Deep Learning 101, Taiwan’s pioneering and highest deep learning meetup, launched on 2016/11/11 @ 83F, Taipei 101

AI是一條孤獨且充滿惶恐及未知的旅程，花俏絢麗的收費課程或活動絕非通往成功的捷徑。
衷心感謝當時來自不同單位的AI同好參與者實名分享的寶貴經驗；如欲移除資訊還請告知。
由 TonTon Huang Ph.D. 發起，及其當時任職公司(台灣雪豹科技)無償贊助場地及茶水點心。

“Deep Learning 101”
台灣最高的深度學習 / 人工智慧 (AI) Meetup .
AI
http://DeepLearning101.TWMAN.ORG

去 YouTube 訂閱 | Facebook | 回 GitHub Pages | 到 GitHub 點星 | 網站 | 到 Hugging Face Space 按愛心

大語言模型	語音處理	自然語言處理	電腦視覺
Large Language Model	Speech Processing	Natural Language Processing, NLP	Computer Vision

▶ 手把手帶你一起踩 AI 坑

手把手帶你一起踩 AI 坑：https://www.twman.org/AI

- **避開 AI Agent 開發陷阱：常見問題、挑戰與解決方案**：探討多種 AI 代理人工具的應用經驗與挑戰，分享實用經驗與工具推薦。
- **白話文手把手帶你科普 GenAI**：淺顯介紹生成式人工智慧核心概念，強調硬體資源和數據的重要性。
- **大型語言模型直接就打完收工？**：回顧 LLM 領域探索歷程，討論硬體升級對 AI 開發的重要性。
- **檢索增強生成(RAG)不是萬靈丹之優化挑戰技巧**：探討 RAG 技術應用與挑戰，提供實用經驗分享和工具建議。
- **大型語言模型 (LLM) 入門完整指南：原理、應用與未來**：探討多種 LLM 工具的應用與挑戰，強調硬體資源的重要性。
- **什麼是大語言模型，它是什麼？想要嗎？(Large Language Model，LLM)**：探討 LLM 的發展與應用，強調硬體資源在開發中的關鍵作用。
- **Diffusion Model 完全解析：從原理、應用到實作 (AI 圖像生成)**；深入探討影像生成與分割技術的應用，強調硬體資源的重要性。
- **ASR/TTS 開發避坑指南：語音辨識與合成的常見挑戰與對策**：探討 ASR 和 TTS 技術應用中的問題，強調數據質量的重要性。
- **那些 NLP 踩的坑**：分享 NLP 領域的實踐經驗，強調數據質量對模型效果的影響。
- **那些語音處理踩的坑**：分享語音處理領域的實務經驗，強調資料品質對模型效果的影響。
- **手把手學深度學習安裝環境**：詳細介紹在 Ubuntu 上安裝深度學習環境的步驟，分享實際操作經驗。

文章目錄

- Diffusion model (擴散模型)
- Digital Human (虛擬數字人)

CV

Computer Vision (電腦視覺)

AnomalyDetection

Anomaly Detection，異常檢測

-2025-09-24：FS-SAM2: Adapting Segment Anything Model 2 for Few-Shot Semantic Segmentation via Low-Rank Adaptation；FS-SAM2靠SAM2+LoRA 實現效能與效率雙優

- 2025-09-20：MOCHA: Multi-modal Objects-aware Cross-arcHitecture Alignment；將大模型多模態語意注入YOLO，少樣本檢測性能大漲
- 2025-07-16：CostFilter-AD：Enhancing Anomaly Detection through Matching Cost Filtering；刷新無監督異常檢測上限！ CostFilter-AD：首個即插即用的代價濾波for異常檢測範式
- 2025-06-13：One-to-Normal：Anomaly Personalization；少樣本異常辨識新突破，擴散模型協助精準偵測
- 2025-06-06：CVPR2025, *DualAnoDiff*：Dual-Interrelated Diffusion Model for Few-Shot Anomaly Image Generation；以大模型檢測工業品異常，復旦騰訊優圖新演算法入選CVPR 2025
- 2025-05-15：**AdaptCLIP**: Adapting CLIP for Universal Visual Anomaly Detection；Github；騰訊開源 AdaptCLIP 模型刷新多領域SOTA
- 2025-05-05：Detect, Classify, Act: Categorizing Industrial Anomalies with Multi-Modal Large Language Models；DeepWiki；數據集
- 2025-04-27：**AnomalyCLIP**: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection；DeepWiki
- 2025-04-26：PaDim；DeepWiki
- 2025-04-12：Anomaly-Aware CLIP, **AA-CLIP**: Enhancing Zero-shot Anomaly Detection via Anomaly-Aware CLIP；DeepWiki
- 2025-03-25：**Dinomaly**：The Less Is More Philosophy in Multi-Class Unsupervised Anomaly Detection；無監督異常檢測（Unsupervised Anomaly Detection，UAD）

ObjectDetection

Object Detection (目標偵測)

- AAAI2025, Multi-clue Consistency Learning to Bridge Gaps Between General and Oriented Object in Semi-supervised Detection；Github；AAAI2025 一個遙感半監督目標偵測（半監督旋轉目標偵測）方法
- 2025-07-24：OV-DINO；開源工業開放詞彙目標偵測
- 2025-06-18：CountVid: Open-World Object Counting in Videos；牛津大學開源類別無關的影片目標計數，影片中也能「指哪數哪」
- 2025-06-15：GeoPix；像素級遙感多模態大模型
- 2025-05-23：VisionReasoner；偵測、分割、計數、問答全拿下？對標Qwen2.5-VL！ VisionReasoner用強化學習統一視覺感知與推理
- 2025-03-14：Falcon: A Remote Sensing Vision-Language Foundation Model；DeepWiki

Segmentation

Segmentation (圖像分割)

- Perceive Anything Model：Recognize, Explain, Caption, and Segment Anything in Images and Videos；對標SAM2 + LLM融合版！港中文開源感知一切模型與百萬級影像描述資料集：辨識、解釋、描述、分割一體化輸出
- RemoteSAM：Towards Segment Anything for Earth Observation
- InstructSAM：A Training-Free Framework for Instruction-Oriented Remote Sensing Object Recognition；DeepWiki
- RESAnything: Attribute Prompting for Arbitrary Referring Segmentation；Project
- CVPR 2025, Segment Any Motion in Videos, Segment Any Motion in Videos；Github
- CVPR 2025 Highlight, Exact: Exploring Space-Time Perceptive Clues for Weakly Supervised Satellite Image Time Series Semantic Segmentation；Github；Exact：基於遙感影像時間序列弱監督學習的作物提取方法
- MatAnyone：視訊摳圖MatAnyone來了，一次指定全程追踪，髮絲級還原
- Meta Segment Anything Model 2 (SAM 2)
    - 60行程式碼訓練/微調Segment Anything 2
    - CLIPSeg：Image Segmentation Using Text and Image Prompts：Huggingface Space
        - 哥廷根大學提出CLIPSeg，能同時作三個分割任務的模型
        - SAM與CLIP強強聯手，實現22000類的分割與識別
- SAMURAI
    - 無需訓練或微調即可得到穩定、準確的追蹤效果！ KF + SAM2 解決快速移動或自遮擋的物件追蹤問題
    - 經典卡爾曼濾波器改進影片版「分割一切」，網友：好優雅的方法
- Grounded SAM 2: Ground and Track Anything in Videos
    - Grounded-Segment-Anything
- SAM2Long：大幅提升SAM 2性能！港中文提出SAM2Long，複雜長視頻的分割模型
- SAM2-Adapter：SAM 2無法分割一切？ SAM2-Adapter：首次讓SAM 2在下游任務適應調校！
- SAM2Point：可提示3D 分割研究里程碑！ SAM2Point：SAM2加持可泛化任3D場景、任意提示！
- Optical Character Recognition，光學文字識別

OCR

Optical Character Recognition (光學文字識別)
針對物件或場景影像進行分析與偵測

- 2025-08-18：DianJin-OCR-R1；點金OCR-R1，模糊蓋章、跨頁表格、文字幻覺全拿下！
- 2025-07-30：dots.ocr；本地部署1.7B参数超强OCR大模型dots.ocr
- 2025-06-16：OCRFlux；DEMO；OCRFlux：一個基於LLM的複雜佈局與跨頁合併的PDF文檔解析
- 2025-06-05：MonkeyOCR；Document Parsing with a Structure-Recognition-Relation Triplet Paradigm
- 2025-05-21：PaddleOCR 3.0；OCR精準度躍升13%，支援多語種、手寫體與高精準度文件解析
- 2025-03-05：PP-DocBee：百度推出文件影像理解PP-DocBee
- 2025-03-03：olmocr：🚀本地部署最强OCR大模型olmOCR！支持结构化精准提取复杂PDF文件内容！
- 2025-02-05：MinerU：將PDF轉換為機器可讀格式的神器
- 2024-12-15：markitdown
- 2024-09-22：OCR2.0时代-GOT来啦！
- 2024-09-11：GOT-OCR-2.0模型开源
- 2024-08-20：萬物皆可AI化！剛開源就有12000人圍觀的OCR 掃描PDF 開源工具！還可轉換為MarkDown！
- advancedliteratemachinery/OCR/OmniParser
- 2024-10-29：Alibaba出品:OmniParser通用文檔複雜場景下OCR抽取
- RapidOCR
- 12個流行的開源免費OCR項目
- 用PaddleOCR的PPOCRLabel來微調醫療診斷書和收據
- TableStructureRec: 表格結構辨識推理庫來了：https://github.com/RapidAI/TableStructureRec

Diffusion model (擴散模型)

