

LayoutLMv3: 통합 텍스트 및 이미지 마스킹을 사용한 문서 AI 사전 교육

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

탕차오 Lv
마이크로소프트 리서치 아시아
tengchaolv@microsoft.com

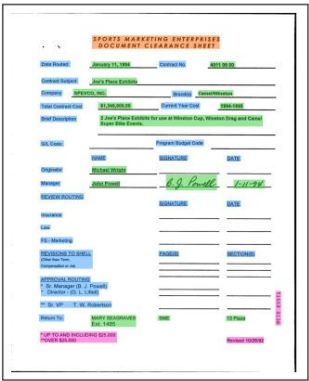
레이 추이
마이크로소프트 리서치 아시아
lecu@microsoft.com

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

후루 웨이
Microsoft Research Asia
fuwei@microsoft.com

추상적인

자가 감독 사전 교육 기술은 Document AI에서 놀라운 발전을 이루었습니다. 대부분의 다중 모드 사전 훈련된 모델은 마스킹된 언어 모델링 목표를 사용하여 텍스트 양식에 대한 양방향 표현을 학습 하지만 이미지 양식에 대한 사전 훈련 목표가 다릅니다. 이러한 불일치는 다중 모드 표현 학습에 어려움을 더합니다. 본 논문에서는 통합된 텍스트 및 이미지 마스킹을 사용하여 문서 AI용 멀티모달 트랜스포머를 사전 훈련하기 위한 LayoutLMv3를 제안합니다. 또한 LayoutLMv3은 텍스트 단어의 해당 이미지 패치가 마스킹되는지 여부를 예측하여 크로스 모달 정렬을 학습하기 위해 단어 패치 정렬 목표로 사전 훈련됩니다. 단순한 통합 아키텍처 및 훈련 목표 덕분에 LayoutLMv3은 텍스트 중심 및 이미지 중심 문서 AI 작업 모두를 위한 사전 훈련된 범용 모델입니다. 실험 결과는 LayoutLMv3이 양식 이해, 영수증 이해, 문서 시각적 질문 응답을 포함한 텍스트 중심 작업뿐만 아니라 문서 이미지 분류 및 문서 레이아웃 과 같은 이미지 중심 작업에서도 최첨단 성능을 달성함을 보여줍니다. 분석. 코드와 모델은 <https://aka.ms/layoutlmv3>에 공개적으로 사용할 수 있습니다.



(a) FUNSD에 대한 텍스트 중심 양식 이해



(b) PubLayNet의 이미지 중심 레이아웃 분석

그림 1: 문서 AI 작업의 예.

CCS 개념

- 응용 컴퓨팅 → 문서 분석; 컴퓨팅 방법론 → 자연어 처리.

키워드

문서 ai, layoutlm, 다중 모드 사전 교육, 버전 및 언어

ACM 참조 형식:

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu 및 Furu Wei. 2022. LayoutLMv3: 통합 텍스트 및 이미지 마스킹을 사용한 문서 AI 사전 교육. 2022년 10월 10-14일, 포르투갈 리스보아에서 제30회 멀티미디어에 관한 ACM 국제 회의(MM '22) 절차에서. ACM, New York, NY, USA, 10페이지. <https://doi.org/10.1145/3503161.3548112>

마이크로소프트 리서치 인턴십 기간 동안의 기여. 교신저자: Lei Cui 및 Furu Wei.

이 저작물의 전체 또는 일부를 개인용 또는 강의실용으로 디지털 또는 하드 카피로 만들 수 있는 권한은 저자가 영리 또는 상업적 이익을 위해 만들거나 배포하지 않고 사본에 이 통지와 첫 페이지에 전체 인용이 포함되어 있는 경우 무료로 부여됩니다. . ACM이 아닌 다른 사람이 소유한 이 저작물의 구성 요소에 대한 저작권은 존중되어야 합니다. 인용으로 초록이 허용됩니다. 달리 복사하거나 재게시하거나 서버에 게시하거나 목록에 재배포하려면 사전에 특정 허가 및/또는 수수료가 필요합니다. . permission@acm.org에서 권한을 요청합니다.

MM '22, 2022년 10월 10-14일, 포르투갈 리스보아
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10... \$15.00
<https://doi.org/10.1145/3503161.3548112>

1. 소개

최근 몇 년 동안 사전 훈련 기술은 문서 이해 작업에서 놀라운 발전을 달성함으로써 Document AI 커뮤니티에서 큰 파장을 일으키고 있습니다 [2, 13-15, 17, 25, 28, 31, 32, 40, 41, 50, 52, 54-56]. 그림 1에서 볼 수 있듯이 사전 훈련된 Document AI 모델은 산업 응용 및 학술 연구에 중요한 스캔 양식 및 학술 논문과 같은 다양한 문서에 대한 레이아웃을 구문 분석하고 주요 정보를 추출할 수 있습니다 [8].

자가 감독 사전 훈련 기술은 재구성 사전 훈련 목표의 성공적인 적용으로 인해 표현 학습에서 급속한 발전을 이루었습니다. NLP 연구에서 BERT는 컨텍스트를 기반으로 임의로 마스킹된 단어 토큰의 원래 어휘 ID를 예측하여 양방향 표현을 학습하는 "마스킹된 언어 모델링"(MLM)을 처음 제안했습니다 [9]. 대부분의 고성능 다중 모달 사전 훈련된 문서 AI 모델은 텍스트 양식에 대해 BERT에서 제안한 MLM을 사용하지만 그림 2에 표시된 것처럼 이미지 양식에 대한 사전 교육 목표가 다릅니다. 예를 들어 DocFormer는 CNN 디코더를 통해 이미지 픽셀을 재구성하는 방법을 학습합니다 [2], 문서 레이아웃과 같은 높은 수준의 구조보다는 시끄러운 세부 사항을 학습하는 경향이 있습니다 [43, 45]. SelfDoc은 더 작은 어휘 [6, 18]에서 이산 기능을 분류하는 것보다 더 노이즈가 많고 학습하기 어려운 마스킹된 영역 기능을 회귀할 것을 제안합니다 [6, 18]. 이미지(예: 조밀한 이미지 픽셀 또는 연속 영역 기능)와 텍스트(예: 이산

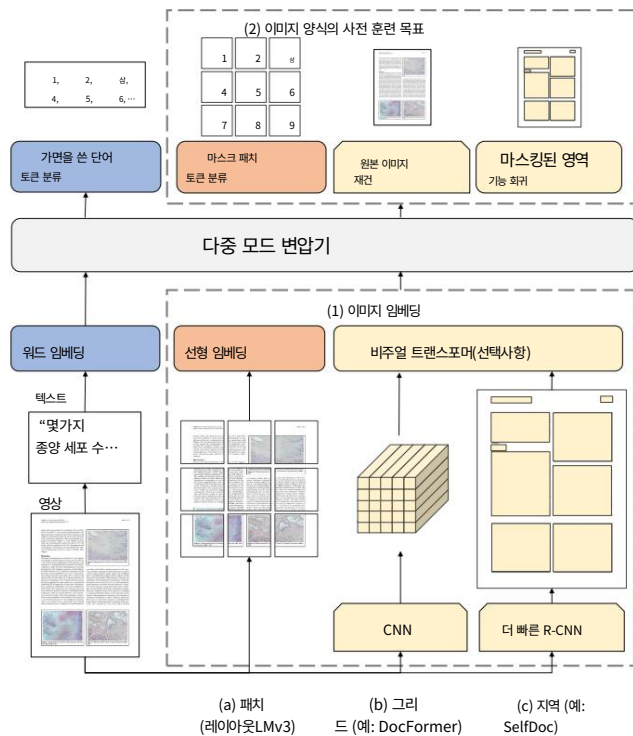


그림 2: (1) 이미지 임베딩에 대한 기존 작업(예: DocFormer [2] 및 SelfDoc [31])과의 비교: LayoutLMv3는 선택 패치를 사용하여 CNN의 계산 병목 현상을 줄이고 교육에서 영역 감독의 필요성을 제거합니다. 객체 감지; (2) 이미지 양식에 대한 사전 교육 목표: LayoutLMv3는 원시 픽셀 또는 영역 기능 대신 마스크된 패치의 개별 이미지 토큰을 재구성하여 잡음이 많은 세부 정보가 아닌 높은 수준의 레이아웃 구조를 캡처하는 방법을 학습합니다.

토큰) 목표는 다중 모드 표현 학습에 필수적인 교차 모드 정렬 학습에 어려움을 더합니다.

텍스트 및 이미지 형식의 사전 훈련 목표의 불일치를 극복 하고 다중 모드 표현 학습을 용이하게 하기 위해 통합 텍스트 및 이미지 마스크 목표 MLM 및 MIM 을 사용하여 문서 AI용 다중 모드 변환기를 사전 훈련하는 LayoutLMv3를 제안 합니다. 그림 3에서 볼 수 있듯이 LayoutLMv3는 텍스트 형식의 마스크된 단어 토큰을 재구성하고 이미지 형식의 마스크된 패치 토큰을 대칭적으로 재구성하는 방법을 학습합니다. DALL-E [43] 및 BEiT [3]에서 영감을 받아 개별 VAE의 잠재 코드에서 대상 이미지 토큰을 얻습니다. 문서의 경우 각 텍스트 단어는 이미지 패치에 해당합니다. 이 크로스 모드 정렬을 학습하기 위해 텍스트 단어의 해당 이미지 패치가 마스크되는지 여부를 예측하는 WPA(Word-Patch Alignment) 목표를 제안합니다.

