variety of NLP tasks in recent years. Despite the widespread use of pre-training models for NLP applications, they almost exclusively focus on text-level manipulation, while neglecting layout and style information that is vital for document image understanding. In this paper, we propose the **LayoutLM** to jointly model interactions

1 INTRODUCTION

Document AI, or Document Intelligence[1], is a relatively new research topic that refers techniques for automatically reading, understanding, and analyzing business documents. Business documents are files that provide details related to a company's internal and external transactions, which are shown in Figure 1. They may be digital-born, occurring as electronic files, or they may be in scanned

Our paper: LayoutLMv2 extends this approach by incorporating image features and spatially aware attention mechanisms, further improving extraction accuracy and document reasoning \cite{b2}.

Literature reference: [2] Y. Xu, Y. Li, T. Zhang, et al., "LayoutLMv2: Multimodal Pre-training for Visually-rich Document Understanding," in *Proc. ACL*, 2021.

Link: Paper (arXiv): https://arxiv.org/abs/2012.14740

PDF: https://arxiv.org/pdf/2012.14740.pdf

LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding

Yang Xu[1*], Yiheng Xu[2*], Tengchao Lv[2*], Lei Cui[2], Furu Wei[2], Guoxin Wang[3], Yijuan Lu[3], Dinei Florencio[3], Cha Zhang[3], Wanxiang Che[1], Min Zhang[4], Lidong Zhou[2]
[1]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology
[2]Microsoft Research Asia [3]Microsoft Azure AI [4]Soochow University
[1]{yxu,car}@ir.hit.edu.cn,
[2]{v-yixu,v-telv,lecu,fuwei,lidongz}@microsoft.com,
[3]{guow,yijlu,dinei,chazhang}@microsoft.com [4]minzhang@suda.edu.cn

2022

Abstract

Pre-training of text and layout has proved effective in a variety of visually-rich document understanding tasks due to its effec- applications. Distinct from conventional information extraction tasks, the VrDU task relies on not only textual information but also visual and layout information that is vital for visually-rich docu-

Our paper: LayoutLMv3 unified text and image masking during pretraining, demonstrated improved generalization across various document understanding tasks \cite{b3}.

Literature reference: [3] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking," *arXiv preprint* arXiv:2204.08387, 2022.

Link: Paper (arXiv): https://arxiv.org/abs/2204.08387

PDF: https://arxiv.org/pdf/2204.08387.pdf

LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking

Yupan Huang*
Sun Yat-sen University
huangyp28@mail2.sysu.edu.cn

Tengchao Lv
Microsoft Research Asia
tengchaolv@microsoft.com

Lei Cui
Microsoft Research Asia
lecu@microsoft.com

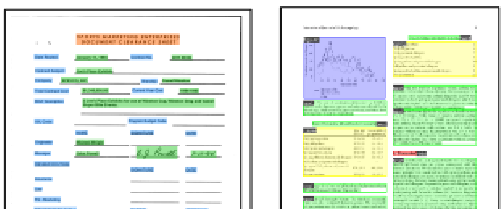
Yutong Lu
Sun Yat-sen University
luyutong@mail.sysu.edu.cn

Furu Wei
Microsoft Research Asia
fuwei@microsoft.com

19 Jul 2022

ABSTRACT

Self-supervised pre-training techniques have achieved remarkable progress in Document AI. Most multimodal pre-trained models use a masked language modeling objective to learn bidirectional representations on the text modality, but they differ in pre-training objectives for the image modality. This discrepancy adds difficulty to multimodal representation learning. In this paper, we propose **LayoutLMv3** to pre-train multimodal Transformers for Document