- 2025-05-28：視覺理解&生成大一統模型 Jodi；alphaXiv
- 2025-05-27：AnomalyAny；CVPR2025 | 突破資料瓶頸！ Stable Diffusion 協助視覺異常檢測，無需訓練即可產生真實多樣異常樣本
- 2025-05-23：HivisionIDPhotos，智慧證件照產生神器；AI證件照，摳圖、換背景、任意尺寸
- 2025-05-19：Index-AniSora；Aligning Anime Video Generation with Human Feedback；B站開源SOTA動畫影片生成模型Index-AniSora！
- 2025-04-26：Insert Anything；DeepWiki
- 2025-04-24：字節Phantom：1280x720影片生成革命！位元組Phantom模型實測：10G顯存效果不輸某靈付費版
- 2025-04-22：MAGI-1：Sand AI 創業團隊推出了全球首個自回歸影片生成大模型MAGI-1，該模型有哪些效能亮點？
- 2025-04-22：SkyReels V2：全球首個無限時長影片生成！新擴散模式引爆兆市場，電影級理解，全面開源
- 2025-04-14：FramePack：不是可靈用不起，而是FramePack更有性價比！開源專案：6G顯存跑13B模型，支援1分鐘影片產生
- 2025-04-14：fantasy-talking：解讀最新基於Wan2.1的音訊驅動數位人FantasyTalking
- 2025-04-05：SkyReels-A2；DeepWiki；SkyReels-A2：用AI重新定义视频创作的未来！
- 2025-03-10：HunyuanVideo-I2V：騰訊開源HunyuanVideo-I2V圖生視訊模型+LoRA訓練腳本，社群部署、推理實戰教學來吧
- 2025-02-25：Wan-Video：超越Sora！阿里萬相大模型正式開源！全模態、全尺寸大模型開源
- 2025-02-14：FlashVideo：來自位元組的視訊增強全新開源演算法，102秒產生1080P視頻
- 2025-01-28：Sana：[ICLR 2025 Oral] Efficient High-Resolution Image Synthesis with Linear Diffusion Transformer；比FLUX快100倍！英偉達聯手MIT、清華開源超快AI影像產生模型
- Flux
    - Flux.1-canny-dev：https://huggingface.co/black-forest-labs/FLUX.1-Canny-dev/

    - Flux.1-depth-dev：https://huggingface.co/black-forest-labs/FLUX.1-Depth-dev/
    - Flux.1-fill-dev：https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev/
    - Flux.1-redux-dev：https://huggingface.co/black-forest-labs/FLUX.1-Redux-dev/
        - 2024-11-26：Flux官方重繪+擴圖+風格參考+ControlNet
        - 2024-11-25：最新flux_fill_inpaint模型體驗。
- 2024-12-17：Leffa：Leffa：Meta AI 開源精確控制人物外觀和姿勢的圖像生成框架，在生成穿著的同時保持人物特徵
- 2024-11-29：PuLID, Pure and Lightning ID Customization via Contrastive Alignment：https://github.com/balazik/ComfyUI-PuLID-Flux
    - 2024-11-07：搞定ComfyUI-PuLID-Flux節點只要這幾步！附一鍵壓縮包
    - 2024-10-08：一文搞懂PuLID FLUX人物換臉&風格遷移
- 2024-11-26：MagicQuill：https://huggingface.co/spaces/AI4Editing/MagicQuill
    - MagicQuill，登上Huggingface趨勢榜榜首的AI P圖神器
- 2024-11-26：OOTDiffusion：https://huggingface.co/spaces/levihsu/OOTDiffusion
    - 開源AI換裝神器OOTDiffusion
- 2024-11-24：Comfyui Impact Pack
    - Comfyui 最強臉部修復工具Impact Pack
- 2024-11-05：ComfyUI OmniGen @ 北京人工智慧研究院：https://huggingface.co/spaces/Shitao/OmniGen
    - ComfyUI 影像生成模型OmniGen，人物一致性處理的也太好了
    - 全能影像生成模型OmniGen：告別ControlNet、ipadapter等插件，僅憑提示即可控制影像生成與編輯

Digital Human (虛擬數字人)

- HeyGem：開源數位人克隆神器
- Duix：全球首個真人數位人，開源了
- Linly-Talker：an intelligent AI system that combines large language models (LLMs) with visual models to create a novel human-AI interaction method.
- EchoMimicV2：[CVPR 2025] EchoMimicV2: Towards Striking, Simplified, and Semi-Body Human Animation
- Hallo3：[CVPR 2025] Highly Dynamic and Realistic Portrait Image Animation with Diffusion Transformer Networks
- MimicTalk：[NeurIPS 2024] MimicTalk: Mimicking a personalized and expressive 3D talking face in minutes
- JoyGen：Audio-Driven 3D Depth-Aware Talking-Face Video Editing
- Latentsync
- MuseTalk

Image Recognition (圖像識別)

- ViT（Vision Transformer）解析：https://github.com/google-research/vision_transformer
- 2040張圖片訓練出的ViT，準確率96.7%，連遷移表現都令人驚訝
- Swin Transformer: 用CNN的方式打敗CNN：https://github.com/microsoft/Swin-Transformer
- EfficientNetV2震撼發布！更小的模型，更快的訓練：https://github.com/d-li14/efficientnetv2.pytorch

Document Understanding (文件理解)

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, Seunghyun Park, "OCR-free Document Understanding Transformer", arXiv preprint, arXiv:2111.15664, 2022.

Document Layout Analysis (文件結構分析)

Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, Weining Li, "A unified toolkit for Deep Learning Based Document Image Analysis", arXiv preprint, arXiv:2103.15348, 2021.

▼ **LayoutLM series**

- **arXiv-2020**# LayoutLM Paper: LayoutLM: Pre-training of Text and Layout for Document Image Understanding

Authors: Yiheng Xu,Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou

Public: arXiv:1912.13318, 2020

github

論文筆記

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 between text and layout information across scanned document images, which is beneficial for a great number of real-world document image understanding tasks such as information extraction from scanned documents. Furthermore, we also leverage image features to incorporate words' visual information into LayoutLM. To the best of our knowledge, this is the first time that text and layout are jointly learned in a single framework for document-level pre-training. It achieves new state-of-the-art results in several downstream tasks, including form understanding (from 70.72 to 79.27), receipt understanding (from 94.02 to 95.24) and document image classification (from 93.07 to 94.42).

在近年來，預訓練技術在多種自然語言處理（NLP）任務中取得了成功。儘管預訓練模型在NLP應用中被廣泛使用，但它們幾乎僅專注於文本層級的操作，忽略了版面和風格信息，而這些對於文檔圖像理解至關重要。

因此，提出了一種名為LayoutLM的模型，它能夠在掃描文檔圖像中共同建模文本和版面信息之間的交互作用，這對於許多現實世界的文檔圖像理解任務（如從掃描文檔中提取信息）非常有益。此外，該論文還利用圖像特徵將文字的視覺信息融入到LayoutLM中。

值得注意的是，這是首次在單一框架中聯合學習文本和版面的文檔層級預訓練。該模型在幾個下游任務上取得了新的最先進結果，包括表格理解（從70.72%提高到79.27%）、收據理解（從94.02%提高到95.24%）和文檔圖像分類（從93.07%提高到94.42%）等。

總的來說，LayoutLM的提出填補了NLP預訓練模型忽略版面和風格信息的空白，並在文檔圖像理解領域取得了重要的突破和成果。

Introduction

1. 理解商業文件是一項非常具有挑戰性的任務。有些公司通過耗時且昂貴的人工登打工作從商業文件中提取數據，同時需要手動定制或配置。

2. 目前的文件分析方法通常基於深度神經網絡，早期的嘗試通常專注於檢測和分析文件的某些部分，如表格區域。[7] 是首先提出基於卷積神經網絡（CNN）的PDF文件表格檢測方法。此後，[21、24、29] 也利用更先進的Faster R-CNN模型[19]或Mask R-CNN模型[9]來進一步提高文件版面分析的準確性。此外，[28] 提出了一種端到端、多模態、完全卷積網絡，從文件圖像中提取語義結構，充分利用來自預訓練自然語言處理模型的文本嵌入。最近，[15] 提出了一種基於圖卷積網絡（GCN）的模型，以結合文本和視覺信息，用於文件人工智慧領域。

3. 但這樣的方法，有兩個限制： (1) They rely on a few human-labeled training samples without fully exploring the possibility of using large-scale unlabeled training samples. (2) They usually leverage either pre-trained CV models or NLP models, but do not consider a joint training of textual and layout information. Therefore, it is important to investigate how self-supervised pre-training of text and layout may help in the document AI area

- 提出的方法 LayoutLM，一種簡單而有效的文本和佈局預訓練方法，用於文檔圖像理解任務。受 BERT 模型[4]的啟發，輸入文本信息主要由文本嵌入和位置嵌入表示，LayoutLM 進一步添加了兩種類型的輸入嵌入：（1）二維位置嵌入，表示文檔中標記的相對位置； (2) 將掃描的令牌圖像嵌入到文檔中。LayoutLM的架構如圖2所示。

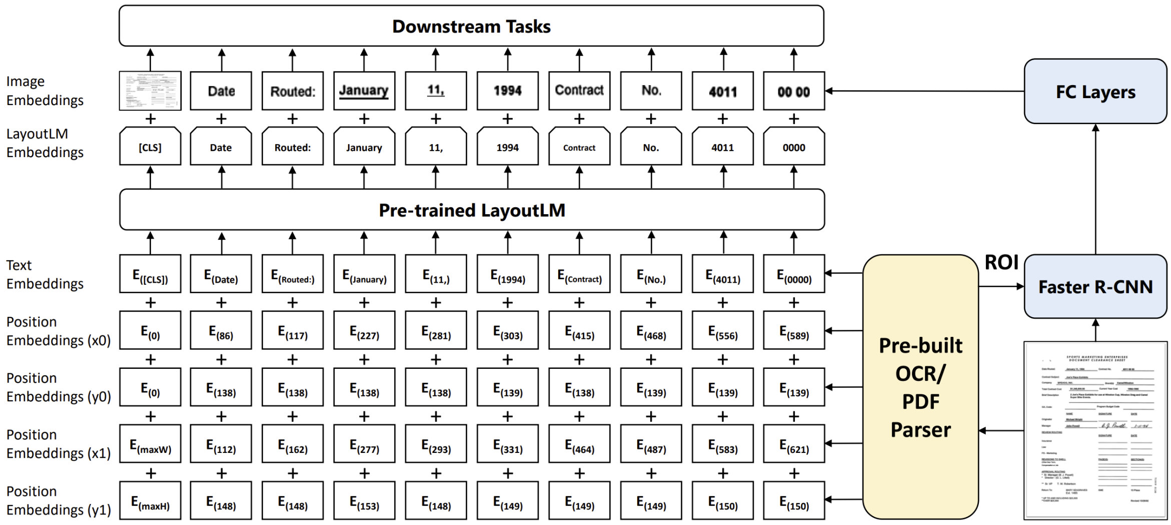

Downstream Tasks
Image Embeddings
Date
Routed:
January
11,
1994
Contract
No.
4011
00 00
FC Layers
LayoutLM Embeddings
[CLS]
Date
Routed:
January
11,
1994
Contract
No.
4011
0000
Pre-trained LayoutLM
ROI
Faster R-CNN
Text Embeddings
E([CLS]) E(Date) E(Routed:) E(January) E(11,) E(1994) E(Contract) E(No.) E(4011) E(0000)
Position Embeddings (x0)
E(0) E(86) E(117) E(227) E(281) E(303) E(415) E(468) E(556) E(589)
Position Embeddings (y0)
E(0) E(138) E(138) E(138) E(138) E(139) E(138) E(139) E(139) E(139)
Position Embeddings (x1)
E(maxW) E(112) E(162) E(277) E(293) E(331) E(464) E(487) E(583) E(621)
Position Embeddings (y1)
E(maxH) E(148) E(148) E(153) E(148) E(149) E(149) E(149) E(150) E(150)
Pre-built OCR/ PDF Parser

Figure 2: An example of LayoutLM, where 2-D layout and image embeddings are integrated into the original BERT architecture. The LayoutLM embeddings and image embeddings from Faster R-CNN work together for downstream tasks.

