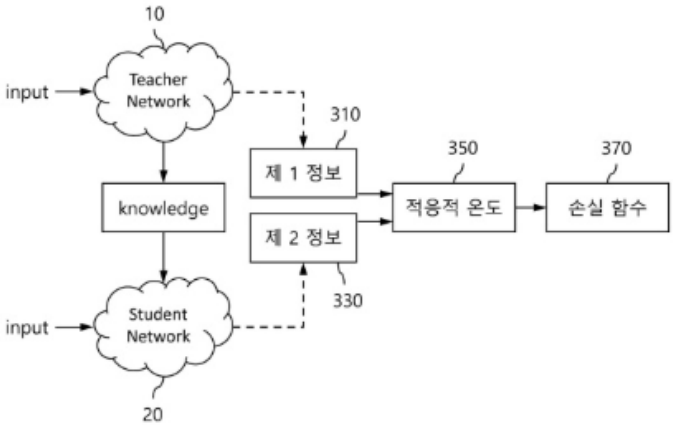
엔트로피 기반 적응형 지식 증류 학습 기술

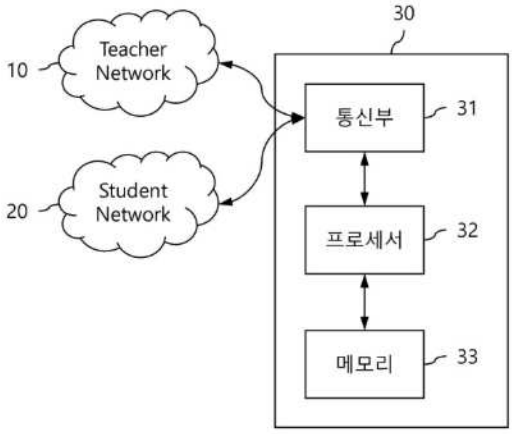
연구개발자: 반도체시스템공학과 한태희 교수

I 기술 개요

01 기술 요약



[적응적 온도 파라미터]



[지식 증류를 이용하는 신경망 학습 장치를 도시한 블록도]

- 본 기술은 복잡한 교사 모델의 지식을 경량화된 학생 모델에 전달하는 지식 증류 학습 방법에 관한 것으로, 교사 및 학생 네트워크로부터 도출된 정보 엔트로피 간의 거리를 기반으로 증류 손실의 온도 파라미터를 적응적으로 변화시키고, 초기의 불안정한 교사 모델 예측은 소프트 레이블로, 학습이 진행된 후의 신뢰할 수 있는 예측은 하드 레이블을 통해 전달하는 것을 특징으로 함

02 지식재산권 현황

No	발명의 명칭	출원번호	출원일
1	멀티코어 프로세서에서 가상 채널을 활용한 네트워크 캐싱 시스템	2023-0057824	2023.05.03
2	지식 증류를 이용하는 신경망의 학습 방법 및 장치	2022-0126921	2022.10.05

