이미지 분류 문제를 위한 focal calibration loss 기반의 지식증류 기법

강지연[1], 이재원[2], 이상민[3]
[1]광운대학교 인공지능응용학과 석사과정
[2]광운대학교 인공지능응용학과 석사과정
[3]광운대학교 정보융합학부 교수

wldus623@gmail.com, jcircle67@gmail.com, smlee@gmail.com

Focal Calibration Loss-Based Knowledge Distillation for Image Classification

Ji-Yeon Kang[1], Jae-Won Lee[2], Sang-Min Lee[3]
[1]Dept. of Artificial Intelligence Applications, Kwangwoon University
[2]Dept. of Artificial Intelligence Applications, KwangWoon University
[3]Dept. of Information Convergence, KwangWoon University

요 약

최근 몇 년 간 딥러닝 기반 모델의 규모와 복잡성이 증가하면서 강력하고, 높은 정확도가 확보되지만 많은 양의 계산 자원과 메모리가 필요하기 때문에 모바일 장치나 임베디드 시스템과 같은 리소스가 제한된 환경에서의 배포에 제약사항이 생긴다. 복잡한 딥러닝 모델의 배포 및 운영 시 요구되는 고성능 컴퓨터 자원의 문제점을 해결하고자 사전 학습된 대규모 모델로부터 가벼운 모델을 학습시키는 지식증류 기법이 제안되었다. 하지만 현대 딥러닝 기반 모델은 높은 정확도 대비 훈련 데이터에 과적합 되는 과잉 확신(overconfidence) 문제에 대한 대책이 필요하다. 본 논문은 효율적인 경량화를 위한 미리 학습된 모델의 과잉 확신을 방지하고자 초점 손실(focal loss)을 이용한 모델 보정 기법을 언급하며, 다양한 손실 함수 변형에 따라서 지식증류의 성능이 어떻게 변화하는지에 대해 탐구하고자 한다.

1. 서론

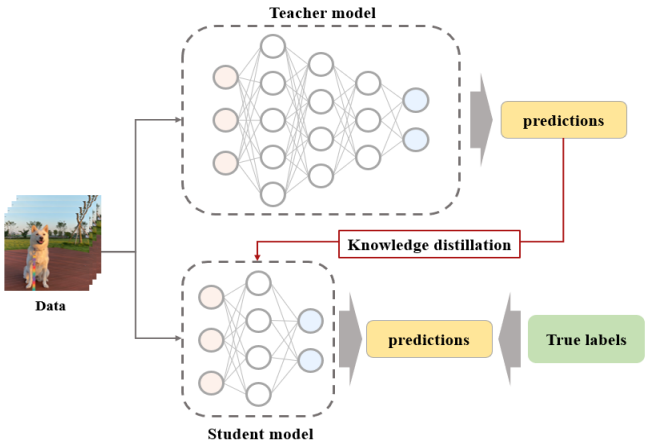
지난 몇 년 동안 딥러닝은 컴퓨터 비전(Krizhevsky et al., 2012), 강화 학습(Silver et al., 2016; Ashok et al. , 2018; Lai et al., 2020) 및 자연어 처리(Devlin et al., 2019). 를 포함한 많은 최신 기술의 도움으로 강력한 GPU 또는 TPU 클러스터에서 수천 개의 레이어로 구성된 매우 심층적인 모델을 쉽게 훈련할 수 있다. 대규모 심층 모델은 높은 정확도와 강력한 성능을 보이지만 많은 양의 계산 자원과 대규모의 메모리가 필요하기 때문에 이를 실시간 애플리케이션, 특히 비디오 감시 및 자율 주행 자동차나 모바일 기기와 같이 리소스가 제한된 장치에 배포하는 것은 큰 과제이다. 이를 해결하기 위하여 모델 가지치기, 양자화, 지식증류 등 여러가지 모델 경량화 방법론들이 제시되고 있다. 그 중 지식증류(knowledge distillation)[1]는 미리 학습된 거대한 신경망 모델로부터 상대적으로 작은 네트워크로 지식을 전달하여 학습하는 기술이다. 이하 미리 학습된 거대한 신경망 모델을 선생 모델, 상대적으로 작은 네트워크를 학생 모델로 칭한다. 본 논문에서는 이러한 지식증류의 손실 함수 변형을 통해 지식증류의 성능을 비교한다.

