지식 증류 모델 수렴 제어와 거짓 정보 정제

남찬희[1], 서지원[2]
[1]한양대학교 컴퓨터 소프트웨어학과 석사과정
[2]한양대학교 컴퓨터 소프트웨어학과 교수
namch01@hanyang.ac.kr, seojiwon@hanyang.ac.kr

Control of Model Convergence and False Information Filtering on Knowledge Distillation

Chanhee Nam[1], Jiwon Seo[2]
[1]Dept. of Computer Science, Hanyang University
[2]Dept. of Computer Science, Hanyang University

요 약

지식 증류 기법에서 학생 모델은 Label Smoothing을 거친 교사 모델의 결괏값을 통해 일반 라벨 데이터로는 학습할 수 없는 정보를 학습한다. 이때, Label Smoothing은 학생 모델의 분포가 교사 모델의 분포에 수렴하지 못하게 하거나 교사로부터 옳지 않은 정보를 학생에게 학습시키는 악영향도 존재한다. 이에 본 연구에서는 학생 모델의 분포가 교사 모델의 원래 분포에 잘 수렴하도록 하기 위한 두 가지 기법과 데이터의 신뢰성을 향상시키기 위한 기법 총 세 가지 기법을 제안한다. 세 가지 기법을 각각 적용했을 때, CIFAR-100 데이터셋에서 최대 0.64%의 Top-1 Accuracy 향상을 보였다.

1. 서론

지식 증류 기법은 학생 모델이 교사 모델로부터 일반 라벨 데이터에서 얻을 수 없는 정보를 추출하는 방법이다. 기존 연구는 주로 교사 모델의 유용한 추가 정보를 학생 모델에 학습시키는 데 집중해왔다. 그러나 이 기법에는 두 가지 주요 문제가 있다. 첫째, Label Smoothing으로 인해 학생 모델이 교사 모델의 본래 분포에 제대로 수렴하지 못할 수 있다. 둘째, 교사 모델로부터 신뢰할 수 없는 정보를 가져올 수 있다.

본 연구에서는 이러한 문제들을 해결하기 위해 세 가지 방법을 제시한다: 1) Epoch Adaptive Temperature, 2) Data Adaptive Temperature, 3) Logits Pruning. 이를 통해 지식 증류 과정의 효율성과 정확성을 향상시키고자 한다.

2. 배경 지식

2.1 지식 증류 기법

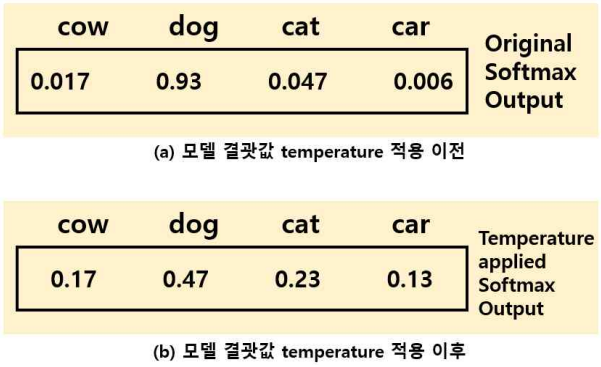
지식 증류 기법(Knowledge Distillation)은 Geoffery Hinton[1]이 제안한 방법으로, 큰 모델(교사 모델)이 가진 정보를 작은 모델(학생 모델)에 전달하는 기술이다. Hinton은 큰 모델이 보유한, 일반적인 데이터 라벨에서는 얻을 수 없는 정보를 'Dark Knowledge'라고 정의했다. 지식 증류 기법의 핵심은 교사 모델의 출력값에 Label Smoothing을 적용하여 교사 모델의 분포를 일반화하는 것이다. 이를 통해 학생 모델은 라벨 간 관계처럼 일반 라벨 데이터로는 학습할 수 없는 정보를 습득할 수 있게 된다.

