시험성적서 위·변조 방지를 위한 문서 엔터티 구조화 - PureChain 통합 프레임워크

안제혁[1], 김동성[2], 이재민*

금오공과대학교 IT융복합공학과[1,2,*]

{wpgur2018[1], dskim[2], ljmpaul*}@kumoh.ac.kr

A Document Entity Structuring and Blockchain-Integrated Framework for Preventing Tampering in Test Certificates – PureChain

Je-Hyeok Ahn[1], Dong-Seong Kim[2], and Jae-Min Lee*

Kumoh National Institute of Technology

Dept. of IT Convergence Eng.[1,2,*]

요 약

이미지 기반 시험성적서의 수작업 검증은 대량 처리 시 오류와 지연을 야기하는 주요 원인이다. 본 논문에서는 문서 레이아웃 이해 모델인 LayoutLMv3를 한국어 시험성적서 데이터셋에 맞게 Fine-tuning하여, 인증서의 주요 엔터티를 추출하고 JSON 형태로 구조화하는 방법을 제안한다. 추출된 Entity 데이터의 무결성 보장을 위해 SHA-256 해시 값을 생성하고, 원본 이미지, Entity 시각화 이미지, 그리고 구조화된 JSON 파일을 하나의 폴더로 통합하여 IPFS에 업로드한다. 최종적으로 데이터가 저장된 IPFS(InterPlanetary File System)의 CID(Content Identifier)를 PureChain에 기록하여 데이터의 위·변조를 방지하는 통합 프레임워크를 구현했다. Fine-tuning 결과, 제안 모델은 Accuracy 약 99.8%, F1-score 0.988의 높은 정확도를 달성했으며, OCR-free 계열 모델인 Donut 대비 한국어 도메인에서 높은 적합성을 보였다. 본 시스템은 공급업체, 시험기관, 수요기관을 잇는 투명한 검증 플랫폼을 제공함으로써 조달 프로세스의 신뢰성과 추적 가능성을 크게 향상시킬 수 있다.

Ⅰ. 서론

2012년 발생한 고리 1호기 정전 및 원전 부품 위조 사건은 한국 원자력 산업의 구조적 부패와 안전관리 시스템의 취약성을 드러낸 중대한 계기였다. 당시 조사 결과, 2,114건의 시험성적서와 62건의 품질 인증 문서가 위조된 것으로 밝혀졌다. 이러한 문제는 기업 간 유착 구조, 독점적 조달 체계, 형식적 감사제도에서 비롯되었으며, 여전히 수천 건의 시험성적서가 PDF나 이미지 형태로 수동 검증되는 비효율적 체계로 남아있다[1]. 기존 관련 연구로, OCR(Optical Character Recognition)-free End-to-End 모델인 Donut은 이미지만으로 JSON 형태의 구조화된 출력을 생성할 수 있지만, 고해상도 문서 처리 비용이 높고 복잡한 테이블 구조 인식에 한계가 있다[2]. 또한, 문서 데이터를 블록체인 네트워크에 저장하여 불변성을 보장하려는 연구들도 활발히 이루어지고 있다. Pure Chain은 기존 블록체인 네트워크보다 향상된 합의 알고리즘과 마이닝 프로세스를 제공하는 BaaS(Blockchain-as-a-Service) 프레임워크이다[3].

본 논문에서는 이러한 한계를 극복하기 위해, 문서 레이아웃 이해 모델인 LayoutLMv3를 한국어 시험성적서에 맞게 미세 조정(Fine-tuning)하여 주요 Entity를 정확히 추출하고, 추출된 데이터를 IPFS와 PureChain에 연동하여 데이터의 무결성과 투명성을 동시에 확보하는 통합 검증 시스템을 제안한다. 이를 통해 수작업 검증의 비효율성을 개선하고 조달 프로세스의 신뢰도를 높이고자 한다[4].

Ⅱ. 딥러닝 기반 문서 구조화 기술 분석

최근 문서 이미지로부터 구조화된 정보를 추출하기 위한 연구는 OCR-Based Transformer와 OCR-Free Transformer으로 구분되며, 파이프라인 차이는 그림 1과 같다. LayoutLM, LayoutLMv2, BROS 등 기존 모델들은 이미지 내 텍스트를 탐지·인식한 후 토큰 단위로 정렬해 문맥적 의미를 해석한다. 이러한 구조는 시각적·언어적 정보를 모두 활용할 수 있