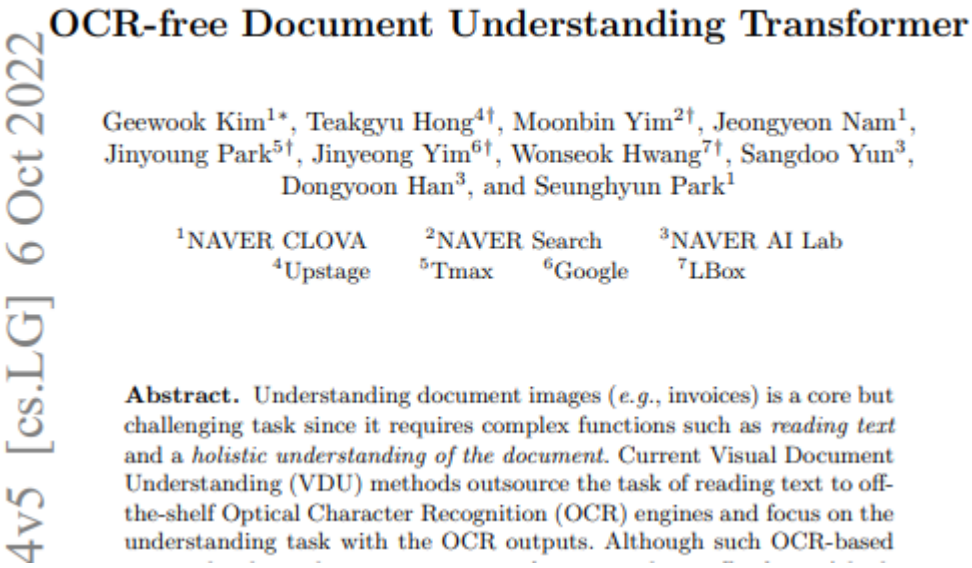


Our paper: In contrast to OCR-dependent approaches, Donut proposed an end-to-end transformer architecture capable of generating structured outputs directly from document images \cite{b4}.

Literature reference: [4] G. Kim, T. Hong, M. Yim, et al., “Donut: Document Understanding Transformer without OCR,” in *Proc. ECCV*, 2022.

Link: Paper (arXiv): https://arxiv.org/abs/2111.15664

PDF: https://arxiv.org/pdf/2111.15664.pdf



4v5 [cs.LG] 6 Oct 2022

OCR-free Document Understanding Transformer

Geewook Kim[1*], Teakgyu Hong[4†], Moonbin Yim[2†], Jeongyeon Nam[1], Jinyoung Park[5†], Jinyeong Yim[6†], Wonseok Hwang[7†], Sangdoo Yun[3], Dongyoon Han[3], and Seunghyun Park[1]

[1]NAVER CLOVA [2]NAVER Search [3]NAVER AI Lab
[4]Upstage [5]Tmax [6]Google [7]LBox

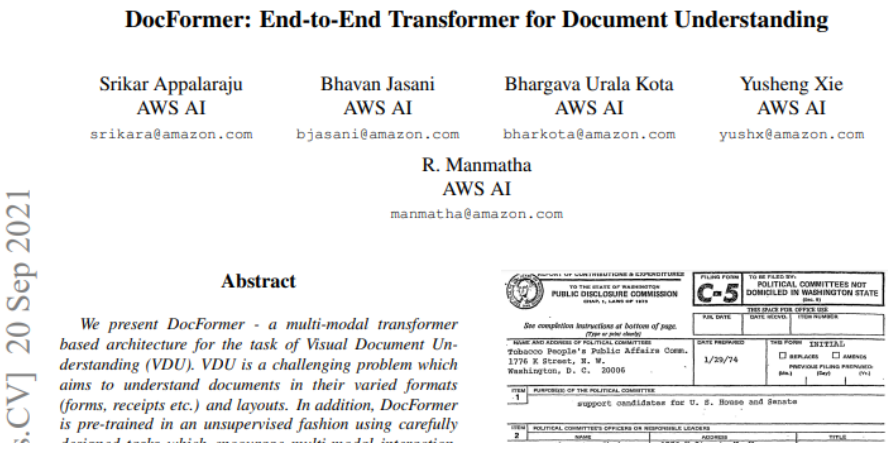
Abstract. Understanding document images (*e.g.*, invoices) is a core but challenging task since it requires complex functions such as *reading text* and a *holistic understanding of the document*. Current Visual Document Understanding (VDU) methods outsource the task of reading text to off-the-shelf Optical Character Recognition (OCR) engines and focus on the understanding task with the OCR outputs. Although such OCR-based

Our paper: Other multimodal models such as DocFormer and StrucTexT further enhanced extraction of key-value pairs and relationships by integrating textual, visual, and structural representations \cite{b5,b6}.