ViT [11] 및 ViLT [22] 에서 영감을 얻은 LayoutLMv3는 페이지 개체 감지와 같은 복잡한 전처리 단계 없이 문서 이미지에서 원시 이미지 패치를 직접 활용 합니다. LayoutLMv3는 통합 MLM, MIM 및 WPA 목표를 사용하여 Transformer 모델에서 이미지, 텍스트 및 다중 모드 표현을 공동으로 학습합니다. 이로 인해 LayoutLMv3는 사전 훈련된 최초의 다중 모드 문서 시가 됩니다.

이미지 임베딩을 위한 CNN이 없는 모델은 매개변수를 크게 절약하고 영역 주석을 제거합니다. 단순한 통합 아키텍처 및 목표 덕분에 LayoutLMv3는 텍스트 중심 작업과 이미지 중심 Document AI 작업 모두를 위한 사전 훈련된 범용 모델이 되었습니다.

사전 훈련된 LayoutLMv3 모델을 텍스트 중심 벤치마크를 포함한 5개의 공개 벤치마크에서 평가했습니다. 양식 이해를 위한 FUNSD [20], 영수증 이해를 위한 CORD [39], 문서 시각적 질문 답변을 위한 DocVQA [38] 및 이미지 중심 벤치마크: 문서 이미지 분류를 위한 RVL-CDIP [16], 문서 레이아웃 분석을 위한 PubLayNet [59]. 실험 결과는 LayoutLMv3가 이러한 벤치마크에서 매개변수 효율성으로 최첨단 성능을 달성한다는 것을 보여줍니다. 또한 LayoutLMv3는 단순하고 깔끔한 아키텍처와 사전 학습 목표를 위해 쉽게 재현할 수 있습니다.

우리의 기여는 다음과 같이 요약됩니다.

- LayoutLMv3는 시각적 특징을 추출하기 위해 사전 훈련된 CNN 또는 Faster R-CNN 백본에 의존하지 않는 Document AI의 첫 번째 멀티모달 모델로, 매개변수를 크게 절약하고 영역 주석을 제거합니다.
- LayoutLMv3는 통합된 이산 토큰 재구성 목표 MLM 및 MIM으로 텍스트와 이미지 다중 모드 표현 학습 간의 불일치를 완화합니다. 또한 크로스 모드 정렬 학습을 용이하게 하기 위해 WPA(Word-Patch Alignment) 목표를 제안합니다. LayoutLMv3는 텍스트 중심 및 이미지 중심 Document AI 작업 모두를 위한 범용 모델입니다. 처음으로 Document AI의 비전 작업에 대한 멀티모달 트랜스포머의 일반성을 보여줍니다. 실험 결과는 LayoutLMv3가 Document AI의 텍스트 중심 작업 및 이미지 중심 작업에서 최첨단 성능을 달성한다는 것을 보여줍니다. 코드와 모델은 <https://aka.ms/layoutlmv3>에서 공개적으로 사용할 수 있습니다.

2 레이아웃LMV3

그림 3은 LayoutLMv3의 개요를 제공합니다.

2.1 모델 아키텍처

LayoutLMv3는 통합된 텍스트-이미지 다중 모드 변환기를 적용하여 교차 모드 표현을 학습합니다. 트랜스포머는 다중 계층 아키텍처를 가지며 각 계층은 주로 다중 헤드 자체 주의 및 위치에 따라 완전히 연결된 피드 포워드 네트워크로 구성됩니다 [49]. Transformer의 입력은 텍스트 임베딩 $Y = y_1$ 및 이미지 임베딩 $X = x_1$: 시퀀스의 연결입니다. 여기서 y 와 x 는 각각 텍스트와 이미지의 시퀀스 길이입니다.

Transformer를 통해 마지막 레이어는 텍스트 및 이미지 컨텍스트 표현을 출력합니다.

텍스트 임베딩. 텍스트 임베딩은 단어 임베딩과 위치 임베딩의 조합입니다. 고성능 OCR 툴킷으로 문서 이미지를 사전 처리하여 텍스트 콘텐츠 및 해당 2D 위치 정보를 얻었습니다. 사전 훈련된 모델 RoBERTa [36]의 단어 임베딩 행렬로 단어 임베딩을 초기화합니다. 위치 임베딩에는 1D 위치 및 2D 레이아웃 위치 임베딩이 포함되며, 여기서 1D 위치는 텍스트 시퀀스 내의 토큰 인덱스를 나타내고 2D 레이아웃 위치는 텍스트의 경계 상자 좌표를 나타냅니다.

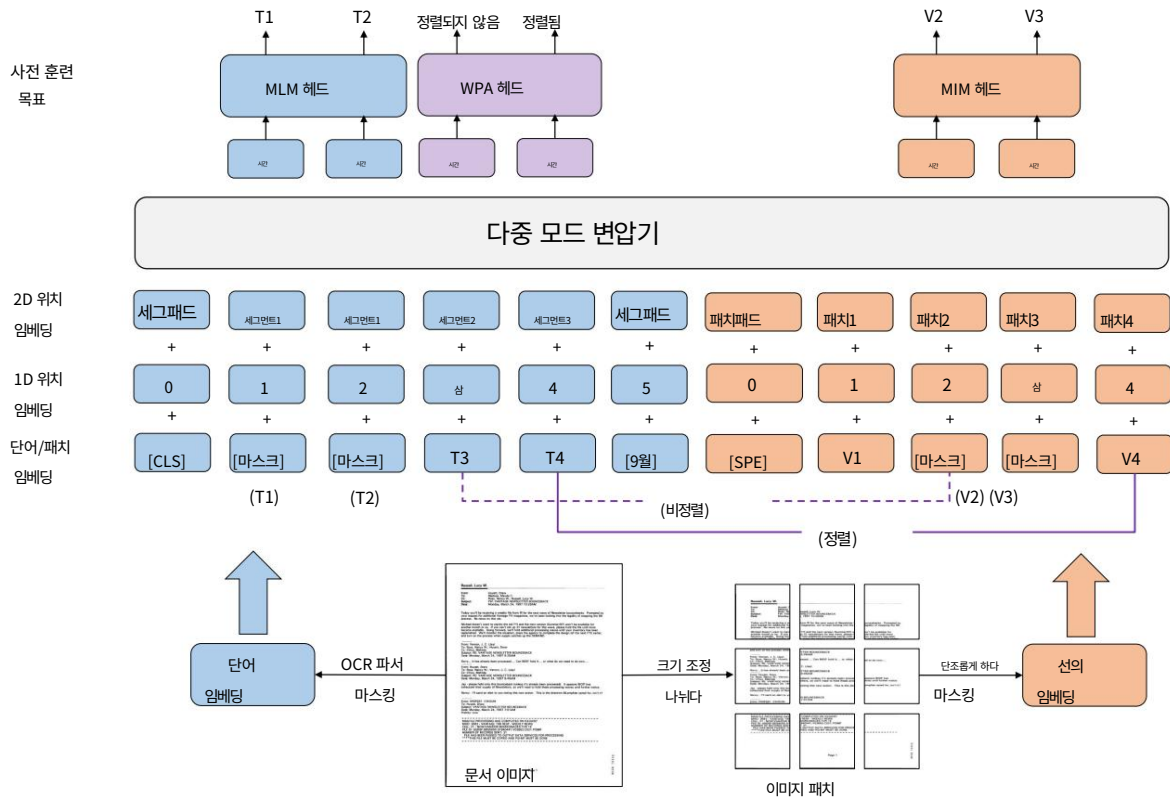


그림 3: LayoutLMv3의 아키텍처 및 사전 훈련 목표. LayoutLMv3은 통합된 텍스트 및 이미지 마스킹 목표를 가진 문서 AI용 사전 훈련된 다중 모드 변압기입니다. 입력 문서 이미지와 해당 텍스트 및 레이아웃 위치 정보가 주어지면 모델은 패치 및 단어 토큰의 선형 프로젝션을 입력으로 사용하여 컨텍스트화된 벡터 표현으로 인코딩합니다. LayoutLMv3은 MLM(Masked Language Modeling) 및 MIM(Masked Image Modeling)의 개별 토큰 재구성 목표로 사전 훈련됩니다. 또한 LayoutLMv3은 텍스트 단어의 해당 이미지 패치가 마스킹되는지 여부를 예측하여 교차 모달 정렬을 학습하기 위해 WPA(Word-Patch Alignment) 목표로 사전 훈련됩니다. "세그"는 세그먼트 수준 위치를 나타냅니다. "[CLS]", "[MASK]", "[SEP]" 및 "[SPE]"는 특별한 토큰입니다.

순서. LayoutLM에 이어 모든 좌표를 이미지 크기로 정규화하고 임베딩 레이어를 사용하여 x 축, y 축, 너비 및 높이 기능을 개별적으로 임베딩합니다 [54]. LayoutLM 및 LayoutLMv2는 각 단어에 위치가 있는 단어 수준 레이아웃 위치를 채택합니다. 대신, 단어가 일반적으로 동일한 의미론적 의미를 표현하기 때문에 세그먼트의 단어가 동일한 2D 위치를 공유하는 세그먼트 수준 레이아웃 위치를 채택합니다 [28].

