CLIP2FL 기반 Long-tail 연합학습을 위한 균형 지식 증류(BKD) 기법

전준, 백민우, 이상금*
*국립한밭대학교
20237142@edu.hanbat.ac.kr, bmw5779@gmail.com, *sangkeum@hanbat.ac.kr

Balanced Knowledge Distillation (BKD) for Long-Tail Federated Learning Based on CLIP2FL

Jun Jeon, Baek Minu, Sangkeum Lee*
*Hanbat National University

요 약

본 연구는 연합학습(Federated Learning) 환경에서 발생하는 Long-tail 분포 문제를 완화하기 위해 균형 지식 증류(BKD) 기법을 제안한다. BKD 는 대규모 사전학습 모델인 Contrastive Language-Image Pre-training(CLIP)을 교사로 활용하되, 지식 증류 과정에서 교사 출력 분포를 클래스별로 재가중함으로써 소수 클래스의 상대적 기여도를 증폭시킨다. 이를 통해 손실 함수를 직접 수정하지 않고도 학습 안정성을 유지하며 불균형 문제를 완화한다. 실험 결과, STL-10 에서는 모든 불균형 조건에서 일관된 성능 향상을 보였으나, CIFAR-10 에서는 교사–학생 간 데이터 특성 불일치로 인해 제한적인 효과가 관찰되었다. 본 연구는 지식 증류 신호의 균형을 직접 제어하는 새로운 연합학습 접근을 제시한다.

Ⅰ. 서 론

연합학습(Federated Learning)은 데이터 프라이버시를 보장하면서 분산된 데이터를 활용할 수 있는 학습 패러다임으로 주목받고 있다 [1]. 특히 무선 센서 네트워크와 같은 분산 환경에서 머신러닝의 적용이 확대됨에 따라 [2], 연합학습의 중요성은 더욱 증가하고 있다. 그러나 실제 환경에서는 클라이언트 간 Long-tail 분포로 인해 소수 클래스에 대한 과소적합과 다수 클래스 중심의 편향이 발생하며, 이는 전역 모델 성능과 수렴 안정성을 저하시킨다 [3]. 이러한 문제를 완화하기 위해 지식 증류(Knowledge Distillation) [4] 기반 접근이 연구되고 있으며, 특히 CLIP2FL 은 대규모 비전-언어 모델인 Contrastive Language-Image Pre-training (CLIP) [5]을 교사로 활용하여 클라이언트의 로컬 모델 학습을 보조하는 프레임워크를 제안하였다 [6]. CLIP 은 대규모 이미지-텍스트 쌍으로 사전학습되어 풍부한 의미적 표현을 제공하므로, 데이터가 제한된 클라이언트에서도 일반화 성능 향상에 기여할 수 있다. 그러나 CLIP2FL 은 교사 모델의 출력 분포를 그대로 사용하기 때문에, Long-tail 환경에서 다수 클래스에 편향된 지식이 학생 모델에 전이되는 한계가 존재한다. 본 연구는 이러한 한계를 극복하기 위해 교사 출력 분포를 클래스별 학습 난이도에 따라 재조정하는 균형 지식 증류(BKD) 기법을 제안하며, 소수 클래스에 대한 학습 신호를 강화함으로써 보다 균형 잡힌 지식 전이를 유도한다.

Ⅱ. 본론

1. 균형 지식 증류 (Balanced Knowledge Distillation)

BKD 의 설계 철학은 손실 함수를 급격히 변경하지 않고, 교사 모델의 출력 분포를 완만하게 조정하여 학습 안정성을 유지하면서 클래스 불균형을 완화하는 데 있다. 이는 Focal Loss 와 같은 직접적인 손실 수정 방식에서 발생할 수 있는 훈련 불안정성을 회피하기 위한 접근이다 [7]. BKD 는 로컬 분류 손실, 클래스별 평균 손실에 기반한 가중치 계산, 교사 출력 재가중, Kullback-Leibler (KL) 발산 기반 지식 증류 손실, 그리고 최종 손실 함수의 가중 합으로 구성된다. 클래스 가중치 λ_c 는 학습이 어려운 클래스일수록 더 큰 값을 가지며, 이를 통해 재정규화된 교사 분포는 소수 클래스의 기여도를 상대적으로 증폭시킨다. 최종 손실은 분류 손실과 지식 증류 손실의 가중 합으로 정의되며, 예비 실험을 통해 α=0.5 를 사용하였다. BKD 알고리즘은 다음과 같은 5 단계로 구성된다.

1.1. 로컬 분류 손실

클라이언트 k의 로컬 데이터 $\mathcal{D}_k = \{(x_i, y_i)\}$에 대해 학생 모델 $f(\cdot;\theta_k)$의 기본 분류 손실은 다음과 같이 정의된다.

$$\mathcal{L}_{\mathrm{CE}}^{(k)} = -\mathbb{E}_{(x,y)\sim\mathcal{D}_k} \sum_{c=1}^{C} 1(y=c)\log\, p_{k,c}(x)$$

여기서 $p_{k,c}(x)$는 클래스 c에 대한 학생 모델의 예측 확률이다.

