JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 28 No. 4, pp. 31-39, April 2023
https://doi.org/10.9708/jksci.2023.28.04.031

Knowledge Distillation based-on Internal/External Correlation Learning

Hun-Beom Bak*, Seung-Hwan Bae**

*M. S. candidate, Vision & Learning Lab, Inha University, Incheon, Korea
**Associate Professor, Vision & Learning Lab, Dept. of Computer Engineering, Inha University, Incheon, Korea

[Abstract]

In this paper, we propose an Internal/External Knowledge Distillation (IEKD), which utilizes both external correlations between feature maps of heterogeneous models and internal correlations between feature maps of the same model for transferring knowledge from a teacher model to a student model. To achieve this, we transform feature maps into a sequence format and extract new feature maps suitable for knowledge distillation by considering internal and external correlations through a transformer. We can learn both internal and external correlations by distilling the extracted feature maps and improve the accuracy of the student model by utilizing the extracted feature maps with feature matching. To demonstrate the effectiveness of our proposed knowledge distillation method, we achieved 76.23% Top-1 image classification accuracy on the CIFAR-100 dataset with the "ResNet-32×4/VGG-8" teacher and student combination and outperformed the state-of-the-art KD methods.

▸ **Key words**: Knowledge distillation, model compression, transformer, correlation learning, Image classification

[요 약]

본 논문에서는 이종 모델의 특징맵 간 상관관계인 외부적 상관관계와 동종 모델 내부 특징맵 간 상관관계인 내부적 상관관계를 활용하여 교사 모델로부터 학생 모델로 지식을 전이하는 Internal/External Knowledge Distillation (IEKD)를 제안한다. 두 상관관계를 모두 활용하기 위하여 특징맵을 시퀀스 형태로 변환하고, 트랜스포머를 통해 내부적/외부적 상관관계를 고려하여 지식 증류에 적합한 새로운 특징맵을 추출한다. 추출된 특징맵을 증류함으로써 내부적 상관관계와 외부적 상관관계를 함께 학습할 수 있다. 또한 추출된 특징맵을 활용하여 feature matching을 수행함으로써 학생 모델의 정확도 향상을 도모한다. 제안한 지식 증류 방법의 효과를 증명하기 위해, CIFAR-100 데이터 셋에서 "ResNet-32×4/VGG-8" 교사/학생 모델 조합으로 최신 지식 증류 방법보다 향상된 76.23% Top-1 이미지 분류 정확도를 달성하였다.

▸ **주제어**: 지식 증류, 모델 압축, 트랜스포머, 상관관계 학습, 이미지 분류

• First Author: Hun-Beom Bak, Corresponding Author: Seung-Hwan Bae
*Hun-Beom Bak (h2hb0307@inha.edu), Vision & Learning Lab, Inha University
**Seung-Hwan Bae (shbae@inha.ac.kr), Vision & Learning Lab, Dept. of Computer Engineering, Inha University
• Received: 2023. 03. 07, Revised: 2023. 03. 24, Accepted: 2023. 04. 03.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

최근에 신경망이 발전함에 따라 컨볼루션 네트워크로 구성된 깊은 신경망이 이미지 분류[1], 객체 검출[2] 같이 다양한 컴퓨터 비전 분야에서 높은 성능을 달성하고 있다. 하지만 신경망 모델을 자원 환경 제약이 GPU 서버보다 많은 임베디드 기기에 활용하기 위해서는 고성능의 경량화된 모델이 필요하다. 가지치기[3] 및 양자화[4]를 통해 경량화된 모델을 획득할 수 있으나, 정확도가 감소하는 문제점이 있다. 감소된 정확도를 복원하기 위해서 미리 학습된 고성능의 교사 모델을 경량화된 학생 모델이 모방하도록 하는 지식 증류(Knowledge distillation) 방법이 주목받고 있다.

지식 증류는 [5]에서 제안된 방법으로, 교사 모델과 학생 모델의 소프트 맥스 출력인 소프트 라벨 간의 손실을 최소화함으로써 학생 모델이 교사 모델을 모방할 수 있도록 학습하는 방법이다. 소프트 라벨뿐만 아니라, 교사 모델과 학생 모델의 특징맵 간 거리를 최소화하는 지식 증류 방법[6]이 제안되었다. 교사 모델과 학생 모델 간 네트워크 구조가 다를 때, 최소화할 특징맵을 임의로 선택하면 이종 모델간 semantic 정보가 다르기 때문에 정확도가 하락할 수 있다. 정확도 하락을 방지하기 위해 교사 모델과 학생 모델 간 이종 모델의 상관관계(Correlation)인 외부적(External) 상관관계를 가중치로 사용하여 이종 모델 간 특징맵의 유사도를 최대화할 수 있는 SemCKD [7] 방법이 제안되었다.

그러나, 두 모델 간의 특징맵 사이의 거리를 최소화하는 방법으로는 내부적(Internal) 상관관계를 활용하지 못하는 문제점이 있다. 내부적 상관관계는 동종 모델 내부 특징맵 간 상관관계이며 diversity 정보가 포함되어 있다. Diversity는 모델 내부의 두 레이어(Layer) 간 상관관계이며, 두 레이어 간 무관계성과 특징맵이 풍부한 정보를 포함하고 있음을 나타낸다[8]. 이러한 diversity로 알 수 있는 풍부한 특징맵 정보는 CNN (Convolutional Neural Network) 모델의 성능을 보장한다[9]. 따라서 잘 학습된 교사 모델의 두 레이어 간의 상관관계 또한 학생 모델이 모방할 수 있도록 외부적 상관관계 또한 활용해야 한다[10].

