모델 앙상블과 지식 증류의 융합을 통한 모델 학습의 최적화

이 름 양성욱 지도교수 황원준

연구 배경

최근 인공지능 기술의 발전으로 복잡한 네트워크 구조를 갖는 딥러닝 모델들이 더욱 다양하고 깊이 있는 학습을 하게 되었다. 그러나, 이러한 발전이 있음에도 고성능을 요구하는 시스템에서는 여전히 모델의 크기와 계산 비용이 큰 제약으로 작용한다. 이러한 문제를 해결하고자 지식 증류(Knowledge Distillation) 기법이 널리 사용되고 있다. 지식 증류는 기존의 크고 복잡한 심층 신경망 모델로부터 핵심적인 정보를 추출하고 이를 상대적으로 작고 간단한 모델에 전달하여 학습시키는 기법이다.
지식 증류에서 나타날 수 있는 문제점들은 여러가지가 존재하지만, 대표적으로는 선생 모델과 학생 모델 사이의 크기 차이에 의해 나타나는 문제가 있다. 교사 모델이 상대적으로 복잡하고 높은 성능을 가지는 데에 비해 학생 모델은 상대적으로 간소화된 구조를 가지고 있어, 교사 모델로부터 증류한 지식을 전부 수용하여 재현하기 어렵기 때문에 나타난다. 이러한 점을 해결하고자 연구된 지식 증류 네트워크가 바로 Teacher Assistant Knowledge Distillation, TA 모델을 추가한 지식 증류 네트워크이다.
TA 모델은 지식 증류에서 선생 모델과 학생 모델의 지식 증류 과정을 보조하는 역할로서 선생 모델과 학생 모델의 크기 간에 큰 간극이 존재할 때 이를 보완하기 위해 사용된다. 실험 결과 TA 모델을 사용한 지식 증류의 결과 성능이 그렇지 않았을 때보다 더욱 좋은 수치를 나타내는 것이 입증되었다.
그러나 이러한 지식 증류 네트워크에서 한계는 존재하는데, 그것은 이 네트워크에서 정보 전달이 단방향으로 이루어진다는 것이다. 특히 TA의 수가 여러 개인 상황에서 이 문제는 두드러질 수 있다. 선생 모델이 하위 모델에게 넘겨줄 모든 지식 중에는 잘못된 정보도 존재하기 마련이다. 잘못된 정보로 학습된 바로 하위의 TA 모델은 잘못된 학습을 바탕으로 기존 잘못된 정보와 함께 또 다른 잘못된 정보를 생성해 낼 것이고, 이러한 현상이 계속 이어지며 가장 하위 모델인 학생 모델에까지 영향을 미치게 된다. 이러한 문제점을 마치 잘못된 정보가 눈사태처럼 불어난다고 하여 “Error Avalanche”, 오류 눈사태라 지칭한다.
이러한 문제점을 보다 효율적으로 해결하고자 연구를 진행해보려 한다.

제안 기법

우선 문제 해결을 위한 핵심을 “가용 가능한 모델의 재사용 " 이라고 생각하였다. 비슷한 문제를 해결하기 위해 Ajou CVPR Lab에서 고안된 지식 증류 네트워크인 Densely Guided Knowledge Distillation, DGKD역시 하위 모델에게 지식 증류를 진행할 시에 가용 가능한 모든 상위 모델과 지식 증류를 진행한다. 이는 Error Avalanche 문제를 해결하는 데에 좋은 성능을 보이지만, 복잡한 구조를 가진다는 단점이 존재한다. 각 한 번의 지식 증류마다 시행되는 Loss 연산은 DGKD 네트워크에서 많은 빈도로 진행되고, 이는 시간적, 구조적 복잡도로 이루어진다.
모델 앙상블을 이용한 지식 증류 구조는 가용 가능한 모델을 최대로 이용하면서도 구조적 복잡도를 최소한으로 하고자 구상되었다. 학생 역할의 하위 모델보다 상위에 있는 모든 모델은 그 출력값들이 평균 앙상블 기법에 의해 한 개의 출력으로 계산된다. 합쳐진 모델들은 일종의 한 개의 pseudo 모델로 간주될 수 있다. 그 말인 즉 학생 역할의 하위 모델은 상위의 모든 모델과 지식 증류를

