OCR 기반 표 이미지의 의미적 단어 임베딩을 이용한 분류 기법에 관한 연구

박선우, 임완수*

성균관대학교

sunp0909@skku.edu, *wansu.lim@skku.edu

A Study on Table Image Classification Method Using OCR-Based Semantic Word Embeddings

Park Sun Woo, Lim Wan Su*

Sungkyunkwan Univ.

요 약

본본 논문은 표 이미지에 포함된 텍스트 정보를 활용하여 표의 유형을 분류하는 OCR 기반 표 이미지 분류 기법을 제안한다. 기존의 표 이미지 분류 연구들은 주로 시각적 특징이나 레이아웃 구조를 활용하는 방식에 초점을 맞추어 왔으나, 이러한 접근은 표의 외형이 유사한 경우 의미적 차이를 충분히 반영하지 못하는 한계를 가진다. 특히 실제 문서 환경에서 표의 분류 기준은 시각적 형태보다 표에 포함된 텍스트의 의미적 내용에 의해 결정되는 경우가 많다. 이를 해결하기 위해 본 연구에서는 표 이미지로부터 OCR을 통해 텍스트를 추출한 후, 단어 단위로 의미 임베딩을 수행하여 표 전체를 하나의 문서 벡터로 표현하는 방식을 제안한다. 한국어와 영어가 혼재된 표 환경을 고려하여 다국어 단어 임베딩 모델을 활용하며, 평균 풀링을 통해 고정 길이의 문서 표현을 생성한다. 생성된 문서 벡터는 다층 퍼셉트론(MLP) 기반 분류기에 입력되어 표의 유형을 예측한다. 실험 결과, 제안한 방법은 복잡한 시각적 특징 추출 없이도 표의 의미적 정보를 효과적으로 반영하여 안정적인 분류 성능을 보임을 확인하였다.

Ⅰ. 서 론

표는 문서, 보고서, 논문, 재무 자료 등 다양한 형태의 문서에서 정보를 구조적으로 요약하고 전달하는 핵심적인 요소로 활용되고 있다. 최근 문서 자동 분석 및 지능형 문서 처리 기술에 대한 관심이 증가함에 따라, 표 이미지를 자동으로 분류하거나 이해하는 기술의 중요성 또한 커지고 있다. 이러한 표 분류 기술은 문서 검색, 정보 추출, 문서 요약 등의 전처리 단계로 활용될 수 있다.

기존의 표 이미지 분류 연구들은 주로 합성곱 신경망(CNN)이나 비전 트랜스포머(ViT)와 같은 시각적 딥러닝 모델을 활용하여 표의 레이아웃 구조나 시각적 패턴을 분석하는 방식에 초점을 맞추어 왔다. 이러한 방법들은 격자 구조나 셀 배치와 같은 외형적 특징을 학습하는 데에는 효과적이지만, 표에 포함된 텍스트의 의미적 내용을 직접적으로 반영하는 데에는 한계를 가진다. 특히 서로 다른 의미를 갖는 표들이 유사한 시각적 구조를 가지는 경우, 시각 기반 접근법만으로는 정확한 분류가 어렵다.

실제 문서 환경에서는 표에 포함된 단어, 수치, 키워드와 같은 텍스트 의미 정보가 표의 유형을 결정하는 핵심 요소로 작용하는 경우가 많다. 이에 본 논문에서는 표 이미지 분류 문제를 시각적 특징 추출 문제가 아닌, OCR을 기반으로 한 텍스트 의미 표현 문제로 재정의한다[1].

본 논문에서는 OCR을 통해 추출한 텍스트를 단어 단위로 임베딩하고, 평균 풀링을 통해 표 전체를 대표하는 문서 벡터를 생성한 후, 이를 경량 MLP 분류기로 분류하는 방법을 제안한다. 제안한 방법은 구조가 단순하고 계산 비용이 낮으며, 시각적 모델에 의존하지 않고도 표의 의미적 특성을 효과적으로 반영할 수 있다는 장점을 가진다.

