Document AI with LLM

심규호

고려대학교
KOREA UNIVERSITY

LayTextLLM: A Bounding Box is Worth One Token - Interleaving Layout and Text in a Large Language Model for Document Understanding

DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding

Dongsheng Wang*, **Natraj Raman***, **Mathieu Sibue***
Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, Xiaomo Liu
JPMorgan AI Research
first.last@jpmorgan.com

Document AI with LLM

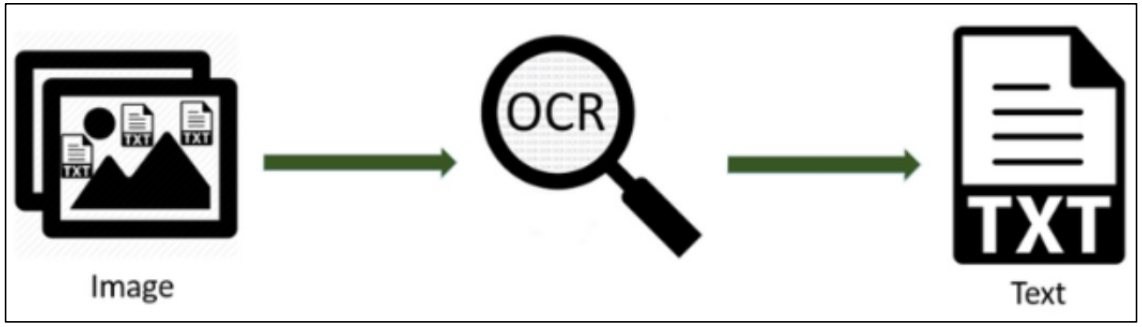
Intro

- Document AI (Intelligent Document Processing)
  - Document를 자동으로 **1.읽어오기 2. 이해하기 3. 분석하기**
  - Challenge:
    - Layout과 format의 *다양성*, scan된 document의 *퀄리티*, 구조의 *복잡성*
  - 태스크 - **DocVQA**(Document Visual Question Answering), **KIE**(Key Information Extraction), **VRDU**(Visually Rich Document Understanding)

Document AI with LLM

OCR (Optical Character Recognition)

- **OCR**
  - Optical Character Recognition 광학 문자 인식
    - 이미지 및 문서에서 텍스트 및 데이터를 추출하고 구조화된 데이터로 변환
  - Input: 이미지
  - Output: Block 단위 정보 – text + bounding box 좌표 [x_1,y_1,x_2,y_2]



Document AI with LLM

Human-Readable way

sion and utilization of document layouts. The proposed layout instruction tuning strategy consists of two components: Layout-aware Pre-training and Layout-aware Supervised Fine-tuning. To capture the characteristics of document layout in Layout-aware Pre-training, three groups of pre-training tasks, corresponding to document-level, region-level and segment-level information, are introduced. Furthermore, a novel module called layout chain-of-thought (LayoutCoT) is devised to enable LayoutLLM to focus on regions relevant to the question and generate accurate answers. LayoutCoT is effective for boosting the performance of document understanding. Meanwhile, it brings a certain degree of interpretability, which could facilitate manual inspection and correction. Experiments on standard benchmarks show that the proposed LayoutLLM significantly outperforms existing methods that adopt open-source 7B LLMs/MLLMs for document understanding.

1. Introduction

Document AI [7], including its document understanding tasks such as document VQA [33, 47] and document visual information extraction [18, 19, 37], is currently a hot topic in both academia and industry. In recent years, document pre-trained models [2, 8, 12, 13, 16, 17, 23, 25, 26, 32, 38, 52, 54, 55, 59] have achieved excellent performance in document AI downstream tasks. However, due to the necessity for fine-tuning on corresponding downstream task data, it is challenging to directly adapt such pre-trained models for *zero-shot* document understanding. In this paper, *zero-shot* refers to not using training sets of downstream tasks.

Figure 1. LLMs/MLLMs for document understanding. The LayoutLLM is an LLM/MLLM based method that integrates a document pre-trained model as encoder. It is trained by the newly proposed layout instruction tuning strategy which consists of Layout-aware Pre-training and Layout-aware Supervised Fine-tuning.

Recently, large language models (LLMs) such as ChatGPT [35] and LLaMA [49, 50], or multimodal large language models (MLLMs) like GPT-4V [1, 36, 56], have shown remarkable zero-shot capabilities across various applications. For Document AI, as shown in Fig. 1 (a), (I) directly prompting LLMs with document text [15, 39] and (II) training document-based MLLMs [3, 57, 60] have also achieved promising results under the zero-shot setting [3, 39, 57, 60].

It is widely accepted that document layout information is vital for document understanding [2, 8, 12, 13, 16, 17, 23, 25, 26, 32, 38, 41, 52, 54, 55, 59]. However, it is difficult to convey document layout information by directly feeding text to LLMs. As Fig. 1(a)(I) shows, representing documents as either flattened plain text or layout text such as text with coordinates [15, 39, 44, 64] is often used for LLMs.

*Equal contribution.