이미지 임베딩. Document AI의 기존 멀티모달 모델은 CNN 그리드 기능 [2, 56]을 추출하거나 Faster R-CNN [44]과 같은 객체 감지기에 의존하여 이미지 임베딩을 위한 영역 기능 [14, 31, 40, 54]을 추출합니다. 계산 병목 현상이 심하거나 지역 감도가 필요합니다. ViT [11] 및 ViLT [22]에서 영감을 받아 이미지 패치를 다중 모드 Transformer에 공급하기 전에 이미지 패치의 선형 프로젝션 기능으로 문서 이미지를 나타냅니다. 구체적으로 문서 이미지의 크기를 x 로 조정하고 이미지를 $I \in \mathbb{R}$ 로 표시합니다. 여기서 I 는 각각 이미지의 채널 크기, 너비 및 높이입니다. 그런 다음 이미지를 일련의 균일한 x 패치로 분할하고 이미지 패치를 차수에 선형으로 투영하고 길이가 $= /$ 인 일련의 벡터로 변환합니다. 그런 다음 추가

예비 실험에서 2D 위치 임베딩을 사용하여 개선을 관찰하지 않았기 때문에 각 패치에 대한 학습 가능한 1D 위치 임베딩. LayoutLMv3은 이미지 특징을 추출하기 위해 CNN에 의존하지 않는 Document AI의 첫 번째 멀티모달 모델입니다.

우리는 LayoutLMv2[56]를 따르는 텍스트 및 이미지 형식에 대한 self-attention 네트워크에서 바이어스 용어로 의미론적 1D 상대 위치 및 공간 2D 상대 위치를 삽입합니다.

2.2 사전 훈련 목표 LayoutLMv3은

MLM, MIM 및 WPA 목표로 사전 훈련되어 자기 감독 학습 방식으로 다중 모달 표현을 학습합니다. LayoutLMv3의 전체 사전 학습 목표가 정의됨

으로 = + + .

목표 I: 가변 언어 모델링(MLM). 언어 측면에서 MLM은 BERT [9]의 마스킹된 언어 모델링과 LayoutLM [54] 및 LayoutLMv2 [56]의 마스킹된 시각적 언어 모델링에서 영감을 받았습니다. 텍스트 토큰의 30%를 마스킹합니다.

포아송 분포(= 3) [21, 27]에서 추출한 스펠 길이를 사용한 스펠 마스크 전략. 사전 훈련 목표는 이미지 토큰 X의 손상된 시퀀스의 컨텍스트 표현을 기반으로 올바른 마스크된 텍스트 토큰 y의 로그 우도를 최대화 하고 마스크된 위치를 나타내는 것입니다. 다음과 같이 Transformer 모델의 매개변수를 표시 하고 후속 교차 엔트로피 손실을 최소화합니다. 및 텍스트 토큰 Y , 어디 " , "

() = ∑_{l=1}^L 로그 yℓ | 엑스 " , " , " (1)

레이아웃 정보를 변경하지 않고 유지하므로 이 목표는 모델이 레이아웃 정보와 텍스트 및 이미지 컨텍스트 간의 대응 관계를 학습하도록 촉진합니다.

목표 II: 마스크 이미지 모델링(MIM). 모델이 문맥 텍스트 및 이미지 표현에서 시각적 콘텐츠를 해석하도록 장려하기 위해 BEiT [3]의 MIM 사전 훈련 목표를 다중 모드 변환기 모델에 적용합니다. MIM 목표는 블록 단위 마스크 전략 [3]으로 약 40% 이미지 토큰의 백분율을 무작위로 마스크하는 MLM 목표와 대칭입니다. MIM 목표는 주변 텍스트 및 이미지 토큰의 컨텍스트에서 마스크된 이미지 토큰 x를 재구성하기 위해 교차 엔트로피 손실에 의해 구동됩니다.

() = ∑_{l=1}^L 로그 엑스 | 엑스 " , " , " (2)

이미지 토큰의 라벨은 시각적 어휘 [43]에 따라 조밀한 이미지 픽셀을 개별 토큰으로 변환할 수 있는 이미지 토큰라이저에서 가져옵니다. 따라서 MIM은 시끄러운 낮은 수준의 세부 정보가 아닌 높은 수준의 레이아웃 구조 학습을 용이하게 합니다.

목표 III: 워드 패치 정렬(WPA). 문서의 경우 각 텍스트 단어는 이미지 패치에 해당합니다. MLM 및 MIM을 사용하여 텍스트 및 이미지 토큰을 무작위로 마스크하므로 텍스트 및 이미지 양식 간에 명시적인 정렬 학습이 없습니다.

따라서 우리는 텍스트 단어와 이미지 패치 간의 세분화된 정렬을 학습하기 위한 WPA 목표를 제안합니다. WPA 목표는 텍스트 단어의 해당 이미지 패치가 마스크되는지 여부를 예측하는 것입니다. 특히 해당 이미지 토큰도 마스크 해제된 경우 마스크 해제된 텍스트 토큰에 정렬된 레이블을 할당합니다.

그렇지 않으면 정렬되지 않은 레이블을 할당합니다. 모델이 마스크된 텍스트 단어와 이미지 패치 간의 대응 관계를 학습 하지 못하도록 WPA 손실을 계산할 때 마스크된 텍스트 토큰을 제외합니다. 문맥 텍스트와 이미지를 입력하고 이진 교차 엔트로피 손실 과 함께 이진 정렬/비정렬 레이블을 출력하는 2계층 MLP 헤드를 사용합니다.

() = - ∑_{l=1}^L log p(zℓ | 엑스 " , " , ") (3)

여기서 - " 마스크되지 않은 텍스트 토큰의 수, zℓ 은 이진수 ℓ 위치에 있는 언어 토큰의 레이블.

3가지 실험

3.1 모델 구성 LayoutLMv3의 네트워크 아키텍처

텍처는 공정한 비교를 위해 LayoutLM [54] 및 LayoutLMv2 [56]의 네트워크 아키텍처를 따릅니다. LayoutLMv3에는 기본 및 대형 모델 크기를 사용합니다. LayoutLMv3BASE는 12계층

12헤드 셀프 어텐션, 숨겨진 크기 = 768, 피드포워드 네트워크의 중간 크기 3,072가 있는 트랜스포머 엔코더.

LayoutLMv3LARGE는 16-head self-attention, hidden size = 1,024 및 4,096 중간 크기의 피드포워드 네트워크를 갖춘 24계층 Transformer 인코더를 채택합니다. 텍스트 입력을 사전 처리하기 위해 최대 시퀀스 길이 = 512인 BPE(Byte-Pair Encoding) [46]로 텍스트 시퀀스를 토큰화합니다. 시작과 끝에 [CLS] 및 [SEP] 토큰을 추가합니다. 각 텍스트 시퀀스.

텍스트 시퀀스의 길이가 [PAD] 토큰을 추가하는 것보다 짧은 경우. 이러한 특수 토큰의 경계 상자 좌표는 모두 0입니다. 이미지 삽입을 위한 매개변수는 다음과 같습니다. × × = 3 × 224 × 224, = 16, = 196.

우리는 메모리 비용을 줄이고 훈련 절차를 가속화하기 위해 분산 및 혼합 정밀도 훈련을 채택합니다. 또한 기밀기 누적 메커니즘을 사용하여 샘플 배치를 여러 미니 배치로 분할하여 큰 배치 크기에 대한 메모리 제약을 극복합니다. 또한 메모리 비용을 줄이기 위해 문서 레이아웃 분석을 위한 그래디언트 체크포인트링 기술을 사용합니다. 교육을 안정화하기 위해 CogView [10]에 따라 주의 계산을 softmax QK = softmax Q √로 변경합니다.

—= —= K - 최대 —= 케이 × , 어디 32입니다.

3.2 LayoutLMv3 사전 교육 다양한 문서 작업에 대한

범용 표현을 배우기 위해 대규모 IIT-CDIP 데이터 세트에서 LayoutLMv3를 사전 교육합니다. IIT CDIP Test Collection 1.0은 약 1,100만 개의 문서 이미지를 포함하고 4,200만 페이지로 분할할 수 있는 대규모 스펠 문서 이미지 데이터 세트입니다 [26]. 우리는 LayoutLMv3를 교육하는 데 1,100만 개만 사용합니다. 우리는 LayoutLM 모델 [54, 56]에 따라 이미지 확대를 수행하지 않습니다. 텍스트 임베딩 레이어와 함께 멀티모달 트랜스 포머 인코더의 경우 LayoutLMv3은 사전 훈련된 RoBERTa [36]가중치에서 초기화됩니다. 우리의 이미지 토큰라이저는 DiT의 사전 훈련된 이미지 토큰라이저, 자체 감독 사전 훈련된 문서 이미지 트랜스포머 모델 [30]에서 초기화됩니다. 이미지 토큰의 어휘 크기는 8,192입니다. 우리는 나머지 모델 매개변수를 동형으로 초기화했습니다. 500,000단계에 대해 배치 크기가 2,048인 Adam 옵티마이저 [23]를 사용하여 Lay outLMv3를 사전 교육합니다. 우리는 1 2, (1, 2) = (0.9, 0.98)의 가중치 감쇠를 사용합니다. LayoutLMv3BASE 모델의 경우 1-4의 학습률을 사용 하고 처음 4.8% 단계에 걸쳐 학습률을 선형으로 워밍업합니다. LayoutLMv3LARGE의 경우 학습률과 워밍업 비율은 각각 5-5와 10%입니다.

3.3 Multimodal 작업에 대한 미세 조정 우리는 LayoutLMv3를

일반적인 자체 감독 사전 교육 접근 방식과 비교하고 사전 교육 방식으로 분류합니다.

