Journal of KIIT. Vol. 24, No. 7, pp. 95-104, Jul. 31, 2026. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2026.24.7.95

신뢰도 기반 가중치와 레이블 교란을 결합한 지식 증류 학습법

고도영*[1], 김태훈**, 안준호*[2], 김남기*[3]

A Knowledge Distillation Method Combining Prediction Confidence-based Weighting and Label Perturbation

Doyoung Koh*[1], Taehoon Kim**, Junho Ahn*[2], and Namgi Kim*[3]

This work was supported by Kyonggi University's Graduate Research Assistantship 2026, Institute of Information & Communications Technology Planning & Evaluation(IITP)-ICAN(ICT Challenge and Advanced Network of HRD) grant funded by the Korea government(Ministry of Science and ICT)(IITP-2024-00436954), MSIT(Ministry of Science and ICT), Korea, under the National Program for Excellence in SW supervised by the IITP(Institute of Information & communications Technology Planning & Evaluation)(2021-0-01393), and National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2026-25489409)

요 약

본 연구는 지식 증류에서 모든 샘플을 동일하게 취급하는 기존 방식의 한계를 지적하고, 교사 모델의 예측 신뢰도 분포를 기반으로 샘플별 학습 기여도를 차등 부여하는 가중 지식 증류(WKD)를 제안한다. 신뢰도 분포의 중앙값과 IQR로 가우시안 가중치를 설계하여 dark knowledge가 풍부한 중간 신뢰도 샘플에 높은 비중을, 극단적 신뢰도 샘플에 낮은 비중을 부여한다. 여기에 레이블 교란을 결합해 고신뢰도에 편중된 분포를 분산시킴으로써 가중치 전략의 실효성을 높인다. CNN 및 Transformer 기반 교사 모델 4종을 대상으로 실험한 결과, 10% 교란 조건에서 Acc@5 92.31%를 보이며 기존 지식 증류 대비 성능 향상을 확인하였다. 소거 실험을 통해 신뢰도 기반 가중치와 레이블 교란이 단독 적용보다 결합 적용에서 더 효과적으로 작용함을 확인하였다.

Abstract

This study proposes Weighted Knowledge Distillation (WKD) to overcome the limitation of conventional Knowledge Distillation (KD), which treats all training samples equally regardless of their informational value. Using the median and IQR of the teacher's prediction confidence distribution, a Gaussian weighting function assigns higher weights to mid-confidence samples, where dark knowledge is most abundant, and lower weights to extreme-confidence ones. Label perturbation is added to mitigate the teacher's high-confidence bias and let the weighting operate effectively. Across four CNN- and Transformer-based teachers, WKD achieves an Acc@5 of 92.31% under 10% perturbation, improving over standard KD. Ablation studies show that confidence-based weighting and label perturbation are individually limited but complementary when combined.

Keywords

knowledge distillation, confidence-based weighting, label perturbation, dark knowledge

* 경기대학교 AI컴퓨터공학부(*[3] 교신저자)
- ORCID[1]: https://orcid.org/0009-0007-0713-3665
- ORCID[2]: https://orcid.org/0000-0003-2916-6410
- ORCID[3]: https://orcid.org/0000-0002-0077-6576
** 경기대학교 컴퓨터과학과 석사과정
- ORCID: https://orcid.org/0009-0004-0494-3177

· Received: May 15, 2026, Revised: Jun. 26, 2026, Accepted: Jun. 29, 2026
· Corresponding Author: Namgi Kim
Dept. of AI Computer Science and Engineering, Kyonggi University, 154-42 Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, South Korea
Tel.: +82-31-249-9662, Email: ngkim@kyonggi.ac.kr

Ⅰ. 서 론

심층 신경망은 모델 규모가 커질수록 성능이 향상되는 경향이 있으나, 대규모 모델은 높은 연산량과 메모리를 요구하여 자원이 제한된 환경에 배포하기 어렵다. 지식 증류[1]-[6]는 대규모 교사 모델이 보유한 지식을 소규모 학생 모델로 전이하는 대표적인 모델 압축 기법으로, 컴퓨터 비전과 자연어 처리를 비롯한 다양한 분야에서 모델 경량화를 위해 사용되고 있다. 교사 모델의 softmax 출력으로부터 소프트 레이블(Soft label)을 생성한다. 소프트 레이블에는 정답 클래스에 대한 정보뿐만 아니라, 오답 클래스들 사이의 상대적 유사도, 이른바 dark knowledge가 내포되어 있다. 학생 모델은 이를 통해 원-핫 레이블만으로는 습득하기 어려운 구조적 정보를 학습할 수 있다.

그런데 교사 모델이 모든 샘플에 대해 동일한 수준의 dark knowledge를 제공하는 것은 아니다. 교사 모델이 거의 확신에 가까운 예측을 내리는 샘플의 경우, 소프트 레이블의 확률 분포가 원-핫 레이블에 수렴하기 때문에 클래스 간 관계 정보가 희박하다. 반대로, 교사 모델조차 판별에 혼란을 보이는 샘플에서는 소프트 레이블 자체가 불안정하여 학생 모델에게 노이즈에 가까운 신호를 전달할 가능성이 있다. 결국 교사 모델의 dark knowledge가 가장 풍부하게 담기는 영역은 교사가 적절한 수준의 불확실성을 보이는 중간 신뢰도 구간의 샘플이라 할 수 있다. 그럼에도 불구하고 기존의 지식 증류 프레임워크는 이러한 샘플 간 정보량의 차이를 고려하지 않고, 모든 샘플에 동일한 비중을 부여한다. 그 결과, dark knowledge가 풍부한 유익한 샘플과 정보가 부족하거나 오히려 학습에 방해가 되는 샘플이 손실 함수에 동등하게 기여한다. 이는 비효율적 구조를 초래한다.