Literature reference: [5] S. Appalaraju, S. S. Datta, Y. N. Murthy, and R. T. Iyer, “DocFormer: End-to-End Transformer for Document Understanding,” *arXiv preprint* arXiv:2106.11539, 2021.

Link: Paper (arXiv): https://arxiv.org/abs/2106.11539

PDF: https://arxiv.org/pdf/2106.11539.pdf



DocFormer: End-to-End Transformer for Document Understanding

Srikar Appalaraju
AWS AI
srikara@amazon.com

Bhavan Jasani
AWS AI
bjasani@amazon.com

Bhargava Urala Kota
AWS AI
bharkota@amazon.com

Yusheng Xie
AWS AI
yushx@amazon.com

R. Manmatha
AWS AI
manmatha@amazon.com

Abstract

We present DocFormer - a multi-modal transformer based architecture for the task of Visual Document Understanding (VDU). VDU is a challenging problem which aims to understand documents in their varied formats (forms, receipts etc.) and layouts. In addition, DocFormer is pre-trained in an unsupervised fashion using carefully

Literature reference: [6] Z. Li, W. Chen, X. Xu, et al., "StrucTexT: Structured Text Understanding with Multi-Modal Transformers," *arXiv preprint* arXiv:2108.02923, 2021.

Link: Paper (arXiv): https://arxiv.org/abs/2108.02923

PDF: https://arxiv.org/pdf/2108.02923.pdf

Our paper: While these studies demonstrate strong capabilities in information extraction, they typically do not address the transformation of extracted data into normalized relational database schemas, which is still missing in practice.

Several widely known benchmark datasets have supported the development of models that understand documents. For example, the FUNSD dataset provides annotated forms for entity and relation extraction \cite{b7}.

Literature reference: [7] G. Jaume, H. K. Ekenel, and J. Thiran, "FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents," in *Proc. ICDAR Workshops*, 2019.

Link: Paper (arXiv): https://arxiv.org/abs/1905.13538

PDF: https://arxiv.org/pdf/1905.13538

Our paper: The CORD dataset offers structured annotations for receipt parsing and line-item extraction \cite{b8}.

Literature reference: [8] S. Park, S. Shin, B. Lee, et al., “CORD: A Consolidated Receipt Dataset for Post-OCR Parsing,” in *Proc. ICDAR*, 2019.

Link: Paper: https://arxiv.org/abs/1911.07936

PDF: https://arxiv.org/pdf/1911.07936

Our paper: The DocVQA benchmark evaluates documents through visual question answering, measuring the ability of models to query document content based on layout and textual information \cite{b9}.

Literature reference: [9] M. Mathew, D. Karatzas, and C. V. Jawahar, “DocVQA: A Dataset for VQA on Document Images,” in *Proc. WACV*, 2021.

Paper (arXiv): https://arxiv.org/abs/2007.00398

PDF: https://arxiv.org/pdf/2007.00398

Our paper: Although these datasets provide valuable evaluation resources, they often focus on specific document types and may not fully represent diverse real-world business documents, especially when the documents are noisy or messy.

Another key direction involves table detection and structural reconstruction, which is critical for converting document information into structured formats. Deep learning approaches such as TableNet and DeepDeSRT demonstrate effective table detection and structure recognition from document images \cite{b10,b11}.