- [T] 텍스트 형식: BERT [9] 및 RoBERTa [36]는 Transformer 아키텍처의 텍스트 정보만 사용하는 일반적인 사전 훈련된 언어 모델입니다. 우리는 FUNSD와 RVL을 사용합니다
- LayoutLM [54]의 RoBERTa의 CDIP 결과 및 LayoutMv2[56]의 BERT 결과. 우리는 RoBERTa의 CORD 및 DocVQA 결과를 재현하고 보고합니다. • [T+I] 텍스트 및 레이아웃 양식: LayoutLM은 BERT [54]의 임베딩에 단어 수준의 공간 임베딩을 추가하여 레이아웃 정보를 통합합니다. StructuralLM은 세그먼트 레벨 레이아웃 정보를 활용합니다 [28]. BROS는 상대 레이아웃을 인코딩합니다.

표 1: CORD[39], FUNSD[20], RVL-CDIP[16] 및 DocVQA[38] 데이터세트에 대한 기존 공개 모델과의 비교.

"T/L/I"는 "텍스트/레이아웃/이미지" 형식을 나타냅니다. "R/G/P"는 "영역/그리드/패치" 이미지 임베딩을 나타냅니다. 우리는 모든 값을 곱합니다 + UDoc 논문[14]에서 CORD 분할은 더 나은 가독성 큰 점수와 직접 비교할 수 없습니다. 을 위해 100이 아닌 훈련/테스트에 대한 626/247 영수증입니다. 다른 작업에서 채택한 공식 800/100 교육/테스트 영수증. 따라서 점수 + 는 다

‡ 로 표시된 모델은 DocVQA를 교육하는 데 더 많은 데이터를 사용하여 더 높은 점수를 받을 것으로 예상됩니다. 예를 들어, TILT는 더 많은 QA 데이터 세트에 대한 지도 교육 단계를 하나 더 도입합니다 [40]. StructuralLM은 교육[28]에서 검증 세트를 추가로 사용합니다.

모델	매개변수 양식 이미지 임베딩			FUNSD 코드 RVL-CDIP DocVQA			
				F1 ↑	F1 ↑ 정확도 ↑	ANLS ↑	
버트베이스 [9]	1억 1천만	티	없음	60.26	89.68	89.81	63.72
로베르타베이스 [36]	1억 2500만	티	없음	66.48	93.54	90.06	66.42
브로스베이스 [17]	1억 1천만	타+엘	없음	83.05	95.73	-	-
LiLTBASE [50]	-	타+엘	없음	88.41	96.07	95.68*	-
레이아웃LMBASE [54]	160M	T+L+I (R) ResNet-101(미세 조정)		79.27	-	94.42	-
셀프독 [31]	-	T+L+I (R) ResNeXt-101		83.36	-	92.81	-
유독 [14]	272M	T+L+I(R) ResNet-50		87.93	98.94 †	95.05	-
틸트베이스 [40]	230M	T+L+I (R) U-Net		-	95.11	95.25	83.92 ‡
XY레이아웃LMBASE [15]	-	T+L+I (G) ResNeXt-101		83.35	-	-	-
레이아웃LMv2BASE [56]	200M	T+L+I(G) ResNeXt101-FPN		82.76	94.95	95.25	78.08
DocFormerBASE [2]	1억 8300만	T+L+I(G) ResNet-50		83.34	96.33	96.17	-
LayoutLMv3BASE (당사) 133M		T+L+I(P) 선형		90.29	96.56	95.44	78.76
버들라지 [9]	340M	티	없음	65.63	90.25	89.92	67.45
로버트타라지 [36]	3억5500만	티	없음	70.72	93.80	90.11	69.52
레이아웃LMLARGE [54]	343M	타+엘	없음	77.89	-	91.90	-
브로스라지 [17]	340M	타+엘	없음	84.52	97.40	-	-
구조적LMLARGE [28]	3억5500만	타+엘	없음	85.14	-	96.08	83.94 ‡
폼넷 [25]	217M	타+엘	없음	84.69	-	-	-
폼넷 [25]	345M	타+엘	없음	-	97.28	-	-
틸트라지 [40]	780M	T+L+I (R) U-Net		-	96.33	95.52	87.05 ‡
레이아웃LMv2LARGE [56]	426M	T+L+I(G) ResNeXt101-FPN		84.20	96.01	95.64	83.48
DocFormerLARGE [2]	536M	T+L+I(G) ResNet-50		84.55	96.99	95.50	-
레이아웃LMv3LARGE (당사) 368M		T+L+I(P) 선형		92.08	97.46	95.93	83.37

* LiLT는 RVL-CDIP 미세 조정에서 ResNeXt101-FPN 백본과 함께 이미지 기능을 사용합니다.

위치 [17]. LiLT는 사전 학습된 텍스트 모델 [50] 을 사용하여 다양한 언어를 미세 조정 합니다 . FormNet은 양식[25]에서 토큰 간의 공간적 관계를 활용합니다 . • [T+L+I (R)] Faster R-CNN 영역 기능이 있는 텍스트, 레이아웃 및 이미지 양식: 이 작업 라인 은 Faster R-CNN 모델 의 RoI 헤드에서 이미지 영역 기능을 추출합니다 [44]. 그 중 LayoutLM [54] 과 TILT [40]는 OCR 단어의 경계 상자를 영역 제안으로 사용 하고 해당 텍스트 임베딩에 영역 기능을 추가합니다.

SelfDoc [31] 및 UDoc [14]는 문서 개체 제안을 사용 하고 영역 기능을 텍스트 임베딩 과 연결합니다 . • [T+L+I (G)] CNN 그리드 기능이 있는 텍스트, 레이아웃 및 이미지 형 식: LayoutLMv2 [56] 및 DocFormer [2]는 개체 감지 없이 CNN 백본으로 이미지 그리 드 기능을 추출합니다 . XYLayoutLM [15] 은 Lay outLMv2의 아키텍처를 채택 하고 레이아웃 표현을 개선합니다 . • [T+L+I (P)] 선형 패치 기능이 있는 텍스트, 레이아웃 및 이미지 형식: LayoutLMv3은 이미지 패치를 인코딩하기 위해 간단한 선형 임베딩으로 CNN 백본을 대체합니다 .

공개적으로 사용할 수 있는 다중 모달 작업에서 LayoutLMv3를 미세 조정합니다. 가능한 벤치마크. 결과는 표 1에 나와 있습니다.

작업 I: 양식 및 영수증 이해. 양식 및 영수증 이해 작업에는 양식 및 영 수증의 텍스트 콘텐츠를 추출하고 구조화해야 합니다. 작업은 레이블 로 각 단어에 태그를 지정하는 것을 목표로 하는 시퀀스 레이블 지정 문제입니다. 양식 이해 작업과 영수증 이해 작업을 위해 각각 선형 레이 어와 MLP 분류기를 사용하여 각 텍스트 토큰의 마지막 숨겨진 상태 레이블을 예측합니다 .

FUNSD 데이터 세트와 CORD 데이터 세트에 대한 실험을 수행합니다. FUNSD [20]는 RVL -CDIP 데이터 세트[16]에서 샘플링된 데이터 세트를 이해하는 노이즈 스캔 형식입니다 . FUNSD 데이터 세트에는 9,707개의 의미론적 엔터티에 대한 포괄적인 주석이 포함된 199개의 문서가 포 함되어 있습니다 . "질문", "답변", "헤더" 또는 "기타" 중에서 각 의미 개체에 레이블을 할당하기 위해 FUNSD 데이터 세트의 의미 개체 레이블 지정 작업에 중점을 둡니다 . 교육 및 테스트 분할에 는 각각 149개 및 50개의 샘플이 포함됩니다. CORD [39] 는 4개의 카테고리로 정의된 30개의 의 미 레이블이 있는 영수증 키 정보 추출 데이터 세트 입니다 . 여기에는 800개의 교육, 100개의 검 중 및 100개의 테스트 예제에 대한 1,000개의 영수증이 포함되어 있습니다. 공식적으로 제공되는 이미지와 OCR 주석을 사용합니다 . 우리는 1,000에 대해 LayoutLMv3를 미세 조정합니다.

표 2: PubLayNet 검증 세트의 문서 레이아웃 분석 mAP @ IOU[0.50:0.95]. 모든 모델은 비전 양식의 정보만 사용합니다. LayoutLMv3은 비교된 ResNets[14, 59] 및 비전 Transformer[30] 백본보다 성능이 뛰어납니다.

모델	프레임워크 백본 텍스트 제목 목록 표 그림 전체								
퍼블라넷[59]	Mask R-CNN	ResNet-101	91.6	Mask R-CNN	84.0	88.6	96.0	94.9	91.0
디테베이스 [30]	Transformer	93.4	Faster R-CNN	ResNet-50	87.1	92.9	97.3	96.7	93.5
유독 [14]			93.9		88.5	93.7	97.3	96.4	93.9
디테베이스 [30]	캐스케이드 R-CNN	트랜스포머	94.4		88.9	94.8	97.6	96.9	94.5
LayoutLMv3BASE (당사 제품) Cascade R-CNN Transformer								97.0	95.1

FUNSD의 경우 학습 속도가 1 -50이고 배치 크기가 16 이고 CORD의 경우 5 - 5 및 64인 단계입니다.

이 작업에 대한 F1 점수를 보고합니다. 큰 모델 크기의 경우 LayoutLMv3는 FUNSD 데이터 세트에서 92.08의 F1 점수를 달성했으며 이는 StructuralLM[28]에서 제공한 SOTA 결과 85.14를 훨씬 능가합니다 . LayoutLMv3 및 StructuralLM은 세그먼트 수준 레이아웃 위치를 사용하는 반면 다른 작업은 단어 수준 레이아웃 위치를 사용합니다. 세그먼트 수준 위치를 사용하면 FUNSD [28] 의 의미론적 엔터티 라벨링 작업에 도움이 될 수 있으므로 두 가지 유형 의 작업을 직접 비교할 수 없습니다. LayoutLMv3은 또한 기본 및 대형 모델 크기 모두에 대해 CORD 데이터 세트에서 SOTA F1 점수를 달성합니다 . 결과는 LayoutLMv3가 텍스트 중심 양식 및 영수증 이해 작업에 상당한 이점을 줄 수 있음을 보여줍니다.