2.1.1 Temperature 하이퍼 파라미터

Temperature 하이퍼 파라미터는 지식 증류 과정에서 중요한 역할을 하는데, 이 파라미터의 주요 목적은 라벨 데이터에서 직접 학습하기 어려운 라벨 간의 관계를 학습하게 하는 것이다. 교사 모델의 출력에 Label Smoothing을 적용함으로써 이를 가능하게 할 수 있다. **[그림 1]**에서 볼 수 있듯이, Temperature 적용 시 '개'와 '고양이' 같은 라벨(클래스) 간의 연관성이 강조되는 효과를 볼 수 있다.

2.2 지식 증류 기법의 분류

지식 증류 기법은 교사 모델의 정보를 학생 모델에 전달하는 방식에 따라 크게 두 종류로 나뉜다. 첫 번째는 교사 모델의 결과(Logits)만을 활용하는 방법[2]이고, 두 번째는 Feature Distillation[3]으로, 교사 모델의 결과(Logits)와 함께 추출된 데이터 특



(그림 1) Label Smoothing 기법 적용을 통한 라벨 간 연관성 강화

징도 학습에 활용한다. 이외에도 다양한 접근 방식이 존재한다. Decoupled Knowledge Distillation[4]은 교사 모델의 결과를 Target Class와 Non-Target Class로 분리하고, Non-Target Class 학습에 더 높은 비중을 부여한다. Teacher Assistant Knowledge Distillation[5]은 교사 모델과 학생 모델 사이에 중간 크기의 모델을 도입하여, 직접적인 지식 전달 대신 중간 모델을 통한 단계적 학습을 구현한다.

2.3 기존 연구의 한계

이러한 연구들은 지식 증류 기법을 크게 발전시켰으나 다음과 같은 한계점을 가지고 있다. 첫째, Label Smoothing에 관한 연구가 부족하다는 점이다. 둘째, 교사 모델의 정보 중 잠재적으로 악영향을 끼칠 수 있는 부분을 제거하지 않는다는 문제가 있다. 본 연구에서는 이러한 한계를 극복하고자, 교사 모델의 정보를 효율적으로 학습하면서 동시에 잠재적 악영향을 사전에 차단하는 방법을 제안한다.

3. 제안 기법

지식 증류 기법은 일반적인 라벨 데이터 학습과 더불어 교사 모델의 정보를 학습하는 별도의 과정을 필요로 한다. 이러한 특성으로 인해 다음 두 가지 주요 문제가 발생한다. 첫째는 Label Smoothing의 양면성 문제인데, Label Smoothing은 라벨 간의 관계와 같은 중요한 정보를 학습하기 위해 필수적이지만 학생 모델이 교사 모델의 일반화된 분포에 너무 빠르게 수렴하게 되면 손실 함수 값이 낮아져 교사 모델의 본래 분포를 온전히 학습하지 못하는 상황이 발생하고, 이는 학생 모델의 성능을 제한하는 요인이 될 수 있다. 둘째는 신뢰할 수 없는 정보 전달 문제이다. Label Smoothing 과정에서 교사 모델의 정보에 일부 변화가 가해지면 학습에 악영향을 미칠 수 있는 정보까지도 학생 모델에 전달될 가능성이 있다. 이는 학생 모델의 학습 품질을 저하시키고, 잘못된 정보를 기반으로 한 예측을 야기할 수 있다.

이러한 문제들을 해결하기 위해 본 논문에서는 세 가지 방법을 제안한다:

1. Epoch Adaptive Temperature (EAT): 학습 과정 중 Label Smoothing을 동적으로 제어하여 학생 모델이 교사 모델의 본래 분포를 더 효과적으로 학습할 수 있도록 한다.
2. Data Adaptive Temperature (DAT): 데이터의 특성에 따라 Label Smoothing을 조절하여 학습의 효율성을 높인다.
3. Logits Pruning (LP): 교사 모델로부터 전달되는 정보 중 신뢰할 수 없는 부분을 식별하고 제거하여, 학생 모델에 전달되는 정보의 품질을 향상시킨다.

본 연구에서는 이러한 방법들을 통해 지식 증류 과정에서 발생하는 문제들을 완화하고, 학생 모델의 학습 효율성과 성능을 개선하고자 한다.