Document AI with LLM

Approaches

- OCR-Free MLLM
  - Image info를 visual signals로 받아들임
  - 복합적인 **Vision Backbone** 구조
  - image 해상도와 관련한 이슈
  - 상당한 컴퓨팅 비용
  - Text와 구조에 대한 semantic 요소들에 대한 이해가 떨어짐
  - Donut(Eccv 2022), Pix2Struct(ICML 2023), mPLUG-DocOwl1.5(EMNLP 2024), Qwen2-VL(CVPR 2024), TextMonkey(CoRR 2024)
- off-the-shelf OCR+LLM
  - **LLM의 일반화 능력**
  - LLM에 대한 입력으로 **Bounding box(text+x,y coordinates)**
  - spatial layouts as **"lightweight visual info"**
  - 상대적으로 적은 컴퓨팅 비용
  - DocLLM(ACL 2024), LayoutLLM(CVPR 2024), LaytextLLM(arxiv 2024), DocLayLLM(CoRR 2024), LMDX(arxiv 2024)

Document AI with LLM

LLM-based approach

- Approaches
  - Document Understanding을 위한 OCR-Based LLMs
    - Off-the-shelf OCR에서 추출된 텍스트들을 바탕으로 LLM의 Reasoning
    - Coordinate-as-tokens
    - DocLLM(ACL 2024), LayoutLLM(CVPR 2024), LaytextLLM(arxiv 2024), DocLayLLM(CoRR 2024), LMDX(arxiv 2024)

[x_min, y_min, x_max, y_max]

HARRISBURG 78|09

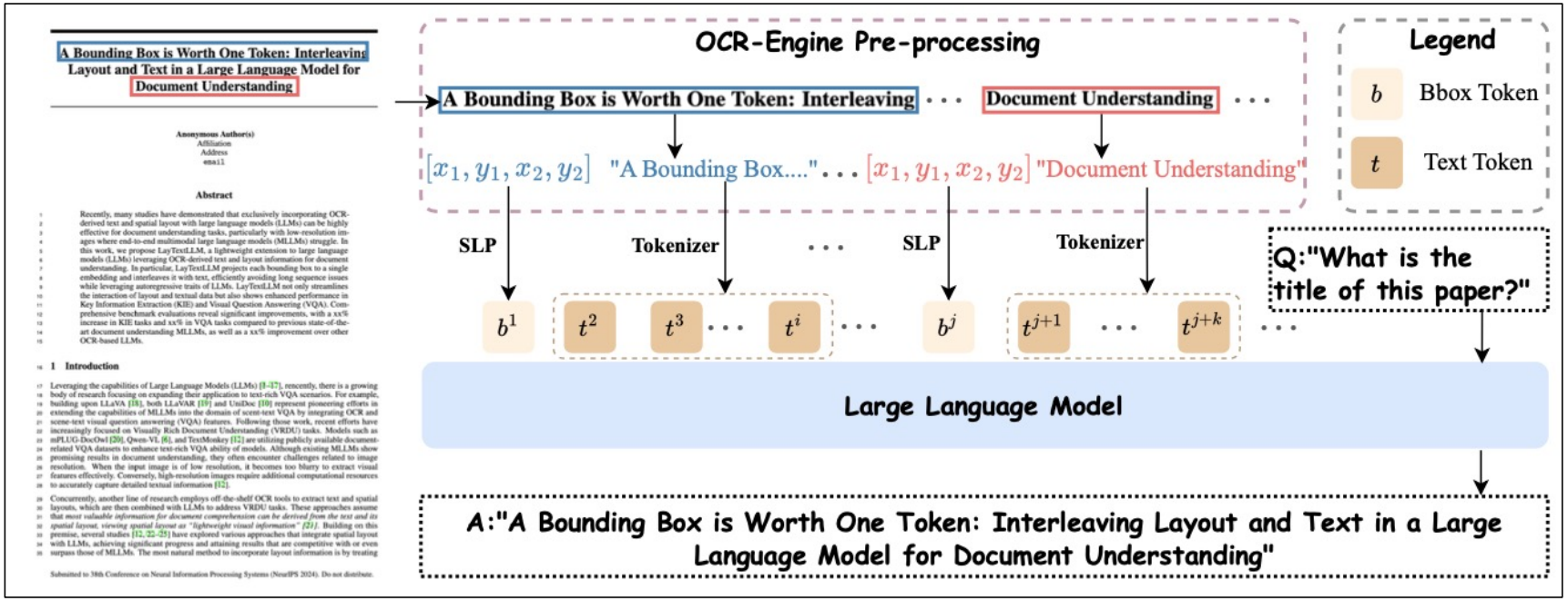
LayTextLLM: A Bounding Box is Worth One Token - Interleaving Layout and Text in a Large Language Model for Document Understanding

Jinghui Lu[*1] **Haiyang Yu**[*2] **Yanjie Wang**[*1] **Yongjie Ye**[1] **Jingqun Tang**[1]
Ziwei Yang[1] **Binghong Wu**[1] **Qi Liu**[1] **Hao Feng**[1] **Han Wang**[1] **Hao Liu**[1] **Can Huang**[†1]
[1]ByteDance Inc. [2]Fudan University
lujinghui@bytedance.com, hyyu20@fudan.edu.cn
{wangyanjie.prince, yeyongjie.ilz, tangjingqun}@bytedance.com
{yangziwei.1221, wubinghong, liuqi.nero}@bytedance.com
{fenghao.2019, wanghan.99, haoliu.0128, can.huang}@bytedance.com

Document AI with LLM

LayTextLLM: A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding

- Overview
  - Spatial Layout Projector → 각 bounding box를 single embedding에 담아냄
  - Training objectives – Layout-aware Next Token Prediction & Shuffled-OCR SFT