EXPERIMENTS

Pre-training Dataset

- 預訓練模型的性能很大程度上取決於數據集的規模和質量。
- 在 IIT-CDIP Test Collection 1.0 上進行了預訓練，其中包含超過 600 萬份文檔，其中包含超過 1100 萬張掃描文檔圖像。此外，每個文檔都有其相應的文本和元數據存儲在 XML 文件中。文本是對文檔圖像應用OCR產生的內容。元數據描述文檔的屬性，例如唯一標識和文檔標籤。儘管元數據包含錯誤和不一致的標籤，但這個大型數據集中的掃描文檔圖像非常適合預訓練我們的模型

Fine-tuning Dataset

FUNSD 數據集。我們在 FUNSD 數據集上評估我們的方法，以理解嘈雜的掃描文檔中的形式。該數據集包括 199 個真實的、完全註釋的掃描表單，其中包含 9,707 個語義實體和 31,485 個單詞。這些形式被組織為相互鏈接的語義實體列表。每個語義實體包括唯一標識符、標籤（即問題、答案、標題或其他）、邊界框、與其他實體的鏈接列表以及單詞列表。數據集分為 149 個訓練樣本和 50 個測試樣本。我們採用詞級F1分數作為評價指標。 We adopt the ***word-level F1 score*** as the evaluation metric.

Task-Specific

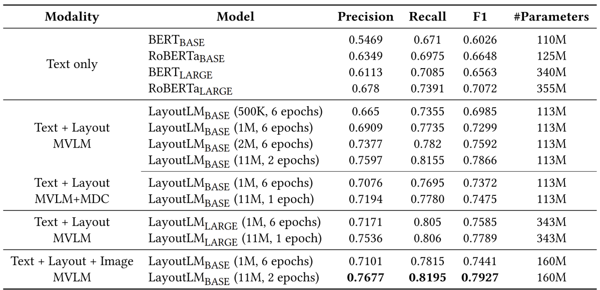
Receipt Understanding. This task requires filling several predefined semantic slots according to the scanned receipt images. For instance, given a set of receipts, we need to fill specific slots (i.g., company, address, date, and total). Different from the form understanding task that requires labeling all matched entities and keyvalue pairs, the number of semantic slots is fixed with pre-defined keys. Therefore, the model only needs to predict the corresponding values using the sequence labeling method.

資料標註時,需要標註 公司, 地址, 日期, 金額

Result

Form Understanding. We evaluate the form understanding task on the FUNSD dataset. The experiment results are shown in Table 1. We compare the LayoutLM model with two SOTA pre-trained NLP models: BERT and RoBERTa [16]. The BERT BASE model achieves 0.603 and while the LARGE model achieves 0.656 in F1. Compared to BERT, the RoBERTa performs much better on this dataset as it is trained using larger data with more epochs. Due to the time limitation, we present 4 settings for LayoutLM, which are 500K document pages with 6 epochs, 1M with 6 epochs, 2M with 6 epochs as well as 11M with 2 epochs. It is observed that the LayoutLM model substantially outperforms existing SOTA pre-training baselines. With the BASE architecture, the LayoutLM model with 11M training data achieves 0.7866 in F1, which is much higher than BERT and RoBERTa with the similar size of parameters. In addition, we also add the MDC loss in the pre-training step and it does bring substantial improvements on the FUNSD dataset. Finally, the LayoutLM model achieves the best performance of 0.7927 when using the text, layout, and image information at the same time. In addition, we also evaluate the LayoutLM model with different data and epochs on the FUNSD dataset, which is shown in Table 2. For different data settings, we can see that the overall accuracy is monotonically increased as more epochs are trained during the pre-training step. Furthermore, the accuracy is also improved as more data is fed into the LayoutLM model. As the FUNSD dataset contains only 149 images for fine-tuning, the results confirm that the pre-training of text and layout is effective for scanned document understanding especially with low resource settings.

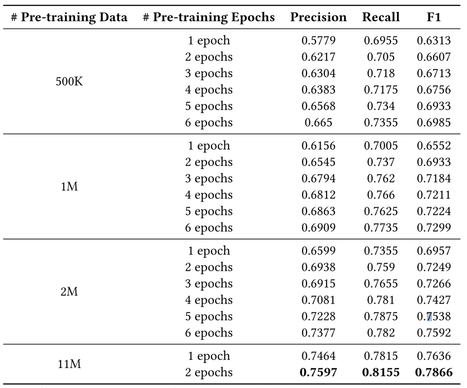
1



Modality	Model	Precision	Recall	F1	#Parameters
Text only	$\text{BERT}_{\text{BASE}}$	0.5469	0.671	0.6026	110M
	$\text{RoBERTa}_{\text{BASE}}$	0.6349	0.6975	0.6648	125M
	$\text{BERT}_{\text{LARGE}}$	0.6113	0.7085	0.6563	340M
	$\text{RoBERTa}_{\text{LARGE}}$	0.678	0.7391	0.7072	355M
Text + Layout MVLM	$\text{LayoutLM}_{\text{BASE}}$ (500K, 6 epochs)	0.665	0.7355	0.6985	113M
	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.6909	0.7735	0.7299	113M
	$\text{LayoutLM}_{\text{BASE}}$ (2M, 6 epochs)	0.7377	0.782	0.7592	113M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	0.7597	0.8155	0.7866	113M
Text + Layout MVLM+MDC	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.7076	0.7695	0.7372	113M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 1 epoch)	0.7194	0.7780	0.7475	113M
Text + Layout MVLM	$\text{LayoutLM}_{\text{LARGE}}$ (1M, 6 epochs)	0.7171	0.805	0.7585	343M
	$\text{LayoutLM}_{\text{LARGE}}$ (11M, 1 epoch)	0.7536	0.806	0.7789	343M
Text + Layout + Image MVLM	$\text{LayoutLM}_{\text{BASE}}$ (1M, 6 epochs)	0.7101	0.7815	0.7441	160M
	$\text{LayoutLM}_{\text{BASE}}$ (11M, 2 epochs)	**0.7677**	**0.8195**	**0.7927**	160M

Table 1: Model accuracy (Precision, Recall, F1) on the FUNSD dataset

2



# Pre-training Data	# Pre-training Epochs	Precision	Recall	F1
500K	1 epoch	0.5779	0.6955	0.6313
	2 epochs	0.6217	0.705	0.6607
	3 epochs	0.6304	0.718	0.6713
	4 epochs	0.6383	0.7175	0.6756
	5 epochs	0.6568	0.734	0.6933
	6 epochs	0.665	0.7355	0.6985
1M	1 epoch	0.6156	0.7005	0.6552
	2 epochs	0.6545	0.737	0.6933
	3 epochs	0.6794	0.762	0.7184
	4 epochs	0.6812	0.766	0.7211
	5 epochs	0.6863	0.7625	0.7224
	6 epochs	0.6909	0.7735	0.7299
2M	1 epoch	0.6599	0.7355	0.6957
	2 epochs	0.6938	0.759	0.7249
	3 epochs	0.6915	0.7655	0.7266
	4 epochs	0.7081	0.781	0.7427
	5 epochs	0.7228	0.7875	0.7538
	6 epochs	0.7377	0.782	0.7592
11M	1 epoch	0.7464	0.7815	0.7636
	2 epochs	**0.7597**	**0.8155**	**0.7866**

Table 2: $\text{LayoutLM}_{\text{BASE}}$ **(Text + Layout, MVLM) accuracy with different data and epochs on the FUNSD dataset**

CONCLUSION

We present LayoutLM, a simple yet effective pre-training technique with text and layout information in a single framework. Based on the Transformer architecture as the backbone, LayoutLM takes advantage of multimodal inputs, including token embeddings, layout embeddings, and image embeddings. Meanwhile, the model can be easily trained in a self-supervised way based on large scale unlabeled scanned document images. We evaluate the LayoutLM model on three tasks: form understanding, receipt understanding, and scanned document image classification. Experiments show that LayoutLM substantially outperforms several SOTA pre-trained models in these tasks. For future research, we will investigate pre-training models with more data and more computation resources. In addition, we will also train LayoutLM using the LARGE architecture with text and layout, as well as involving image embeddings in the pre-training step. Furthermore, we will explore new network architectures and other self-supervised training objectives that may further unlock the power of LayoutLM.

- **arXiv-2021**# LayoutLMv2

Paper: LayoutLMv2: Multi-modal Pre-training for Visually-Rich Document Understanding

Authors: Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, Lidong Zhou

Public: arXiv:1912.13318, 2020arXiv:2012.14740, 2021

github

Abstract

Pre-training of text and layout has proved effective in a variety of visually-rich document understanding tasks due to its effective model architecture and the advantage of large-scale unlabeled scanned/digital-born documents. We propose LayoutLMv2 architecture with new pre-training tasks to model the interaction among text, layout, and image in a single multi-modal framework. Specifically, with a two-stream multi-modal Transformer encoder, LayoutLMv2 uses not only the existing masked visual-language modeling task but also the new text-image alignment and text-image matching tasks, which make it better capture the cross-modality interaction in the pre-training stage. Meanwhile, it also integrates a spatial-aware self-attention mechanism into the Transformer architecture so that the model can fully understand the relative positional relationship among different text blocks. Experiment results show that LayoutLMv2 outperforms LayoutLM by a large margin and achieves new state-of-the-art results on a wide variety of downstream visually-rich document understanding tasks, including FUNSD (0.7895 to 0.8420), CORD (0.9493 to 0.9601), SROIE (0.9524 to 0.9781), Kleister-NDA (0.8340 to 0.8520), RVL-CDIP (0.9443 to 0.9564), and DocVQA (0.7295 to 0.8672).

Introduction

The contributions of this paper are summarized as follows: • We propose a multi-modal Transformer model to integrate the document text, layout, and visual information in the pre-training stage, which learns the cross-modal interaction end- to-end in a single framework.