본 연구는 이러한 문제 인식에서 출발하여, 교사 모델의 예측 신뢰도 분포를 통계적으로 분석하고 이에 기반하여 샘플별 학습 기여도를 차등적으로 조절하는 가중 지식 증류 방법(WKD, Weighted Knowledge Distillation)을 제안한다. 구체적으로, 전체 훈련 데이터의 신뢰도 분포에서 중앙값과 사분위 범위(IQR, Interquartile Range)를 추출하고, 이를 파라미터로 하는 가우시안 가중치 함수를 설계하여 중간 신뢰도 구간의 샘플에 높은 가중치를, 극단적 신뢰도 구간의 샘플에 낮은 가중치를 부여한다. 한편, 학습 데이터에 의도적으로 노이즈를 추가하는 것이 일반화 성능을 높인다는 연구[7]-[9]가 있다. 본 연구에서는 이에 착안하여 의도적 레이블 교란을 결합하여, 교사 모델의 신뢰도 분포를 적절히 변화시킴으로써 가중치 전략의 실효성을 높인다. 다양한 교사 모델 아키텍처와 레이블 교란 비율에서 수행한 실험과 소거 실험을 통해, 신뢰도 기반 가중치와 레이블 교란이 각각 독립적으로 효과가 제한적이나 결합 시 상호 보완적으로 작용하여 기존 지식 증류 대비 강건하고 우수한 성능을 달성함을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 지식 증류와 레이블 교란 기반 정규화에 대한 관련 연구를 정리하고, 3장에서는 제안하는 신뢰도 기반 가중 지식 증류의 설계와 레이블 교란의 결합 전략을 상술한다. 4장에서는 실험 환경, 교사 모델의 신뢰도 분포 분석, 주요 결과 및 소거 실험을 제시하며, 5장에서 결론과 향후 과제를 논의한다.

Ⅱ. 관련 연구

2.1 지식 증류

지식 증류는 앙상블 모델의 출력을 소규모 모델로 전이하는 개념을 제안한 데서 비롯되었으며[1], 이후 온도 스케일링을 적용한 소프트 레이블 기반의 학습 프레임워크로 체계화되었다[2]. 이후 교사의 중간층 특징을 전이하는 특징 기반 방법[3], 샘플 간 관계 구조를 전이하는 관계 기반 방법[4] 등 지식의 표현 형태를 다양화하는 방향으로 발전해왔다. 최근에는 로짓 기반 증류의 손실 함수를 정답 클래스와 비정답 클래스로 분리하여 각각의 역할을 분석하거나[5], 강한 교사 모델에서의 증류 시 예측 간 상관관계 보존에 초점을 맞추는 연구[6]가 등장하였으나, 이들 역시 개별 훈련 샘플이 담고 있는 dark knowledge의 양이 균일하지 않다는 점, 그리고

이에 따라 샘플별로 학습 기여도를 조절할 필요가 있다는 점은 상대적으로 주목받지 못하였다. 표 1은 기존 지식 증류 방법론들과 제안 방법을 비교한 것으로, 기존 방법들은 지식의 표현 형태 다양화에 집중하여 개별 샘플의 정보량 차이를 고려하지 못한다는 공통적 한계를 지닌다. 본 연구는 이 공백에 주목하여, 교사 모델의 예측 신뢰도를 기준으로 샘플의 정보량을 평가하고 학습 비중을 차등화하는 접근을 취한다.

2.2 의도적 레이블 교란을 활용한 정규화

훈련 레이블에 의도적인 변형을 가하여 과적합을 억제하는 연구도 활발히 이루어져 왔다. Label Smoothing[7]은 원-핫 레이블의 확률 질량을 균등 분포와 혼합하여 모델의 과신을 억제하는 정규화 기법이다. DisturbLabel[8]은 매 훈련 반복에서 일정 비율의 샘플 레이블을 임의의 오답 클래스로 교체하는 방식으로, 서로 다른 노이즈 조건에서 훈련된 다수의 모델을 암묵적으로 앙상블하는 효과를 가진다. DirectionalDisturbLabel[9]은 이를 개선하여, 모델이 높은 확신을 보이는 샘플에 한정하여 방향성 있는 교란을 수행한다. 이러한 기법들은 단독 학습 환경에서의 정규화 효과가 입증되었으나, 지식 증류 환경에서 교사 모델의 소프트 레이블과 어떠한 상호작용을 일으키는지에 대한 분석은 부재하다. 본 연구에서는 레이블 교란이 교사 모델의 신뢰도 분포를 변화시키는 효과에 착안하여, 이를 신뢰도 기반 샘플 가중치와 결합하는 전략을 제시한다.