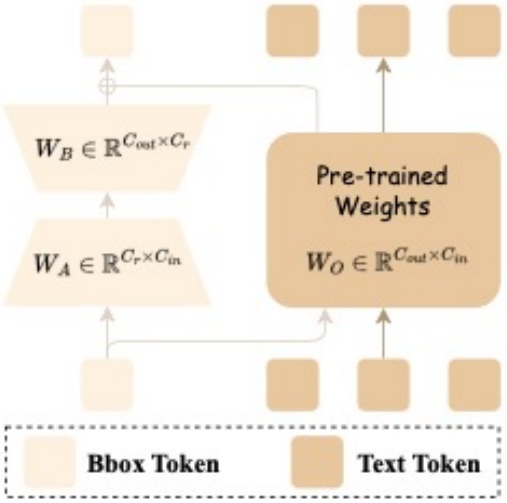
Document AI with LLM

LayTextLLM: A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding

- Model Architecture
  - Spatial Layout Projector (SLP)
    - OCR-추출된 자표들을 bounding box token들로 변환
  - P-LoRA (Partial Low-Rank Adaptation) on Bounding Box Tokens → *parameter overhead를 줄이는 동시에 LLM의 내제 지식 보존*

$$[x_1, y_1, x_2, y_2]$$

$$z = W \cdot c + b,$$



$$\hat{x}_t = W_0 x_t + B_0$$
$$\hat{x}_b = W_0 x_b + W_B W_A x_b + B_0$$
$$\hat{x} = [\hat{x}_b, \hat{x}_t]$$

Document AI with LLM

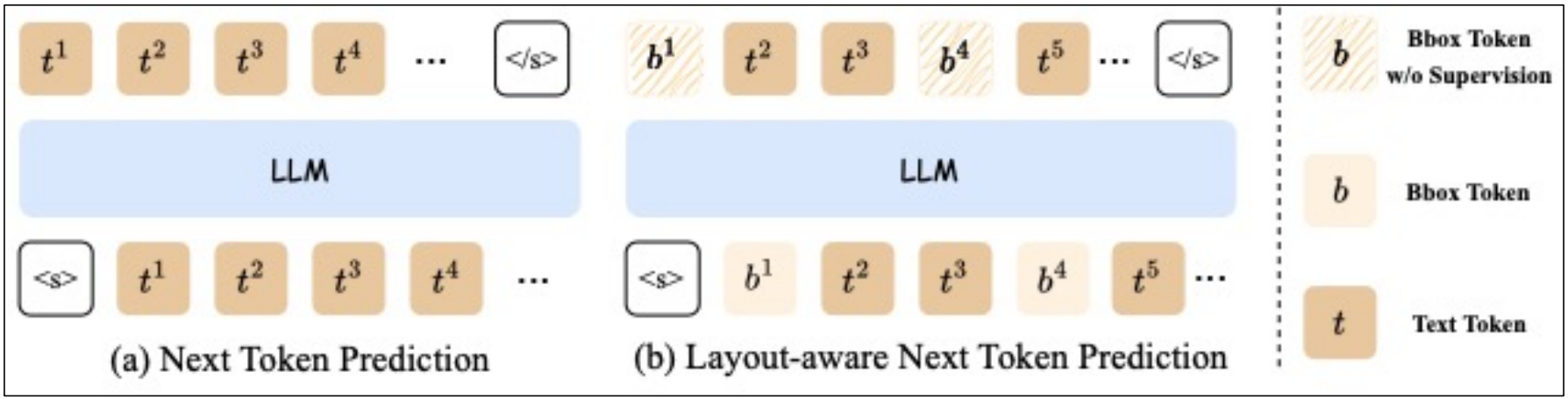
LayTextLLM: A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding

- Training Procedure - Pretraining

$$\mathcal{L} = -\frac{1}{T}\sum_{i=1}^{T} \log P\left(t^i \mid t^1, t^2, \ldots, t^{i-1}\right)$$

  - Layout-aware Next Token Prediction
    - Spatial layouts & textual modal 간의 **alignment** 강화
    - Text token들에 대한 loss만 계산

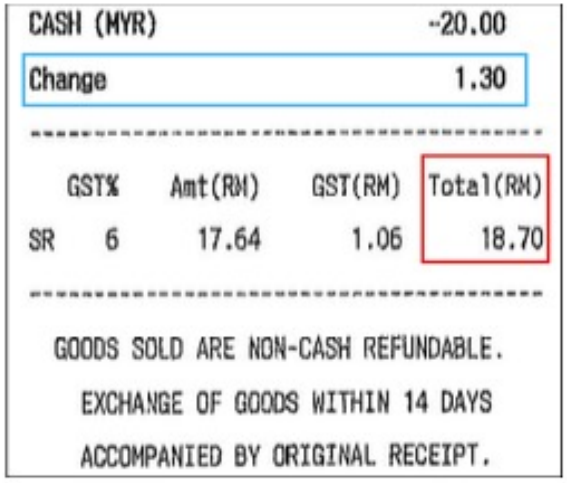
    ➔ Freeze the LLMs and only update parameters of SLP and P-LoRA



Document AI with LLM

LayTextLLM: A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding

- Training Procedure – SFT
  - Shuffled-OCR Supervised Fine-tuning
    - *Inductive bias & errors in text order* from the OCR engine
    - Generalizability & Robustness 향상
  - Unfreeze all parameters



→ *"... Change, 1.30, GST%, Amt(RM), GST(RM), Total(RM), SR, 6, 17.64, 1.06, 18.70 ..."*

Q: What is the value of the field **Change** ?

