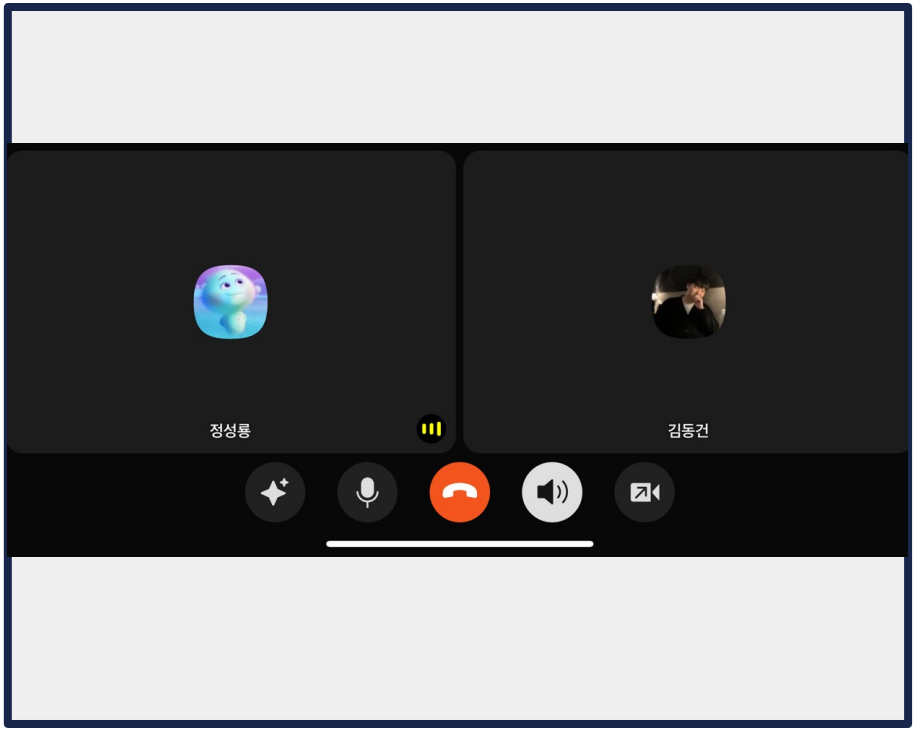
CUAI Model Compression 1팀

2022.03.11

발표자 : 정성룡

스터디원 소개 및 만남 인증

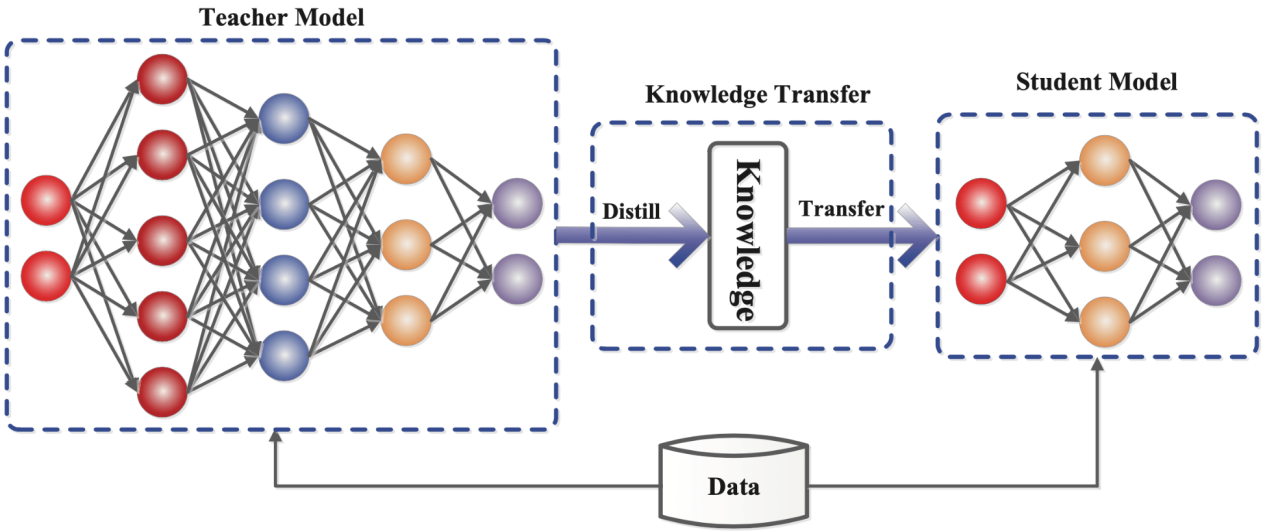


팀원 1 : 김동건 (AI학과)

팀원 2 : 정성룡 (AI학과)

목차

Knowledge distillation



Knowledge Distillation이란 크고 복잡한 **Teacher 모델**이 학습한 **지식을** 작고 간단한 **Student 모델**로 효과적으로 **전달하는 방법**

Dark Knowledge란?

딥러닝 모델의 **지식 증류(Knowledge Distillation)** 과정에서 **Dark Knowledge**는 **정답 데이터(label)와는 다른 추가적인 정보**를 의미

• **Teacher 모델**은 단순히 정답을 예측하는 것이 아니라, **출력 확률 분포**를 생성합니다.

• 이 확률 분포는 **클래스 간의 유사성이나 관계를 반영**하며, 이는 **정답(label)만 제공하는 것보다 더 풍부한 정보**를 담고 있습니다.

• **Student 모델**은 이 확률 분포를 학습함으로써, 단순한 정답 학습이 아니라 **더 깊은 지식**을 얻을 수 있습니다.

👉 **Dark Knowledge는 학생 모델이 더 일반화된 학습을 하도록 도와주는 중요한 요소**

Dark Knowledge란?

An example of hard and soft targets

cow	dog	cat	car	
0	1	0	0	original hard targets
10^{-6}	.9	.1	10^{-9}	output of geometric ensemble

문제 설정 : LLM에서의 KD 과정에서 발생하는 문제

1. KL-Divergence의 비대칭성으로 인한 Mode Averaging 및 Mode Collapse 문제

2. Teacher model이 지나치게 강한 Confidence를 가지는 문제

3. Teacher과 Student의 토크나이저가 다른 경우, 분포 매칭의 불가능성

4. 모델 학습(Teacher Forcing)방식과 추론(Autoregressive) 방식의 차이로 인한 Exposure Bias 문제

프로젝트 주제(연구 방향)

1. KL-Divergence의 비대칭성으로 인해 Mode Averaging 및 Mode Collapse 현상이 발생

2. Teacher 모델이 지나치게 강한 Confidence를 가지는 문제
   - 이는 Capacity Gap을 유발하며, Dark Knowledge의 활용을 제대로 하지 못하는 원인이 될 수 있음

👉 LLM에서 발생하는 위 문제를 해결하기 위한 연구 진행