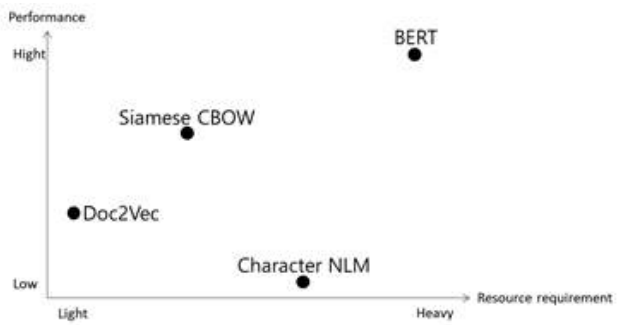
■ 연구과제 요약문

과제명(기간)	인공지능 기술을 통한 문서 분류 자동화 시스템 구축 (2019.04.01. ~ 2020.03.31.)
연구책임자	조 성 준 (zoon@snu.ac.kr)
개요	- 문서 형태의 비정형 데이터로부터 텍스트 정보를 수집하고, 해당 문서 내용에 기반한 카테고리를 도출하여 문서를 자동으로 분류한다. - 대상 문서는 영어 및 한글로 된 논문 (NIPS 20년치, CVPR 25년치) 및 wired.com, Tech Meme, zdnet 등의 뉴스이고 - 대상 내용의 키워드는 machine learning (머신러닝), artificial intelligence (인공지능), 빅데이터 이다.
연구개발 결과	- 전체적으로 현존 Language Modeling SOTA 모델인 BERT가 가장 좋은 성능을 보임. 그러나 pre-trained model 기반 결과이며, 임베딩 벡터 차원수를 768차원이라는 비교적 큰 숫자에 맞춰야만 하는 한계가 있음. Training 단계부터 재학습을 원할 시, 시간과 장비, 두 리소스가 모두 받쳐줘야 함 - 비교적 shallow model인 Siamese CBOW가 그 다음으로 좋은 성능을 보임. 영어 뉴스나 한글 논문의 경우 Doc2Vec이 Unweighted Siamese CBOW와 비슷하거나 조금 나은 성능을 보임. 계산 시간과 resource allocation, 그리고 한글&영어 동시 분석 관점에서, Doc2Vec이 Siamese CBOW의 좋은 대체제로 보임 - 영어 논문과 같이 길이가 길고 사용하는 단어의 양이 방대한 경우, 최적화 시 문맥적 고려가 필수적이며, 문장 표현을 문서 단위로 aggregate할 때 weighting scheme이 중요한 요소로 판단됨 
활용분야 및 기대효과	- 급변하는 기술 트렌드 분석 및 분류를 위하여 반도체 설계 엔지니어링 지식 데이터베이스 개발의 일환으로 DB 구축 목적의 주요 기능인 연관 검색 엔진으로 활용할 수 있다. 엔지니어링 지식 DB 개발의 선행 연구로 기술문서 카테고리 분류 자동화 시스템이 필수적이다. 기존의 단순 키워드 기반 분류가 아닌, 인공지능,머신러닝을 적극 활용하여 대규모의 문서 카테고리 분류가 필요함. 이를 통해 정확하고 빠르고 또한 새로운 키워드 등장에도 유연하게 대처할 수 있다. 문서 분류 비용의 절감 효과도 있다.