III. 제안 방법

3.1 개요

본 연구에서 제안하는 가중 지식 증류는 교사 모델이 각 훈련 샘플에 대해 보유한 dark knowledge의 양이 균일하지 않다는 관찰에서 출발한다. 교사 모델의 예측 신뢰도를 dark knowledge의 간접 지표로 활용하고, 전체 데이터셋의 신뢰도 분포를 통계적으로 분석하여 샘플별 학습 기여도를 차등 부여하는 것이 핵심 아이디어이다. 나아가 레이블 교란을 결합함[10]으로써 교사 모델의 신뢰도 분포 변화를 유도하고, 이러한 분포 변화가 신뢰도 기반 가중치 전략의 효과에 미치는 영향을 분석한다.

표 1. 기존 지식 증류 방법론들과 제안 방법의 비교
Table 1. Comparison of previous knowledge distillation methods and the proposed method

Paper	Knowledge / key concept	Target dataset	Teacher	Student	Performance (Representative)
Buciluǎ et al. [1]	Pseudo-data (MUNGE)	8 Tabular datasets (e.g., ADULT, COVTYPE)	Ensemble selection	SIngle neural net	1000x compression
Hinton et al. [2]	Soft targets	MNIST, speech, JFT	Ensemble of DNNs	Single DNN	60.8 (1.9↑ vs. Student)
Romero et al. [3]	Feature maps	CIFAR-10/100, SVHN	Maxout CNN	FitNets	91.61 (1.43↑ vs. Teacher)
Park et al. [4]	Structural relations	CIFAR-100, CUB-200	ResNet-50, etc.	VGG11, etc.	74.66 (3.40↑ vs. Student)
Zhao et al. [5]	Decoupled soft targets	CIFAR-100, ImageNet	ResNet32x4, etc.	ResNet8x4, etc.	76.32 (3.82↑ vs. Student)
Huang et al. [6]	Pearson correlation	CIFAR-100, ImageNet	ResNet-34, etc.	ResNet-18, etc.	72.07 (0.86↑ vs. KD)
Ours	Label perturbation + WKD	Food-101 (10% noise)	EfficientNet-B0, etc.	SqueezeNet1_1	Overall improvements in Acc, F1-Score, etc. (vs. 0% KD)

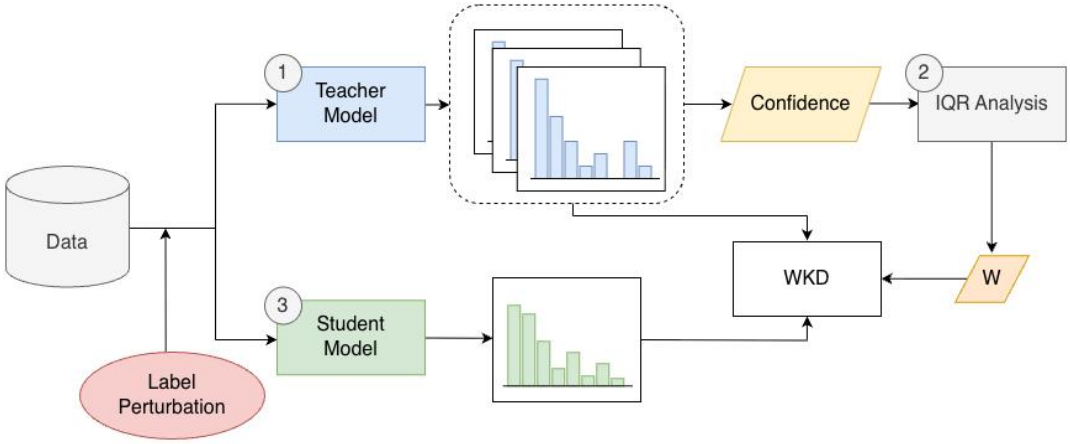

그림 1. 신뢰도 기반 가중치와 레이블 교란을 결합한 지식 증류의 전체 파이프라인
Fig. 1. Overall pipeline of confidence-based weighted knowledge distillation with label perturbation

제안 방법의 전체 구조는 그림 1과 같으며, 구체적인 절차는 다음과 같다. 먼저 훈련 데이터에 레이블 교란을 적용한 후 교사 모델을 훈련(①)한다. 훈련된 교사 모델로부터 온도 스케일링된 softmax를 통해 소프트 레이블을 생성한다. 여기서 온도 스케일링이란, 모델의 출력 로짓을 1보다 큰 온도 매개변수 (τ)로 나누어 출력 확률 분포를 부드럽게 만듦으로써, 정답 외의 클래스들에 대한 교사 모델의 dark knowledge를 더 잘 반영하도록 하는 기법이다. 이와 함께 기본 softmax 출력의 최대 확률값으로 샘플별 예측 신뢰도를 산출한다. 다음으로, 전체 데이터의 신뢰도 분포에 대해 IQR 분석(②)을 수행하고, 이를 기반으로 한 가우시안 가중치 함수에 입력하여 샘플별 가중치를 계산한다. 마지막으로 학생 모델의 훈련 단계(③)에서는 소프트 레이블과의 쿨백-라이블러(KL, Kullback–Leibler) 발산 및 교란된 레이블과의 크로스 엔트로피 손실 모두에 앞서 도출한 가중치를 적용하여 학습을 진행한다.

