IT BizNews

고문서 해석에 사용되는 딥러닝 기술 ②

최태우 기자 승인 2019.02.18 10:34

지난 글에 이어 이번 글에서는 인쇄물의 판독에 사용되는 인공지능(AI) 기술, 또 인쇄물을 넘어 손글씨까지 판독 가능한 문서 분석 기술에 대해 알아본다.

스웨덴 룰레오 공과대학교(Luleå University of Technology)의 마커스 레비츠키(Marcus Liwicki) 교수에 따르면 몇 년 전 역사학자들은 수백만 페이지에 달하는 인쇄물을 스캔이미지로 디지털화를 진행했지만 "전부 다 전사하기란 불가능했다"고 말한 바 있다.

특정 정치 인사에 관심이 있던 어느 학자는 딥러닝 기반의 OCR 도구를 사용해 기계 판독이 가능한 고문서 버전을 쿼리할 수 있으며 그 정치 인사를 언급한 모든 내용을 찾을 수 있었다고 말하기도 했다.

그러나 활자 인쇄 외에도 삽화나 여백에 기록된 내용, 워터마크가 담긴 문서는 많다. "GPU는 디지털 인문학 연구의 전 과정을 혁신시켰다"고 말한 레비츠키 교수는 이러한 고문서의 특징을 분석할 수 있는 딥러닝 도구를 개발하고 있다.

그가 진행하는 프로젝트(HisDoc)는 뉴럴 네트워크를 사용해 출판시기, 사용된 글꼴과 같은 특정 문서 내의 고급 기능을 식별하면서 각 페이지를 분석해 글자와 이미지의 유무를 판별하는데 사용된다.

시편과 기도서를 통합한 14세기 초 문서에 그려진 원숭이 간의 대결을 그린 삽화 [source=Flickr]

레비츠키는 이번 프로젝트에서 엔비디아 GPU 클러스터를 사용해 8만개 워터마크 데이터베이스에서 트레이닝 시켰다. 워터마크는 출판된 지역과 시기가 동일하다는 단서가 된다. 학자들이 워터마크가 일치하는 문서 판독에 심혈을 기울이는 이유가 여기에 있다.

인쇄물을 넘어 손글씨까지 판독하는 AI

물론, 역사 기록은 인쇄물로만 생성되는 것은 아니다. 학자들이 관심을 보이는 여러 문서는 손글씨로 기록되어 있다. 인쇄물과 달리 손글씨는 판독에 어려움이 따른다. 작가가 축약어를 사용하기도 하고, 페이지마다 글씨체가 조금씩 달라지기도 한다. 인쇄물은 활자가 똑바른 수평을 이루지만 손글씨는 그 형태가 각각 다르다.

여기에서도 뉴럴 네트워크가 효과적인 전사 도구로 사용될 수 있다. 이탈리아 로마 트레 대학교(Roma Tre University) 소속의 파올로 메리알도(Paolo Merialdo), 도나텔라 피르마니(Donatella Firmani), 엘리나 니에두(Elena Nieddu) 연구원은 딥러닝을 사용해 바티칸 비밀 문서고에 보관돼 있던 12세기 교황 서신을 전사했다.

이들 연구원은 엔비디아 쿼드로 GPU와 컨볼루션 뉴럴 네트워크를 사용해 96% 정확도로 손글씨 문자를 인식식하고 라틴어 모델 기반으로 각 단어에 가장 알맞은 문자를 결정하는 시스템을 개발했다.

여기에서 한 발 더 나아가, 우크라이나 이고르 시코르스키 키예프 공과대학교(Igor Sikorsky Kyiv Polytechnic Institute)의 학생 연구원은 키예프의 성 소피아 성당 돌담에 그려진 중세 그래피티를 해석할 수 있는 뉴럴 네트워크를 개발했다. 그결과 엔비디아 GPU로 구동되는 딥러닝 모델은 개별 문자 인식 작업에서 99% 정확도를 보였다.

문서 분석은 전세계적으로 열리는 기술 컨퍼런스의 관심주제는 물론, CVPR/NeurIPS와 같은 주요 컴퓨터비전, 머신러닝학회에서도 최신 연구 결과가 발표되는 등 매우 활발하게 연구되는 분야 중 하나다.

글 : 이샤 살리안(Isha Salian) / 과학·인공지능 라이터 / 엔비디아

최태우 기자 taewoo@itbiznews.com

인쇄하기