Q: What is the value of the filed **Total(RM)** ?

Document AI with LLM

LayTextLLM: A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding

- Experiments - Datasets
  - Pre-training data
    - IIT-CDIP Test Collecdtion 1.0
    - DocBank
  - SFT data
    - Document Dense Descriptioin(DDD) & Layout-aware SFT data
    - VQA - DocVQA, InfoVQA, ChartQA, VisualMRC
    - KIE - SROIE, CORD, FUNSD. POIE

Document AI with LLM

LayTextLLM: A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding

- Experiments - Implementation
  - Llama2-7B-base
    - $\text{LayTextLLM}_{\text{zero}}$, $\text{LayTextLLM}_{\text{vqa}}$, $\text{LayTextLLM}_{\text{all}}$
      - **$\text{LayTextLLM}_{\text{zero}}$:** Document Dense Descriptioin(DDD), Layout-aware SFT data
      - **$\text{LayTextLLM}_{\text{vqa}}$: +** DocVQA, InfoVQA
      - **$\text{LayTextLLM}_{\text{all}}$: +** KIE datasets

Document AI with LLM

LayTextLLM: A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding

- Experiments – Baseline Models
  - OCR-free Baselines
    - UniDoc, DocPedia, Monkey, InternVL, InternLMXComposer2, TextMonkey, TextMonkey$_{+}$
  - OCR-based Baselines – OCR-derived text as input
    - Llama2-7B-base, Llama2-7B-chat
    - Llama2-7B-base$_{coor}$, Llama2-7B-chat$_{coor}$
    - Davinici-003-175B$_{coor}$
    - DocLLM, LayoutLLM, LayoutLLM$_{CoT}$

Document AI with LLM

LayTextLLM: A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding

- Experiments - Metrics
  - OCR-free Baselines
    - Accuracy
  - OCR-based baselines – OCR-derived text as input
    - Accuracy
    - F1
    - ANLS (Average Noramalized Levenshtein Similarity)
    - CIDEr (Consensus-based Image Description Evaluation)

Document AI with LLM

LayTextLLM: A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding

- Experiments – Comparison w/SOTA OCR-free MLLMs
  - $\text{LayTextLLM}_{\text{zero}}$ significantly outperforms → zero-shot 능력
  - SRIOE, POIE
    - MLLM low-resolution images에 의한 낮은 성능 → LayTextLLM의 Robustness

	Document-Oriented VQA			KIE			
	DocVQA	InfoVQA	Avg	FUNSD	SROIE	POIE	Avg
Metric	*Accuracy %*						
OCR-free							
UniDoc [10]	7.7	14.7	11.2	1.0	2.9	5.1	3.0
DocPedia [9]	47.1*	15.2*	31.2	29.9	21.4	39.9	30.4
Monkey [55]	50.1*	25.8*	38.0	24.1	41.9	19.9	28.6
InternVL [56]	28.7*	23.6*	26.2	6.5	26.4	25.9	19.6
InternLM-XComposer2 [15]	39.7	28.6	34.2	15.3	34.2	49.3	32.9
TextMonkey [12]	64.3*	28.2*	46.3	32.3	47.0	27.9	35.7
TextMonkey_{+} [12]	66.7*	28.6*	47.7	42.9	46.2	32.0	40.4
text + polys							
LayTextLLM_{zero} (Ours)	72.1	35.7	53.9	**47.5**	**86.4**	68.9	**67.6**
LayTextLLM_{vqa} (Ours)	**77.2***	**42.1***	**59.7**	48.8	75.7	**70.6**	65.0

Table 1: Comparison with SOTA OCR-free MLLMs. * indicates the training set used.

Document AI with LLM

LayTextLLM: A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding

- Experiments – Comparison w/SoTA OCR-based methods
  - $\text{LayTextLLM}_{\text{zero}}$ 의 **zero-shot 환경**에도 불구하고 DocLLM에 못지 않는 성능
  - coordinate-as-tokens(coor)를 사용할 때 과도한 토큰 수 발생

	Document-Oriented VQA			KIE			
	DocVQA	VisualMRC	Avg	FUNSD	CORD	SROIE	Avg
Metric	*ANLS % / CIDEr*			*F-score %*			
Text							
Llama2-7B-base	34.0	182.7	108.3	25.6	51.9	43.4	40.3
Llama2-7B-chat	20.5	6.3	13.4	23.4	51.8	58.6	44.6
Text + Polys							
$\text{Llama2-7B-base}_{coor}$ [29]	8.4	3.8	6.1	6.0	46.4	34.7	29.0
$\text{Llama2-7B-chat}_{coor}$ [29]	12.3	28.0	20.1	14.4	38.1	50.6	34.3
$\text{Davinci-003-175B}_{coor}$ [29]	-	-	-	-	92.6	95.8	-
DocLLM [25]	69.5*	264.1*	166.8	51.8*	67.6*	91.9*	70.3
LayTextLLM_{zero} (Ours)	65.5	200.2	132.9	47.2	77.2	83.7	69.4
LayTextLLM_{vqa} (Ours)	75.6*	179.5	127.6	52.6	70.7	79.3	67.5
LayTextLLM_{all} (Ours)	**77.2***	**277.8***	**177.6**	**64.0***	**96.5***	**95.8***	**85.4**

Table 2: Comparison with other OCR-based methods. * indicates the training set used.