전체 절차는 다음과 같다. 먼저 훈련 데이터에 레이블 교란을 적용한 후 교사 모델을 훈련한다. 훈련된 교사 모델로부터 온도 스케일링된 softmax를 통해 소프트 레이블을 생성하고, 기본 softmax 출력의 최대 확률값으로 샘플별 예측 신뢰도를 산출한다. 이 신뢰도를 IQR 기반 가우시안 가중치 함수에 입력하여 샘플별 가중치를 계산한다. 학생 모델의 훈련 단계에서는 소프트 레이블과의 KL 발산 및 교란된 레이블과의 크로스 엔트로피 손실 모두에 이 가중치를 적용한다. 제안 방법의 전체 구조는 그림 1과 같다.

3.2 신뢰도 기반 가중치 설계

교사 모델 T의 softmax 출력에서 각 샘플 i에 대해 가장 높은 클래스 확률을 예측 신뢰도 c_i로 정의한다.

$$c_i = \max_k \frac{\exp(z_{T,i,k})}{\sum_j \exp(z_{T,i,j})} \tag{1}$$

식 (1)에서 $z_{T,i,k}$는 i번째 샘플의 클래스 인덱스 k에 대한 로짓이다. c_i가 1에 가까울수록 교사 모델은 해당 샘플을 높은 확신으로 분류하고 있으며, 이 경우 소프트 레이블의 확률 분포가 원-핫 레이블에 수렴하여 클래스 간 관계 정보가 희박해진다. 반대로 c_i가 $1/K$에 가까울수록 교사 모델은 클래스 간 판별에 혼란을 보이고 있으며, 이때의 소프트 레이블은 학생 모델에게 노이즈에 가까운 신호를 전달할 가능성이 있다.

본 연구에서는 단순히 특정 신뢰도 구간을 이진적으로 선별하는 대신, 전체 데이터셋의 신뢰도 분포로부터 중앙값($\tilde{c}$)과 사분위 범위($IQR = Q_3 - Q_1$)를 추출하여 연속적인 가우시안 가중치 함수를 설계한다.

$$w_i = \exp\left(-0.5\left(\frac{c_i - \tilde{c}}{1.5 \times IQR}\right)^2\right) \quad (2)$$

식 (2)는 가우시안 분포의 너비를 결정하는 파라미터를 $1.5 \times IQR$로 설정함으로써, 중앙값 부근의 샘플에 가장 높은 가중치가 집중되고, 양극단으로 갈수록 가중치가 점진적으로 감소하는 구조를 형성한다. 이때, 중앙값과 IQR은 데이터에 의존적으로 결정되므로, 교사 모델이나 데이터셋이 변경되더라도 별도의 하이퍼파라미터 탐색 없이 자동으로 적응하는 이점이 있다.

이 설계의 의도는 다음과 같다. 첫째, 신뢰도가 지나치게 낮은 샘플($c_i \ll \tilde{c}$)은 교사 모델의 판별력이 부족하여 소프트 레이블 자체가 불안정하므로 가중치를 낮추어 학습의 혼선을 방지한다. 둘째, 신뢰도가 과도하게 높은 샘플($c_i \gg \tilde{c}$)은 소프트 레이블이 원-핫에 가까워 dark knowledge가 희박하므로 비중을 줄여 정규화 효과를 유지한다. 셋째, 신뢰도가 중앙값 부근인 샘플($c_i \approx \tilde{c}$)은 교사 모델이 적절한 수준의 불확실성을 보이면서 클래스 간 상관관계 정보를 가장 조밀하게 담고 있는 영역이므로 학습 기여를 극대화한다.

3.3 레이블 교란의 적용

신뢰도 기반 가중치의 효과를 극대화하기 위한 보조 전략으로 레이블 교란을 결합한다. 전체 훈련 샘플 중 비율 η에 해당하는 샘플을 무작위로 선정하여, 해당 샘플의 정답 레이블 $\hat{y}$를 다음의 조건부 확률에 따라 교란된 레이블 y로 변경한다.

$$P(\hat{y} = k | y = j) = \begin{cases} 1-\eta & \text{if } k = j \\ \dfrac{\eta}{K-1} & \text{if } k \neq j \end{cases} \quad (3)$$

식 (3)에서 K는 총 클래스의 수이며, 선택된 샘플은 원래 클래스를 제외한 $K-1$개의 클래스 중 하나로 균등 확률에 의해 재할당된다. 레이블 교란은 교사 모델의 학습 과정에 노이즈를 도입하여 과적합을 억제하는 정규화 효과를 가진다. 그림 2는 ConvNeXt-Tiny[11] 교사 모델에 대해 레이블 교란별 신뢰도 분포를 시각화한 것이다. 교란이 없는 경우 신뢰도가 1.0 부근에 극단적으로 편중되어 있으나, 교란 비율이 증가할수록 분포가 분산되는 경향을 보인다. 이러한 분포 변화는 3.2절의 가우시안 가중치가 유효하게 작동할 수 있는 환경을 조성하며, 구체적인 통계량과 분석은 4.2절에서 상술한다.