이 논문에서는 지식 증류 과정에서 내부적 상관관계와 외부적 상관관계를 모두 활용하기 위해 트랜스포머(Transformer)[11] 기반의 Internal/External Knowledge Distillation (IEKD)를 제안한다. 제안하는 방법은 교사 모델과 학생 모델의 특징맵을 시퀀스 형태로 임베딩한다. 임베딩 된 특징맵을 트랜스포머 인코더를 통해 동일 모델 간 internal 상관관계와 디코더를 통해 이종 모델 간 external 상관관계를 활용하여 지식 증류를 위한 특징맵을 추출한다. 기존 이종 모델 간 특징맵 사이의 거리를 측정하는 손실 함수뿐만 아니라, 추가적으로 트랜스포머를 통하여 추출된 교사와 학생 모델의 특징맵 간 거리를 측정하는 손실 함수를 추가하여 지식 증류된 학생 모델의 정확도 향상을 도모한다.

본 연구의 기여를 정리하면 다음과 같다: (1) 트랜스포머를 사용하여 내부적(Internal) 상관관계와 외부적(External) 상관관계를 모두 활용하여 지식 증류를 위한 특징맵을 추출하는 Internal/External Knowledge Distillation (IEKD)를 제안한다. (2) 교사와 학생 모델의 중간 레이어의 출력 특징맵을 시퀀스 형태로 변형 가능한 임베딩 함수를 제안한다. (3) 이미지 분류 데이터 셋인 CIFAR-100 [12] 데이터 셋에서 제안된 IEKD와 다른 논문의 정확도를 비교 분석을 진행하였다. 제안된 IEKD는 CNN 기반의 다양한 교사 모델과 학생 조합에서 다른 연구보다 높은 정확도를 달성하였다. 모델 절제 실험을 통하여 IEKD의 모든 구성 요소를 사용하였을 때, 정확도가 가장 높게 나옴으로써, 제안한 구성 요소들의 효율성을 증명한다.

본 논문은 2장에서 제안한 지식 증류 방법과 관련된 기존 연구들에 대해 설명한다. 3장에서는 본 논문에서 제안하는 IEKD에 대해 설명한다. 4장에서는 CIFAR-100 데이터 셋을 통해 최신 지식 증류 연구와 정확도를 비교 및 절제 실험을 수행한다. 5장에서 본 논문의 결론에 대해 작성하는 것으로 구성된다.

II. Related Works

이 장에서는 본 연구와 연관된 기존 연구에 대해 기술한다.

1. Knowledge Distillation

지식 증류는 [5]에서 제안된 방법으로, 경량화된 학생 모델이 고성능의 교사 모델의 성능을 달성하기 위해 사용되며, 객체 검출[13], 음성 분류[14] 등 다양한 분야에서 사용된다. 지식 증류의 메커니즘은 학생 모델이 교사 모델의 예측 확률을 소프트 라벨 (soft label)로 학습한다. 하지만 두 모델 간 capacity 차이가 큰 경우, 학습이 실패하는 문제점이 있다. 이를 위해서는 소프트 라벨을 타겟 클래스와 논 타겟 클래스로 분리하여 지식 증류를 수행[15]하거나, 교사 모델의 분류기를 그대로 재사용[16]하거나, 교사 모델과 학생 모델 간 동일한 데이터 증강을 적용하는 function matching [17] 연구나 관계성을 증류하는 연구[18]가 제안되었다.

2. Feature-Map Based Knowledge Distillation

소프트 라벨 기반의 지식 증류 방법은 교사 모델과 학생 모델 간의 성능 격차가 존재했다. 이 격차를 줄이기 위해 소프트 라벨만이 아니라 교사와 학생의 중간 레이어의 특징맵과 특징맵의 변형들을 사용하는 지식 증류 방법들 [6-7, 10, 19-23]이 제안되었다. FitNet [6]에서는 특징맵 간 거리를 최소화하는 지식 증류 방법이 제안되었다. AT [19]는 spatial attention map을 증류하는 attention transfer를 제안하였다. SP [20]는 같은 입력 샘플에 대해 교사 모델과 학생 모델이 동일한 activation을 같도록 하는 지식 증류 방법을 제안한다. VID [21]는 교사 모델과 학생 모델 간 mutual information을 최대화할 수 있는 지식 증류 방법을 제안한다. HKD [22]는 교사 모델의 flow를 학생 모델이 학습하도록 한다. SemCKD [7]는 교사 모델과 학생 모델의 특징맵 간 어텐션 메커니즘을 통해 산출한 external 상관관계를 이용한다. TaT [23]는 spatial한 pixel끼리 one-to-one으로 매칭하는 대신, one-to-many로 매칭 시키는 방법을 제안한다. 위에서 제안된 방법들은 교사-학생 관계인 이종 모델끼리의 상관관계만 활용하고 있으며, 모델 내부의 상관관계는 포착할 수 없다. ICKD [10]는 동일 모델 내부 특징맵의 채널 간 상관관계를 증류하여 교사 모델로부터 동종 모델의 레어어간 상관관계인 diversity를 학생 모델이 학습하지만 external 관계성을 활용하지는 못한다. 위 방법들과 달리 제안된 IEKD는 트랜스포머를 활용하여 external 상관관계와 internal 상관관계를 모두 활용할 수 있다. External 상관관계를 통해 학생 모델이 교사 모델의 특징맵을 모방하도록 학습할 뿐만 아니라, internal 상관관계를 활용함으로써, 동종 모델의 내부 레이어 간 상관관계도 학습하여 결과적으로 학생 모델 정확도를 향상한다.