Document AI with LLM

LayTextLLM: A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding

- Experiments – Comparison w/LayoutLLM
  - KIE – LayTextLLM이 OCR-based 결과들에 효과적으로 reasoning
  - VQA – strongly related to vision info
    - Vision Encoder가 있는 LayoutLLM이 우세 + CoT

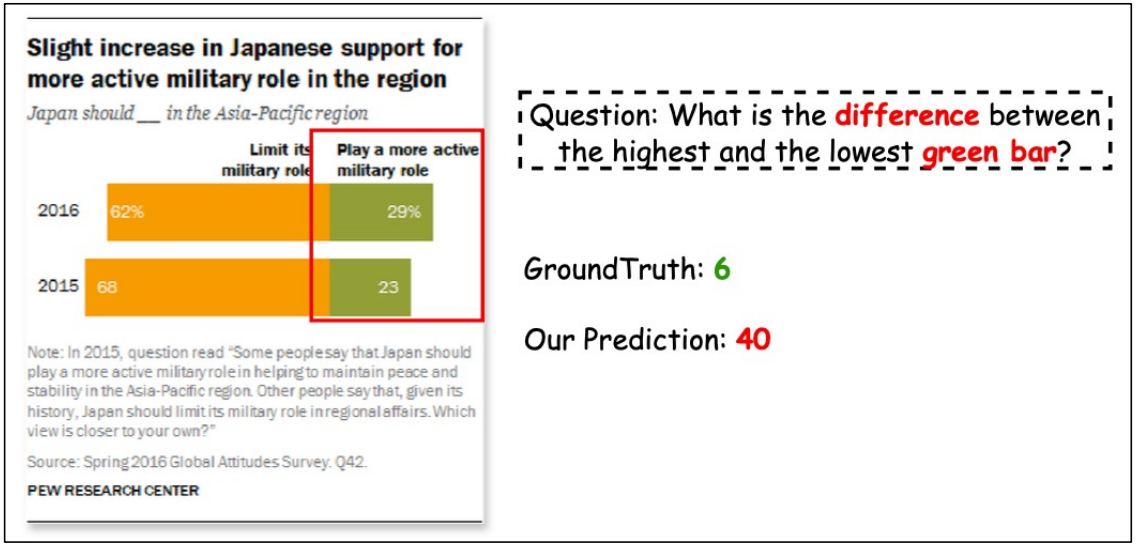
	Document-Oriented VQA			KIE			
	DocVQA	VisualMRC	Avg	FUNSD^{-}	CORD^{-}	SROIE^{-}	Avg
Metric	*ANLS %*						
Visual + Text + Polys							
LayoutLLM [27]	72.3	-	-	74.0	-	-	-
LayoutLLM$_{CoT}$ [27]	74.2	**55.7**	**64.9**	79.9	63.1	72.1	71.7
Text							
Llama2-7B-base	34.0	25.4	29.7	42.1	46.7	60.6	49.8
Llama2-7B-chat	20.5	9.9	15.2	15.1	20.0	35.6	23.5
Text + Polys							
Llama2-7B-base$_{coor}$ [29]	8.4	6.7	7.5	4.3	33.0	47.2	28.1
Llama2-7B-chat$_{coor}$ [29]	12.3	12.2	12.2	11.9	6.4	39.4	19.2
LayTextLLM$_{zero}$ (Ours)	65.5	37.4	51.5	72.0	45.5	82.0	66.5
LayTextLLM$_{all}$ (Ours)	**77.2***	41.7*	59.5	**81.0***	**82.5***	**96.1***	**86.5**

Table 3: Comparison with LayoutLLM. $^{-}$ indicates that the cleaned test set used in Luo et al. [27].

Document AI with LLM

LayTextLLM: A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding

- Conclusion
  - Limitations
    - ChartQA – visual cues



	ChartQA
OCR-free	
UniDoc [10]	10.9
DocPedia [9]	46.9*
Monkey [55]	54.0*
InternVL [56]	45.6*
InternLM-XComposer2 [15]	51.6*
TextMonkey [12]	58.2*
TextMonkey$_+$ [12]	**59.9***
text + polys	
LayTextLLM$_{zero}$ (Ours)	22.8
LayTextLLM$_{vqa}$ (Ours)	23.4*
LayTextLLM$_{all}$ (Ours)	35.4*

Document AI with LLM

LayTextLLM: A Bounding Box is Worth One Token: Interleaving Layout and Text in a Large Language Model for Document Understanding

- Conclusion
  - Document comprehension 능력을 굉장히 효율적으로 향상시킴
    - Architecture - Spatial Layout Projector & P-LoRA
    - Pretraining - Layout-aware Next Token Prediction
    - Fine-tuning - Shuffled OCR Fine-tuning

DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding

Dongsheng Wang*, **Natraj Raman***, **Mathieu Sibue***
Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, Xiaomo Liu
JPMorgan AI Research
first.last@jpmorgan.com