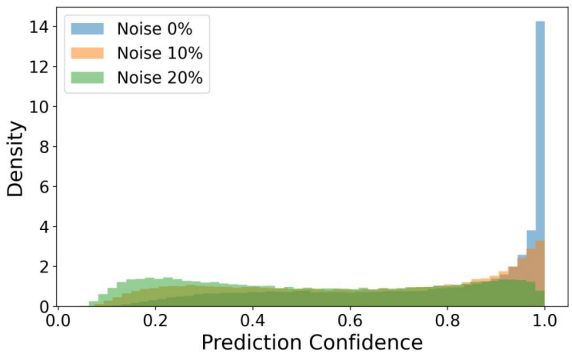

그림 2. 레이블 교란 비율별 신뢰도 분포
Fig. 2. Confidence distribution by noise rate

3.4 목적 함수

학생 모델의 훈련에 사용되는 최종 목적 함수 L_{total}은 다음과 같이 정의된다.

$$L_{total} = \frac{1}{N}\sum_{i=1}^{N} w_i \begin{bmatrix} (1-\alpha)L_{CE,i}(y_i, \sigma(z_{S,i})) \\ +\alpha\tau^2 L_{KL,i}(\sigma(z_{T,i}/\tau), \sigma(z_{S,i}/\tau)) \end{bmatrix} \quad (4)$$

식 (4)에서 L_{CE}는 교란된 레이블 $\hat{y}_i$와 학생 모델의 출력 분포 간의 크로스 엔트로피 손실, L_{KL}은 교사 모델의 소프트 레이블과 학생의 softmax 분포 간의 KL 발산이다. α는 두 손실 항의 균형 계수, τ는 온도 파라미터, w_i는 식 (2)에서 산출한 샘플별 가중치이다. 이 구조를 통해 dark knowledge가 풍부한 중간 신뢰도 샘플의 기여는 강화되고, 정보가 부족하거나 불안정한 극단적 신뢰도 샘플의 영향은 억제된다.

IV. 실험 및 결과

4.1 실험 환경

제안 방법의 유효성을 검증하기 위해 101개의 음식 카테고리로 구성된 Food-101 데이터셋[12]을 사

용하였다. Food-101은 학습용 75,750장과 평가용 25,250장의 이미지를 포함한다. 교사 모델 아키텍처의 다양성을 확보하기 위해 CNN(Convolutional Neural Network) 계열의 ConvNeXt-Tiny, EfficientNet-B0[13], ResNeXt-50[14]과 Transformer 계열의 ViT-B/16[15]을 사용하였으며, 학생 모델로는 경량 아키텍처인 SqueezeNet1_1[16]을 채택하였다. 레이블 교란은 Symmetric noise 방식으로 0%, 10%, 20%의 세 가지로 적용하였으며, 결과의 안정성을 확보하기 위해 복수의 난수 시드로 독립 실험을 수행하고, 그 평균값을 최종 결과로 보고하였다.

교사 모델은 ImageNet[17] 사전학습 가중치를 초기값으로 사용하되, 분류층만 AdamW 옵티마이저[18](학습률 1×10^{-3}, 가중치 감쇠 0.05)로 10에포크 동안 미세 조정하였다. 학생 모델은 AdamW(학습률 1×10^{-4}, 가중치 감쇠 0.05)로 50에포크 동안 훈련하였다. 온도 파라미터 $\tau=3.0$, 균형 계수 $\alpha=0.7$, 배치 사이즈는 128로 설정하였으며, mixed precision(bf16)을 적용하였다. 모든 평가는 레이블 교란이 적용되지 않은 원본 테스트셋에서 Acc@1, Acc@5, F1-Score, Precision, Recall의 5가지 지표로 수행하였다.

4.2 교사 모델의 신뢰도 분포 분석

제안 방법의 핵심 전제, 즉 레이블 교란이 교사 모델의 신뢰도 분포를 변화시킨다는 가설을 검증하기 위해, ConvNeXt-Tiny 교사 모델에 대해 레이블 교란 비율별 신뢰도 분포를 분석하였다. 전체 훈련 데이터에 대해 교사 모델의 softmax 출력에서 최대 클래스 확률을 추출하고, 그 분포를 그림 2에 시각화 하였다. 각 조건에서의 통계량을 표 2에 정리하였다.

교란이 없는 경우(0%) 교사 모델의 신뢰도는 중앙값 0.8683로 대다수 샘플이 고신뢰도 영역에 극단적으로 편중되어 있다. 이는 소프트 레이블이 원-핫 레이블에 가까워 dark knowledge가 희박한 상태임을 의미한다. 10% 교란 환경에서는 중앙값이 0.6634로 이동하고 IQR이 0.4201에서 0.5155로 확대되어, 중간 신뢰도 구간에 분포하는 샘플이 증가하였다. 이는 3.2절에서 설계한 가우시안 가중치가 유효하게 적용할 수 있는 조건이 형성되었음을 의미한다. 반면 20% 환경에서는 중앙값이 0.5157까지 하락하여 교사 모델의 판별력 자체가 저하된 양상을 보인다.