작업 II: 문서 이미지 분류. 문서 이미지 분류 작업은 문서 이미지의 범주를 예측하는 것을 목 표로 합니다.

특수 분류 토큰 ([CLS])의 출력 숨겨진 상태를 MLP 분류기로 공급하여 클래스 레이블을 예측합니다.

RVL-CDIP 데이터 세트 에 대한 실험을 수행합니다 . 이것은 16개 범주로 레이 블이 지정된 IIT-CDIP 컬렉션의 하위 집합입니다 [16]. RVL-CDIP 데이터 세트에 는 400,000개의 문서 이미지가 포함되어 있으며 그 중 320,000개는 훈련 이미 지, 40,000개는 검증 이미지, 40,000개는 테스트 이미지입니다. Microsoft Read API를 사용하여 텍스트 및 레이아웃 정보를 추출합니다 . 우리는 배치 크기 가 64이고 학습률이 2-5인 20,000단계에 대해 LayoutLMv3를 미세 조정합니 다.

평가 메트릭은 전체 분류 정확도입니다. Lay outLMv3는 이전 작업보다 훨씬 작은 모델 크기로 더 좋거나 비슷한 결과를 얻습니다. 예를 들어, Lay outLMv2 와 비교하여 LayoutLMv3는 기본 모델과 큰 모델 크기에서 각각 0.19%와 0.29%의 절대적인 개선을 달성했으며 훨씬 간단한 이미지 임베딩(예: 선형 대 ResNeXt101-FPN)을 사용합니다.

결과는 간단한 이미지 임베딩이 이미지 중심 작업에서 바람직한 결과를 얻을 수 있음을 보여 줍니다.

작업 III: 시각적 질문 답변을 문서화합니다. 문서 시각적 질문 답변은 문서 이미지와 질문을 입력으로 받아 답변을 출력하는 모델이 필요합니다 [38]. 모델이 이진 분류기로 각 텍스트 토큰의 마지막 숨겨진 상태를 분류하여 시작 및 끝 위치를 예측하는 추출 QA 문제로 이 작 업을 공식화 합니다 .

우리는 문서 이미지에 대한 시각적 질문 답변을 위한 표준 데이터 세트인 DocVQA 데이터 세트에 대한 실험을 수행합니다 [38]. DocVQA 데이터 세트의 공식 파티션은 훈련/검증/테스트 세트에 대해 각각 10,194/1,286/1,287개의 이 이미지와 39,463/5,349/5,188개의 질문으로 구성됩니다. 훈련 세트에서 모델을 훈 련시키고, 테스트 세트에서 모델을 평가하고, 공식 평가 웹사이트에 제출하여 결과 를 보고합니다. Microsoft 읽기 API를 사용합니다.

LayoutLMv2에서와 같이 이미지에서 텍스트 및 경계 상자를 추출하고 휴리스틱 을 사용하여 추출된 텍스트에서 주어진 답변을 찾습니다.

배치 크기 128, 학습률 3-5 , 워밍업 비율 0.048 로 100,000단계에 대해 LayoutLMv3BASE를 미세 조정합니다 .

LayoutLMv3LARGE 의 경우 단계 크기, 배치 크기, 학습률 및 워밍업 비율은 각각 200,000, 32, 1-5 및 0.1입니다.

일반적으로 사용되는 편집 거리 기반 메트릭 ANLS (Average Normalized Levenshtein Similarity라고도 함)를 보고합니다. LayoutLMv3BASE 는 LayoutLMv2BASE 의 ANLS 점수를 78.08에서 78.76으로 개선하고 훨씬 간단한 이미 지 임베딩(즉, ResNeXt101-FPN에서 선형 임베딩으로)을 제공합니다. LayoutLMv3LARGE 는 LayoutLMv3BASE 보다 절대 ANLS 점수 4.61을 더 얻습니다 .

결과는 LayoutLMv3가 문서 시각적 질문 응답 작업에 효과적인을 보여줍니다.

3.4 비전 작업에 대한 미세 조정 다중 모드 도메인에서 시

각적 도메인으로의 LayoutLMv3 일반성을 입증하기 위해 LayoutLMv3를 문 서 레이아웃 분석 작업 으로 전환합니다 . 이 작업은 경계 상자와 표, 그림, 텍스 트 등과 같은 범주를 제공하여 구조화되지 않은 디지털 문서의 레이아웃을 감지하 는 것입니다. 이 작업은 다운스트림 애플리케이션을 위해 문서를 기계가 읽을 수 있 는 형식으로 구문 분석하는 데 도움이 됩니다 . 우리는 이 작업을 텍스트 임베딩이 없는 객체 감지 문제로 모델링하며 이는 기존 작업에서 효과적입니다 [14, 30, 59].

우리는 Detec tron2[53]를 사용하여 구현된 FPN [34] 과 함께 Cascade R-CNN 검출 기 [4] 의 기능 백본으로 LayoutLMv3를 통합합니다 . 우리는 LayoutLMv3 기본 모델의 레이어 4, 6, 8 및 12와 같은 다양한 Transformer 레이어에서 단일 스케일 기능을 추출 하는 표준 방식을 채택합니다 . 해상도 수정 모듈을 사용하여 단일 스케일 기능을 다중 스 케일 FPN 기능으로 변환합니다[1, 30, 33].

우리는 PubLayNet 데이터 세트 [59]에 대한 실험을 수행합니다. 데이터 세트 에는 텍스트, 제목, 목록, 그림 및 테이블의 5개 문서 레이아웃 범주에 걸쳐 경계 상자 및 다각형 분할 로 주석이 달린 연구 논문 이미지가 포함되어 있습니다. 공식 분할에는 335,703개의 훈련 이미지, 11,245개의 검증 이미지 및 11,405개의 테 스트 이미지가 포함됩니다.

훈련 분할에서 모델을 훈련하고 표준 관행 [14, 30, 59] 에 따라 검증 분할에서 모델을 평 가합니다. 우리는 1,000개의 워밍업 단계와 DiT [30] 에 이어 0.05의 가중치 감쇠가 있는 AdamW 옵티마이저를 사용하여 60,000단계에 대해 모델을 훈련합니다 .

LayoutLMv3는 비전 및 언어 양식의 입력으로 사전 훈련되었으므로 경험적으로 더 큰 배치 크기 32와 더 낮은 학습률 2 - 4를 사용합니다 . 일관성을 유지하기 위 해 미세 조정 단계에서 뒤집기 또는 크롬 핑 확대 전략을 사용하지 않습니다.

표 3: 일반적인 텍스트 중심 작업(FUNSD 및 CORD에 대한 양식 및 영수증 이해) 및 이미지 중심 작업(RVL-CDIP의 문서 이미지 분류 및 PubLayNet의 문서 레이아웃 분석)에 대한 이미지 임베딩 및 사전 훈련 목표에 대한 절제 연구. 모든 모델은 학습률 3 - 4로 150,000단계 동안 100만 개의 데이터에 대해 BASE 크기 로 훈련되었습니다.

#	영상 포함시키다	매개변수	사전 훈련 목표	FUNSD 코드 RVL-CDIP PubLayNet			
				F1 ↑	F1 ↑	정확도 ↑	지도 ↑
1 없음		1억 2500만 MLM		88.64	96.27	95.33	해당 없음 손실 발산
2 선형		1억2600만 MLM		89.39	96.11	95.00	94.38
3 선형		1억 3200만 MLM+MIM		89.19	96.30	95.42	
4 선형		133M MLM+MIM+WPA 89.78			96.49	95.53	94.43

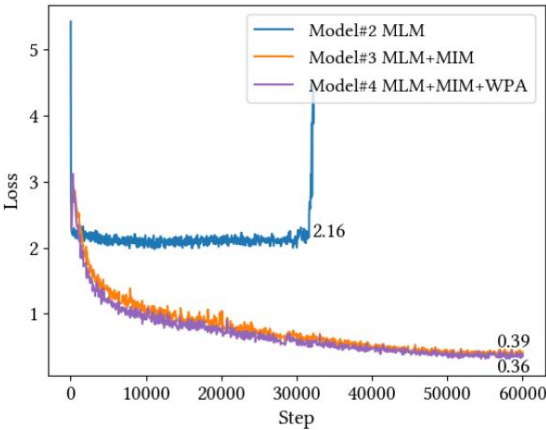


그림 4: PubLayNet 데이터 세트에서 LayoutMv3의 제거된 모델을 미세 조정하는 손실 수렴 곡선. 모델 #2의 손실은 수렴되지 않았습니다. MIM 목표를 통합하면 손실이 정상적으로 수렴됩니다. WPA 대물렌즈는 손실을 더욱 감소시킵니다. 색상으로 가장 잘 보입니다.

우리의 사전 훈련 단계와 함께. 우리는 자기 주의 네트워크에서 상대적인 위치를 DiT로 사용하지 않습니다.