III. Methodology

이 장에서는 기존의 지식 증류 방법과 본 논문에서 제안하는 IEKD의 각 구성 요소들(Transformer, distillation loss, feature matching)에 대해 자세히 설명한다.

1. Vanilla Knowledge Distillation

본 장에서는 [5]에서 제안된 지식 증류에 대해 설명한다. 지식 증류는 교사 모델의 소프트 맥스 $\sigma(\cdot)$의 출력인 소프트 라벨과 학생 모델의 예측 확률 사이의 손실을 최소화한다. 일반적으로 손실 함수는 학생 모델의 예측 확률과 라벨 $\mathbf{y}_{GT}$ 간 cross-entropy loss (CE)와 학생과 교사 모델 출력 사이의 Kullback-Leibler divergence (KL)로 구성되어 있으며 식 (1)처럼 정의된다.

$$L_{KD} = \tau^2 \mathrm{KL}(\sigma(\mathbf{g}_t/\tau), \sigma(\mathbf{g}_s/\tau)) + \mathrm{CE}(\mathbf{g}_s, \mathbf{y}_{GT}) \quad (1)$$

식 (1)에서 $\mathbf{g}_s$와 $\mathbf{g}_t$는 교사 모델과 학생 모델의 소프트맥스 직전 레이어 출력이며, τ는 temperature factor이다.

2. Internal/External Knowledge Distillation

본 장에서는 트랜스포머[11]를 사용하여, internal 상관관계와 external 상관관계를 활용할 수 있는 트랜스포머 기반 지식 증류 방법에 대해 설명하며 자세한 구조는 Fig. 1의 (a)에 나타나 있다. Fig. 1의 (a)에서 교사 모델의 특징맵과 교사 모델의 시퀀스 특징맵 및 internal/external 상관관계를 모두 고려한 특징맵은 초록색으로 묘사하였다. 학생 모델의 특징맵과 학생 모델의 시퀀스 특징맵 및 internal/external 상관관계를 모두 고려한 특징맵은 노란색으로 묘사하였다. 각 모델의 특징맵을 시퀀스 형태의 특징맵으로 임베딩하여 트랜스포머에 입력함으로써, 두 상관관계를 고려한 특징맵을 획득한다. 트랜스포머에서 추출된 교사/학생 모델의 특징맵 간 손실 함수를 최소화함으로써 지식 증류를 수행한다.

트랜스포머는 자연어 처리(NLP)를 위해 제안된 구조로, 인코더-디코더 구조로 구성된 모델이다. 인코더에서 입력 시퀀스를 입력받아 셀프 어텐션을 통해 내부적인 (internal) 상관관계를 고려한다. 디코더는 셀프 어텐션과 멀티 헤드 어텐션으로 구성되어 있으며, 셀프 어텐션을 통해 타겟 시퀀스를 입력 받아 내부적 상관관계를 고려한다. 인코더의 출력과 타겟 시퀀스의 셀프 어텐션 출력 사이의 멀티 헤드 어텐션을 통해 두 시퀀스 간 외부적인 상관관계를 고려한 특징을 추출한다. IEKD는 교사 모델과 학생 모델의 특징맵들을 시퀀스 형태로 변환하여 트랜스포머를 적용시켜 지식 증류를 위한 특징맵을 추출한다.

트랜스포머에 각 모델의 특징맵을 입력하기 위한 시퀀스 형태로 변환하기 위해, 식 (2)처럼 교사 모델과 학생 모델의 특징맵 $F_{t_i} \in \mathbb{R}^{B\times C_{t_i}\times H_{t_i}\times W_{t_i}}$와 $F_{s_j} \in \mathbb{R}^{B\times C_{s_j}\times H_{s_j}\times W_{s_j}}$를 함수 $\psi(\cdot)$를 통해 시퀀스 매트릭스 $V_{t_i} \in \mathbb{R}^{B\times E}$와 $V_{s_j} \in \mathbb{R}^{B\times E}$로 임베딩 한다. $\psi(\cdot)$는 Fig. 2처럼 1×1 컨볼루션 레이어와 ReLU 활성화 함수, L2 normalize layer, fully connected 레이어로 구성되어 있다. i 와 j는 교사 모델과 학생 모델 레이어의 인덱스를 나타내며, B는 배치