2. 관련 연구

2.1 지식증류

본 개념은 사전에 훈련된 선생 모델의 지식을 더 작고 비교 가능한 학생 모델로 효과적으로 전달하는 것을 목표로 한다. 전통적인 방법은 고정된 온도 매개변수를 사용해여 Kullback-Leibler 발산 손실을 최소화 하고 두 모델의 출력 분포를 일치시키는 것이다.

지식증류 성능을 향상시키기 위해서 다양한 형태의 지식증류 방법이 설계되었으며 로짓(logit) 기반, 표현 기반, 그리고 관계 기반 등 크게 세 가지 유형으로 분류된다.



(그림 1) 지식증류 기본, 선생 모델과 학생 모델은 동일한 이미지를 입력으로 받으며, 학습 데이터에 대한 예측 확률 분포를 각각 생성한다. 학생 모델은 선생 모델의 지식을 기반으로 예측을 수행하며, 이 예측 결과를 평가한다.

일반적인 지식증류의 손실 함수는 KL-divergence 와 cross-entropy 를 포함된다. KL-divergence 는 선생 모델의 예측 값과 학생 모델의 예측 값의 차이를 계산하고, cross-entropy 는 학생 모델의 예측 값과 실제 정답 값의 차이를 계산한다. 본 연구에서는 이러한 손실 함수에 Focal calibration term 을 추가하여 선생 모델이 학생 모델에게 지식을 증류할 때 정제된 지식을 전달할 수 있도록 한다.

2.2 Focal calibration

딥러닝은 훈련 과정에서 모델의 신뢰도와 정확성 사이의 불일치로 인해 예측을 신뢰하기 어려운 경우가 있다. 네트워크가 정확하게 보정되고, 확신을 가지게 하기 위해 여러가지 모델 보정 방법론이 제시되었다. Temperature 스케일링[3]은 Platt 스케일링의 단일 매개변수를 변형하여 보정된 확률 얻는다.

Focal calibration 은 모델이 얼마나 데이터를 잘 분류하는지에 따라 미니 배치의 개별 샘플에서 생성된 손실 구성 요소에 초점손실(focal loss) 가중치를 추가로 적용한다. 네트워크는 과적합이 발생하는 경우, 정확성과는 무관하게 예측에 대해 높은 확신을 가질 때 보정 오류와 강한 연관성이 나타난다. [6] 이 문제의 원인 중 하나는 네트워크가 미니 배치의 개별 샘플을 얼마나 잘 분류하든지 상관없이 교차 엔트로피 목표가 전체 미니 배치에 걸쳐 소프트맥스 분포와 실제 원-핫 인코딩 간의 차이를 최소화하려는 것이다. Focal calibration 은 이와 같은 문제를 해결하고자 적용된다.

3. 실험 설계

3.1 실험 환경 및 구성

실험은 Intel Core i7-10700 2.90GHz CPU 와 GeForce RTX 3070 GPU 를 탑재한 PC 에서 진행하였다. 실험에 사용한 모델은 총 6 개로 Resnet-56 Resnet-110 을 Resnet-20 의 선생 모델로, Resnet32x4 와 Wide Resnet40-2 를 Shufflenet-V1 의 선생 모델로 설정하였다.

실험에 활용한 데이터는 CIFAR-100 를 사용하였으며 각 이미지는 32x32 픽셀 크기의 컬러 이미지와 100 개의 클래스로 구성된다.