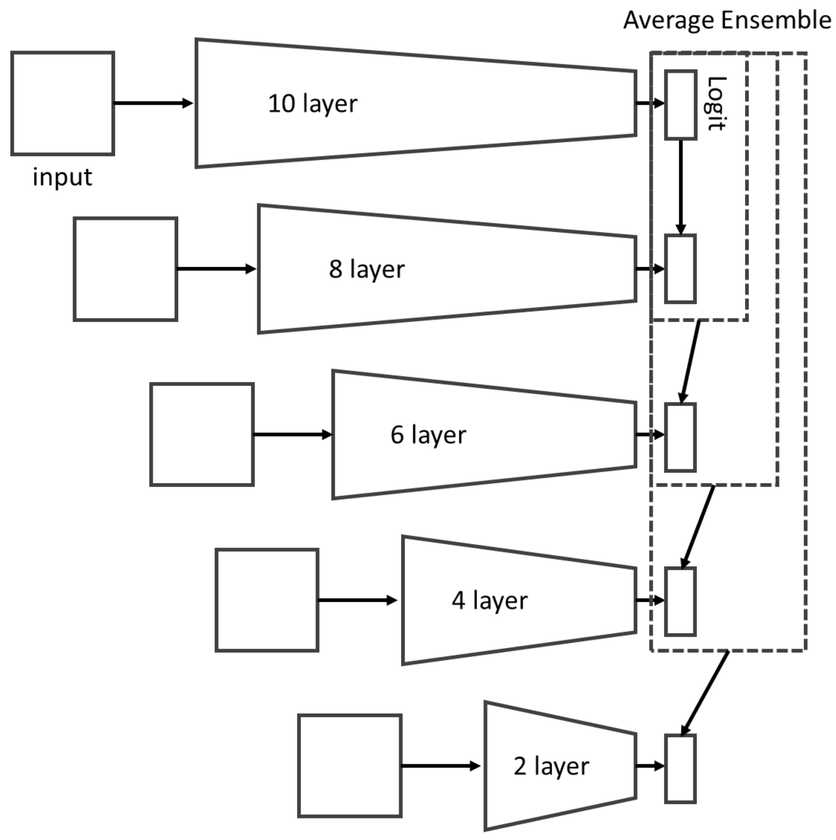

Fig 1. 모델 앙상블 기법을 적용한 지식 증류 네트워크

시행해야 하는 대신 하나의 pseudo 모델과만 지식 증류를 시행할 수 있게 된다. 이러한 구조적인 특성은 기존보다 Distillation Loss 연산의 횟수를 줄이는 데에 효과적이며, 최종 Loss 연산에서도 여러 개의 Loss 값이 영향을 미치는 대신 그 계산 구조가 매우 단순해지게 된다.

실험 및 결론

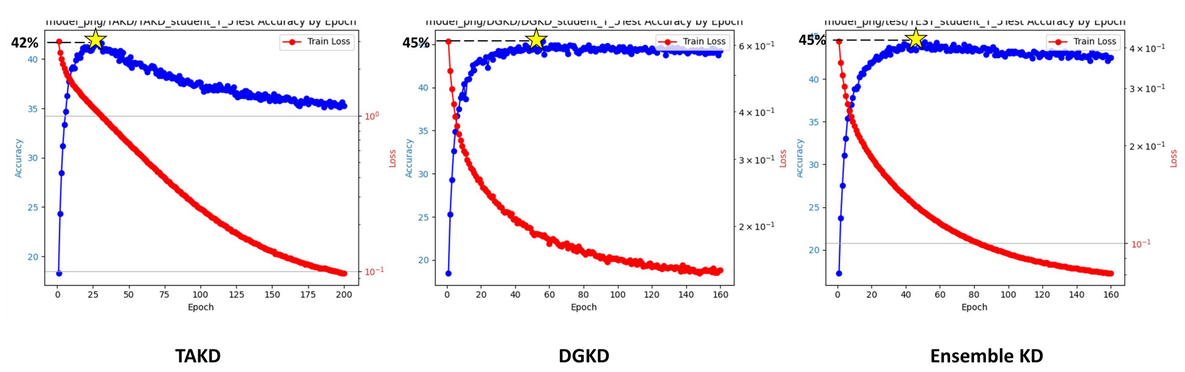

Fig 2. 여러 지식 증류 네트워크에서 학생 모델의 성능 측정

다음은 각각 TAKD, DGKD, Model Ensemble KD 네트워크를 이용하여 CNN 10layer 최상위 모델로부터 CNN 2layer 하위 모델을 훈련시킨 결과를 나타낸다. 결과, TAKD는 학생 모델이 약 42%의 정확도를 보이는 반면 모델 앙상블이 적용된 KD는 학생 모델의 성능을 DGKD와 비슷하게 약 45%정도 까지 끌어올리는 모습을 보여준다.

결론 및 향후 연구

준수한 성능을 보이는 모델 앙상블 KD 이지만 구조의 특성상 가용 가능한 모든 모델을 사용하므로 모델 자체의 연산의 수는 여전히 비용이 매우 높다. 따라서 향후 네트워크를 더욱 발전시키기 위해 모델 자체의 연산을 줄일 수 있는 다른 형태의 구조를 고안하는 것이 시간적 비용 절감의 핵심일 것 이다. 또한 본 지식 증류에서 사용된 모델 앙상블 기법은 상대적으로 단순한 평균 앙상블 기법을 사용하였으므로, 더욱 발전된 형태의 모델 앙상블 기법을 사용한다면 상위 모델의 역할을 맡는 pseudo 모델의 성능이 올라가므로 자연스럽게 성능이 증가할 가능성이 존재한다.