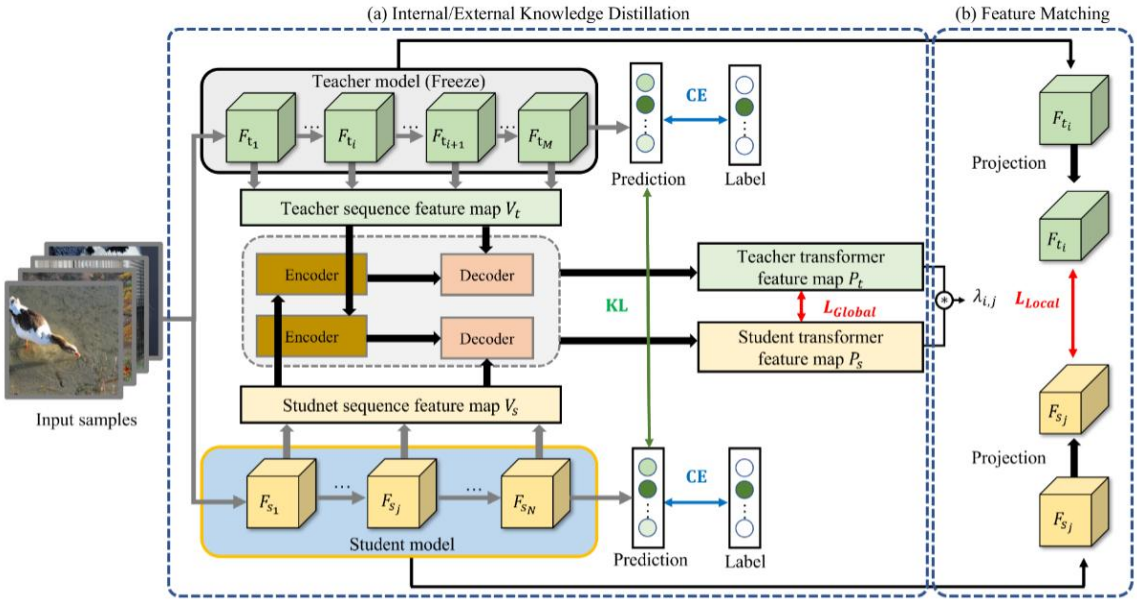

Fig. 1. Proposed framework for internal and external knowledge distillation

사이즈를 나타낸다. C_{t_i}와 W_{t_i}, H_{t_i}는 교사 모델의 i번째 특징맵의 채널의 수, 너비, 높이를 나타내며, C_{s_j}와 W_{s_j}, H_{s_j}는 학생 모델의 j번째 특징맵의 채널의 수, 너비, 높이를 나타낸다.

$$V_{t_i} = \psi(F_{t_i}) \quad (2)$$
$$V_{s_j} = \psi(F_{s_j})$$

식 (2)에서 획득한 V_{t_i} 와 V_{s_j}를 쌓아서 시퀀스 특징맵 $V_t \in \mathbb{R}^{B\times M\times E}$와 $V_s \in \mathbb{R}^{B\times N\times E}$를 만들 수 있다. M과 N은 교사 모델과 학생 모델의 레이어 수를 나타낸다. E는 임베딩 매트릭스의 차원을 나타낸다.

두 상관관계를 고려한 특징맵을 획득하기 위해 V_s는 트랜스포머 인코더 Enc에 입력되고, V_t는 트랜스포머 디코더 Dec의 셀프 어텐션에 입력된다. 트랜스포머 디코더의 멀티 헤드 어텐션에 트랜스포머 인코더 Enc의 출력과 디코더 Dec의 셀프 어텐션 출력을 사용하여, internal 상관관계와 external 상관관계를 모두 고려된 교사 모델의 특징맵 $P_t \in \mathbb{R}^{B\times M\times E}$를 획득할 수 있다.

$$P_t = Dec(Enc(V_s), \mathrm{V}_t) \quad (3)$$

같은 방법으로 식 (4)처럼 internal 상관관계와 external 상관관계를 모두 고려된 학생 모델의 특징맵 $P_s \in \mathbb{R}^{B\times N\times E}$를 얻을 수 있다.

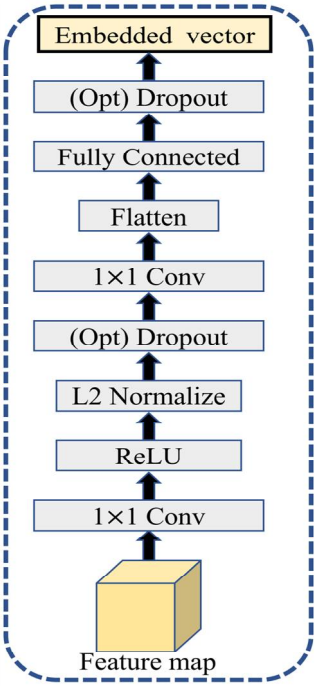

Fig. 2. A network architecture for feature embedding

$$P_s = Dec(Enc(V_t), \mathrm{V}_s) \quad (4)$$

P_s와 P_s의 전치 행렬 간 내적과 P_t와 P_t의 전치행렬 간 내적의 평균 제곱 오차(Mean Square Error, MSE)를 손실 함수 L_{Global}로 식 (5)처럼 사용한다.

$$L_{Global} = \| P_t \cdot P_t^{\mathrm{T}} - P_s \cdot P_s^{\mathrm{T}} \|^{2_2} \quad (5)$$

3. Feature Matching

본 장에서는 Fig. 1의 (b)에서 확인 할 수 있듯이 교사/학생 모델의 특징맵을 투사하여 차원 및 공간적인 형태를 일치시키고, external 상관관계를 가중치로 사용하여 두 모델의 특징맵 간 거리를 최소화하는 SemCKD [7] 기반의 feature matching에 대해 설명한다. SemCKD는 이종 모델의 특징맵 간의 external 상관관계를 특징맵 간 손실 함수의 가중치로 사용하여 상관관계가 높은 특징맵끼리 유사해지도록 한다. 교사 모델과 학생 모델 사이의 external 상관관계를 구하기 위해 앞선 3.2장에서의 P_t와 P_s를 사용한다. 교사 모델의 i번째 레이어과 학생 모델의 j번째 레이어의 external 상관관계 $\boldsymbol{\lambda}_{i,j}$를 계산하기 위해 식 (6)을 사용한다.

$$\boldsymbol{\lambda}_{i,j} = \frac{\exp(p_{t_i} \cdot p_{s_j}^{\mathrm{T}})}{\sum_{i=1}^{N}\sum_{j=1}^{M}\exp(p_{t_i} \cdot p_{s_j}^{\mathrm{T}})} \quad (6)$$

$\boldsymbol{\lambda}_{i,j}$는 두 레이어의 트랜스포머로부터 추출된 특징맵이 얼마나 유사한지 나타내며, $p_{t_i} \in \mathbb{R}^{B\times E}$와 $p_{s_j} \in \mathbb{R}^{B\times E}$는 교사 모델의 i번째 레이어의 해당하는 트랜스포머 출력 특징맵과 학생 모델의 j번째 레이어의 해당하는 트랜스포머 출력 특징맵을 나타낸다.