표 2. ConvNeXt-Tiny의 신뢰도 통계량
Table 2. Confidence statistics of ConvNeXt-Tiny

	0%	10%	20%
Median	0.8683	0.6634	0.5157
IQR	0.4201	0.5155	0.5091
Mean	0.7623	0.6284	0.5316

4.3 주요 결과

표 3는 레이블 교란 비율 0%, 10%, 20%에서 제안 방법(WKD)과 기존 지식 증류(KD)의 성능을 비교한 결과이다. 4개의 교사 모델에 대해 Acc@1, Acc@5, F1-Score, Precision, Recall의 5가지 지표를 측정하였으며, 동일한 교사 모델과 교란 비율에서 두 방법을 비교하여 제안 방법의 효과를 확인할 수 있다.

레이블 교란이 없는 환경(0%)에서는 지식 증류와 가중 지식 증류가 전반적으로 유사한 성능을 보였다. 이는 4.2절에서 분석한 바와 같이 교란이 없는 교사 모델의 신뢰도가 1.0 근처에 극단적으로 편중되어, 가우시안 가중치를 적용하더라도 샘플 간 차별화가 이루어지기 어려운 분포 특성에 기인한다.

10% 교란 환경에서는 제안 방법이 네 개의 교사 모델 전반에서 대부분의 평가 지표에 대해 기존 지식 증류를 상회하는 결과를 보였다. Acc@1 기준으로 ConvNeXt-Tiny, EfficientNet-B0, ViT-B/16, ResNeXt-50 모두에서 제안 방법이 기존 지식 증류보다 우수하였으며, 특히 EfficientNet-B0와 ResNeXt-50에서는 5개 지표 전반에서 일관된 성능 향상이 나타났다. 이러한 결과는 제안 방법이 특정 아키텍처에 종속되지 않고 CNN 및 Transformer 기반 교사 모델 모두에 적용 가능함을 보여준다.

20% 교란 환경에서는 교사 모델에 따라 우열이 혼재하여 일관된 경향이 관찰되지 않았다. 이는 4.2절에서 확인한 바와 같이 20% 교란 시 교사 모델의 신뢰도 중앙값이 0.52까지 하락하여 판별력 자체가 저하되었기 때문으로, 과도한 레이블 교란은 오히려 가중치 전략의 효과를 저해함을 시사한다.

표 3. 레이블 교란 비율별 가중 지식 증류와 지식 증류의 성능 비교
Table 3. Performance comparison of weighted knowledge distillation and knowledge distillation by label perturbation rate

Teacher	Noise	Method	Acc@1	Acc@5	F1-Score	Precision	Recall
ConvNeXt-Tiny	0%	WKD	72.91% (-0.21)	92.19% (-0.13)	72.75% (-0.41)	73.34% (-0.5)	72.91% (-0.21)
		KD	**73.12%**	**92.32%**	**73.16%**	**73.84%**	**73.12%**
	10%	WKD	**72.89%** (+0.05)	**92.31%** (+0.11)	**72.87%** (+0.03)	73.61% (-0.05)	**72.89%** (+0.05)
		KD	72.84%	92.20%	72.84%	**73.66%**	72.84%
	20%	WKD	**71.82%** (+0.02)	**91.70%** (+0.11)	**71.87%** (+0.21)	**72.81%** (+0.22)	**71.82%** (+0.02)
		KD	71.80%	91.59%	71.66%	72.59%	71.80%
EfficientNet-B0	0%	WKD	70.53% (-0.06)	**90.63%** (+0.09)	70.43% (-0.09)	71.12% (-0.08)	70.53% (-0.06)
		KD	**70.59%**	90.54%	**70.52%**	**71.20%**	**70.59%**
	10%	WKD	**70.60%** (+0.16)	**90.47%** (+0.06)	**70.53%** (+0.19)	**71.46%** (+0.23)	**70.60%** (+0.16)
		KD	70.44%	90.41%	70.34%	71.23%	70.44%
	20%	WKD	**69.62%** (+0.11)	**90.17%** (+0.09)	69.47% (-0.06)	70.38% (-0.11)	**69.62%** (+0.11)
		KD	69.51%	90.08%	**69.53%**	**70.51%**	69.51%
ViT-B/16	0%	WKD	**72.87%** (+0.09)	**92.20%** (+0.11)	**72.84%** (+0.2)	**73.76%** (+0.28)	**72.87%** (+0.09)
		KD	72.78%	92.09%	72.64%	73.48%	72.78%
	10%	WKD	**72.66%** (+0.14)	**92.12%** (+0.12)	**72.65%** (+0.17)	73.50% (-0.04)	**72.66%** (+0.14)
		KD	72.52%	92.00%	72.48%	**73.54%**	72.52%
	20%	WKD	**71.44%** (+0.03)	**91.34%** (+0.02)	**71.49%** (+0.05)	72.46% (-0.07)	**71.44%** (+0.03)
		KD	71.41%	91.32%	71.44%	**72.53%**	71.41%
ResNeXt-50	0%	WKD	71.18% (-0.23)	91.16% (-0.01)	71.06% (-0.24)	71.88% (+0.12)	71.18% (+0.23)
		KD	**71.41%**	**91.17%**	**71.30%**	**72.20%**	**71.41%**
	10%	WKD	**71.12%** (+0.49)	**91.06%** (+0.03)	**71.11%** (+0.58)	**72.08%** (+0.61)	**71.12%** (+0.49)
		KD	70.63%	91.03%	70.53%	71.47%	70.63%
	20%	WKD	**69.90%** (+0.15)	**90.06%** (+0.12)	**69.78%** (+0.13)	70.81% (-0.17)	**69.90%** (+0.15)
		KD	69.75%	89.94%	69.65%	**70.98%**	69.75%