Document AI with LLM

DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding

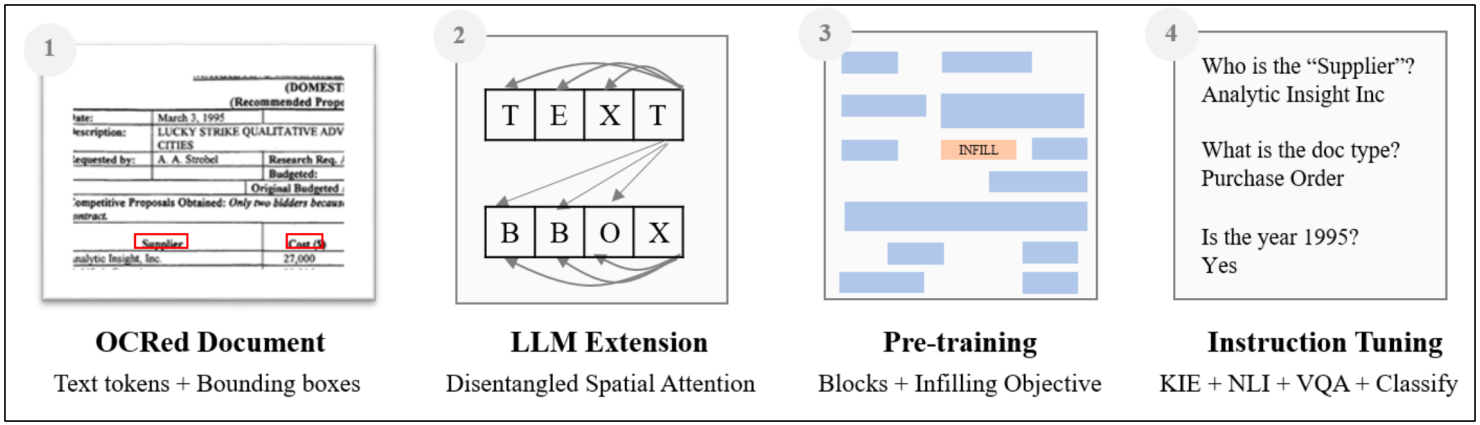
- Approaches
  - *Textual* and *Spatial* Modal들의 교점에 풍부한 semantic 요소
  - *Visual encoder들에 대한 비용 부담을 피하면서* Intrinsic Multi-modality 를 담아냄
  - It is sufficient to merely Include the spatial layout structure
    - Spatial layout info는 OCR에서 추출된 bounding box info (x, y coordinates)

Document AI with LLM

DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding

- Model Architecture
  - Disentangled Spatial Attention
    - Layout과 text 간의 상호관계를 disentangled manner로 계산

      e.g. LayoutLMv2는 단순히 position embedding layer 추가
  - Infilling objective(Fill in the Middle) into the visual document contexts
  - Instruction Tuning – KIE, NLI, VQA, classification



Document AI with LLM

DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding

- Model Architecture – Disentangled Spatial Attention
  - (1), (2) 는 흔히 아는 정통 attention mechanism
  - (3) 은 DocLLM의 **Disentangled Spatial Attention Mechanism**
    - Encode the **bounding boxes** into hidden vectors → $\mathbf{S} \in \mathbb{R}^{T \times d}$
    - *Text-to-Text, Text-to-Spatial, Spatial-to-Text, Spatial-to-Spatial*

      ➔ Incorporate **Layout Information** of the text tokens & Enables **Selective Focus**

$$\mathbf{Q}^t = \mathbf{H}\mathbf{W}^{t,q}, \quad \mathbf{K}^t = \mathbf{H}\mathbf{W}^{t,k}, \quad \mathbf{A}^t_{i,j} = \mathbf{Q}^t_i \mathbf{K}^{t\intercal}_j \tag{1}$$

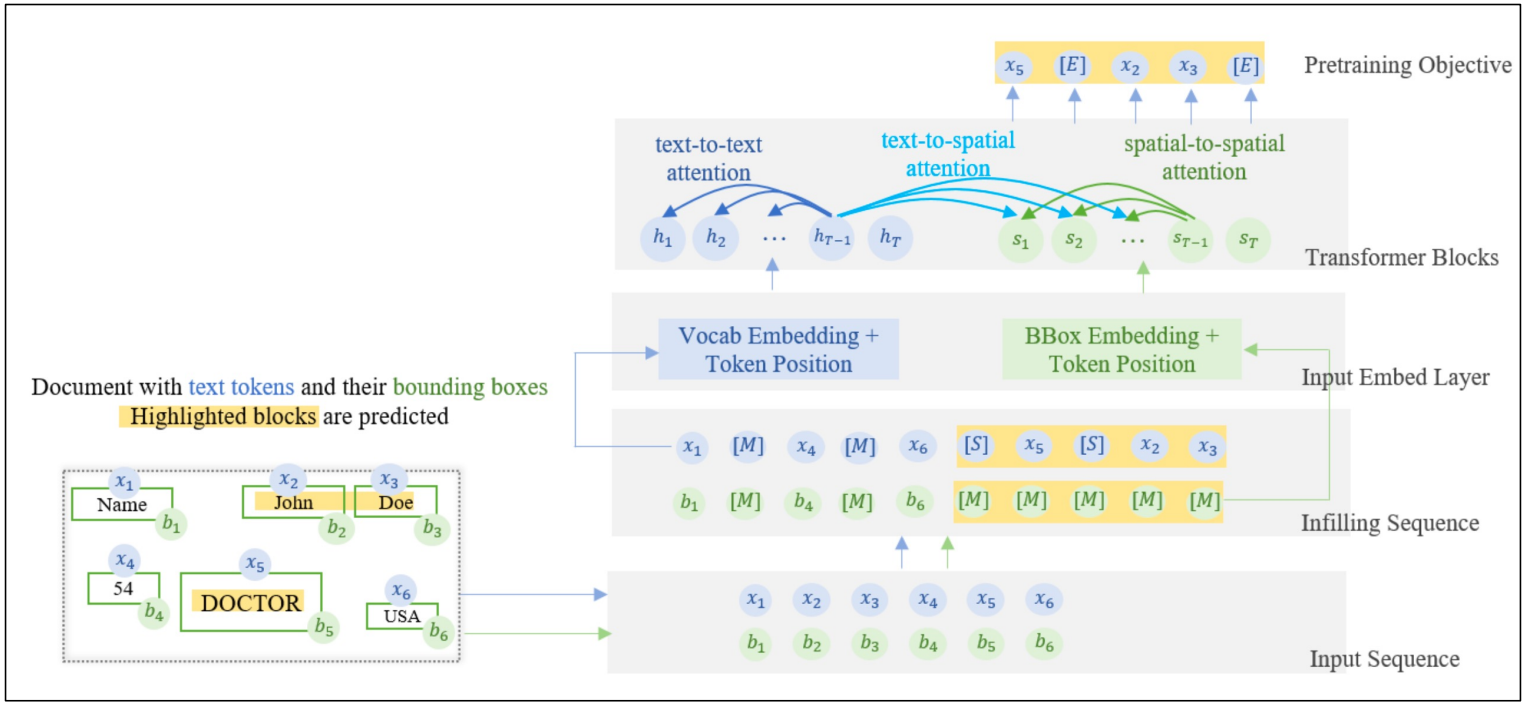
$$\mathbf{V}^t = \mathbf{H}\mathbf{W}^{t,v}, \quad \mathbf{H}' = softmax(\frac{\mathbf{A}^t}{\sqrt{d}})\mathbf{V}^t. \tag{2}$$