엔트로피 기반 적응형 지식 증류 학습 기술

03 기술의 우수성

- **적응적 온도 스케줄링**

  -기존 고정 온도 파라미터의 한계를 극복하고, 정보 엔트로피 기반의 거리를 활용해 온도 파라미터를 실시간 최적화

- **경량 모델 성능 극대화**

  -이진화 신경망(BNN) 등 극한으로 경량화된 학생 모델의 정확도 손실을 최소화하고 성능을 향상

- **온라인 지식 증류 최적화**

  -학습 초기(고온/소프트 레이블)부터 후기(저온/하드 레이블)까지 지식 전달의 강도를 적응적으로 조절하여 학습 안정성 및 효율을 극대화

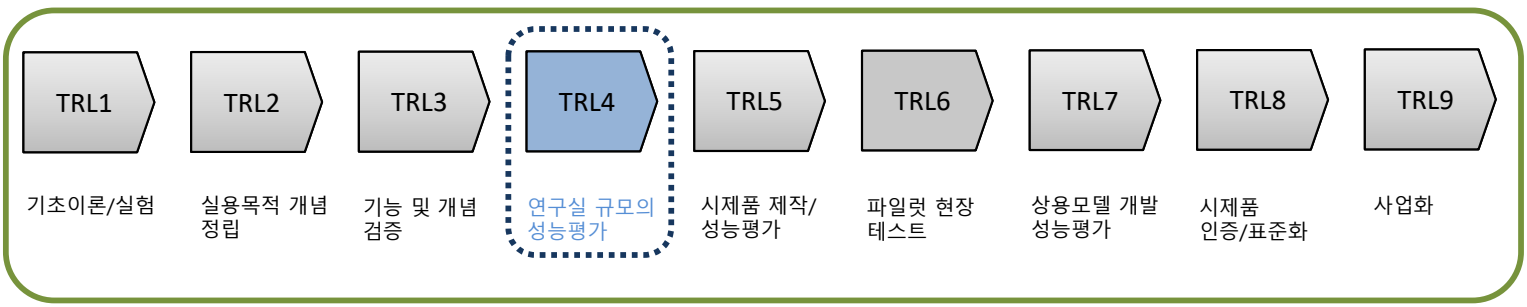
- **양자화 격차 해소**

  -교사-학생 네트워크 간의 양자화 격차를 인지하고 적절한 성능 차이를 유지하도록 유도하여 지식 전달의 품질 확보

- **손실 함수 정교화**

  -쿨백-라이블러 발산(KL divergence)과 정보 엔트로피를 활용한 선형 결합 손실 함수를 통해 KD 학습의 안정성을 확보

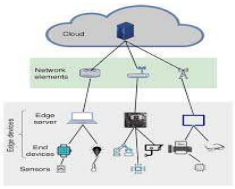
04 기술 개발 완성도

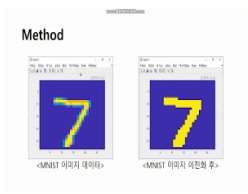


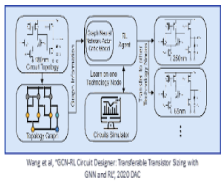
엔트로피 기반 적응형 지식 증류 학습 기술

II 기술 동향

01 기술응용분야







[엣지 AI 디바이스]

스마트폰, IoT 센서 등 저사양/저전력 환경의 AI 모델 경량화/배포

[자율주행 로봇]

실시간 처리와 저지연이 필수적인 임베디드 시스템용 고속 AI칩 개발

[클라우드/서버 경량화]

LLM 등 대규모 모델의 서비스 비용 절감

[이진화 신경망 활용]

메모리/계산 속도가 중요한 초경량 이진화 신경망 성능 향상

[AI 반도체 설계]

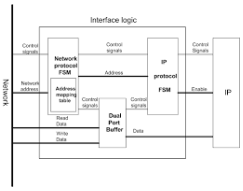
경량화된 신경망 구조에 최적화된 저전력 AI 칩 개발 및 IP 확보

02 기술 동향

[~2015]

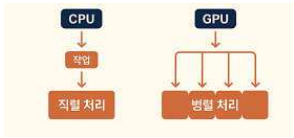
멀티코어 증가

[2016~2020]



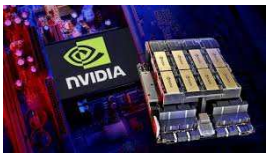
NoC 구조 확산

[2020~현재]



AI·병렬 연산 확대

[향후 전망]



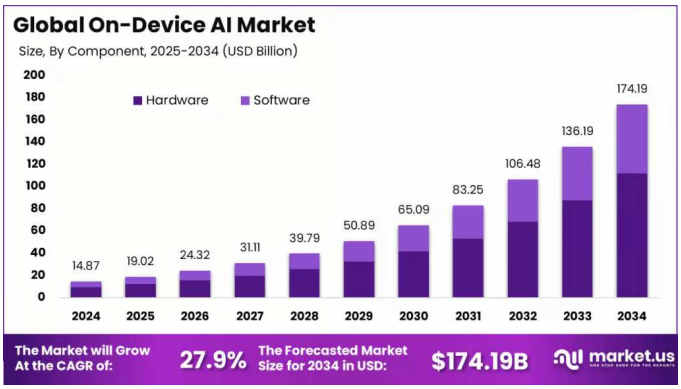
고성능 CPU·AI 칩 중심

인공지능 시장은 거대 모델의 운영 효율성 및 비용 부담 증가로 인해 신경망 압축 및 경량화 기술이 핵심 경쟁 분야로 부상하고 있고, 적응적 온도 파라미터를 통해 지식 증류의 근본적인 한계를 해결하여, 저사양 엣지 디바이스에서도 고성능 AI 모델을 구현할 수 있는 기술적 우위를 제공함

엔트로피 기반 적응형 지식 증류 학습 기술

III 시장 동향

01 시장규모



- 글로벌 온디바이스 AI 시장은 2024년 148억 7천만 달러에서 2034년 1,741억 9천만 달러로 성장할 것으로 예상되며 , 2025년부터 2034년까지 연평균 성장률 27.9% 규모로 성장할 것으로 예상됨

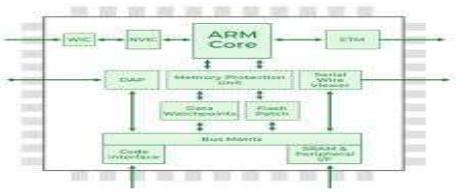
02 주요 시장 참여자

[NVIDIA 社 GPU, TensorRT, CUDA 제품]

- 클라우드 및 자율주행용 고성능 AI 칩 제조사. AI 모델의 경량화 및 최적화 도구를 제공

[Qualcomm 社 Snapdragon 기술]

- 모바일 및 엣지 디바이스용 AI 칩 선두 주자. 온디바이스 AI에 최적화된 경량 모델 기술 개발 주도



[ARM 社 ARM Architecture 기술]

- 모바일 및 임베디드 시스템의 핵심 아키텍처 제공. 저전력 AI 연산 및 경량화 기술 표준화에 기여

기술 이전 상담 및 문의

산학사업팀 담당자 김민경 / 031-290-5087 / mikkypat@skku.edu