경계 상자 의 평균 평균 정밀도 (MAP) @ IOU(intersection over union) [0.50:0.95]를 사용하여 성능을 측정하고 결과를 표 2에 보고합니다. ResNets [14, 59] 및 동시 비전 과 비교합니다. 트랜스포머 [30] 백본 . LayoutMv3는 모든 메트릭에서 다른 모델을 능가하여 전체 mAP 점수 95.1을 달성했습니다. LayoutMv3는 "제목" 범주에서 높은 이득을 얻습니다. 제목은 일반적으로 다른 범주보다 훨씬 작고 텍스트 콘텐츠로 식별할 수 있기 때문에 이러한 개선은 LayoutMv3 사전 교육에 언어 양식을 통합한 덕분입니다 . 이러한 결과는 LayoutMv3의 일반성과 우수성을 보여줍니다.

3.5 절제 연구 표 3에서 우리는 이

이미지 임베딩과 사전 훈련 목표 의 효과를 연구합니다 . 먼저 MLM 목표로 사전 훈련된 텍스트 및 레이아웃 정보를 사용하는 기본 모델 #1을 구축합니다 . 그런 다음 모델 #2로 표시된 기본 모델 의 이미지 임베딩으로 선형 투영 이미지 패치를 사용합니다 . 모델을 추가로 사전 훈련합니다.

MIM 및 WPA 대물렌즈가 있는 #2는 단계별로 새 모델을 #3 및 #4로 각각 나타냅니다.

그림 4에서는 배치 크기가 16이고 학습률 이 2-4 인 PubLayNet 데이터 세트에서 미세 조정했을 때 모델 #2, #3 및 #4의 손실을 시각화합니다. 모델 #2를 교육하려고 했습니다. {1 4, 2 4, 4 4}의 학습 속도와 {16, 32}의 배치 크기가 결합되었지만 모델 #2의 손실은 수렴되지 않았고 PubLayNet의 mAP 점수는 0에 가깝습니다.

선형 이미지 임베딩의 효과. 이미지 임베딩이 없는 모델 #1이 일부 작업에서 좋은 결과를 달성한 것을 관찰했습니다 . 이는 텍스트 및 레이아웃 정보를 포함한 언어 양식이 문서 이해에 중요한 역할을 한다는 것을 시사합니다. 그러나 결과는 여전히 만족스럽지 않습니다. 또한 모델 #1은 비전 양식 없이 일부 이미지 중심 문서 분석 작업을 수행할 수 없습니다. 예를 들어 경계 상자가 이미지와 긴밀하게 통합되어 있기 때문에 PubLayNet의 문서 레이아웃 분석 작업 에 비전 형식이 중요합니다 . 선형 이미지 임베딩 의 간단한 설계 와 적절한 사전 훈련 목표가 결합되어 이미지 중심 작업 뿐만 아니라 일부 텍스트 중심 작업도 지속적으로 개선할 수 있습니다.

MIM 사전 훈련 목표의 효과. 선형 귀 이미지 임베딩을 모델 #2에 대한 입력으로 텍스트 임베딩과 단순히 연결하면 CORD 및 RVL-CDIP의 성능이 저하되는 반면 PubLayNet의 손실은 분기됩니다. 우리는 모델이 이미지 양식과 관련된 사전 훈련 목표 없이 선형 패치 임베딩에서 의미 있는 시각적 표현을 학습하지 못했다고 추측합니다 .

MIM 목표는 입력 이미지 패치의 일부를 무작위로 마스킹하고 출력 에서 재구성 하여 모델의 마지막 레이어까지 이미지 정보를 보존함으로써 이 문제를 완화합니다 [22]. 모델 #3과 모델 #2의 결과를 비교하면 MIM 목표는 CORD 및 RVL-CDIP에 이점이 있습니다. 단순히 선형 이미지 임베딩을 사용하면 FUNSD가 개선되었으므로 MIM은 FUNSD에 더 이상 기여하지 않습니다. 교육 에 MIM 목표를 통합함으로써 그림 4 와 같이 PubLayNet을 미세 조정할 때 손실이 수렴되고 바람직한 mAP 점수를 얻습니다. 결과는 MIM이 교육을 정규화하는 데 도움이 될 수 있음을 나타냅니다. 따라서 MIM은 PubLayNet의 문서 레이아웃 분석과 같은 비전 작업에 매우 중요합니다.

WPA 사전 훈련 목표의 효과. 표 3의 모델 #3과 #4를 비교하면 WPA 목표가 모든 작업을 지속적으로 개선한다는 것을 알 수 있습니다. 또한 WPA 목표는 그림 4의 PubLayNet에서 비전 작업의 손실을 줄입니다. 이러한 결과는 교차 모달 표현 학습뿐만 아니라 이미지 표현 학습에서도 WPA의 효과를 확인합니다.

매개변수 비교. 이 표는 16×16 패치 프로젝션(#1 $\rightarrow$ #2)에 대한 이미지 임베딩 통합이 0.6M 매개변수만 도입함을 보여줍니다. 매개변수는 CNN 백본의 매개변수(예: ResNet-101의 경우 44M)에 비해 무시할 수 있습니다. MIM 헤드와 WPA 헤드는 사전 훈련 단계에서 6.9M 및 0.6M 매개변수를 도입합니다. 이미지 임베딩에 의해 도입된 매개변수 오버헤드는 50,265의 텍스트 어휘 크기에 대해 39.2M 매개변수가 있는 MLM 헤드에 비해 미미합니다. 토큰이 저는 MIM의 레이블을 생성하기 위한 독립 실행형 모듈이지만 Transformer 백본에 통합되지 않았기 때문에 매개변수를 계산할 때 이미지 토큰라이저를 계산하지 않았습니다.

4 관련 작업

Multimodal self-supervised pre-training 기술은 문서 레이아웃 및 이미지 표현 학습의 성공적인 적용으로 인해 문서 인텔리전스에서 급속한 발전을 이루었습니다 [2, 13–15, 17, 25, 28, 31, 32, 40, 41, 50, 52, 54–56]. LayoutLM 및 후속 작업은 텍스트의 공간 좌표를 인코딩하여 레이아웃 표현 학습을 결합합니다 [17, 25, 28, 54]. 그런 다음 다양한 작업에서 CNN과 Transformer [49] 셀프 어텐션 네트워크를 결합하여 이미지 표현 학습을 결합합니다. 이러한 작업은 CNN 그리드 기능[2, 56]을 추출하거나 영역 기능을 추출하기 위해 개체 감지기에 의존합니다 [14, 31, 40, 54]. 이는 무거운 계산 병목 현상을 설명하거나 영역 감도가 필요합니다. 자연 이미지 시각 및 언어 사전 훈련(VLP) 분야에서 연구 작업은 미리 정의된 객체 클래스의 한계를 없애기 위해 영역 기능 [5, 47, 48]에서 그리드 기능 [19]으로 이동하는 것을 보았습니다. 지역 감도. ViT(Vision Transformer) [11]에 영감을 받아 CNN의 약점을 극복하기 위해 CNN이 없는 VLP에서도 최근의 노력이 있었습니다.

여전히 대부분은 시각적 기능을 학습하기 위해 별도의 셀프 어텐션 네트워크에 의존합니다. 따라서 계산 비용이 감소하지 않습니다 [12, 29, 57]. 가벼운 선형 레이어로 시각적 기능을 학습하고 모델 크기와 실행 시간을 크게 줄이는 ViLT는 예외입니다 [22]. ViLT에서 영감을 받은 LayoutLMv3는 CNN 없이 이미지 임베딩을 활용하는 Document AI의 첫 번째 멀티모달 모델입니다.

재구성 사전 훈련 목표는 표현 학습에 혁명을 일으켰습니다. NLP 연구에서 BERT는 먼저 양방향 표현을 학습하기 위해 "마스크된 언어 모델링"(MLM)을 제안하고 광범위한 언어 이해 작업에 대한 최신 기술을 발전시켰습니다 [9]. CV 분야에서 MIM(Masked Image Modeling)은 가시적인 맥락에서 마스크된 콘텐츠 컨디셔닝을 예측하여 풍부한 시각적 표현을 학습하는 것을 목표로 합니다. 예를 들어, ViT는 마스크된 패치의 평균 색상을 재구성하여 ImageNet 분류에서 성능 향상을 가져옵니다 [11]. BEiT는 개별 VAE에서 학습한 시각적 토큰을 재구성하여 이미지 분류 및 의미 분할에서 경쟁력 있는 결과를 얻습니다 [3]. DiT는 BEiT를 문서 이미지로 확장하여 문서 레이아웃 분석으로 확장합니다[30].

MLM 및 MIM에서 영감을 받아 시각 및 언어 분야의 연구자들은 다중 모달 표현 학습을 위한 재구성 목표를 탐구했습니다. 대부분의 잘 수행되는 시각 및 언어 사전 훈련(VLP) 모델은 BERT가 제안한 텍스트 모달리티에 대한 MLM을 사용하지만 이미지 모달리티에 대한 사전 훈련 목표는 다릅니다. 다양한 이미지 임베딩에 대응하는 MIM에는 MRM(마스크된 영역 모델링), MGM(마스크된 그리드 모델링) 및 마스크된 패치 모델링의 세 가지 변형이 있습니다.

(MPM). MRM은 원래 영역 특징을 회귀시키거나 [5, 31, 48] 마스크된 영역에 대한 개체 레이블을 분류하는 데 효과적인 것으로 입증되었습니다. MGM은 또한 가려진 그리드 기능에 대한 시각적 사전의 매핑 인덱스를 예측하는 것이 목적인 SOHO에서도 탐색되었습니다 [19]. 패치 수준 이미지 임베딩을 위해 Visual Parsing [57]은 간단한 선형 이미지 인코더에는 적용되지 않는 self-attention 이미지 인코더의 주의 가중치에 따라 시각적 토큰을 마스크하도록 제안했습니다. ViLT [22] 및 METER [12]는 ViT [11] 및 BEiT [3]과 유사한 MPM을 활용하려고 시도합니다. ViT [11] 및 BEiT[3]는 이미지 패치에 대한 시각적 어휘의 평균 색상 및 이산 토큰을 각각 재구성하지만 다운스트림에서 성능이 저하되었습니다. 작업. 우리의 LayoutLMv3는 먼저 선형 패치 이미지 임베딩을 위한 MIM의 효과를 보여줍니다.