특징맵 간 거리를 최소화하기 위해, 식 (7)처럼 $\boldsymbol{\lambda}_{i,j}$로 두 특징맵 F_{t_i}와 F_{s_j} 간의 손실 함수를 스케일링 함으로써, external 상관관계가 높은 특징맵끼리 지식 증류가 수행되도록 한다.

$$L_{Local} = \sum_{i=1}^{N}\sum_{j=1}^{M}\boldsymbol{\lambda}_{i,j} \parallel \phi(F_{t_i}) - \phi(F_{s_j}) \parallel^{2_2} \quad (7)$$

특징맵 F_{t_i}와 F_{s_j}의 채널과 공간적(Spatial) 크기를 일치시키기 위해 다층 퍼셉트론과 적응적 풀링(Adaptive Pooling)으로 구성된 변환 함수 $\phi(\cdot\ )$을 통과시킨다.

전체적인 손실 함수 $L_{Distill}$은 아래 식 (8)과 같다

$$L_{Distill} = L_{KD} + \zeta L_{Global} + \beta L_{Local} \quad (8)$$

이 때, 식 (8)에서 β와 ζ은 손실 함수 L_{Local}과 L_{Global}를 스케일링하는 각 scaling factor이다.

IV. Experiments

이번 장에서는 이 논문에서 제안한 IEKD 방법의 성능 평가를 진행하였으며, 이를 위한 실험 설정과 타 지식 증류 방법과 비교 결과(Comparison result)에 대해 논의한다. 또한, 제안한 IEKD의 각 구성 요소에 대한 절제 실험(Ablation study)과 scaling factor ζ와 트랜스포머 레이어 개수 변화에 따른 학생 모델의 Top-1 정확도를 분석하여 민감도 분석(Sensitivity Analysis)에 대해서도 논의한다.

1. Experiment setting

본 논문에서 제안하는 IEKD 기법의 효율성을 증명하기 위해 이미지 분류 데이터 셋인 CIFAR-100 [12]을 활용하여 실험을 수행하였다. CIFAR-100은 100개의 클래스와 60,000장의 샘플로 구성되어 있다. 각 클래스는 32×32 해상도의 학습 샘플 500장과 평가 샘플 100장으로 구성되어 있다. 본 논문에서는 모델의 정확도를 비교하기 위해, 주어진 평가 샘플을 통해 Top-1 정확도(Top-1 Accuracy)를 측정한다. Top-1 정확도는 전체 샘플 중 학습한 모델의 가장 높은 클래스 예측 값 중 실제 클래스와 일치한 샘플의 비율이다. 교사 모델과 학생 모델 및 다른 지식 증류 방법의 Top-1 정확도는 SemCKD [7]와 ICKD [10] 및 TaT [23] 논문의 실험 결과를 인용하였으며, IEKD는 CIFAR-100 데이터 셋에서 1회 측정하였다.

IEKD 기법은 [7]을 베이스라인으로 하여 구현되었다. 제안하는 IEKD 기법에서 internal 상관관계와 external 상관관계를 활용하기 위한 트랜스포머의 인코더와 디코더는 각각 6개의 레이어와 8개의 헤드로 구성된다. 임베딩 매트릭스의 차원 E는 16으로 설정하였다. 공정한 비교 평가를 위해 타 논문과 동일한 딥러닝 라이브러리(PyTorch [24]) 및 하이퍼 파라미터로 모델을 학습한다. 따라서 배치 사이즈 B는 64로 설정하였으며, 식 (1)에서 temperature factor τ는 4로 설정하였다. 식 (8)에서 L_{Local}을 scaling 하기 위한 factor β는 400으로 설정하였다. 학습률(Learning rate)은 0.05로 설정하였으며, 150, 180, 210 에포크(Epoch)마다 학습률을 10분의 1로 감소시켰다. 최적화 기법은 stochastic gradient descent (SGD)을 사용하였으며, 240 에포크로 학습을 진행하였다.

본 논문에서는 베이스라인과 공정한 정확도 비교 평가를 위해서, 베이스라인의 비교 평가 실험에 사용된 CNN 기반의 ResNet [1], VGG [25], MobilenetV2 [26], WRN-40-2 [27], ShuffleNetV2 [28] 네트워크를 교사 및 학생 모델로 사용하였다. VGG-8/VGG-13과 ResNet-8×

4/ResNet-32×4와 같이 교사와 학생이 유사한 구조의 조합 2개와 5개 이종 모델 조합으로 설정하였다. 각 교사 모델과 학생 모델의 조합에서 L_{Global}를 scaling 하기 위한 factor ζ는 학생 모델이 가장 높은 Top-1 정확도를 달성하였을 때의 값을 실험적으로 확인한 값이다. 이는 Table 1에서 확인 할 수 있으며, 각 모델의 특징맵의 채널 수에 따라 임베딩 함수 내의 컨볼루션 레이어의 입/출력 채널 수가 변하여 트랜스포머의 입력 특징맵에 영향을 주기 때문에 교사/학생 모델의 구조에 따라 ζ값의 차이가 발생하였다.