3.1 Epoch Adaptive Temperature(EAT)

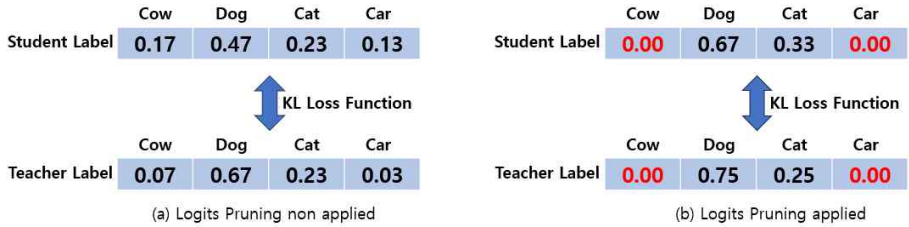
Label Smoothing 기법에서 Temperature를 높게 조정할수록 교사 모델의 출력값이 균등(Uniform)하게 변하지만, Temperature를 낮게 조정할수록 교사 모델의 출력값을 그대로 반환한다. 이는 레이블에 대한 하이퍼 파라미터가 교사 모델의 분포를 반영하는데 민감하게 적용될 수 있음을 의미하므로 Temperature 값을 급작스럽게 변화시키는 것은 학생 모델의 성능을 낮출 수 있다.

본 논문에서는 교사 모델과 학습 모델의 분포 차이가 크게 나는 학습 초기의 경우, 높은 Temperature를 통해서 교사 모델의 일반화된 분포를 학습하게 한다. 이후, Temperature를 학습 과정에서 점진적으로 감소시켜 교사 모델의 본래 분포를 학습시키는 EAT 기법을 제시한다.

3.2 Data Adaptive Temperature (DAT)

DAT는 일반적으로 단일 Temperature를 적용하는 지식 증류 기법과 달리 데이터 단위로 Temperature를 적용한다. 각 데이터는 Loss 값의 크기가 다르므로 교사 모델의 분포를 보다 정확하게 학습하도록 세분화하여 적용한다.

학습 과정에서 학생과 교사 모델의 결괏값에 대한 KL Divergence Loss를 측정하고, Loss 값의 크기에 따라 Temperature의 크기를 부여한다. 다시 말하여, 학습 과정에 주입된 데이터의 배치 단위로 Loss 값의 상위 10%와 하위 10%에 대해 Temperature를



(그림 2) Logits Pruning 의 적용 방법 도식화, (a)와 (b)는 Logit Pruning 적용 전, 후를 표현

높고 낮게 적용한다. 이를 통해 데이터별로 교사 모델의 분포가 세밀하게 학습이 되도록 유도한다.

3.3 Logits Pruning (LP)

Temperature로 교사와 학생 모델의 결괏값을 평탄화하게 되면 라벨 간의 관계성을 효과적으로 전달할 수 있지만, 라벨(클래스) 간의 존재하지 않은 연관성을 학생 모델에게 주입할 수도 있다. 예를 들어, **[그림 1]**에서 Label Smoothing 후 결과에서 소와 자동차의 잘못된 연관성이 더 증가하는 것이 확인된다. 따라서, 본 논문에서는 교사 모델에서 비교적 신뢰할 수 없는 낮은 결괏값을 가지는 라벨을 가지치기해서 잘못된 유사성을 학습하지 않도록 하는 LP를 제안한다.

가지치기의 방법으로는 logits값에서 하위 20%의 값에 낮은 값을 할당하여 Softmax 이후에 라벨값이 0이 되도록 유도한다. 학생 모델도 마찬가지로 같은 라벨들에 할당을 적용한다. **[그림 2]**의 (a)와 (b)는 각각 LP를 적용하기 전, 후를 표현한 것이다. (a)의 경우, Label Smoothing을 통해 발생한 잘못된 라벨 간의 연관성(e.g. 'dog'와 'car')을 그대로 학습하지만 (b)는 연관성이 없는 라벨에 대해서는 학습이 되지 않기 때문에 잘못된 정보를 학습하지 않게 된다.

$$\textbf{KL Loss Function: } D_{KL}(P||Q) = \sum_{x \in \chi} P(x) log \frac{P(x)}{Q(x)}$$