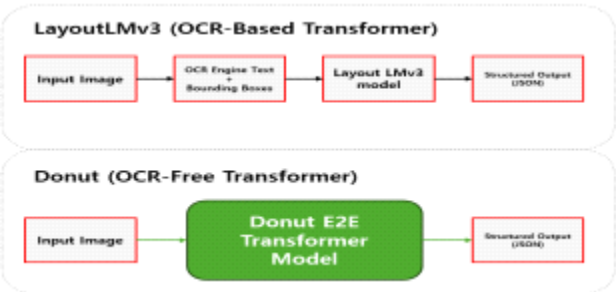


그림 1 OCR 기반 모델과 OCR-Free 모델의 파이프라인 비교

으나, OCR 오류 전파, 도메인 적응성 한계, 복잡한 파이프라인 문제를 가진다. 이를 해결하기 위해 제안된 Donut(Document Understanding Transformer)은 OCR 절차를 생략한 End-to-End 문서 이해 모델로, Swin Transformer와 BART 기반 구조를 통해 이미지로부터 직접 JSON 형태의 구조화 출력을 생성한다. 이를 통해 Donut은 OCR 오류를 제거하고 다양한 언어와 문서 도메인에 유연하게 대응할 수 있다. 본 논문에서는 한국어 시험성적서에 대한 Donut 모델의 적용성을 검증하기 위해 840장의 시험성적서 이미지를 전처리하여 Fine-tuning을 수행하였다. 10 epoch 동안 loss는 9.0에서 0.06까지 안정적으로 감소했으며, val_loss는 0.1233으로 수렴하였다. 그러나 테스트 결과 생성된 JSON의 주요 엔터티 7개 중 정답 라벨과 일치한 경우는 없었다. 이는 Donut의 사전학습 목표에 기반하여 복잡한 표 문서에서 읽기 순서가 뒤섞이며 정보 손실이 발생하기 때문이다. 또한, 디코더 출력 길이 제한 및 토크나이저 특성으로 숫자나 특수기호가 왜곡되는 현상도 확인되었다. 이러한 결과는 Donut이 비정형 문서의 전반적 레이아웃 이해에는 효과적이지만, 시험성적서와 같은 정형 문서의 필드 인식에는 한계가 있음을 보여준다. 따라서 본 연구에서는 정확한 위치 정보(bounding box)를 기반으로 엔터티를 예측할 수 있는 LayoutLMv3 모델을 채택하여 문서 구조화를 수행하였다.

Ⅲ. AI - Blockchain 기반 시험성적서 위·변조 방지 시스템

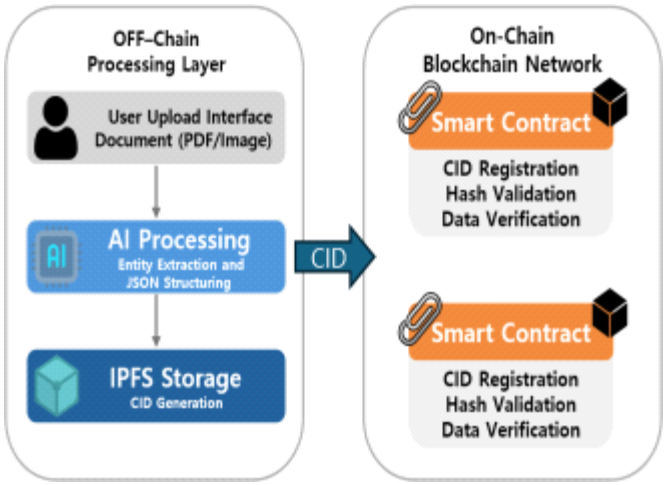


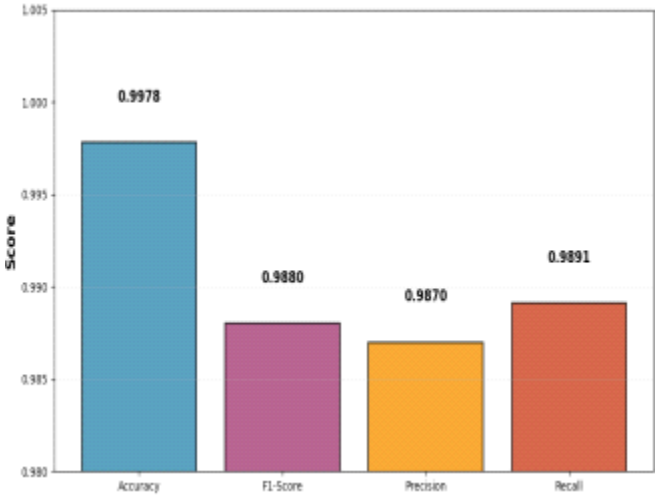
그림 2 AI - Blockchain 기반 시험 성적서 위·변조 방지 시스템