표 4. 신뢰도 기반 가중치와 레이블 교란의 소거 실험
Table 4. Ablation study on the effects of confidence-based weighting and label perturbation

Setting	Acc@1	Acc@5	F1-Score	Precision	Recall
w/o Label Perturbation	70.53%	**90.63%**	70.43%	71.12%	70.53%
w/o Weighting	70.44%	90.41%	70.34%	71.23%	70.44%
w/o Label Perturbation, Weighting	70.59%	90.54%	70.52%	71.20%	70.59%
Ours	**70.60%**	90.47%	**70.53%**	**71.46%**	**70.60%**

4.4 구성 요소별 소거 실험

4.3절에서 제안 방법이 10% 교란 환경에서 가장 효과적으로 작동함을 확인하였다. 본 절에서는 이러한 성능 개선이 레이블 교란과 신뢰도 가중치의 결합에서 비롯된 것인지, 혹은 개별 요소만으로도 달성 가능한 것인지를 검증한다. 10% 환경에서 성능 개선이 가장 안정적으로 나타난 EfficientNet-B0를 대표 교사 모델로 선정하고, 레이블 교란 유무와 신뢰도 가중치 유무에 따른 네 가지 조합을 비교하여 결과를 표 4에 정리하였다.

레이블 교란 없이 신뢰도 가중치만 적용한 경우는 기존 지식 증류 대비 오히려 소폭 하락하는 결과를 보였다. 4.2절에서 확인한 바와 같이 교란이 없는 교사 모델의 신뢰도는 1.0 부근에 극단적으로 편중되어 있어, 가우시안 가중치가 대다수 샘플에 낮은 가중치를 부여하게 되며 이로 인해 유효한 학습 신호까지 억제되는 부작용이 발생한 것으로 판단된다.

신뢰도 가중치 없이 레이블 교란만 적용한 경우도 기존 대비 뚜렷한 개선을 보이지 않았다. 이는 레이블 교란이 신뢰도 분포를 중간 구간으로 분산

시키는 효과를 가지지만, 모든 샘플에 동일한 비중을 부여하는 기존 지식 증류 프레임워크에서는 분산된 분포의 이점을 활용하지 못하기 때문으로 판단된다.

반면, 두 요소를 결합한 제안 방법에서는 대부분의 평가 지표에 걸쳐 가장 우수한 결과가 나타났다. 이 결과는 4.2절의 신뢰도 분포 분석과 종합하여 다음과 같이 해석할 수 있다. 레이블 교란은 교사 모델의 신뢰도 분포를 고신뢰도 편중 상태에서 중간 신뢰도 구간이 확대된 형태로 변화시킨다. 이렇게 분산된 분포 위에서 가우시안 가중치가 적용되면, dark knowledge가 풍부한 중간 신뢰도 샘플의 기여는 강화되고 극단적 샘플의 영향은 억제되는 선별 효과가 발생한다. 즉, 레이블 교란은 가중치 전략이 작동할 수 있는 분포적 환경을 조성하고, 가중치 전략은 교란 환경에서 발생하는 노이즈의 부정적 영향을 억제하는 상호 보완적 관계에 있다.

Ⅴ. 결론 및 향후 과제

본 연구에서는 지식 증류 과정에서 교사 모델이 각 훈련 샘플에 대해 제공하는 dark knowledge의 양이 균일하지 않다는 점에 착안하였다. 이를 바탕으로 샘플별 학습 기여도를 차등 조절하는 가중 지식 증류를 제안하였다. 교사 모델의 예측 신뢰도 분포를 IQR로 분석하여 가우시안 가중치 함수를 설계하고, 레이블 교란을 결합하였다.

교사 모델의 신뢰도 분포를 시각화한 결과, 레이블 교란이 없는 교사 모델은 신뢰도가 1.0 부근에 극단적으로 편중되어 dark knowledge가 희박하였다. 반면, 10% 교란 환경에서는 중간 신뢰도 구간이 확대되어 가우시안 가중치가 유효하게 작용할 수 있는 분포가 형성됨을 확인하였다. 네 가지 교사 아키텍처에서의 실험 결과, 10% 레이블 교란 조건에서 제안 방법이 기존 지식 증류 대비 전반적으로 우수한 성능을 달성하였으며, 이러한 경향은 CNN 및 Transformer 기반 교사 모델 모두에서 공통적으로 관찰되었다. 소거 실험을 통해 신뢰도 가중치와 레이블 교란이 개별적으로는 성능 개선에 기여하지 못하거나 오히려 하락을 유발하나, 결합 시 상호 보완적 시너지를 발생시킴을 확인하였다. 이는 레이블 교란이 가중치 전략에 적합한 분포적 환경을 조성하고, 가중치 전략이 교란 환경의 노이즈를 억제하는 구조적 상보성을 형성하기 때문으로 해석된다.