3.2 손실함수

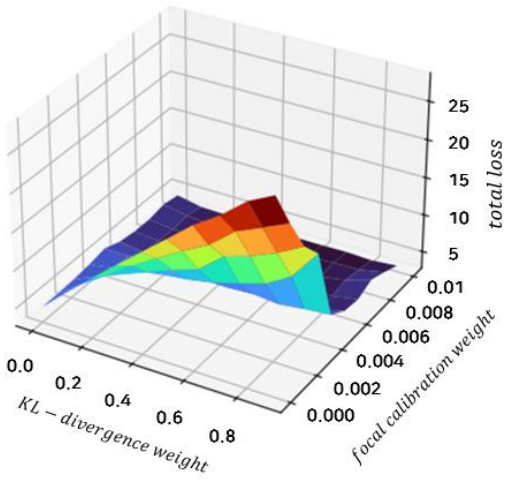
$$L_{total} = \alpha T^2 * KL\left(\sigma\left(\frac{Z_s}{T}\right), \sigma\left(\frac{Z_t}{T}\right)\right) + \beta L_c(\sigma(Z_s), y) + \gamma L_{fl}(\sigma(Z_s), y)$$

위 식은 기존 지식증류 손실함수에 초점 손실을 더한 식이다. 식에 포함된 알파, 베타, 감마 값들을 조절하여 전체 손실 값들을 비교하고자 한다.

각 가중치의 범위는 다음과 같다. 알파는 0 에서 0.1, 베타는 1-알파, 감마는 0 에서 0.01 로 각각 범위를 정하고, 알파와 감마는 각 범위를 10 개의 구간으로 나누어 총 100 개의 파라미터 조합을 만들어 가중치에 따른 total loss 의 값을 비교해본다.

알파가 0.5, 베타가 0.5, 감마가 0 값을 갖는 예시 같은 경우에는 focal calibration 을 사용하지 않는 기존 지식증류의 결과값을 도출해낼 수 있고, 감마 값을 점차 추가해 나간다면 focal calibration 정도에 따라 지식 증류 전체의 손실 값을 비교해보는 것 가능하다.

4. 실험 결과



(그림 2) 가중치 정도에 따른 total loss 값의 변화

위 그림 2 는 손실 함수의 구성에 따라 변하는 학생 모델의 정확도 값을 보여준다.

붉을 색상을 보일수록 손실 값이 높은 경우를 의미하며, 어두운 푸른 계열의 색상을 가질수록 낮은 손실 값을 가짐

을 의미한다.

따라서 위 그림에서 기존 지식증류 결과 값인 focal calibration weight 값이 0 을 가질 때보다 0 보다 조금 더 큰 값을 점차 값을 증가해 감에 따라 전체 손실 값이 감소하는 것을 확인할 수 있다.

<표 1> CIFAR-100 데이터에서의 학생 모델 top-1 정확도

Teacher Acc	**RN-56**	**RN-110**	**RN32x4**	**WRN40-2**
	71.63	71.77	77.55	75.10
Student Acc	**RN-20**	**RN-20**	**SN-V1**	**SN-V1**
	67.51	67.51	71.67	71.67

5. 결론

본 연구는 이미지 분류 문제에서 지식증류 기법의 강건함을 알아보고자 손실함수 변형을 통해 여러가지 실험 결과를 확인해보았다. 그 결과 focal calibration wright 가 전체 손실 값에 많은 영향을 끼치는 결과를 보였고, 전체 손실 값을 감소시키는 것을 확인했다. 하지만, 감마의 값이 너무 증가하였을 때 손실 값이 수렴성을 아예 찾지 못하거나 음수 값을 가지게 되는 경우가 발생하여 이를 해결 하고자 하는 추가적인 연구가 필요하다.

참고문헌

[1] Knowledge Distillation from A Stronger Teacher 에 [16] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015. 1, 3, 5, 6, 7, 8, 14

[2] Wang, X., Chen, Y., & Zhu, W. (2021). A survey on curriculum learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9), 4555-4576.

[3] Xiang, L., Ding, G., & Han, J. (2020). Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16 (pp. 247-263). Springer International Publishing.

[4] Jin, X., Peng, B., Wu, Y., Liu, Y., Liu, J., Liang, D., ... & Hu, X. (2019). Knowledge distillation via route constrained optimization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1345-1354).

[5] Guo, C et al. On calibration of modern neural networks

[6] Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., & Dokania, P. (2020). Calibrating deep neural networks using focal loss. Advances in Neural Information Processing Systems, 33, 15288-15299.