본 논문에서 제안하는 AI - Blockchain 기반 시험성적서 위·변조 방지 시스템은 그림 2와 같다.. 시스템은 Off-Chain Processing Layer와 On-Chain Blockchain Network의 두 부분으로 이루어진다. 먼저 Off-Chain Layer에서는 사용자가 시험성적서(PDF 또는 이미지)를 업로드하면, Fine-Tuned LayoutLMv3 모델이 문서에서 인증번호·업체명·제품명·날짜 등의 주요 엔터티 7개를 추출하고 이를 구조화된 JSON 형식으로 변환한다. 그 결과물은 IPFS 에 저장되어 CID(Content Identifier)가 생성된다. 이후 On-Chain Network에서는 스마트 컨트랙트를 통해 CID와 SHA-256 해시값이 PureChain에 기록된다. 이를 통해 데이터의 무결성과 추적성이 보장되며, 검증 노드는 CID를 기반으로 문서의 진위 여부를 검증할 수 있다. 시험성적서 이미지로부터 엔터티를 자동 추출하기 위해 LayoutLMv3 모델을 학습시키기 전에 데이터 전처리 과정을 수행하였다. 먼저, 원본 시험성적서 이미지를 수집한 후 PaddleOCR을 이용하여 각 문서 내 단어 단위 텍스트와 해당 영역의 바운딩 박스(bounding box)를 추출하였다. 이후 Label Studio를 통해 각 텍스트 영역에 대한 수동 라벨링을 수행하여 엔터티별 주석(annotation)을 생성하였다. 이때 OCR 결과와 주석 데이터를 IOU(Intersection over Union) 기준으로 정렬하여 NER(Named Entity Recognition) 학습에 필요한 라벨을 자동으로 매칭하였으며, 문서 내 논리 구조에 따라 엔터티 단위를 분리하는 Split Logic을 적용하였다. 최종적으로 이러한 과정을 통해 텍스트, 위치, 라벨 정보가 통합된 JSONL 형식의 학습 데이터를 생성하였으며, 모델 학습 형식에 맞게 Train data는 170개, Test data는 43개로 구성하여 학습을 진행하였다.

Ⅳ. 시험 및 평가

본 논문에서 학습은 AMD Ryzen 7 7700 (8-Core, 16 Threads, 3.8 GHz) 프로세서, 32 GB RAM, 및 NVIDIA GeForce RTX 4070 (12 GB VRAM, CUDA 12.6) 환경에서 수행되었다.

LayoutLMv3 Fine-tuned 로그 결과를 그림 3과 같이 나타내었다. 평가 결과는 Accuracy, Precision, Recall, F1-Score의 네 가지 주요 지표를 기준으로 측정하였다. 모델의 정확도(Accuracy) 는 0.9978로 매우 높은 수준을 보였으며, 이는 대부분의 문서에서 올바른 엔터티를 정확히 식별했음을 의미한다. F1-Score 는 0.9880, Precision 은 0.9870, Recall 은 0.9891로 나타나, Precision과 Recall 간의 균형 또한 우수하게 유지되었다. 이러한 결과는 LayoutLMv3가 문서 내 텍스트의 시각적 배치 정보와 언어적 문맥을 효과적으로 학습했음을 보여준다.또한, 해당 환경에서 모



델의 평가 실행시간(eval_runtime)은 1.7138 초, 초당 처리 샘플 수(eval_samples_per_second)는 25.09개로 측정되었으며, 이는 고성능 GPU 환경에서의 실시간 추론에도 충분히 적합한 효율성을 보였다.

Ⅴ. 결론 및 향후 연구

본 논문에서는 LayoutLMv3 - Blockchain 융합 기반 시험성적서 위·변조 방지 시스템을 구현하여 문서의 자동 구조화와 추적 가능성을 확보하였다. 또한 Donut 모델과 비교하였을 때, 약 630개 적은 데이터셋으로도 엔터티 추출의 정밀도와 효율성 측면에서 우수한 성능을 보였다. 또한, PureChain에 데이터를 연동하여, 위·변조를 방지하고, 낮은 TPS를 통해 기존 블록체인 네트워크가 가지는 확장성 및 효율성 문제를 해결하였다. 향후에는 주요 엔터티를 중심으로 LLM 및 엔지니어가 정보를 보다 효율적으로 처리할 수 있는 구조로 확장하고, 스마트컨트랙트를 통해 엔터티 기반 위·변조 여부를 자동 검증하는 시스템으로 발전시키고자 한다.

ACKNOWLEDGMENT

본 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 지역지능화혁신인재양성사업 (IITP-2025-RS-2020-II201612, 33%)과 2025년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업 (No. RS-2025-25431637, 33%)과 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업의 연구결과로 수행되었음 (IITP-2025-RS-2024-00438430, 34%).

참 고 문 헌

[1] P.-A Speed, "South Korea's nuclear power industry: recovering from scandal", Journal of World Energy Law & Business, vol. 13, no. 1, pp. 47-57, Mar. 2020.

[2] G.-W Kim, T.-G Hong, M.-B Yim, J.-Y Nam, J.-Y Park, J.-Y Yim, W.-S Hwang, S.-D Yun, D.-Y Han, and S.-H Park, "OCR-free Document Understanding Transformer", in Proc. European Conference on Computer Vision (ECCV), Tel Aviv, Israel, pp. 498-517, Oct. 2022.

[3] I.-S. Igboanusi, C.-A. Nnadiekwe, J.-U. Ogbede, D.-S. Kim, and A. Lensky, "BOMS: blockchain-enabled organ matching system", Scientific Reports, vol. 14, no. 15936, pp. 1-17, Jul. 2024.

[4] M.-R.-R. Ansori, Allwinnaldo, A.-R. Naufal, I.-S. Igboanusi, J.-M. Lee, and D.-S. Kim, "HADES: Hash-Based Audio Copy Detection System for Copyright Protection in Decentralized Music Sharing", IEEE Transactions on Network and Service Management, vol. 20, no. 3, pp. 2845-2853, Sep. 2023.