다중 모드 모델에서 시각 및 언어(VL) 정렬 학습을 위해 다양한 교차 모드 목표가 추가로 개발됩니다.

이미지-텍스트 일치는 거친 VL 정렬을 학습하는 데 널리 사용됩니다 [2, 5, 19, 22, 56]. 세분화된 VL 정렬을 학습하기 위해 UNITER는 문맥화된 이미지 임베딩을 단어 임베딩으로 전송하는 최소 비용을 계산하는 최적 전송에 기반한 단어 영역 정렬 목표를 제안합니다 [5].

ViLT는 이 목표를 패치 수준 이미지 임베딩으로 확장합니다 [22]. 자연 이미지와 달리 문서 이미지는 텍스트 단어와 이미지 영역 간의 명시적인 미세한 정렬 관계를 의미합니다. 이 관계를 사용하여 UDoc은 Contrastive learning과 유사성 증류를 사용하여 동일한 영역에 속하는 이미지와 텍스트를 정렬합니다 [14]. LayoutLMv2는 원시 이미지의 일부 텍스트 라인을 커버하고 각 텍스트 토큰이 커버되는지 여부를 예측합니다 [56]. 대조적으로 우리는 MIM의 마스크 작업을 자연스럽게 활용하여 효과적이고 통합된 방식으로 정렬/비정렬 쌍을 구성합니다.

5 결론 및 향후 작업

이 백서에서는 LayoutLM에 대한 모델 아키텍처 및 사전 교육 목표를 재설계하는 문서 AI용 다중 모드 변환기를 사전 교육하기 위한 LayoutLMv3를 제시합니다. Document AI의 기존 멀티모달 모델과 달리 LayoutLMv3는 시각적 특징을 추출하기 위해 사전 훈련된 CNN 또는 Faster R-CNN 백본에 의존하지 않으므로 매개변수를 크게 절약하고 영역 주석을 제거합니다. 우리는 통합된 텍스트 및 이미지 마스크 사전 훈련 목표인 마스크된 언어 모델링, 마스크된 이미지 모델링 및 단어 패치 정렬을 사용하여 다중 모달 표현을 학습합니다. 광범위한 실험 결과는 간단한 아키텍처와 통합된 목표로 텍스트 중심 및 이미지 중심 Document AI 작업 모두에 대해 LayoutLMv3의 일반성과 우수성을 보여주었습니다. 향후 연구에서는 모델이 더 많은 훈련 데이터를 활용하여 SOTA 결과를 더욱 향상시킬 수 있도록 사전 훈련된 모델을 확장하는 방법을 조사할 것입니다. 또한 Document AI 산업에서 더 많은 실제 비즈니스 시나리오를 촉진하기 위해 몇 가지 샷 및 제로 샷 학습 기능을 탐색할 것입니다.

6 승인

유익한 토론과 영감을 주신 Yiheng Xu에게 감사드립니다. 이 작업은 보조금 NO.2016ZT06D2110에 따라 NSFC(U1811461) 및 광둥 소개 혁신 및 기업가 팀을 위한 프로그램의 지원을 받았습니다.

표 4: EPHOIE 테스트 세트에서 중국 F1 점수의 시각적 정보 추출.

모델	과목 시험시간 이름 학교 #시험 #좌석만 #학생 성적 점수 평균											
BiLSTM+CRF [24]	98.51	100.0	98.87	98.80	75.86	72.73	94.04	84.44	98.18	69.57	89.10	
GCN 기반 [35]	98.18	100.0	99.52	100.0	88.17	86.00	97.39	80.00	84.44	94.44	81.82	92.55
그래프IE [42]	94.00	100.0	95.84	97.06	82.19	93.07	85.33			94.44	76.19	90.26
트리에 [58]	98.79	100.0	99.46	99.64	88.64	85.92	97.94	84.32	88.73	97.02	80.39	93.21
비에스 [51]	99.39	100.0	99.67	99.28	91.81	99.29	89.47			98.35	86.27	95.23
StrucTexT [32]	99.25	100.0	99.47	99.83	97.98	95.43	98.29	97.33		99.25	93.73	97.95
LayoutLMv3-ChineseBASE (당사) 98.99		100.0	99.77	99.20	100.0	100.0	98.82	99.78		98.31	97.27	99.21

부록

A.1 중국어 사전 훈련의 LayoutLMv3 중국어어

LayoutLMv3. 영어뿐만 아니라 중국어에서도 LayoutLMv3의 효과를 입증하기 위해 LayoutLMv3-중국어 모델을 기본 크기로 사전 훈련합니다.

중국어로 된 5천만 문서 페이지에서 학습됩니다. 우리는 공개적으로 사용 가능한 디지털 태생 문서를 다운로드하고 이러한 문서를 처리하기 위해 Common Crawl(<https://commoncrawl.org/>) 의 원칙에 따라 대규모 중국어 문서를 수집합니다. 텍스트 임베딩 레이어와 함께 멀티모달 트랜스포머 인코더 의 경우 LayoutLMv3-Chinese는 XLM-R [7]의 미리 훈련된 가중치에서 초기화됩니다 . 나머지 모델 매개 변수를 무작위로 초기화했습니다. 기타 학습 설정은 LayoutLMv3와 동일합니다.

시각적 정보 추출에 대한 미세 조정. 시각적 정보 추출(VIE)은 문서 이미지에서 핵심 정보를 추출해야 합니다. 작업은 미리 정의된 레이블로 각 단어에 태그를 지정하는 것을 목표로 하는 시퀀스 레이블 지정 문제입니다. 선형 레이어를 사용하여 각 텍스트 토큰의 마지막 숨겨진 상태 레이블을 예측합니다 .

EPHOIE 데이터 세트에 대한 실험을 수행합니다. 에 포이에 [51] 다양한 레이아웃과 배경을 가진 시험지 헤드 로 구성된 시각 정보 추출 데이터셋입니다 . 여기에는 15,771개의 중국어 텍스트 인스턴스에 대한 포괄적인 주석이 있는 1,494개의 이미지가 포함되어 있습니다. EPHOIE 데이터 세트에 대한 토큰 수준 엔터티 레이블 지정 작업에 중점을 두어 미리 정의된 10개의 범주 중에서 각 문자에 레이블을 할당합니다. 훈련 및 테스트 세트에는 각각 1,183개 및 311개의 이미지가 포함되어 있습니다. 우리는 100세대에 대해 LayoutLMv3-중국어에 미세 조정합니다 . 배치 크기는 16이고 학습률은 5 - 5 이며 첫 번째 Epoch에 대한 선형 워밍업이 있습니다.

우리는 이 작업에 대한 F1 점수를 보고하고 표 4에 결과를 보고합니다. LayoutLMv3-Chinese는 대부분의 메트릭에서 우수한 성능을 보여주고 99.21%의 SOTA 평균 F1 점수를 달성합니다. 결과는 보여줍니다
LayoutLMv3은 중국어 VIE 작업에 상당한 이점을 제공합니다.

참조

[1] Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek 등. 2021. Xcit: 교차 공분산 이미지 변환기. NeurIPS에서.
[2] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie 및 R. Man matha. 2021. DocFormer: 문서 이해를 위한 종단 간 변환기. ICCV에서.
[3] Hangbo Bao, Li Dong, Songhao Piao 및 Furu Wei. 2022. BEiT: BERT 프리 이미지 변환기 교육. ICLR에서.
[4] Zhaowei Cai와 Nuno Vasconcelos. 2018. Cascade r-cnn: 고품질 객체 감지에 대한 탐구. CVPR에서.
[5] 첸 첸, 리 린지에, 리쥔 위, 아흐메드 엘 콜리, 파이살 아흐메드, 저 간, 유 쉹, 징징 류. 2020. Uiter: 범용 이미지-텍스트 표현 학습. ECCV에서.