$$\begin{aligned} \mathbf{Q}^s = \mathbf{S}\mathbf{W}^{s,q}, \qquad \mathbf{K}^s = \mathbf{S}\mathbf{W}^{s,k}, \\ \mathbf{A}_{i,j} = \mathbf{Q}^t_i \mathbf{K}^{t\intercal}_j + \lambda_{t,s}\mathbf{Q}^t_i \mathbf{K}^{s\intercal}_j \\ + \lambda_{s,t}\mathbf{Q}^s_i \mathbf{K}^{t\intercal}_j + \lambda_{s,s}\mathbf{Q}^s_i \mathbf{K}^{s\intercal}_j, \end{aligned} \tag{3}$$

Document AI with LLM

DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding

- Pretraining
  - Infilling objective → Fill in the Middle



Document AI with LLM

DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding

- Instruction Tuning
  - Tasks – VQA(Visual Question Answering), NLI(Natural Language Inference), KIE(Key Information Extraction), CLS(Document Classification)

Task	Template type	Prompt template	Expected response
VQA	Extraction	`{document} {question}`	`answer` annotation
NLI	MCQ	`{document} "{statement}"`, Yes or No?	`answer` annotation
KIE	Extraction	`{document}` What is the value for the `"{key}"`?	Associated `value` annotation
	MCQ	`{document}` What is `"{value}"` in the document? Possible choices: `{keys}`. *(where* `keys` *is a subset of all the key names in the dataset in random order)*	Associated `key` annotation
	Internal classification	`{document}` What is `"{value}"` in the document?	Associated `key` annotation
CLS	MCQ	`{document}` What type of document is this? Possible choices: `{classes}`. *(where* `classes` *is a subset of all the classes in the dataset in random order)*	`class` annotation
	Internal classification	`{document}` What type of document is this?	`class` annotation

Document AI with LLM

DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding

- Datasets
  - Pre-training
    - IIT-CDIP Test Collection 1.0(~16M), DocBank(~500K)

    ➔ 16.7M pages & 3.8B tokens
  - Instruction Tuning
    - VQA – DocVQA, WTQ, VisualMRC, DUDE
    - NLI - TabFact
    - KIE – KLC, CORD, FUNSD, DeepForm, PWC, SROIE, VRDU
    - CLS – RVL-CDIP
    - + BuDDIE

Document AI with LLM

DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding

- Evaluation – Setup
  - Model Configuration
    - DocLLM-1B (Based on Falcon-1B), DocLLM-7B (Based on Llama2-7B)
  - Settings
    - SDDS (Same Datasets, Different Splits), STDD (Same Tasks, Different Datasets)

      STDD → Industry's use cases에서 태스크는 크게 변하지 않지만 document 속성들은 자주 변동
  - Baselines
    - SoTA LLMs w/Zero-shot prompts + OCR-extracted text(w/o spatial info)

      → GPT4 + OCR, Llama2 + OCR
    - DocAI LLMs

      → mPLUG-DocOwl, UReader

Document AI with LLM

DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding

- Evaluation - SDDS
  - 16개 중 12개 최고 성능
  - GPT4가 VQA 5개 중 3개 최고 성능 → Task 자체의 reasoning & abstraction 에 대한 복잡도
  - DocLLM 1B의 7B version 에 못지 않는 성능