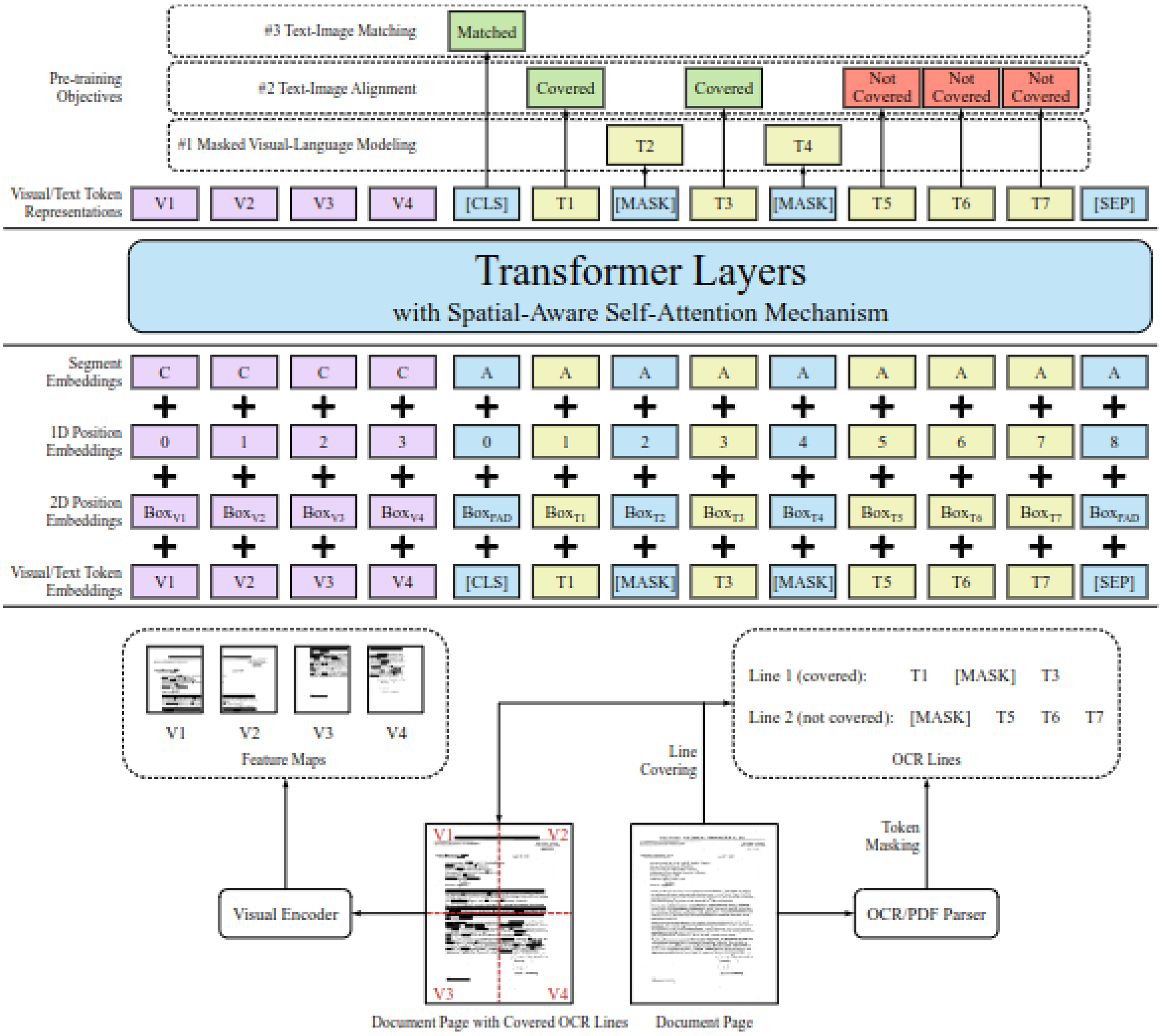
Meanwhile, a spatial-aware self-attention mechanism is integrated into the Transformer architecture.

• In addition to the masked visual-language model, we add text-image alignment and text- image matching as the new pre-training strate- gies to enforce the alignment among different modalities.

• LayoutLMv2 significantly outperforms and achieves new SOTA results not only on the conventional VrDU tasks but also on the VQA task for document images, which demon- strates the great potential for the multi-modal pre-training for VrDU.

Approach

Model Architecture



Pre-training Objectives
#3 Text-Image Matching
Matched
#2 Text-Image Alignment
Covered
Covered
Not Covered
Not Covered
Not Covered
#1 Masked Visual-Language Modeling
T2
T4
Visual/Text Token Representations
V1 V2 V3 V4 [CLS] T1 [MASK] T3 [MASK] T5 T6 T7 [SEP]
Transformer Layers
with Spatial-Aware Self-Attention Mechanism
Segment Embeddings
C C C C A A A A A A A A A
1D Position Embeddings
0 1 2 3 0 1 2 3 4 5 6 7 8
2D Position Embeddings
Box_V1 Box_V2 Box_V3 Box_V4 Box_PAD Box_T1 Box_T2 Box_T3 Box_T4 Box_T5 Box_T6 Box_T7 Box_PAD
Visual/Text Token Embeddings
V1 V2 V3 V4 [CLS] T1 [MASK] T3 [MASK] T5 T6 T7 [SEP]
V1 V2 V3 V4
Feature Maps
Line 1 (covered): T1 [MASK] T3
Line 2 (not covered): [MASK] T5 T6 T7
OCR Lines
Line Covering
Token Masking
Visual Encoder
OCR/PDF Parser
V1 V2 V3 V4
Document Page with Covered OCR Lines
Document Page

Figure 1: An illustration of the model architecture and pre-training strategies for LayoutLMv2

- Text Embedding
  - WordPiece 分詞：LayoutLMv2 使用 WordPiece 方法將 OCR 文本序列進行分詞。WordPiece 方法是一種在自然語言處理中常用的分詞方法。它將文本分割成更小的、更有意義的片段。LayoutLMv2 使用 WordPiece 方法可以更好地捕捉文本中的信息，並提高模型的性能。

```
ti = TokEmb(wi)+PosEmb1D(i)+SegEmb(si)
```

  - Segment embedding：LayoutLMv2 為每個文本片段分配一個特定的 segment si ∈ {[A], [B]}。segment embedding 是用於區分不同文本片段的嵌入。它可以幫助模型更好地理解文本的上下

文。

- Token embedding：token embedding 是表示 token 本身的嵌入。它由一個固定大小的向量表示。token embedding 可以捕捉 token 的語義信息和句法信息。
- Positional embedding：positional embedding 是表示 token 在序列中的位置的嵌入。它由一個可變大小的向量表示。positional embedding 可以幫助模型理解 token 在句子中的位置信息。

LayoutLMv2 的最終文本嵌入是三個嵌入的和。它可以捕捉 token 的語義、句法和位置信息。這些信息可以幫助模型更好地理解文本，並進行多模態文檔理解。

- Visual Embedding

  - CNN 視覺編碼器：LayoutLMv2 使用 ResNeXt-FPN 架構作為 CNN 視覺編碼器。ResNeXt-FPN 是一個深度卷積神經網路，它可以捕捉圖像中的局部和全局特徵。LayoutLMv2 使用 ResNeXt-FPN 將頁面圖像轉換為固定長度的序列。

```
vi = Proj(VisTokEmb (I))i+PosEmb1D(i)+SegEmb([C])
```

  - 視覺嵌入：LayoutLMv2 將 CNN 視覺編碼器的輸出特徵圖平均池化到一個固定大小的圖像。然後，它將圖像攤平成一個視覺嵌入序列。視覺嵌入序列由 W × H 個 token 組成，其中 W 是圖像的寬度，H 是圖像的高度。
  - Positional embedding：LayoutLMv2 為每個視覺 token 添加一個 1D 位置嵌入。1D 位置嵌入可以幫助模型理解視覺 token 在圖像中的位置信息。
  - Segment embedding：LayoutLMv2 將所有視覺 token 都附加到視覺片段 [C]。視覺片段 [C] 用於區分視覺 token 和文本 token。

LayoutLMv2 的最終視覺嵌入是三個嵌入的和。它可以捕捉圖像的局部和全局特徵，以及視覺 token 在圖像中的位置信息。這些信息可以幫助模型更好地理解視覺信息，並進行多模態文檔理解。

- Layout Embedding

  - 布局嵌入層：布局嵌入層是用於將由 OCR 結果表示的軸對齊 token 邊界框表示的空間布局信息嵌入。邊界框包括寬度、高度和角坐標。LayoutLMv2 使用兩個嵌入層分別對 x 軸特徵和 y 軸特徵進行嵌入。

$$\begin{aligned}\mathbf{l}_i = \mathrm{Concat}\big(&\mathrm{PosEmb2D_x}(x_{\mathrm{min}}, x_{\mathrm{max}}, width), \\ &\mathrm{PosEmb2D_y}(y_{\mathrm{min}}, y_{\mathrm{max}}, height)\big)\end{aligned}$$

  - Normalized and discretized coordinates：LayoutLMv2 將所有坐標標準化並離散化到 [0, 1000] 的範圍內。這可以簡化模型的計算，並提高模型的性能。
  - Concatenated bounding box features：LayoutLMv2 將六個邊界框特徵連接起來構建一個 token 級 2D 位置嵌入，也稱為布局嵌入。
  - Visual token embeddings：LayoutLMv2 將視覺 token 嵌入映射回圖像區域，並且不會重疊或遺漏。當計算邊界框時，視覺 token 可以被視為均勻分割的網格。
  - Empty bounding box：LayoutLMv2 將空邊界框 (0, 0, 0, 0, 0, 0) 附加到特殊 token [CLS]、[SEP] 和 [PAD]。

布局嵌入層可以捕捉文本 token 和視覺 token 的空間布局信息。這些信息可以幫助模型更好地理解文檔的布局，並進行多模態文檔理解。

- Multi-modal Encoder with Spatial-Aware Self-Attention Mechanism
    - 多模態編碼器：多模態編碼器是 LayoutLMv2 模型的核心部分。它將文本嵌入、視覺嵌入和布局嵌入連接在一起，並使用自注意力機制來捕捉文本、視覺和布局信息之間的關係。
    - 空間感知自注意力機制：空間感知自注意力機制是 LayoutLMv2 模型的一個新穎的貢獻。它可以捕捉文本 token 和視覺 token 的空間布局信息。這可以幫助模型更好地理解文檔的布局，並進行多模態文檔理解。

$$\begin{aligned}\mathbf{x}_i^{(0)} &= X_i + \mathbf{l}_i, \text{ where} \\ X &= \{\mathbf{v}_0, \ldots, \mathbf{v}_{WH-1}, \mathbf{t}_0, \ldots, \mathbf{t}_{L-1}\}\end{aligned}$$