[6] 조재민, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, Aniruddha Kembhavi. 2020. X-LXMERT: 멀티 모달 트랜스포머를 사용한 페인트, 캡션 및 질문 답변 . EMNLP에서.
[7] Alexis Conneau, Kartikay Khandelwal, 나만 고갈, Vishrav Chaudhary, Guil laume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer 및 Veselin Stoyanov. 2020. 규모에 따른 감독되지 않은 교차 언어 표현 학습 . ACL에서.
[8] Lei Cui, Yiheng Xu, Tengchao Lv 및 Furu Wei. 2021. 문서 IE: 벤치마크, 모델 및 애플리케이션. arXiv 프리프린트 arXiv:2111.08609 (2021).
[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: 언어 이해를 위한 심층 양방향 변환기 사전 훈련. NAACL 에서 .
[10] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang 등. 2021. Cogview: 변환기를 통한 텍스트-이미지 생성 마스터링. NeurIPS에서.
[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. 2021. 이미지는 16x16 단어의 가치가 있습니다: 대규모 이미지 인식을 위한 변환기. ICLR에서.
[12] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng 등. 2021. 엔드 투 엔드 비전 및 언어 변환기 교육에 대한 경험적 연구. arXiv 프리프린트 arXiv:2111.02387 (2021).
[13] ukasz Garncarek, Rafa Powalski, Tomasz Stanis awek, Bartosz Topolski, Piotr Halama, Micha Turski, Filip Graliński. 2021. LAMBERT: 정보 추출을 위한 레이아웃 인식 언어 모델링. ICDAR에서.
[14] Jiuxiang Gu, Jason Kuen, Vlad Morariu, Handong Zhao, Rajiv Jain, Nikolaos Barmpalios, Ani Nenkova 및 Tong Sun. 2021. UniDoc: 문서 이해를 위한 통합 사전 교육 프레임워크. NeurIPS에서.
[15] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, Liqing Zhang. 2022. XLayoutLM: 시각적으로 풍부한 문서 이해를 위한 레이아웃 인식 다중 모드 네트워크를 향하여. CVPR에서.
[16] Adam W Harley, Alex Ufkes, Konstantinos G Derpanis. 2015. 문서 이미지 분류 및 검색을 위한 Deep Convolutional Net의 평가 . ~ 안에 ICDAR.
[17] 홍탁규, 김동현, 자민기, 황원석, 남대현, 박성래 . 2022. BROS: 문서에서 더 나은 핵심 정보 추출을 위해 텍스트 및 레이아웃에 초점을 맞춘 사전 훈련된 언어 모델 . AAAI에서.
[18] Yupan Huang, Hongwei Xue, Bei Liu 및 Yutong Lu. 2021. 양방향 이미지 및 텍스트 생성을 위한 다중 모드 변환기 통합 . ACM 멀티미디어에서.
[19] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu 및 Jianlong Fu. 2021. 즉시 보기: 시각 언어 표현 학습을 위한 엔드 투 엔드 사전 교육. CVPR에서.

[20] Guillaume Jaume, Hazim Kemal Ekenel, Jean-Philippe Thiran. 2019. FUNSD: 잡음이 많은 스캔 문서에서 양식 이해를 위한 데이터 세트. ICDAR에서.
[21] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, Omer Levy. 2020. Spanbert: 범위를 표현하고 예측하여 사전 훈련 개선 . 전산 언어학 협회 거래 8(2020), 64–77.
[22] 김원재, 손보재, 김일두. 2021. Vilt: 컨볼루션 또는 지역 감독이 없는 시각 및 언어 변환기. ICML에서.
[23] 디데릭 P 캄파와 지미 바. 2014. Adam: 확률적 최적화 방법 . arXiv 프리프린트 arXiv:1412.6980 (2014).
[24] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, Chris Dyer. 2016. 명명된 엔터티 인식을 위한 신경 아키텍처. NAACL HLT 에서 .

[25] Chen-Yu Lee, Chun-Liang Li, Timothy Dozat, Vincent Perot, Guolong Su, Nan Hua, Joshua Ainslie, Renshen Wang, Yasuhisa Fujii, Tomas Pfister. 2022. FormNet: 양식 문서 정보 추출에서 순차 모델링을 넘어서는 구조적 인코딩. ACL에서.
[26] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman 및 J. Heard. 2006. 복잡한 문서 정보 처리를 위한 테스트 컬렉션 구축. ~ 안에 SIGIR.

[27] Mike Lewis, Yinhan Liu, 나만 고알, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov 및 Luke Zettlemoyer. 2020. BART: 자연어 생성, 번역 및 이해를 위한 사전 스 간 사전 훈련 노이즈 제거. ACL에서.

[28] Chenliang Li, Bin Bi, Ming Yan, Wei Wang, Songfang Huang, Fei Huang, Luo Si. 2021. StructrallLM: 형태 이해를 위한 구조 사전 훈련. ACL에서.

[29] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, Steven Chu Hong Hoi. 2021. Align before fuse: 운동량 증류를 통한 시각 및 언어 표현 학습. NeurIPS에서.

[30] Junlong Li, Yiheng Xu, Tengchao Lv, Lei Cui, Cha Zhang 및 Furu Wei. 2022. DiT: Document Image Transformer를 위한 자체 감독 사전 교육. arXiv 프리프린트 arXiv:2203.02378 (2022).

[31] Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, Hongfu Liu. 2021. SelfDoc: 자기 감독 문서 표현 학습. CVPR에서.

[32] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, Errui Ding. 2021. StrucTexT: 다중 모드 변환기를 사용한 구조화된 텍스트 이해. ACM 멀티미디어에서.

[33] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, Ross Girshick. 2021. 비전 변환기를 사용한 벤치마킹 감지 전이 학습. arXiv 프리프린트 arXiv:2111.11429 (2021).

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie. 2017. 물체 감지를 위한 기능 피라미드 네트워크. CVPR에서.

[35] Xiaojing Liu, Feiyu Gao, Qiong Zhang 및 Huasha Zhao. 2019. 시각적으로 풍부한 문서에서 다중 모드 정보 추출을 위한 그래프 컨볼 루션. NAACL HLT 에서 .

[36] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer 및 Veselin Stoyanov. 2019. Roberta: 강력하게 최적화된 bert 사전 훈련 접근 방식. arXiv 프리프린트 arXiv:1907.11692 (2019).

[37] Jiasen Lu, Dhruv Batra, Devi Parikh 및 Stefan Lee. 2019. Vilbert: 시각 및 언어 작업을 위한 작업에 구애받지 않는 시각 언어적 표현 사전 교육. ~ 안에 NeurIPS.

[38] Minesh Mathew, Dimosthenis Karatzas 및 CV Jawahar. 2021. Docvqa: 데이터 세트 문서 이미지의 vqa용. WACV에서.

[39] 박승현, 신승, 이바도, 이준엽, 서재홍, 서민준, 이할수. 2019. CORD: 사후 OCR 구문 분석을 위한 통합 영수증 데이터 세트. 신경 정보 처리 시스템의 문서 인터랙션 워크숍에서 .

[40] Rafal Pawalski, ukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michal Pietruszka, Gabriela Pa ka. 2021. 텍스트-이미지-레이아웃 변환기를 사용한 문서 이해에 대한 Full-TILT Boogie 진행 . ICDAR에서.

[41] Subhojeet Pramanik, Shashank Mujumdar, Hima Patel. 2020. 문서 표현 학습을 위한 다중 모드, 다중 작업 학습 기반 사전 훈련 프레임워크를 향하여 . arXiv 프리프린트 arXiv:2009.14457 (2020).

[42] 유지에 첸. 2019. 정보 추출을 위한 그래프 기반 프레임워크. 박사 논문. 매사추세츠 공과 대학.

[43] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever. 2021. 제로샷 텍스트 투 이미지 생성. ICML에서.

[44] Shaoqing Ren, Kaiming He, Ross B. Girshick, Jian Sun. 2015. 더 빠른 R-CNN: 지역 제안 네트워크를 통한 실시간 객체 탐지를 향하여. TPAMI 39, 1137-1149.

[45] Tim Salimans, Andrej Karpathy, Xi Chen, Diederik P. Kingma. 2017. Pixel CNN++: Discretized Logistic Mixture Likelihood 및 기타 수정을 통해 PixelCNN 개선 . ICLR에서.

[46] Rico Sennrich, Barry Haddow, Alexandra Birch. 2016. 뉴럴 머신 하위 단어 단위로 화자 단어의 번역. ACL에서.

[47] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei 및 Jifeng Dai. 2019. VL-BERT: 일반 시각적 언어 표현의 사전 훈련. ICLR에서.

[48] 하오 탄과 모하트 반살. 2019. LXMERT: Cross-Modality Encoder 학습 트랜스포머의 표현. EMNLP에서.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser 및 Illia Polosukhin. 2017. 주의 가 필요합니다. NeurIPS에서.

[50] Jiapeng Wang, Lianwen Jin, Kai Ding. 2022. LiLT: 구조화된 문서 이해를 위한 간단하면서도 효과적인 언어 독립적 레이아웃 변환기 . ACL에서.

[51] Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, Mingxiang Cai. 2021. 실세계에서 강력한 시각적 정보 추출을 향하여 : 새로운 데이터 세트 및 새로운 솔루션. AAAI 에서 .

[52] Te-Lin Wu, Cheng Li, Mingyang Zhang, Tao Chen, Spurthi Amba Hombaiah, Michael Bendersky. 2021. LAMPRET: 문서 이해를 위한 레이아웃 인식 다중 모드 사전 교육. arXiv 프리프린트 arXiv:2104.08405 (2021).

[53] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, Ross Girshick. 2019. 디텍트론2. <https://github.com/facebookresearch/detectron2>.

[54] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou. 2020. Layoutlm: 문서 이미지 이해를 위한 텍스트 및 레이아웃 사전 훈련. KDD 에서 .

[55] Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang 및 Furu Wei. 2021. LayoutXML: 시각적으로 풍부한 다국어 문서 이해를 위한 다중 모드 사전 교육 . arXiv 프리프린트 arXiv:2104.08836 (2021).

[56] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, Lidong Zhou. 2021. LayoutLMv2: 시각적으로 풍부한 문서 이해를 위한 다중 모드 사전 훈련 . ACL에서.

[57] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li 및 Jiebo Luo. 2021. 상호 양식 탐색: 시각 및 언어 사전 교육을 위한 Self-Attention을 사용한 시각적 구문 분석. NeurIPS에서.

[58] Peng Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Jing Lu, Liang Qiao, Yi Niu 및 Fei Wu. 2020. TRIE: 문서 이해를 위한 종단 간 텍스트 읽기 및 정보 추출 . ACM 멀티미디어에서.

[59] Xu Zhong, Jianbin Tang, Antonio Jimeno Yepes. 2019. PubLayNet: 최대 규모 문서 레이아웃 분석을 위한 데이터 세트. ICDAR에서.