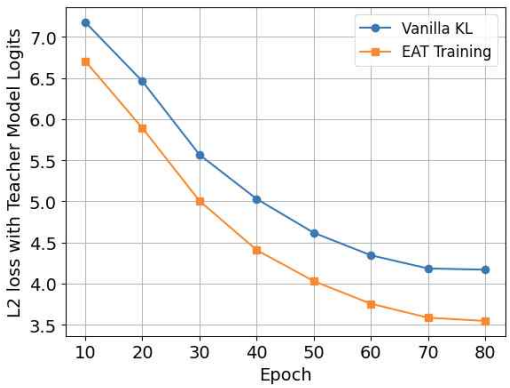
(식 1) 지식 증류에서의 KL Loss 함수

LP 기법은 불필요한 클래스들의 값을 0으로 유도했기 때문에 **[식 1]**의 KL Loss Function에서 볼 수 있듯, 교사 모델의 라벨값인 P(x)와 학생 모델의 라벨 값인 Q(x)가 일치하게 되기 때문에 해당 라벨로부터 오는 Loss가 0이 되고 학습이 되지 않는다는 것을 증명한다.

4. 평가 및 분석

제안된 기법들을 평가하기 위해 CIFAR-100 이미지 분류 태스크를 사용했다. 교사 모델로는 ResNet-110을, 학생 모델로는 ResNet-8, ResNet-20을 사용했다. 교사 모델은 CIFAR-100에 사전학습된 모델을 사용했으며 학생 모델은 학습 Epoch를 80회 진행하였다. Temperature의 경우, 기존 지식 증류 기법에서 많이 활용되는 설정값인 4로 초기화하였다.

앞서 제시한 기법들은 대부분의 지식 증류 기법들에 접목이 가능하다. 실험에서는 DKD (Decoupled Knowledge Distillation)[4] 지식 증류 기법에 추가로 제시한 기법들을 적용했을 때의 성능을 실험했다.



(그림 3) Vanilla 지식 증류 기법과 EAT 기법을 적용했을 때의 L2 Loss 변화 차이

4.1 Epoch Adaptive Temperature 평가

EAT 기법에서 Temperature는 기본값인 4를 기준으로 학습 시작 시에는 +40%(+20%) 이상 값에서 -40%(-20%) 이하값으로 점진적으로 학습을 진행하며 조정한다. 조정하는 단계는 10 Epoch 단위로 설정하였다.

[표 1]을 확인하면 ResNet-8에서는 Temperature를 점진적으로 +40%에서 -40%으로 변화했을 때 가장 Accuracy 상승 폭이 높았다. 이는 DKD 기법보다 0.55%만큼 Top-1 정확도가 상승한 것을 확인할 수 있다. ResNet-20에서는 0.5%만큼 Top-1 정확도가 상승했으며 ResNet-8과 비교 시 Temperature 변화를 상대적으로 적은 비율을 주었을 때 최적의 성능을 보였다.

접목한 EAT 기법이 DKD 기법에 비해 학생 모델의 성능을 향상시키는 것은 학습 과정에서도 확인이

가능하다. **[그림 3]**은 EAT 기법과 DKD 기법의 교사 모델과의 Loss 값의 변화를 분석한 것으로, EAT 기법이 학습 초기 단계부터 교사 모델의 분포에 가까워지려는 경향을 볼 수 있다.

Model	Method	Temperature	Top-1 acc	Top-5 acc
ResNet-8	KD		57.28	85.77
	DKD		58.97	86.26
	DKD+EAT	+40% - -40%	59.52	86.9
	DKD+DAT	5%, -5%	59.47	86.56
ResNet-20	KD		68.92	91.37
	DKD		69.29	91.26
	DKD+EAT	+20% - -20%	**69.79**	91.66
	DKD+DAT	5%, -5%	69.44	91.48
Model	**Method**	**Pruning Ratio (%)**	**Top-1 acc**	**Top-5 acc**
ResNet-8	DKD+LP	0, 0, 20, 40	**59.61**	86.62
ResNet-20	DKD+LP	5, 10, 15, 20	69.64	91.6
ResNet-110 (Teacher)			74.31	92.92

(표 1) EAT, DAT, LP 성능 비교표

4.2 Data Adaptive Temperature 평가

DAT는 데이터별로 교사 모델과 학생 모델의 KL Loss를 측정하고 Loss가 큰 상위 10%의 데이터와 Loss가 작은 하위 10%의 데이터에 대해 Temperature를 서로 다르게 부여하여 학습을 진행하였다.