Table 1. Scaling factor ζ for each teacher and student combinations

Teacher / Student	ζ
ResNet-32×4 / VGG-8	0.1
ResNet-32×4 / VGG-13	0.005
VGG-13 / ShuffleNetV2	0.1
ResNet-32×4 / ShuffleNetV2	0.003
WRN-40-2 / MobileNetV2	0.3
VGG-13 / VGG-8	0.005
ResNet-32×4 / ResNet-8×4	0.1

2. Comparison result

CIFAR-100에서 IEKD와 다양한 지식 증류 방법과 정확도 비교 평가를 진행하였다. 비교 대상에는 KD [5], FitNet [6], AT [19], SP [20], VID [21], HKD [22], SemCKD [7]과 최신 기법인 ICKD [10], TaT [23]이 포함되어 있다. 실험 결과는 Table 2에서 확인 할 수 있다. 제안하는 IEKD 기법으로 학습한 모델이 다른 지식 증류 방법으로 학습한 학생 모델 대비 더 향상된 판별 정확도를 달성하였음을 확인하였다. 기존 KD 대비 “ResNet-32×4/VGG-8”는 정확도가 최대 3.5% 향상된 수치를 보였다. “VGG-13/ShuffleNetV2”에서는 정확도 향상 수치가 1.31%로 가장 낮은 변화폭을 보여주었다.

이는 모델의 internal 상관관계와 external 상관관계를 모두 지식 증류에 사용하는 것이 학생 모델 학습에 도움되는 것을 보여준다. 베이스라인 방법인 SemCKD와 최신 기법인 TaT는 external 상관관계만 사용한다. SemCKD는 이종 모델 간 특징맵끼리 external 상관관계 사용하여 특징맵 간 거리를 최소화한다, TaT는 특징맵의 픽셀 단위 매칭을 위해 학생 모델과 교사 모델의 특징맵에 어텐션 매커니즘을 적용하여 external 상관관계만 사용한다. Intenral 상관관계만 사용한 ICKD는 소프트 맥스 직전 마

Table 2. Top-1 Accuracy (%) of knowledge distillations comparison on CIFAR-100

Teacher	ResNet-32×4	ResNet-32×4	VGG-13	ResNet-32×4
	79.42	79.42	74.64	79.42
Student	VGG-8	VGG-13	ShuffleNetV2	ShuffleNetV2
	70.46 ± 0.29	74.82 ± 0.22	72.60 ± 0.12	72.60 ± 0.12
KD [5]	72.73 ± 0.15	77.17 ± 0.11	75.60 ± 0.21	75.49 ± 0.24
FitNet [6]	72.91 ± 0.18	77.06 ± 0.14	75.44 ± 0.11	75.82 ± 0.22
AT [19]	71.90 ± 0.13	77.23 ± 0.19	75.41 ± 0.10	75.91 ± 0.14
SP [20]	73.12 ± 0.10	77.72 ± 0.33	75.54 ± 0.18	75.77 ± 0.08
VID [21]	73.19 ± 0.23	77.45 ± 0.13	75.22 ± 0.07	75.55 ± 0.18
HKD [22]	72.63 ± 0.12	76.76 ± 0.13	76.24 ± 0.09	76.64 ± 0.05
SemCKD [7]	75.27 ± 0.13	79.43 ± 0.02	76.39 ± 0.12	77.62 ± 0.32
ICKD [10]	NA	NA	NA	NA
TaT [23]	NA	NA	NA	NA
IEKD (ours)	**76.23**	**79.83**	**76.91**	**77.88**

Teacher	WRN-40-2	VGG-13	ResNet-32×4
	75.61	74.64	79.42
Student	MobileNetV2	VGG-8	ResNet-8×4
	65.43 ± 0.29	70.46 ± 0.29	73.09 ± 0.30
KD [5]	68.70 ± 0.22	73.38 ± 0.05	74.42 ± 0.05
FitNet [6]	68.64 ± 0.12	73.63 ± 0.11	74.32 ± 0.08
AT [19]	68.79 ± 0.13	73.51 ± 0.08	75.07 ± 0.03
SP [20]	68.48 ± 0.36	73.53 ± 0.23	74.29 ± 0.07
VID [21]	68.37 ± 0.24	73.63 ± 0.07	74.55 ± 0.10
HKD [22]	69.23 ± 0.16	73.06 ± 0.24	74.86 ± 0.21
SemCKD [7]	69.61 ± 0.05	74.43 ± 0.25	76.23 ± 0.04
ICKD [10]	NA	73.88	75.48
TaT [23]	NA	74.35	75.54
IEKD (ours)	**70.03**	**75.1**	**76.63**

지막 레이어의 특징맵의 내부 채널 상관 정보를 전달한다. 반면 제안된 IEKD는 internal 상관관계와 external 상관관계를 모두 활용하기 때문에 더 좋은 성능을 보여주었다.