原始自注意力機制：原始自注意力機制可以捕捉文本 token 之間的關係。但是，它不能有效地捕捉文本 token 和視覺 token 之間的關係。這是因為原始自注意力機制只考慮了 token 的絕對位置。而文本 token 和視覺 token 的空間布局信息是重要的。

$$\alpha_{ij} = \frac{1}{\sqrt{d_{head}}} \left(\mathbf{x}_i \mathbf{W}^Q\right) \left(\mathbf{x}_j \mathbf{W}^K\right)^{\mathsf{T}}$$

空間感知自注意力機制：空間感知自注意力機制可以捕捉文本 token 和視覺 token 之間的關係。它是原始自注意力機制的擴展。它不僅考慮了 token 的絕對位置，還考慮了 token 的空間布局信息。

$$\alpha'_{ij} = \alpha_{ij} + \mathbf{b}^{(1\mathrm{D})}_{j-i} + \mathbf{b}^{(2\mathrm{D_x})}_{x_j - x_i} + \mathbf{b}^{(2\mathrm{D_y})}_{y_j - y_i}$$

空間感知自注意力機制的工作原理：空間感知自注意力機制首先計算原始自注意力機制的注意力分數。然後，它將注意力分數加上空間信息。空間信息由三個部分組成：

```
* 1D 相對位置信息：1D 相對位置信息表示 token 在序列中的位置。
* 2Dx 相對位置信息：2Dx 相對位置信息表示 token 在 x 軸上的相對位置。
* 2Dy 相對位置信息：2Dy 相對位置信息表示 token 在 y 軸上的相對位置。
```

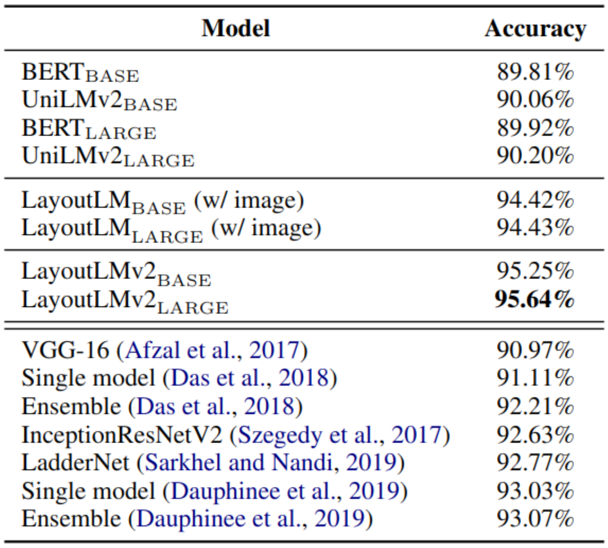
$$\mathbf{h}_i = \sum_j \frac{\exp\left(\alpha'_{ij}\right)}{\sum_k \exp\left(\alpha'_{ik}\right)} \mathbf{x}_j \mathbf{W}^V$$

最後，空間感知自注意力機制使用加權平均的方法來計算輸出向量。空間感知自注意力機制可以有效地捕捉文本 token 和視覺 token 之間的關係。這可以幫助模型更好地理解文檔的布局，並進行多模態文檔理解。

Pre-training Tasks

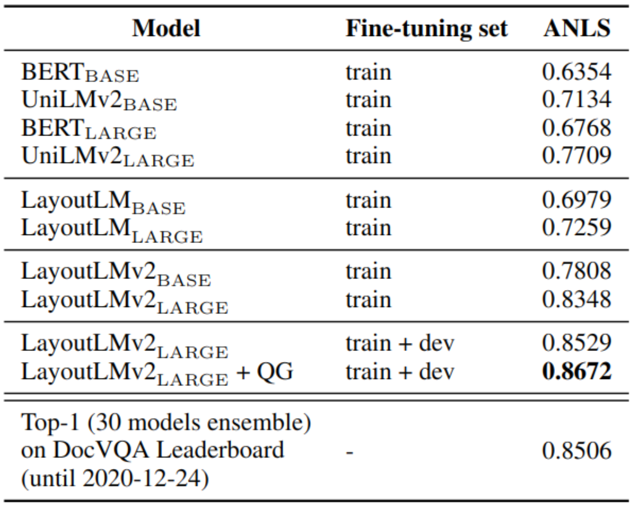
- Masked Visual-Language Modeling (MVLM)：MVLM 是 LayoutLMv2 模型使用的一種預訓練任務。它可以幫助模型學習文本和圖像的跨模態關係。在 MVLM 任務中，模型會被隨機遮蔽一些文本 token，並要求模型恢復這些遮蔽的 token。模型可以利用文本 token 的上下文信息和圖像信息來恢復遮蔽的 token。
- Text-Image Alignment (TIA)：TIA 是 LayoutLMv2 模型使用的另一種預訓練任務。它可以幫助模型學習文本和圖像中空間位置的對應關係。在 TIA 任務中，模型會被隨機選擇一些文本行，並在圖像中遮蔽這些文本行對應的圖像區域。模型可以利用文本 token 的上下文信息來判斷是否有圖像區域被遮蔽。
- Text-Image Matching (TIM)：TIM 是 LayoutLMv2 模型使用的最後一種預訓練任務。它可以幫助模型學習文本和圖像的粗粒度對應關係。在 TIM 任務中，模型會被輸入一個文本序列和一張圖像。模型需要判斷這張圖像是否來自於這個文本序列。

Result



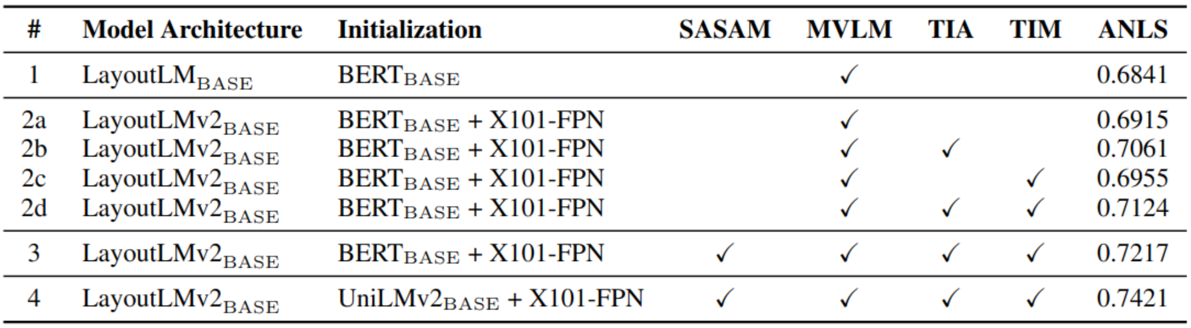
Model	Accuracy
$\text{BERT}_{\text{BASE}}$	89.81%
$\text{UniLMv2}_{\text{BASE}}$	90.06%
$\text{BERT}_{\text{LARGE}}$	89.92%
$\text{UniLMv2}_{\text{LARGE}}$	90.20%
$\text{LayoutLM}_{\text{BASE}}$ (w/ image)	94.42%
$\text{LayoutLM}_{\text{LARGE}}$ (w/ image)	94.43%
$\text{LayoutLMv2}_{\text{BASE}}$	95.25%
$\text{LayoutLMv2}_{\text{LARGE}}$	**95.64%**
VGG-16 (Afzal et al., 2017)	90.97%
Single model (Das et al., 2018)	91.11%
Ensemble (Das et al., 2018)	92.21%
InceptionResNetV2 (Szegedy et al., 2017)	92.63%
LadderNet (Sarkhel and Nandi, 2019)	92.77%
Single model (Dauphinee et al., 2019)	93.03%
Ensemble (Dauphinee et al., 2019)	93.07%

Table 3: Classification accuracy on the RVL-CDIP dataset



Model	Fine-tuning set	ANLS
$\text{BERT}_{\text{BASE}}$	train	0.6354
$\text{UniLMv2}_{\text{BASE}}$	train	0.7134
$\text{BERT}_{\text{LARGE}}$	train	0.6768
$\text{UniLMv2}_{\text{LARGE}}$	train	0.7709
$\text{LayoutLM}_{\text{BASE}}$	train	0.6979
$\text{LayoutLM}_{\text{LARGE}}$	train	0.7259
$\text{LayoutLMv2}_{\text{BASE}}$	train	0.7808
$\text{LayoutLMv2}_{\text{LARGE}}$	train	0.8348
$\text{LayoutLMv2}_{\text{LARGE}}$	train + dev	0.8529
$\text{LayoutLMv2}_{\text{LARGE}}$ + QG	train + dev	**0.8672**
Top-1 (30 models ensemble) on DocVQA Leaderboard (until 2020-12-24)	-	0.8506

Table 4: ANLS score on the DocVQA dataset, "QG" denotes the data augmentation with the question generation dataset.



#	Model Architecture	Initialization	SASAM	MVLM	TIA	TIM	ANLS
1	$\text{LayoutLM}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$		✓			0.6841
2a	$\text{LayoutLMv2}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ + X101-FPN		✓			0.6915
2b	$\text{LayoutLMv2}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ + X101-FPN		✓	✓		0.7061
2c	$\text{LayoutLMv2}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ + X101-FPN		✓		✓	0.6955
2d	$\text{LayoutLMv2}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ + X101-FPN		✓	✓	✓	0.7124
3	$\text{LayoutLMv2}_{\text{BASE}}$	$\text{BERT}_{\text{BASE}}$ + X101-FPN	✓	✓	✓	✓	0.7217
4	$\text{LayoutLMv2}_{\text{BASE}}$	$\text{UniLMv2}_{\text{BASE}}$ + X101-FPN	✓	✓	✓	✓	0.7421

Table 5: Ablation study on the DocVQA dataset, where ANLS scores on the validation set are reported. "SASAM" means the spatial-aware self-attention mechanism. "MVLM", "TIA" and "TIM" are the three pre-training tasks. All the models are trained using the whole pre-training dataset for one epoch with the BASE model size.

CONCLUSION

- LayoutLMv2 is a multi-modal pre-training approach for visually-rich document understanding tasks.

- LayoutLMv2 model not only considers text and layout information, but also integrates image information in the pre-training stage with a single multi-modal framework.
- LayoutLMv2 model uses spatial-aware self-attention mechanism to capture the relative relationship among different bounding boxes.
- LayoutLMv2 model uses new pre-training objectives to enforce the learning of cross-modal interaction among different modalities.
- LayoutLMv2 model has substantially outperformed the SOTA baselines in the document intelligence area on 6 different VrDU tasks, which greatly benefits a number of real-world document understanding tasks.
- For future research, we will further explore the network architecture as well as the pre-training strategies for the LayoutLM family. Meanwhile, we will also investigate the language expansion to make the multi-lingual LayoutLMv2 model available for different languages, especially the non-English areas around the world.