Ⅱ. 본론

2.1 시스템 개요 및 파이프라인

본 연구는 표 이미지에 포함된 텍스트 의미 정보를 이용하여 표의 클래스를 분류하는 방법을 제안한다. 전체 파이프라인은 OCR 기반 텍스트 추출, 텍스트 전처리 및 토큰화, 단어 임베딩 생성, 문서 벡터 구성, 그리고 MLP 분류 단계로 구성된다. 본 파이프라인은 격자선이나 레이아웃과 같은 시각적 특징을 직접 학습하지 않고, 표 내부 텍스트의 의미적 단서를 중심으로 분류를 수행한다는 점에서 기존 이미지 기반 접근과 차별화된다.

2.2 OCR 기반 텍스트 추출 및 전처리

입력 표 이미지는 다국어 OCR 설정을 적용하여 텍스트로 변환된다[1]. OCR 결과에는 줄바꿈, 공백, 특수문자 및 인식 오류가 포함될 수 있으므로, 의미적 임베딩을 위해 정규화 과정을 수행한다. 구체적으로 특수문자를 제거하고 숫자, 영문, 한글 문자만 유지한 뒤 소문자 변환을 적용한다. 이후 공백을 기준으로 토큰화를 수행하여 단어 단위 토큰 집합을 생성하며, 임베딩 사전에 존재하지 않는 토큰은 제외한다.

2.3 단어 임베딩 및 문서 벡터 구성

전처리된 토큰은 사전 학습된 FastText 기반 단어 임베딩 모델을 이용하여 벡터로 변환된다[2]. FastText는 subword 정보를 활용하므로 OCR 과정에서 발생하는 철자 변형이나 미등록 단어에 대해서도 비교적 강인한 표현을 제공한다[2][3]. 한국어와 영어 임베딩을 동시에 사용함으로써 다국어 표 환경을 고려한다.

표 이미지는 단어 벡터들의 집합으로 표현되며, 평균 풀링을 적용하여 고정 길이의 문서 벡터를 생성한다. 이 문서 벡터는 표 전체의 의미적 분포를 요약한 표현으로, 분류기의 입력으로 사용된다.

2.4 MLP 기반 분류기

생성된 문서 벡터는 다층 퍼셉트론(MLP) 기반 분류기에 입력된다. 분류기는 입력층, 두 개의 은닉층, 출력층으로 구성되며, 은닉층에는 ReLU 활성화 함수를 적용한다. 모델은 교차 엔트로피 손실 함수를 최소화하도록 학습되며, Adam 옵티마이저를 통해 파라미터를 최적화한다.

2.5 데이터셋 구성 및 합성 데이터 생성 방식

본 연구에서는 의료 검사 결과 표 양식을 모사한 합성 표 이미지 데이터셋을 구축하였다. 데이터셋은 총 5개 클래스(cdc, EASI, labs, liver, rx)로 구성되며, 클래스당 2,000장씩 총 10,000장의 표 이미지를 생성하였다. 각 클래스는 검사 목적과 항목 구성, 텍스트 패턴이 상이하도록 설계되어 텍스트 의미 정보만으로도 클래스 구분이 가능하도록 구성하였다.

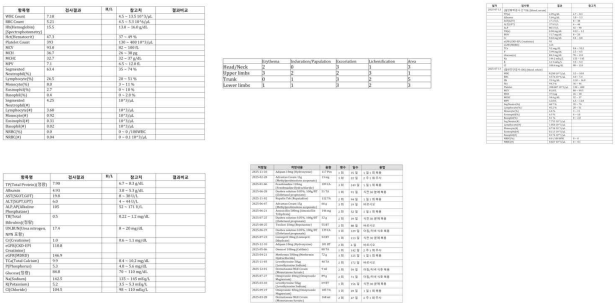

그림 1. 클래스별 합성 표 이미지 예시(cdc, EASI, labs, liver, rx)

2.6 실험 설정

제안한 방법의 분류 성능을 평가하기 위해 데이터셋을 학습/검증/시험 세트로 분할하였다. 분할은 클래스 비율을 유지하는 stratified 방식으로 수행하였으며, 학습/검증/시험 비율은 각각 80%, 10%, 10%로 설정하였다. 평가지표는 다중 클래스 분류의 성능을 대표하는 정확도(Accuracy)를 사용하였다.

또한 OCR 기반 텍스트 추출 및 임베딩 과정에서 동일한 전처리 규칙을 전 데이터에 일관되게 적용하여 실험 재현성을 확보하였다.