[표 1]에서 모델별로 DKD+DAT에서 확인할 수 있듯, 기본 Temperature 값을 기준으로 Loss 값 상위, 하위데이터에 대해 각각 5%, -5%를 적용했다. ResNet-8의 경우에는 DKD보다 0.5% 만큼 더 높은 Top-1 정확도를 기록했으며 ResNet-20은 DKD보다 0.15% 만큼 더 높은 Top-1 정확도를 기록했다. 이를 통해, 데이터별로 Temperature를 다르게 적용하는 것이 학생 모델의 성능을 향상시키는 것을 확인했다.

4.3 Logits Pruning 평가

교사 모델의 결괏값 중에서 라벨값이 가장 낮은 라벨 데이터부터 가지치기하는 LP는 학습이 진행되면서 가지치기 비율을 높일 때 정확도의 상승폭이 가장 컸다. 이는 기존의 증류 기법들은 학습이 진행될수록 클래스 간에 존재하지 않는 연관성을 학습하기 때문에 성능 하락의 원인이 될 수 있다고 사료된다.

[표 1]을 확인하면 LP 옆에 최대 가지치기 비율 4개가 명시되어 있다. 이는 80번의 Epoch를 20개씩 4분할 뒤, 각 분할에 대한 가지치기 비율을 나타내었다. ResNet-8의 경우, 최저 가지치기 비율을 0%, 최대 40%로 설정했을 때 성능이 가장 높았다. 또한, DKD보다 0.64%만큼 Top-1 정확도가 상승하였다. ResNet-20 에서는 0.35%의 향상을 보였다. 즉, 교사의 정보 중 학습에 방해되는 정보가 포함되어 있다는 것 실험적으로 증명하였다.

5. 결론

본 연구는 기존 연구들이 주로 교사 모델에서 학생 모델로의 정보 전달에만 집중했던 것과 다른 측면에서 지식 증류 기법의 효율성과 정확성을 향상시키는 두 가지 주요 접근법을 제시한다. 첫째, 학생 모델이 교사 모델의 분포를 선별적으로 학습하여 학습의 효율성을 높였고, 둘째, 교사 모델의 정보 중 학생 모델에게 악영향을 미칠 수 있는 요소들을 제거하여 학습의 질을 개선하였다.

향후 연구에서는 본 논문에서 제안한 세 가지 기법(EAT, DAT, LP)을 통합 적용하면서, 학습 과정 중 동적으로 최적의 Temperature 변동률과 Pruning Ratio를 결정하는 방법을 탐구할 계획이다. 이를 통해 지식 증류 기법의 성능을 더욱 향상시킬 수 있을 것으로 기대한다.

"본 연구는 과학기술정보통신부 및
정보통신기획평가원의 대학ICT연구센터사업의
연구결과로 수행되었음"
(IITP-No.2024-2021-0-01817,No.RS-2020-II201373,
No.2022-0-00498)

참고문헌

[1] Hinton Geoffrey. "Distilling the knowledge in a neural network." NIPS. City of Montreal, Canada. 2015. 9P.

[2] Cho, Jang Hyun, and Bharath Hariharan. "On the efficacy of knowledge distillation." IEEE/CVF. Long Beach, CA, USA. 2019. 13P.

[3] Sun, Siqi, et al. "Patient knowledge distillation for bert model compression." ACL. Firenze, Italy. 2019. 10P.

[4] Zhao, Borui, et al. "Decoupled knowledge distillation." IEEE/CVF. Louisiana, USA 2022. 10p.

[5] Mirzadeh, Seyed Iman. "Improved knowledge distillation via teacher assistant." AAAI. New York, USA. 2020. 11p