- **arXiv-2021**Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Furu Wei, "LayoutXLM: Multimodal Pre-training for Multilingual Visually-rich Document Understanding", arXiv:2104.08836, 2021

github

Abstract

Multimodal pre-training with text, layout, and image has achieved SOTA performance for visually-rich document understanding tasks recently, which demonstrates the great potential for joint learning across different modalities. In this paper, we present LayoutXLM, a multimodal pre-trained model for multilingual document understanding, which aims to bridge the language barriers for visually-rich document understanding. To accurately evaluate LayoutXLM, we also introduce a multilingual form understanding benchmark dataset named XFUND, which includes form understanding samples in 7 languages (Chinese, Japanese, Spanish, French, Italian, German, Portuguese), and key-value pairs are manually labeled for each language. Experiment results show that the LayoutXLM model has significantly outperformed the existing SOTA cross-lingual pre-trained models on the XFUND dataset. The pre-trained LayoutXLM model and the XFUND dataset are publicly available at this https URL.

Introduction

- Multimodal pre-training for visually-rich Document Understanding (VrDU) has achieved new SOTA performance on several public benchmarks recently

- Meanwhile, we are well aware of the demand from the non-English world since nearly 40% of digital documents on the web are in non-English languages.

- Although a large amount of multilingual text data has been used in these cross-lingual pre-trained models, text-only multilingual models cannot be easily used in the VrDU tasks because they are usually fragile in analyzing the documents due to the format/layout diversity of documents in different countries, and even different regions in the same country.

- We propose LayoutXLM, a multimodal pretrained model for multilingual document understanding, which is trained with large-scale real-world scanned/digital-born documents.

- We also introduce XFUND, a multilingual form understanding benchmark dataset that includes human-labeled forms with key-value pairs in 7 languages (Chinese, Japanese, Spanish, French, Italian, German, Portuguese).
- LayoutXLM has outperformed other SOTA multilingual baseline models on the XFUND dataset, which demonstrates the great potential for the multimodal pre-training for the multilingual VrDU task. The pre-trained LayoutXLM model and the XFUND dataset are publicly available at https://aka.ms/layoutxlm.

Approach

Model Architecture

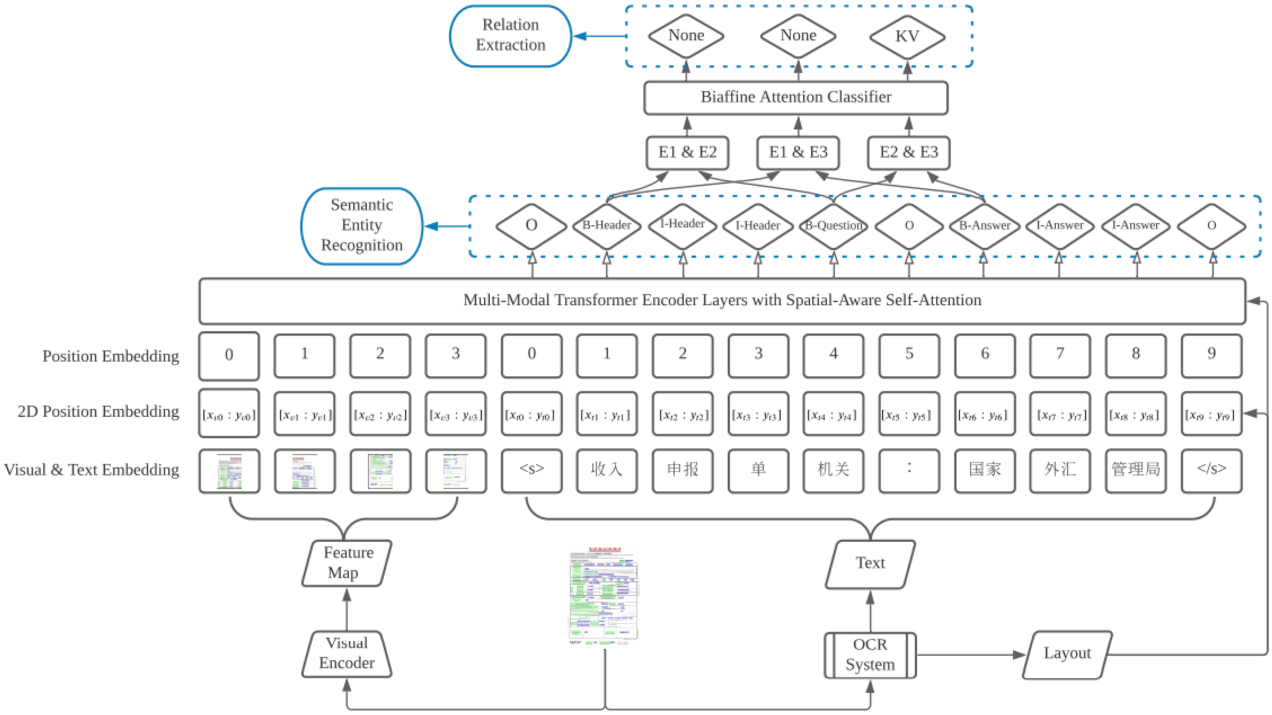


Figure 1: Architecture of the LayoutXLM Model, where the semantic entity recognition and relation extraction tasks are also demonstrated.

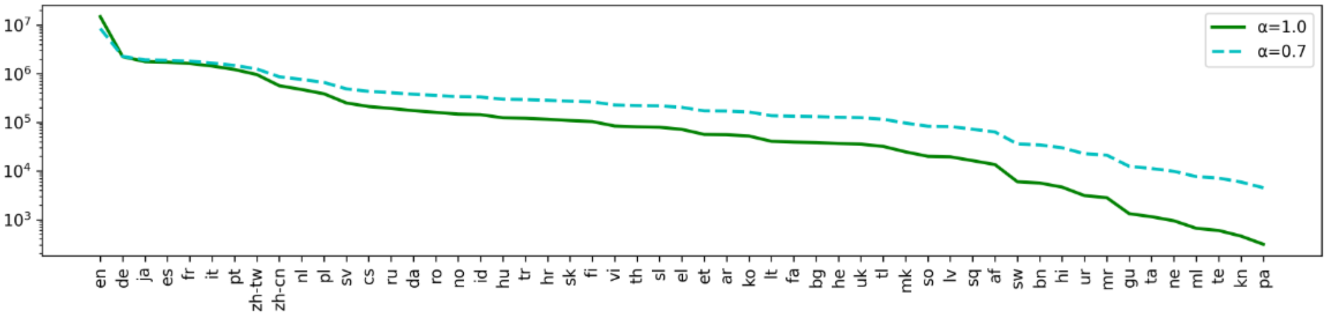


Figure 2: Language distribution of the dataset for pre-training LayoutXLM

- Similar to the LayoutLMv2 framework, we built the LayoutXLM model with a multimodal Trans- former architecture

Pre-training

- In the MMVLM objective, the model is trained to predict a masked text token based on its remaining text context and the whole layout clues. The layout clues are the bounding boxes of the text tokens.
- To obtain the layout clues for multilingual documents, LayoutXLM uses character-level bounding boxes. This is because the definition of the linguistic unit is different from language to language. For example, in English, a word is the basic unit, but in Chinese, a character is the basic unit.
- The MMVLM objective is trained on a dataset of visually-rich documents. The dataset includes documents in 7 languages: Chinese, Japanese, Spanish, French, Italian, German, and Portuguese.
- The Text-Image Alignment (TIA) task is another pre-training task for LayoutXLM. This task is designed to help the model capture the fine-grained alignment relationship between text and image.
- The MMVLM objective is designed to help LayoutXLM learn to predict missing text tokens in a document, given the remaining text context and the layout clues. This is a challenging task, as the layout clues can be noisy and incomplete.
- The TIA task is designed to help LayoutXLM learn to align text and image regions in a document. This is important for tasks such as table detection and question answering.
- The TIM task is designed to help LayoutXLM learn to understand the semantic relationship between text and image in a document. This is important for tasks such as summarization and translation.