2.7 분류 성능 결과

실험 결과, 제안한 OCR 기반 의미 임베딩 분류기는 검증 및 시험 세트에서 모두 100%의 분류 정확도를 달성하였다. 이는 클래스별로 서로 다른 검사 항목 구성과 텍스트 패턴이 존재하는 합성 데이터 환경에서, OCR을 통해 추출된 텍스트 의미 정보만으로도 표의 클래스를 명확히 구분할 수 있음을 보여준다.

다만 본 결과는 클래스별 생성 규칙이 명확한 합성 데이터 기반 평가이며, 실제 문서 환경에서는 양식 다양성, 노이즈, 스캔 품질 변화 등으로 인해 추가적인 검증이 필요하다.

2.8 문서 임베딩의 정성적 분석(t-SNE 시각화)

문서 벡터가 클래스별로 어떤 분포를 형성하는지 확인하기 위해 t-SNE 기반 차원 축소 시각화를 수행하였다[4]. 평균 풀링으로 생성된 문서 임베딩을 2차원 공간으로 투영한 결과를 그림 2에 제시한다.

시각화 결과, 동일 클래스 표들이 임베딩 공간에서 군집을 형성하며, 클래스 간 분리 경향이 나타났다. 이는 제안한 텍스트 기반 문서 임베딩이 클래스별 의미적 특징을 효과적으로 반영하고 있음을 정성적으로 뒷받침한다.

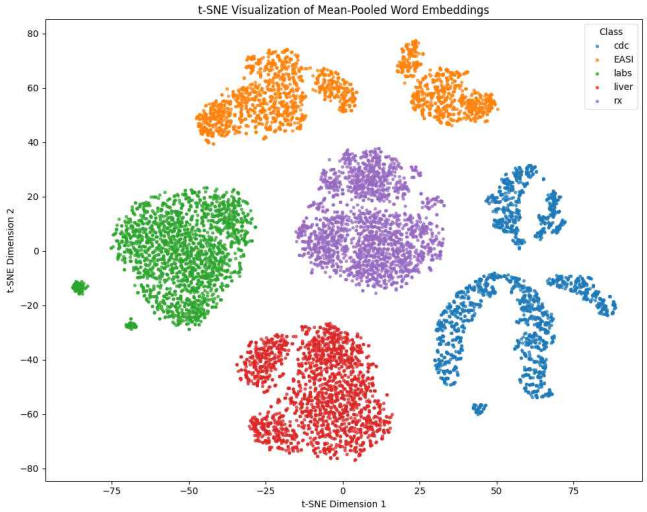

그림 2. 평균 풀링된 문서 임베딩의 t-SNE 시각화 결과

Ⅲ. 결론

본 논문에서는 OCR 기반 텍스트 의미 정보를 활용한 표 이미지 분류 기법을 제안하였다. 제안한 방법은 표 이미지로부터 텍스트를 추출하고 이를 단어 임베딩 기반 의미 공간에서 표현함으로써, 시각적 특징에 의존하지 않고도 표의 유형을 분류할 수 있도록 설계되었다. 평균 풀링과 경량 MLP 분류기를 통해 단순하면서도 효과적인 분류 구조를 구현하였다.

실험 및 시각화 분석 결과, 제안한 방법은 표의 의미적 특성을 효과적으로 반영하며 클래스 간 분리 가능한 표현 공간을 형성함을 확인하였다. 향후 연구에서는 단어 중요도를 고려한 가중 풀링이나 문맥 정보를 반영하는 임베딩 기법을 적용하여 실제 문서 환경에서의 일반화 성능을 향상시키고자 한다.

ACKNOWLEDGMENT

본 연구는 보건복지부의 재원으로 한국보건산업진흥원의 보건의료기술연구개발 사업 지원에 의하여 이루어진 것임.
(No. RS-2025-02223417)

참 고 문 헌

[1] Smith, R. "An overview of the Tesseract OCR engine," Proceedings of the Ninth International Conference on Document Analysis and Recognition, pp. 629-633, 2007.

[2] Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. "Enriching word vectors with subword information," Transactions of the Association for Computational Linguistics, vol. 5, pp. 135-146, 2017.

[3] Mikolov, T., Chen, K., Corrado, G., and Dean, J. "Efficient estimation of word representations in vector space," Proceedings of the International Conference on Learning Representations, 2013.

[4] van der Maaten, L., and Hinton, G. "Visualizing data using t-SNE," Journal of Machine Learning Research, vol. 9, pp. 2579-2605, 2008.