	Dataset	GPT4+OCR – (T) ZS	Llama2+OCR 7B (T) ZS	mPLUG-DocOwl 7B (T+V) SDDS	UReader 7B (T+V) SDDS	DocLLM-1B 1B (T+L) SDDS	DocLLM-7B 7B (T+L) SDDS
VQA	DocVQA	**82.8**	47.4	62.2	65.4	61.4	69.5
	WTQ *(Accuracy)*	**65.4**	25.0	26.9	29.4	21.9	27.1
	VisualMRC *(CIDEr)*	255.1	115.5	188.8	221.7	245.0	**264.1**
	DUDE*	**54.6**	38.1	-	-	42.6	47.2
	BuDDIE	76.4	48.8	-	-	84.5	**86.7**
NLI	TabFact	**77.1**	48.2	60.2	67.6	58.0	66.4
KIE	KLC	45.9	27.8	30.3	32.8	58.9	**60.3**
	CORD	58.3	13.8	-	-	66.9	**67.4**
	FUNSD	37.0	17.8	-	-	48.2	**51.8**
	DeepForm	42.1	20.5	42.6	49.5	71.3	**75.7**
	PWC	18.3	6.8	-	-	25.7	**29.06**
	SROIE	90.6	56.4	-	-	91.0	**91.9**
	VRDU a.-b.*	43.7	18.7	-	-	87.6	**88.8**
	BuDDIE	66.1	10.8	-	-	95.4	**96.0**
CLS	RVL-CDIP	68.2	32.8	-	-	90.9	**91.8**
	BuDDIE	84.9	40.9	-	-	98.3	**99.4**

Document AI with LLM

DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding

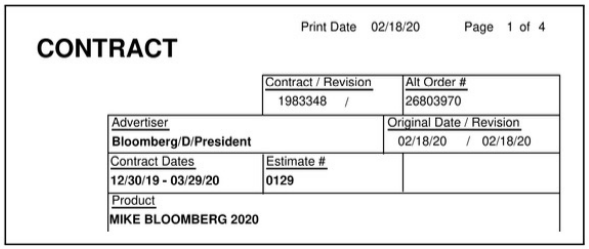
- Evaluation - STDD
  - Llama2를 dataset 5개 중 4개에서 능가
  - Classification Accuracy 상당히 낮음 → 오로지 하나의 CLS dataset 에 훈련

Model	Size	Setting	DocVQA	KLC	BuDDIE		
			VQA	KIE	VQA	KIE	CLS
GPT4+OCR	–	ZS	**82.8**	45.9	**76.4**	66.1	**84.9**
Llama2+OCR	7B	ZS	47.4	27.8	48.4	10.8	40.9
DocLLM-1B	1B	STDD	53.5	40.1	65.5	63.0	20.8
DocLLM-7B	7B	STDD	63.4	**49.9**	73.3	**72.6**	31.1

Document AI with LLM

DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding

- Experiments - Qualitative
  - (a) – Semantic nuances of enterprise documents
  - (b) – Spatial Reasoning ability
  - (c) – *limitation at a counting task → numerical concepts*



CONTRACT

Print Date 02/18/20 Page 1 of 4

Contract / Revision	Alt Order #
1983348 /	26803970

Advertiser	Original Date / Revision
Bloomberg/D/President	02/18/20 / 02/18/20

Contract Dates	Estimate #	
12/30/19 - 03/29/20	0129	

Product
MIKE BLOOMBERG 2020

(a) Prompt: What is the value for the “advertiser”?
DocLLM: Bloomberg/D/President
GPT4+OCR: MIKE BLOOMBERG 2020

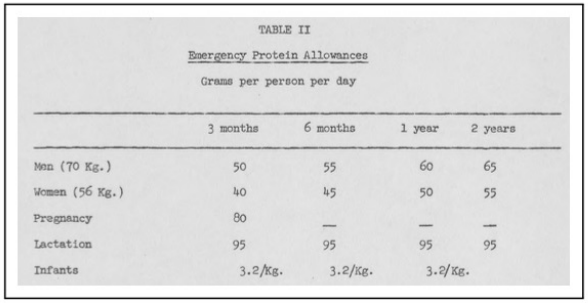


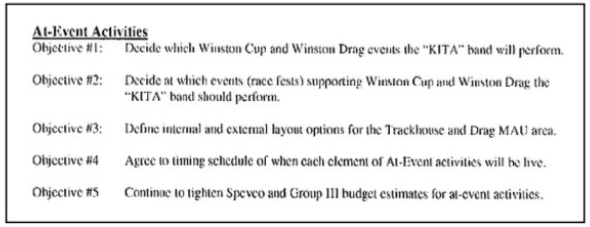
TABLE II

Emergency Protein Allowances

Grams per person per day

	3 months	6 months	1 year	2 years
Men (70 Kg.)	50	55	60	65
Women (56 Kg.)	40	45	50	55
Pregnancy	80	—	—	—
Lactation	95	95	95	95
Infants	3.2/Kg.	3.2/Kg.	3.2/Kg.	

(b) Prompt: What is written under the heading ‘emergency protein allowances’?
DocLLM: Grams per person per day
GPT4+OCR: Men (70 Kg.) 50 55 ...



At-Event Activities

Objective #1: Decide which Winston Cup and Winston Drag events the “KITA” band will perform.

Objective #2: Decide at which events (race fests) supporting Winston Cup and Winston Drag the “KITA” band should perform.

Objective #3: Define internal and external layout options for the Trackhouse and Drag MAU area.

Objective #4 Agree to timing schedule of when each element of At-Event activities will be live.

Objective #5 Continue to tighten Speveo and Group III budget estimates for at-event activities.

(c) Prompt: How many objectives are listed under at-event activities?
DocLLM: 4
GPT4+OCR: 5

Document AI with LLM

DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding

- Conclusion
  - Limitation
    - Context 길이
    - 복잡한 reasoning task 수행
  - Impact
    - 풍부한 정보의 layout이 담겨진 document에 generative reasoning

      → By Lightweight extension to traditional LLMs

Q&A

감사합니다.