1.2. 클래스 가중치 계산

클래스 c에 대한 평균 손실을 다음과 같이 정의한다.

$$\bar{\ell}_c = \mathbb{E}_{(x,y)\sim\mathcal{D}_k, y=c}[-\log p_{k,c}(x)]$$

이를 기반으로 클래스 가중치 λ_c를 산출한다.

$$\lambda_c = \frac{\bar{\ell}_c}{\frac{1}{C}\sum_{j=1}^{C}\bar{\ell}_j}$$

이때 모든 클래스 가중치는 다음 조건을 만족한다.

$$\frac{1}{C}\sum_{c=1}^{C}\lambda_c = 1$$

클래스 가중치는 각 통신 라운드의 로컬 학습 시작 시점에 계산되며, 해당 라운드 동안 고정된다. 초기 학습 단계에서의 손실값 불안정을 완화하기 위해, 첫 5 에폭 동안은 $\lambda_c = 1$로 설정하는 warm-up 기간을 적용하였다.

1.3. 교사 출력 재가중

사전 학습된 교사 모델(CLIP)의 출력 확률 분포를 $q_c(x)$라 할 때, BKD는 이를 클래스 가중치로 재조정한다. 본 연구에서는 CLIP 의 이미지 인코더를 통해 추출된 특징 벡터와 각 클래스의 텍스트 임베딩 간 코사인 유사도를 계산한 후, softmax 를 적용하여 $\tilde{q}_c(x)$ 를 산출한다..

$$\tilde{q}_c(x) = \frac{\lambda_c\, q_c(x)}{\sum_{j=1}^{C}\lambda_j\, q_j(x)}$$

$\tilde{q}(x)$ 는 재정규화된 균형 교사 분포로, 소수 클래스의 기여도가 상대적으로 증폭된다.

1.4. 지식 증류 손실

재가중된 교사 분포 $\tilde{q}(x)$와 학생 모델 출력 $p_k(x)$간의 지식 증류 손실은 KL-발산으로 정의된다. 본 연구에서는 temperature τ=1 을 사용하였으며, 이는 BKD 의 재가중 메커니즘이 이미 분포 조정 역할을 수행하므로 추가적인 softening 이 불필요하기 때문이다.

$$\mathcal{L}_{\mathrm{KD}}^{(k)} = \mathbb{E}_{x\sim\mathcal{D}_k}\sum_{c=1}^{C}\tilde{q}_c(x)\log\frac{\tilde{q}_c(x)}{p_{k,c}(x)}$$

1.5. 최종 학습 손실

각 클라이언트의 최종 학습 손실은 분류 손실과 BKD 기반 지식 증류 손실의 가중 합으로 구성된다.

$$\mathcal{L}_{\mathrm{total}}^{(k)} = \mathcal{L}_{\mathrm{CE}}^{(k)} + \alpha\,\mathcal{L}_{\mathrm{KD}}^{(k)}$$

여기서 α 는 지식 증류 항의 영향력을 조절하는 하이퍼파라미터이다. 제안된 BKD 기법의 실질적인 효과를 검증하기 위해, 다음 장에서는 Long-tail 분포 환경에서 수행된 실험 설계 및 그 결과를 분석한다.

2. 실험 결과

본 실험은 BKD 기법이 Long-tail 연합학습 환경에서 모델 성능을 향상시키는지를 검증하기 위해 수행되었다. STL-10 과 CIFAR-10 데이터셋을 사용하였으며, Imbalance Factor(IF)를 10, 50, 100 으로 설정하여 다양한 불균형 조건을 구성하였다.

구분	IF	Base(%)	BKD(%)
CIFAR-10	10	73.89	73.06
	50	76.38	76.04
	100	83.61	81.19
STL-10	10	42.44	43.25
	50	43.63	45.31
	100	52.34	53.89

Table 1. 데이터셋과 IF 에 따른 정확도 성능 비교

실험 결과, STL-10 에서는 모든 불균형 조건에서 BKD 가 베이스라인 대비 일관된 성능 향상을 보였다. 반면 CIFAR-10 에서는 성능 향상이 제한적이었으며, 이는 교사 모델(CLIP)과 학생 모델 간의 특징 공간 불일치에서 기인하는 것으로 분석된다. CLIP 은 고해상도 자연 이미지 기반의 풍부한 표현을 학습한 반면, CIFAR-10 은 저해상도의 단순한 객체 중심 데이터로 구성되어 있어 효과적인 지식 전이가 제한되었다. 이는 BKD 의 성능이 교사–학생 간 데이터 특성 정합성에 의존함을 시사한다.

Ⅲ. 결론

본 연구는 연합학습 환경의 Long-tail 분포 문제를 완화하기 위해 교사 모델의 지식 증류 신호를 재가중하는 BKD 기법을 제안하였다. 기존 지식 증류 연구와 달리, 본 연구는 지식 증류 신호의 균형을 직접 제어함으로써 클래스 불균형 문제를 다루는 새로운 접근을 제시한다. 실험을 통해 BKD 의 유효성과 함께, 교사–학생 간 데이터 특성 불일치가 성능에 미치는 영향을 확인하였다. 향후 연구로는 데이터 특성에 따라 재가중 전략을 조절하는 적응형 재가중 기법과, 교사–학생 간 표현 정합을 강화하는 방법의 결합이 유망할 것으로 기대된다.

참 고 문 헌

[1] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. AISTATS.

[2] Kim, T., Vecchietti, L. F., Choi, K., Lee, S., & Har, D. (2020). Machine learning for advanced wireless sensor networks: A survey. IEEE Sensors Journal, 21(11), 12379-12397.

[3] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-IID data. arXiv:1806.00582.

[4] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. NeurIPS Workshop.

[5] Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. ICML.

[6] Guo, Z., Zhang, Z., Li, X., & Liu, Y. (2023). CLIP2FL: Federated Learning with Vision-Language Models. NeurIPS.

[7] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. ICCV.