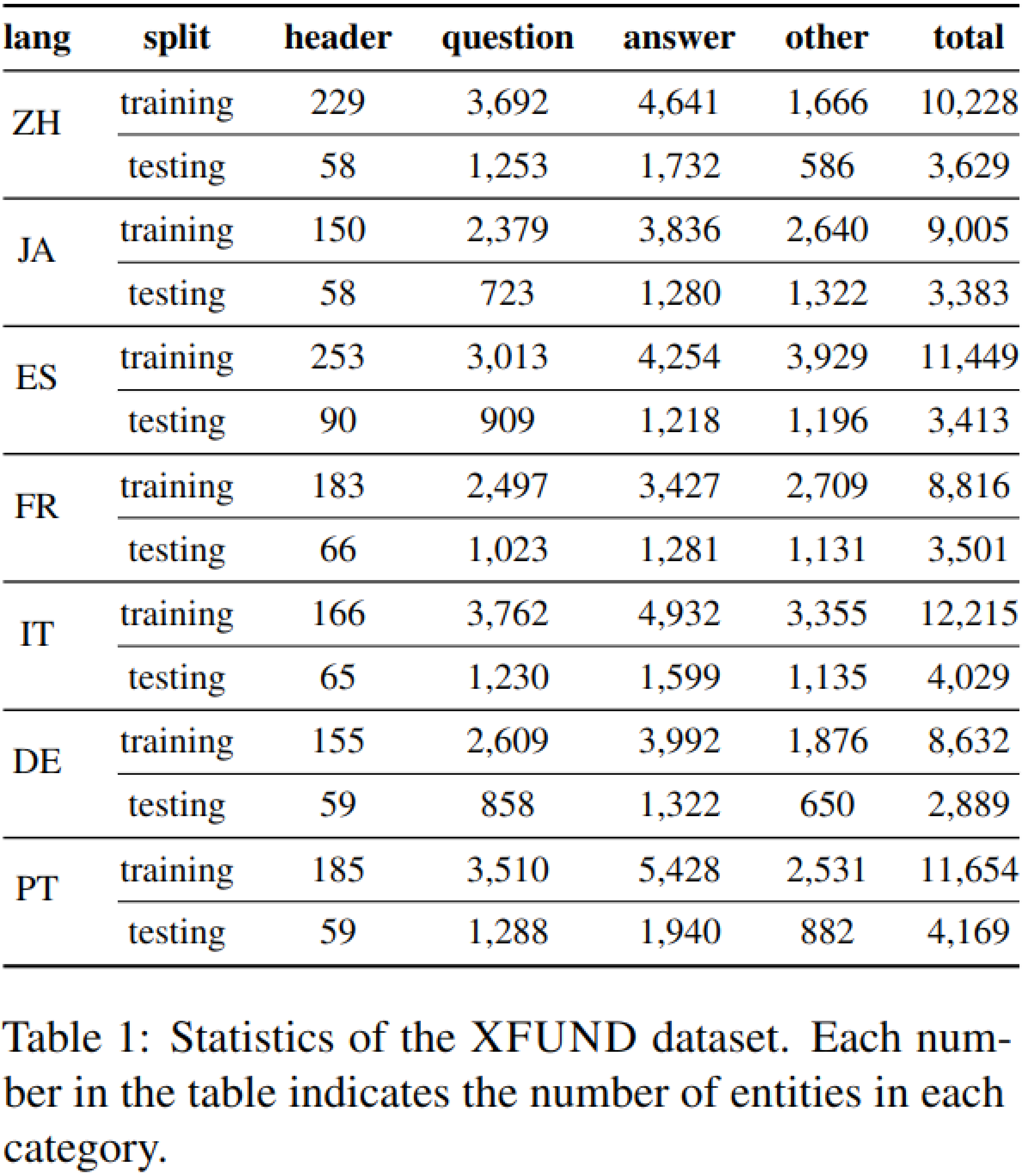
Experiments

Pre-training LayoutXLM

Pre-training LayoutXLM Following the original LayoutLMv2 recipe, we train LayoutXLM models with two model sizes. For theLayoutXLM_BASE model, we use a 12-layer Trans-former encoder with 12 heads and set the hidden size to d = 768. For the LayoutXLM_LARGE model,we increase the layer number to 24 with 16 head sand hidden size to d = 1, 024. ResNeXt101-FPNis used as a visual backbone in both models. Finally, the number of parameters in these two mod-els are approximately 345M and 625M. During the pre-training stage, we first initialize the Trans-former encoder along with text embeddings fromInfoXLM and initialize the visual embedding layer with a Mask-RCNN model trained on PubLayNet. The rest of the parameters are initialized randomly.Our models are trained with 64 Nvidia V100 GPUs

Fine-tuning on XFUND

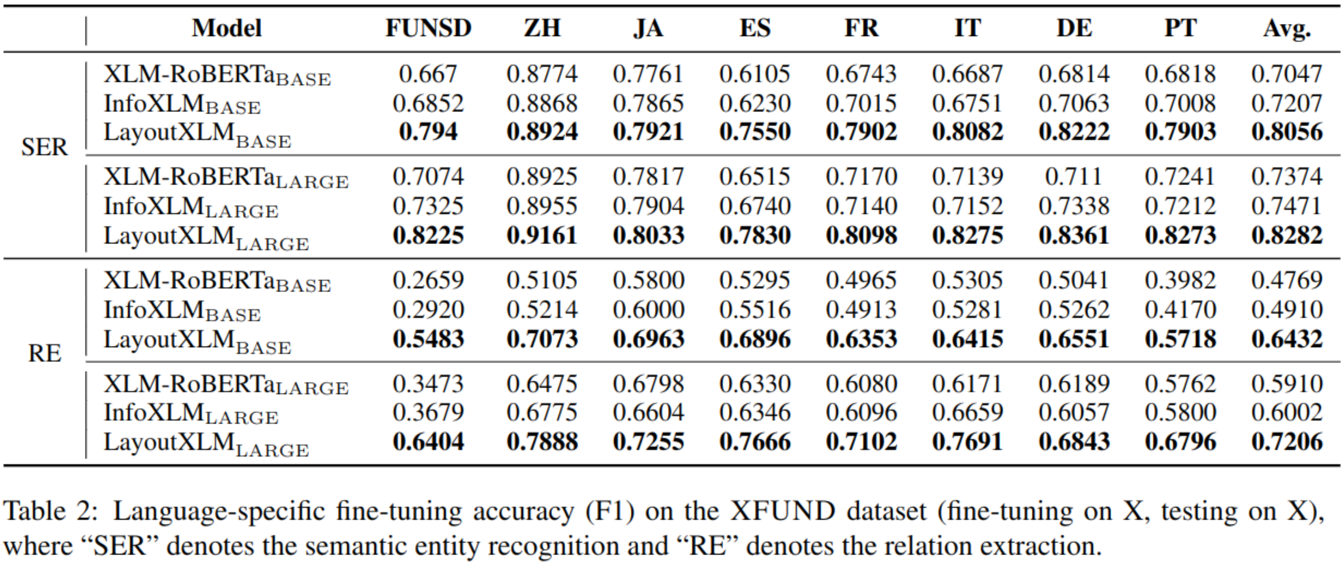
We conduct experi-ments on the XFUND benchmark. Besides the ex-periments of typical language-specific fine-tuning,we also design two additional settings to demon-strate the ability to transfer knowledge among dif-ferent languages, which are zero-shot transfer learn-ing and multitask fine-tuning.



lang	split	header	question	answer	other	total
ZH	training	229	3,692	4,641	1,666	10,228
	testing	58	1,253	1,732	586	3,629
JA	training	150	2,379	3,836	2,640	9,005
	testing	58	723	1,280	1,322	3,383
ES	training	253	3,013	4,254	3,929	11,449
	testing	90	909	1,218	1,196	3,413
FR	training	183	2,497	3,427	2,709	8,816
	testing	66	1,023	1,281	1,131	3,501
IT	training	166	3,762	4,932	3,355	12,215
	testing	65	1,230	1,599	1,135	4,029
DE	training	155	2,609	3,992	1,876	8,632
	testing	59	858	1,322	650	2,889
PT	training	185	3,510	5,428	2,531	11,654
	testing	59	1,288	1,940	882	4,169

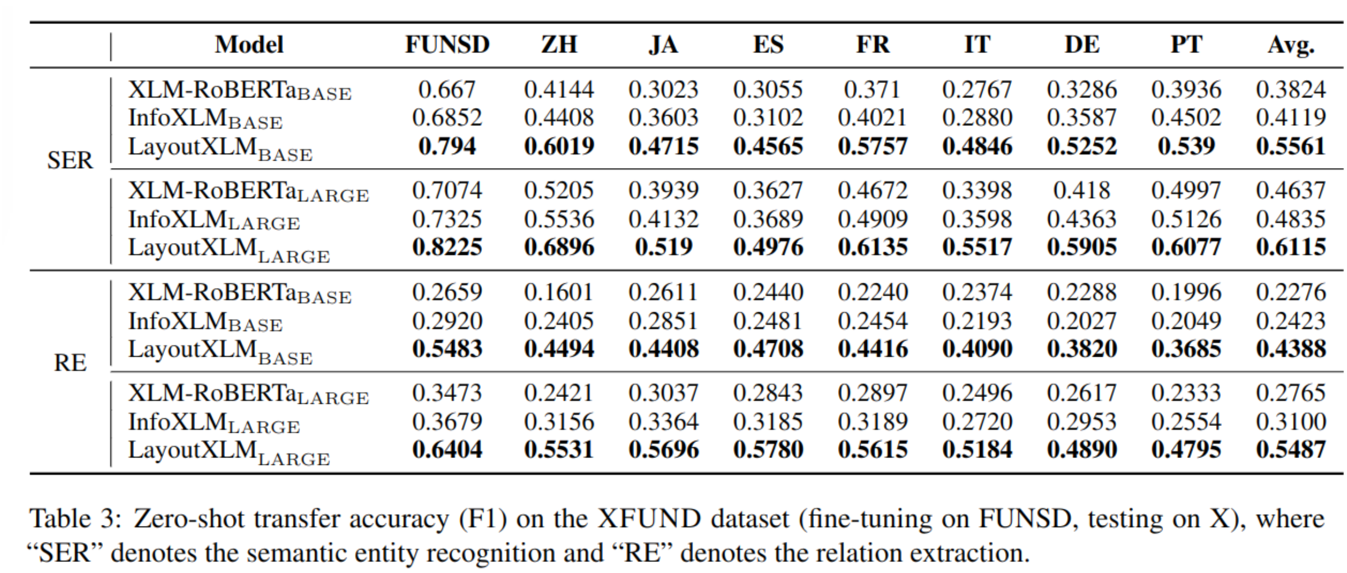
Table 1: Statistics of the XFUND dataset. Each number in the table indicates the number of entities in each category.

Results



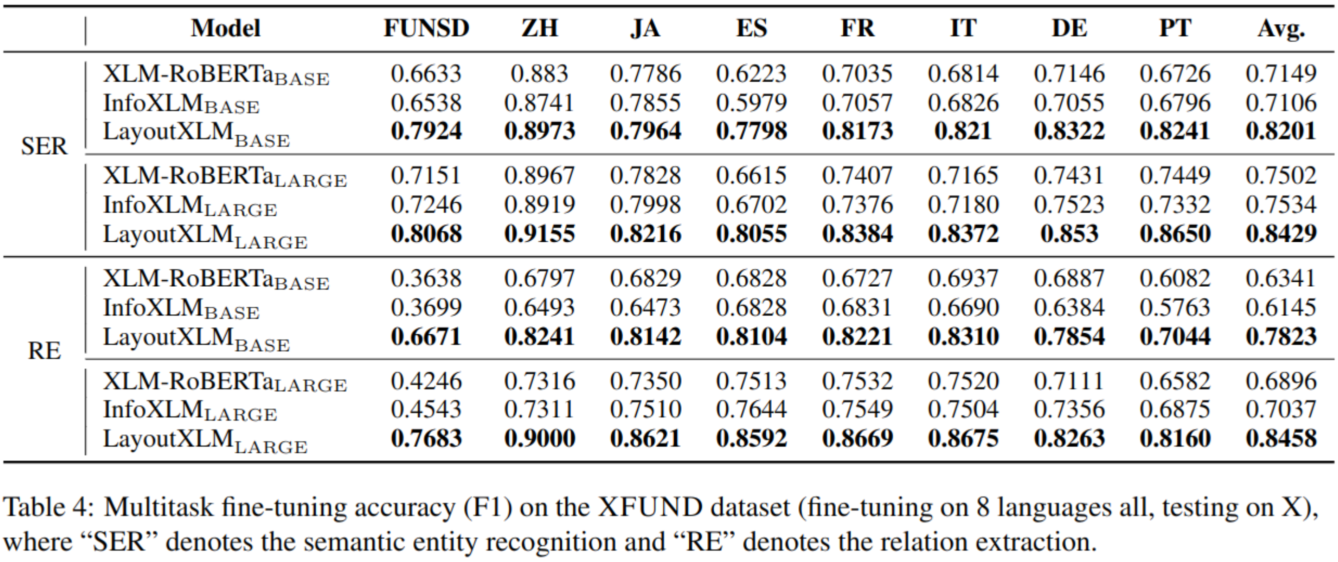
	Model	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
SER	XLM-RoBERTa$_{\text{BASE}}$	0.667	0.8774	0.7761	0.6105	0.6743	0.6687	0.6814	0.6818	0.7047
	InfoXLM$_{\text{BASE}}$	0.6852	0.8868	0.7865	0.6230	0.7015	0.6751	0.7063	0.7008	0.7207
	LayoutXLM$_{\text{BASE}}$	**0.794**	**0.8924**	**0.7921**	**0.7550**	**0.7902**	**0.8082**	**0.8222**	**0.7903**	**0.8056**
	XLM-RoBERTa$_{\text{LARGE}}$	0.7074	0.8925	0.7817	0.6515	0.7170	0.7139	0.711	0.7241	0.7374
	InfoXLM$_{\text{LARGE}}$	0.7325	0.8955	0.7904	0.6740	0.7140	0.7152	0.7338	0.7212	0.7471
	LayoutXLM$_{\text{LARGE}}$	**0.8225**	**0.9161**	**0.8033**	**0.7830**	**0.8098**	**0.8275**	**0.8361**	**0.8273**	**0.8282**
RE	XLM-RoBERTa$_{\text{BASE}}$	0.2659	0.5105	0.5800	0.5295	0.4965	0.5305	0.5041	0.3982	0.4769
	InfoXLM$_{\text{BASE}}$	0.2920	0.5214	0.6000	0.5516	0.4913	0.5281	0.5262	0.4170	0.4910
	LayoutXLM$_{\text{BASE}}$	**0.5483**	**0.7073**	**0.6963**	**0.6896**	**0.6353**	**0.6415**	**0.6551**	**0.5718**	**0.6432**
	XLM-RoBERTa$_{\text{LARGE}}$	0.3473	0.6475	0.6798	0.6330	0.6080	0.6171	0.6189	0.5762	0.5910
	InfoXLM$_{\text{LARGE}}$	0.3679	0.6775	0.6604	0.6346	0.6096	0.6659	0.6057	0.5800	0.6002
	LayoutXLM$_{\text{LARGE}}$	**0.6404**	**0.7888**	**0.7255**	**0.7666**	**0.7102**	**0.7691**	**0.6843**	**0.6796**	**0.7206**