3. Ablation study

제안된 IEKD의 각 구성 요소에 대해 절제 실험(Ablation study)을 진행하였다. 제안한 IEKD를 임베딩(Embedding) 함수 ψ와 3.2장의 트랜스포머(Transformer), L_{Global}로 나누어 실험을 진행하였다. 세 구성 요소가 모두 없는 경우는 베이스라인과 동일하게 지식 증류를 수행하였다. 임베딩 함수가 추가된 경우는 베이스라인의 특징맵을 임베딩 하는 함수를 ψ로 변경하였다. 트랜스포머가 추가된 경우, 임베딩 매트릭스 간 트랜스포머 연산을 통해 산출된 특징맵으로 external 상관관계를 계산하였다. 마지막으로 L_{Global}이 없는 경우는 식 (8)에서 ζ 값을 0으로 설정하였다. 실험을 위해 사용된 교사/학생 모델은 이종 모델 조합에서 선택하였으며 교사모델은 ResNet-32×4을 사용하였고 학생 모델은 VGG-8을 사용하였다. 실험 결과는 Table 3에서 확인할 수 있다. Table 3의 (1)은 베이스라인이다. Table 3의 (2)에서 나타나듯이, 임베딩 함수만 추가되는 경우 베이스라인과 비교할 때 Top-1 정확도가 하락하는 것을 확인하였으며, 베이스라인은 시퀀스 형태의 입력을 사용하지 않기 때문에 정확도가 하락하였다. Table 3의 (3)에서 트랜스포머가 추가되면 정확도가 향상되는 것을 확인하였으며, 이는 트랜스포머를 통해 internal/external 상관관계를 모두 활용하였기 때문이다. Table 3의 (4)에서 L_{Global}까지 모두 적용한 경우, 가장 정확도가 높은 것을 확인하였으며, 이는 트랜스포머를 통해 추출된 특징맵을 증류하는 것이 학생 모델의 정확도 향상에 기여하기 때문이다.

Table 3. Ablation study for components of proposed IEKD on CIFAR-100

	ψ	Transformer	L_{Global}	Top-1 Acc.(%)
(1)	×	×	×	74.95
(2)	○	×	×	72.73
(3)	○	○	×	75.96
(4)	○	○	○	76.23

4. Sensitivity Analysis

민감도 분석(Sensitivity analysis)에서는 트랜스포머의 인코더/디코더 레이어 수와 scaling factor ζ가 학생 모델의 Top-1 정확도에 주는 영향을 분석한다. 실험을 위해 인코더와 디코더의 레이어 수는 동일한 레이어 수로 설정하였으며, 레이어의 수는 {1, 4, 6, 9, 12}개로 실험을 진행하였다. 교사 모델과 학생 모델은 ResNet-32×4와 VGG-8를 사용하였으며, 실험 결과는 Table 4와 같다. 트랜스포머의 레이어가 6개 일 때 가장 높은 Top-1 정확도를 확인할 수 있었다. 트랜스포머의 레이어가 6개까지는 깊어질수록 Top-1 정확도가 상승하는 경향을 보여주었다. 하지만 6개를 넘어가면 Top-1 정확도가 하락하는 경향을 확인하였으며, 최적의 인코더와 디코더의 레이어 수는 6으로 확인되었다.

Table 4. Top-1 accuracy for the different number of layers on CIFAR-100

The number of transformer layers	Top-1 Accuracy (%)
1	73.15
4	75.5
6	76.23
9	75.57
12	75.67

또한 식 (8)에서 scaling factor ζ의 값을 변경하여 Top-1 정확도를 측정하였다. 실험 조건은 위와 같으며 scaling factor β는 400으로 고정하였다. ζ는 0.001부터 1000까지 10의 지수로 설정하였다. 실험 결과는 Table 5와 같다. ζ가 0.001부터 0.1까지는 Top-1 정확도가 증가하는 추세를 보여주었으며, 0.1부터는 Top-1 정확도가 감소하는 추세를 보여주었다. 또한 ζ값이 증가하면 Top-1 정확도가 감소하는 것을 확인하였으며, 최적의 ζ값은 0.1로 확인되었다.

Table 5. Top-1 accuracy for the different ζ values on CIFAR-100

ζ	Top-1 Accuracy (%)
0.001	73.15
0.01	75.5
0.1	76.23
1	75.57
10	75.67
100	71.85
1000	71.86

V. Conclusions

기존 특징맵 기반의 지식 증류 방법은 동종 모델의 특징맵 간 상관관계인 내부적 상관관계와 이종 모델 간 상관관계인 외부적 상관관계를 모두 활용하지 못하는 문제점이

있었다. 이 논문에서는 외부적 상관관계와 내부적 상관관계를 모두 활용 가능한 IEKD를 제안한다. 외부적 상관관계와 내부적 상관관계를 모두 활용하기 위해 트랜스포머로 새로운 특징맵을 추출한다. 이를 위해 교사와 학생 모델의 특징맵을 시퀀스 형태로 변환하는 임베딩 함수도 제안한다. 이미지 분류에서 다양한 교사-학생 모델 조합에서 IEKD가 정확도 비교 실험에서 "ResNet-32×4/VGG-8" 조합으로 최신 논문들보다 높은 Top-1 정확도 76.23%를 달성하였으며, 제안된 외부적/내부적 상관관계를 모두 활용하여 지식 증류를 수행하는 것이 학생 모델의 정확도 향상하는 것을 확인하였다. 또한 절제 실험을 통해 제안된 IEKD가 학생 모델의 성능을 향상할 수 있는 것을 확인하였으며, IEKD가 임베디드 시스템, 자율 주행과 같은 태스크에서 고성능의 경량화된 모델 생성에 이바지할 것으로 기대한다.

ACKNOWLEDGEMENT

This work was supported in part by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. NRF-2022R1C1C1009208) and funded by the Ministry of Education (No.2022R1A6A1A03051705); supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No.2022-0-00448: Deep Total Recall, 10%, No. RS-2022-00155915: Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)).

