산학협력 프로젝트 결과보고서

프로젝트명	딥러닝 모델의 소이넷 포팅	
수행기간	2023. 04. 01 ~ 2023. 11. 29	
지도교수	이미향 교수님	
참여기업명	소이넷	
참여학생인원	총 5 명	
참여학생명단 (남녀구분작성)	남	남
	백승렬 (서명)	김호재 (서명)
	박세훈 (서명)	이해성 (서명)
	박제현 (서명)	(서명)

□ 본 보고서는 2023년도 과학기술정보통신부 및 정보통신기획평가원에서 주관하여 진행하는 'SW중심대학사업'의 결과물입니다.

□ 본 보고서의 내용을 전재할 수 없으며, 인용할 때에는 반드시 과학기술정보통신부와 정보통신기획평가원의 'SW중심대학'의 결과물이라는 출처를 밝혀야 합니다.

별첨3 산합협력 프로젝트 결과보고서

프로젝트명	딥러닝 모델의 소이넷 포팅		
협력기관(국가)	소이넷(한국)	과제책임자	이미향 교수
수행기간	2023. 04. 01 ~ 2023. 11. 29(08개월)	소요예산	2000만원

소요예산 세부내역	- 인건비 2000만원

참여인원	구분	인원수	성명(모두 기재)
	교수	1	이미향
	학부생	5	백승렬, 김호재, 박제현, 박세훈, 이해성
	기업체	2	김용호, 유경석
	계	888	

추진배경	최적화된 추론엔진의 제공

본 프로젝트는 소이넷 회사가 지닌 소이넷 솔루션을 이용하여 시중에서 현재 많이 쓰이고 있는 Yolo모델등의
러 모델들에 대해 레이어 단계에서부터 효율적인 연산을 제공하여 전반적인 성능을 향상시키는 최적화 작업
해보고자 시작되었다.

목표 및 내용	기존보다 빠르고 효율적인 딥러닝 모델 추론엔진 빌드

본 프로젝트는 소이넷 솔루션으로 동일 딥러닝 모델에 대해 기존보다 적은 메모리로 빠
게 동작하는 추론엔진 제공을 통해 비용절감과 실시간 시스템에 적용가능 하도록 함을 목적으로 한다

해당 프로젝트는 기존의 학습된 딥러닝 모델들의 특징들(모델구조, 가중치)을 파일로 불러오고 파일로 불러온
보를 넣어 소이넷 솔루션을 기반으로 하는 cpp,cuda로 작성된 프로그램을 통해 기존보다 빠르고 효율적인
러닝 모델의 추론 연산을 하는 엔진을 빌드하는 방식으로 진행하였다

빌드한 추론 엔진들의 성능을 측정을 위해 기존의 추론 프로그램과 속도, 그리고 메모리 사용량을 측정한 결
, 속도는 2~5배 정도가 빨라졌고, 메모리 사용량은 2~4배정도 줄어들었다.

그림 1> <그림 2>

image image

그림 1>: 기존 프로그램과 해당 프로젝트 결과물의 실행속도 차이

그림 2>: 기존 프로그램과 해당 프로젝트 결과물의 메모리 사용량 차이

기대효과	속도향상 및 메모리 사용량 감소를 통한 최적화된 추론엔진의 제공

결과적으로 최적화된 엔진들을 제공하여 실제 AI 활용에 있어서 드는 비용을 줄일 수 있다
저사양의 기기에서도 잘 돌아갈 수 있다(용량 부담 감소),
실시간 시스템에 응용가능하다.

본인 자필 서명 필수
프로젝트에 참여한 학생 모두 작성