Literature reference: [10] A. Paliwal, D. Vishwanath, R. Rahul, M. Sharma, and L. Vig, “TableNet: Deep Learning Model for End-to-End Table Detection and Tabular Data Extraction from Scanned Document Images,” in *Proc. ICDAR*, 2019.

PDF: https://arxiv.org/pdf/2001.01469

https://www.semanticscholar.org/reader/d7284721854bd9db96a9e442caef0609d4324415

Literature reference: [11] S. Schreiber, S. Agne, I. Wolf, A. Dengel, and S.

Ahmed, “DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images,” in *Proc. ICDAR*, 2017.

PDF: https://www.dfki.de/fileadmin/user_upload/import/9672_PID4966073.pdf

Our paper: Subsequent models, including CascadeTabNet and Table Transformer, improved detection accuracy and structural reconstruction using advanced architectures \cite{b12,b14}.

Literature reference: [12] P. Prasad, A. Sarkar, M. P. K. Reddy, and M. N. S. Swamy, “CascadeTabNet: An Approach for End to End Table Detection and Structure Recognition from Image-Based Documents,” in *Proc. CVPR Workshops*, 2020.

Link: Paper (arXiv): https://arxiv.org/abs/2004.12629

PDF: https://arxiv.org/pdf/2004.12629.pdf

Literature reference: [14] B. Smock and R. Abraham, “Table Transformer: A Transformer-based Approach to Table Detection and Structure Recognition,” *arXiv preprint* arXiv:2110.00061, 2021.

Link: Paper (arXiv): https://arxiv.org/abs/2110.00061

PDF: https://arxiv.org/pdf/2110.00061.pdf

Our paper: Large-scale datasets such as SciTSR and PubTables-1M enabled more comprehensive training and evaluation of table extraction systems \cite{b13,b15}.

Literature reference: [13] Z. Chi, H. Huang, H. Yu, et al., “SciTSR: Extracting Table Structure from Scientific Tables,” in *Proc. ICDAR*, 2019.

https://arxiv.org/pdf/1908.04729
https://www.semanticscholar.org/paper/1fb193f5a0d809f12094cad3a55c299969c87baf

Literature reference: [15] B. Smock, R. Pesala, and R. Abraham, “PubTables-1M: Towards Comprehensive Table Extraction from Unstructured Documents,” in *Proc. CVPR*, 2022.

Link:PDF: https://arxiv.org/pdf/2110.00061

Despite this progress, extracting complex tables with unusual layouts, merged cells, and domain-specific structures remains challenging and does not always produce correct outputs. In addition, classic document image understanding tasks such as classification and retrieval also remain important as a supporting layer for many pipelines; Harley et al. evaluated deep convolutional networks for document image classification and retrieval and showed that CNN-based representations can be effective for document images \cite{b16}.

[16] A. W. Harley, A. Ufkes, and K. G. Derpanis, “Evaluation of Deep Convolutional Nets for Document Image Classification and Retrieval,” in *Proc. ICDAR*, 2015.

Link: https://arxiv.org/abs/1502.07058

https://arxiv.org/pdf/1502.07058

Related to this direction, the ICDAR 2019 Scene Text Visual Question Answering competition (ST-VQA) also evaluates question answering over images that contain text, and it is often used as an OCR/VQA benchmark for reasoning about scene text \cite{b17}.

[17] Y. Huang, Q. Liu, and D. Karatzas, “ICDAR 2019 Competition on Scene Text Visual Question Answering (ST-VQA),” in *Proc. ICDAR*, 2019.

Link: https://arxiv.org/pdf/1907.00490

Our paper: Despite this progress, extracting complex tables with unusual layouts, merged cells, and domain-specific structures remains challenging and does not always produce correct outputs.

In addition to extraction, transforming document-derived data into usable relational databases requires schema reasoning and evaluation of queryability. The Spider dataset introduced a large-scale benchmark for cross-domain text-to-SQL generation \cite{b18}

Literature reference: [18] T. Yu, R. Zhang, K. Yasunaga, et al., “Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic

Parsing and Text-to-SQL Task," in *Proc. EMNLP*, 2018.
Link: Paper (arXiv): https://arxiv.org/abs/1809.08887

https://arxiv.org/pdf/1809.08887

Our paper:, while RAT-SQL improved schema-aware encoding to enhance query generation accuracy \cite{b19}.