향후에는 다양한 도메인과 규모의 데이터셋, 상이한 용량의 학생 모델로 실험을 확대하고, 훈련 에포크에 따라 가중치를 적응적으로 조절하는 동적 전략을 탐구할 계획이다. 나아가, 본 연구에서 제안한 신뢰도 기반 가중치 프레임워크는 교사 모델의 신뢰도 분포를 변화시킬 수 있는 다양한 정규화 기법과의 결합과 로짓 기반 증류 외에 특징 기반 증류로의 확장도 검토할 수 있을 것이다.

References

[1] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression", Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, pp. 535-541, Aug. 2006. https://doi.org/10.1145/1150402.1150464.

[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network", arXiv preprint arXiv:1503.02531, Mar. 2015. https://doi.org/10.48550/arXiv.1503.02531.

[3] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for Thin Deep Nets", Proc. of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, May 2015. https://doi.org/10.48550/arXiv.1412.6550.

[4] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational Knowledge Distillation", Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 3962-3971, Jun. 2019. https://doi.org/10.1109/cvpr.2019.00409.

[5] B. Zhao, Q. Cui, R. Song, Y. Qiu, and J. Liang, "Decoupled Knowledge Distillation", Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 11953-11962, Jun. 2022.

https://doi.org/10.1109/cvpr52688.2022.01165.

[6] T. Huang, S. You, F. Wang, C. Qian, and C. Xu, "Knowledge Distillation from A Stronger Teacher", Proc. of the Advances in Neural Information Processing Systems (NeurIPS), New Orleans, Louisiana, USA, Vol. 35, pp. 33716-33727, Dec. 2022. https://doi.org/10.48550/arXiv.2205.10536.

[7] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 2818-2826, Jun. 2016. https://doi.org/10.1109/cvpr.2016.308.

[8] L. Xie, J. Wang, Z. Wei, M. Wang, and Q. Tian, "DisturbLabel: Regularizing CNN on the Loss Layer", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, Nevada, USA, pp. 4753-4762, Jun. 2016. https://doi.org/10.48550/arXiv.1605.00055.

[9] Y. Kim, H. Lukashonak, P. Tarepakdee, K. Zavalich, and M. Arif, "Disturbing Target Values for Neural Network Regularization", arXiv preprint arXiv:2110.05003, Oct. 2021. https://doi.org/10.48550/arXiv.2110.05003.

[10] D. Koh, T. Kim, S. Lee, and N. Kim, "Weighted Knowledge Distillation with DisturbLabel for Robust Learning", Proc. of the 17th International Conference on Ubiquitous and Future Networks (ICUFN), Milan, Italy, Jul. 2026.

[11] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 11976-11986, Jun. 2022. https://doi.org/10.1109/cvpr52688.2022.01167.

[12] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – Mining Discriminative Components with Random Forests", Proc. of the European Conference on Computer Vision (ECCV), Cham, Switzerland, pp. 446-461, Sep. 2014. https://doi.org/10.1007/978-3-319-10599-4_29.

[13] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", Proc. of the 36th International Conference on Machine Learning (ICML), Long Beach, CA, USA, Vol. 97, pp. 6105-6114, Jun. 2019. https://doi.org/10.48550/arXiv.1905.11946.

[14] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 1492-1500, Jul. 2017. https://doi.org/10.1109/cvpr.2017.634.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", Proc. of the 9th International Conference on Learning Representations (ICLR), Online, May 2021. https://doi.org/10.48550/arXiv.2010.11929.

[16] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-Level Accuracy with 50x Fewer Parameters and <0.5MB Model Size", arXiv preprint arXiv:1602.07360, Feb. 2016. https://doi.org/10.48550/arXiv.1602.07360.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database", Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, pp. 248-255, Jun. 2009. https://doi.org/10.1109/CVPR.2009.5206848.

[18] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization", Proc. of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, May 2019. https://doi.org/10.48550/arXiv.1711.05101.

저자소개

고 도 영 (Doyoung Koh)

2021년 3월 ~ 현재 : 경기대학교
인공지능전공 학부과정
관심분야 : 딥러닝, 인공지능,
효율적 학습

김 태 훈 (Taehoon Kim)

2025년 8월 : 경기대학교
인공지능전공(학사)
2025년 9월 ~ 현재 : 경기대학교
컴퓨터과학과 석사과정
관심분야 : 컴퓨터 비전, 딥러닝,
인공지능

안 준 호 (Junho Ahn)

2006년 2월 : 홍익대학교
컴퓨터공학과(공학사)
2008년 2월 : 연세대학교
컴퓨터공학과(공학석사)
2013년 12월 : University of
Colorado Boulder
컴퓨터과학과(공학박사)
2013년 12월 ~ 2017년 1월 : 한국전자통신연구원(ETRI)
선임연구원
2017년 2월 ~ 2023년 2월 : 한국교통대학교
컴퓨터공학전공 부교수
2023년 3월 ~ 현재 : 경기대학교 AI컴퓨터공학부 부교수
관심분야 : 컴퓨터 비전, 딥러닝, 인공지능

김 남 기 (Namgi Kim)

1997년 2월 : 서강대학교 전산학
(공학사)
2000년 2월 : KAIST 전산학
(공학석사)
2005년 2월 : KAIST 전산학
(공학박사)
2007년 2월 : 삼성전자 통신연구소
책임연구원
2007년 3월 ~ 현재 : 경기대학교 AI컴퓨터공학부 교수
관심분야 : 통신시스템, 네트워크