Table 2: Language-specific fine-tuning accuracy (F1) on the XFUND dataset (fine-tuning on X, testing on X), where "SER" denotes the semantic entity recognition and "RE" denotes the relation extraction.



	Model	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
SER	XLM-RoBERTa$_{\text{BASE}}$	0.667	0.4144	0.3023	0.3055	0.371	0.2767	0.3286	0.3936	0.3824
	InfoXLM$_{\text{BASE}}$	0.6852	0.4408	0.3603	0.3102	0.4021	0.2880	0.3587	0.4502	0.4119
	LayoutXLM$_{\text{BASE}}$	**0.794**	**0.6019**	**0.4715**	**0.4565**	**0.5757**	**0.4846**	**0.5252**	**0.539**	**0.5561**
	XLM-RoBERTa$_{\text{LARGE}}$	0.7074	0.5205	0.3939	0.3627	0.4672	0.3398	0.418	0.4997	0.4637
	InfoXLM$_{\text{LARGE}}$	0.7325	0.5536	0.4132	0.3689	0.4909	0.3598	0.4363	0.5126	0.4835
	LayoutXLM$_{\text{LARGE}}$	**0.8225**	**0.6896**	**0.519**	**0.4976**	**0.6135**	**0.5517**	**0.5905**	**0.6077**	**0.6115**
RE	XLM-RoBERTa$_{\text{BASE}}$	0.2659	0.1601	0.2611	0.2440	0.2240	0.2374	0.2288	0.1996	0.2276
	InfoXLM$_{\text{BASE}}$	0.2920	0.2405	0.2851	0.2481	0.2454	0.2193	0.2027	0.2049	0.2423
	LayoutXLM$_{\text{BASE}}$	**0.5483**	**0.4494**	**0.4408**	**0.4708**	**0.4416**	**0.4090**	**0.3820**	**0.3685**	**0.4388**
	XLM-RoBERTa$_{\text{LARGE}}$	0.3473	0.2421	0.3037	0.2843	0.2897	0.2496	0.2617	0.2333	0.2765
	InfoXLM$_{\text{LARGE}}$	0.3679	0.3156	0.3364	0.3185	0.3189	0.2720	0.2953	0.2554	0.3100
	LayoutXLM$_{\text{LARGE}}$	**0.6404**	**0.5531**	**0.5696**	**0.5780**	**0.5615**	**0.5184**	**0.4890**	**0.4795**	**0.5487**

Table 3: Zero-shot transfer accuracy (F1) on the XFUND dataset (fine-tuning on FUNSD, testing on X), where "SER" denotes the semantic entity recognition and "RE" denotes the relation extraction.



	Model	FUNSD	ZH	JA	ES	FR	IT	DE	PT	Avg.
SER	XLM-RoBERTa$_{\text{BASE}}$	0.6633	0.883	0.7786	0.6223	0.7035	0.6814	0.7146	0.6726	0.7149
	InfoXLM$_{\text{BASE}}$	0.6538	0.8741	0.7855	0.5979	0.7057	0.6826	0.7055	0.6796	0.7106
	LayoutXLM$_{\text{BASE}}$	**0.7924**	**0.8973**	**0.7964**	**0.7798**	**0.8173**	**0.821**	**0.8322**	**0.8241**	**0.8201**
	XLM-RoBERTa$_{\text{LARGE}}$	0.7151	0.8967	0.7828	0.6615	0.7407	0.7165	0.7431	0.7449	0.7502
	InfoXLM$_{\text{LARGE}}$	0.7246	0.8919	0.7998	0.6702	0.7376	0.7180	0.7523	0.7332	0.7534
	LayoutXLM$_{\text{LARGE}}$	**0.8068**	**0.9155**	**0.8216**	**0.8055**	**0.8384**	**0.8372**	**0.853**	**0.8650**	**0.8429**
RE	XLM-RoBERTa$_{\text{BASE}}$	0.3638	0.6797	0.6829	0.6828	0.6727	0.6937	0.6887	0.6082	0.6341
	InfoXLM$_{\text{BASE}}$	0.3699	0.6493	0.6473	0.6828	0.6831	0.6690	0.6384	0.5763	0.6145
	LayoutXLM$_{\text{BASE}}$	**0.6671**	**0.8241**	**0.8142**	**0.8104**	**0.8221**	**0.8310**	**0.7854**	**0.7044**	**0.7823**
	XLM-RoBERTa$_{\text{LARGE}}$	0.4246	0.7316	0.7350	0.7513	0.7532	0.7520	0.7111	0.6582	0.6896
	InfoXLM$_{\text{LARGE}}$	0.4543	0.7311	0.7510	0.7644	0.7549	0.7504	0.7356	0.6875	0.7037
	LayoutXLM$_{\text{LARGE}}$	**0.7683**	**0.9000**	**0.8621**	**0.8592**	**0.8669**	**0.8675**	**0.8263**	**0.8160**	**0.8458**

Table 4: Multitask fine-tuning accuracy (F1) on the XFUND dataset (fine-tuning on 8 languages all, testing on X), where "SER" denotes the semantic entity recognition and "RE" denotes the relation extraction.

Conclusion

In this paper, we present LayoutXLM, a multi- modal pre-trained model for multilingual visually- rich document understanding.

The LayoutXLM model is pre-trained with 30 million scanned and digital-born documents in 53 languages.

Mean- while, we also introduce the multilingual form un- derstanding benchmark XFUND, which includes key-value labeled forms in 7 languages.

Experi- mental results have illustrated that the pre-trained LayoutXLM model has significantly outperformed the SOTA baselines for multilingual document un- derstanding, which bridges the language gap in real- world document understanding tasks.

We make LayoutXLM and XFUND publicly available to advance the document understanding research.

For future research, we will further enlarge the multilingual training data to cover more languages as well as more document layouts and templates.

In addition, as there are a great number of business documents with the same content but in different languages, we will also investigate how to leverage the contrastive learning of parallel documents for the multilingual pre-training.

- **arXiv-2022**Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu and Furu Wei, "LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking", arXiv preprint, arXiv:2204.08387, 2022.

github

PaperWithCode

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 AI with unified text and image masking. Additionally, LayoutLMv3 is pre-trained with a word-patch alignment objective to learn cross-modal alignment by predicting whether the corresponding image patch of a text word is masked. The simple unified architecture and training objectives make LayoutLMv3 a general-purpose pre-trained model for both text-centric and image-centric Document AI tasks. Experimental results show that LayoutLMv3 achieves state-of-the-art performance not only in text-centric tasks, including form understanding, receipt understanding, and document visual question answering, but also in image-centric tasks such as document image classification and document layout analysis. The code and models are publicly available at this https URL. https://aka.ms/layoutlmv3

Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, Seunghyun Park,"OCR-free Document Understanding Transformer",arXiv:2111.15664,2022

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li and Furu Wei, "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models", arXiv preprint, arXiv:2109.10282, 2021

Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang and Furu Wei, "DiT: Self-supervised Pre-training for Document Image Transformer", arXiv preprint, arXiv:2203.02378, 2022.

Text Recognition (文字識別)

Lukas Blecher, Guillem Cucurull, Thomas Scialom and Robert Stojnic, "Nougat: Neural Optical Understanding for Academic Documents", arXiv:2308.13418, 2023. https://facebookresearch.github.io/nougat/ https://www.jiqizhixin.com/articles/2023-08-30-3

Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, Yongdong Zhang, "Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition", arXiv:2103.06495, 2021.

Shancheng Fang, Zhendong Mao, Hongtao Xie, Yuxin Wang, Chenggang Yan, Yongdong Zhang,"ABINet++: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Spotting",arXiv:2211.10578, 2022

Yuliang Liu, Chunhua Shen, Lianwen Jin, Tong He, Peng Chen, Chongyu Liu, Hao Chen, "ABCNet v2: Adaptive Bezier-Curve Network for Real-time End-to-end Text Spotting", arXiv preprint, arXiv:2105.03620, 2021.

Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, Yu-Gang Jiang, "SVTR: Scene Text Recognition with a Single Visual Model", arXiv:2205.00159,2022

DeepFake Detection (深度偽造偵測)

H. Zhao, et al., "Multi-attentional Deepfake Detection", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, Nashville, TN, USA, 2021, pp. 2185-2194.

Sun, Zekun and Han, Yujie and Hua, Zeyu and Ruan, Na and Jia, Weijia, "Improving the Efficiency and Robustness of Deepfakes Detection through Precise Geometric Features", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z. Li, "Face Forgery Detection by 3D Decomposition", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.