Literature reference: [19] B. Wang, R. Shin, X. Liu, O. Polozov, and M. Richardson, "RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers," in *Proc. ACL*, 2020.

Link: Paper (ACL Anthology): https://aclanthology.org/2020.acl-main.677/

https://aclanthology.org/2020.acl-main.677.pdf

Our paper: PICARD further improved SQL validity by applying constrained decoding during query generation \cite{b20}.

Literature reference: [20] T. Scholak, R. Schucher, and P. Bahdanau, "PICARD: Parsing Incrementally for Constrained Auto-Regressive Decoding from Language Models," in *Proc. EMNLP*, 2021.

Link: Paper (arXiv): https://arxiv.org/abs/2109.05093

PDF: https://arxiv.org/pdf/2109.05093.pdf

Our paper: In database research, automated discovery of functional dependencies and normalization has been explored through frameworks such as Metanome and algorithms like TANE, which support schema refinement and relational consistency \cite{b21,b22}.

Literature reference: [21] F. Papenbrock, J. Ehrlich, J. Marten, T. Neubert, J.-P. Rudolph, T. Kruse, J. Schmidl, and F. Naumann, "Functional Dependency Discovery: An Experimental Evaluation of Seven Algorithms," *PVLDB*, vol. 8, no. 10, 2015. *(Metanome ecosystem / dependency discovery line.)*

Link: PVLDB page: https://www.vldb.org/pvldb/vol8/p1082-papenbrock.pdf

Literature reference: [22] J. Huhtala, J. Kärkkäinen, P. Porkka, and H. Toivonen, "TANE: An Efficient Algorithm for Discovering Functional and

Approximate Dependencies," *The Computer Journal*, vol. 42, no. 2, 1999.

https://dm-gatech.github.io/CS8803-Fall2018-DML-Papers/tane.pdf

Our paper: However, these methods generally assume structured and relatively clean input data, which is not always the case for document-derived information and can lead to failures.

Recent studies have also investigated the integration of large language models with external tools and multimodal reasoning. Toolformer demonstrated that language models can learn to invoke external tools during inference \cite{b23}

Literature reference: [23] T. Schick, J. Dwivedi-Yu, R. Dessì, et al., "Toolformer: Language Models Can Teach Themselves to Use Tools," *arXiv preprint* arXiv:2302.04761, 2023.
Link: Paper (arXiv): https://arxiv.org/abs/2302.04761

PDF: https://arxiv.org/pdf/2302.04761.pdf

Our paper:, while ReAct introduced a reasoning-and-action framework for controlled task execution. \cite{b24}

Literature reference: [24] S. Yao, J. Zhao, D. Yu, et al., "ReAct: Synergizing Reasoning and Acting in Language Models," *arXiv preprint* arXiv:2210.03629, 2022.

Paper (arXiv): https://arxiv.org/abs/2210.03629

PDF: https://arxiv.org/pdf/2210.03629.pdf

Our paper: Vision-language models such as LLaVA and BLIP-2 further expanded multimodal reasoning by combining visual encoders with large language models. \cite{b25,b26}

Literature reference: [25] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," *arXiv preprint* arXiv:2304.08485, 2023. *(LLaVA)*

Link: Paper (arXiv): https://arxiv.org/abs/2304.08485

PDF: https://arxiv.org/pdf/2304.08485.pdf

Literature reference: [26] J. Li, D. Li, S. Savarese, and S. Ermon, “BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models,” *arXiv preprint* arXiv:2301.12597, 2023.

Link: Paper (arXiv): https://arxiv.org/abs/2301.12597

PDF: https://arxiv.org/pdf/2301.12597.pdf