REFERENCES

[1] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770-778. June 2016. DOI: 10.1109/CVPR.2016.90

[2] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, June 2017. DOI: 10.1109/TPAMI.2016.2577031.

[3] Han, Song, Huizi Mao, and William J. Dally. "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding." In International Conference on Learning Representations, May 2016.

[4] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali Farhadi. "Xnor-net: Imagenet classification using binary convolutional neural networks.", Computer Vision-ECCV 2016: 14th European Conference, pp. 525-542, October 2016. DOI: 10.1007/978-3-319-46493-0_32

[5] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," arXiv, March 2015. DOI: 10.48550/arXiv.1503.02531

[6] A. Romero, Nicolas Ballas, S. Kahou, Antoine Chassang, C. Gatta, and Yoshua Bengio. "Fitnets: Hints for thin deep nets". In International Conference on Learning Representations, May 2015.

[7] Chen, Defang, et al. "Cross-layer distillation with semantic calibration." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 8. pp. 7028-7036, February 2021. DOI: 10.1609/aaai.v35i8.16865

[8] Li, Hao-Ting, Shih-Chieh Lin, Cheng-Yeh Chen, and Chen-Kuo Chiang. "Layer-level knowledge distillation for deep neural network learning." Applied Sciences, Vol .9, No. 10, May 2019 DOI: 10.3390/app9101966

[9] Han, Kai, et al. "Ghostnet: More features from cheap operations." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1577-1586, June 2020. DOI: 10.1109/CVPR42600.2020.00165

[10] Liu, Li, et al. "Exploring inter-channel correlation for diversity-preserved knowledge distillation." Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8251-8260, October 2021. DOI: 10.1109/ICCV48922.2021.00816

[11] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30, pp, 6000-6010 December 2017. DOI: 10.5555/3295222.3295349

[12] Alex Krizhevsky and Geoffrey Hinton. "Learning multiple layers of features from tiny images." Technical Report, pp. 1-60, 2009

[13] Yang, Zhendong, et al. "Focal and global knowledge distillation for detectors." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4633-4642, June 2022. DOI: 10.1109/CVPR52688.2022.00460

[14] Gong, Yuan, et al. "Cmkd: Cnn/transformer-based cross-model knowledge distillation for audio classification." arXiv preprint arXiv:2203.06760, March 2022. DOI: 10.48550/arXiv.2203.06760

[15] Zhao, Borui, et al. "Decoupled knowledge distillation." Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp. 11943-11952, June 2022. DOI: 10.1109/CVPR52688.2022.01165

[16] Chen, Defang, et al. "Knowledge distillation with the reused teacher classifier." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11923-11932, June

2022. DOI: 10.1109/CVPR52688.2022.01163

[17] Beyer, Lucas, et al. "Knowledge distillation: A good teacher is patient and consistent." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10915-10924, June 2022. DOI: 10.1109/CVPR52688.2022.01065

[18] Park, Wonpyo, et al. "Relational knowledge distillation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3962-3971, June 2019. DOI: 10.1109/CVPR.2019.00409

[19] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer," In International Conference on Learning Representations, April 2017

[20] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," In International Conference on Computer Vision, pp. 1365–1374, November 2019. DOI:10.1109/ICCV.2019.00145

[21] S. Ahn, S. X. Hu, A. C. Damianou, N. D. Lawrence, and Z. Dai, "Variational information distillation for knowledge transfer," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9163–9171, June 2019. DOI: 10.1109/CVPR.2019.00938

[22] N. Passalis, M. Tzelepi, and A. Tefas, "Heterogeneous knowledge distillation using information flow modeling," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2336–2345, June 2020. DOI: 10.1109/CVPR42600.2020.00241

[23] Lin, Sihao, et al. "Knowledge distillation via the target-aware transformer." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10905-10914, June 2022. DOI: 10.1109/CVPR52688.2022.01064

[24] Paszke, Adam, et al. "Pytorch: An imperative style, high-performance deep learning library." Advances in neural information processing systems 32, pp. 8026–8037, December 2019.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In International Conference on Learning Representations, May 2015.

[26] Sandler, Mark, et al. "Mobilenetv2: Inverted residuals and linear bottlenecks." Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510-4520, June 2018. DOI: 10.1109/CVPR.2018.00474

[27] Zagoruyko, Sergey, and Nikos Komodakis. "Wide residual networks." in Proceedings of the British Machine Vision Conference, pp. 87.1-87.12, September 2016. DOI: 10.5244/C.30.87

[28] Ma, Ningning, et al. "Shufflenet v2: Practical guidelines for efficient cnn architecture design." Proceedings of the European conference on computer vision (ECCV), pp. 116-131, September 2018.

Authors

Hun-Beom Bak received the BS degree in Physics from Incheon National University in 2020, and is currently pursuing the MS degree with the Department of Electrical and Computer Engineering at Inha University, Korea. His current research interest include model compression, knowledge distillation and data-free knowledge distillation.

Seung-Hwan Bae received the BS degree in information and communication engineering from Chungbuk National University, in 2009 and the MS and PhD degrees in information and communications from the Gwangju Institute of Science and Technology (GIST), in 2010 and 2015, respectively. He was a senior researcher at Electronics and Telecommunications Research Institute (ETRI) in Korea from 2015 to 2017. He was an assistant professor in the Department of Computer Science and Engineering at Incheon National University, Korea from 2017 to 2020. He is currently an Associate Professor with the Department of Computer Engineering at Inha University, His research interests include object tracking, object detection, generative model learning, continual learning, on-device ML, etc.