다중 신경망으로부터 해석 중심의 적응적 지식 증류

이자윤 [1], 조인휘 [2]
[1] 한양대학교 컴퓨터소프트웨어학과 석사과정
[2] 한양대학교 컴퓨터소프트웨어학과 교수

zihyunmay29@hanyang.ac.kr, iwjoe@ hanyang.ac.kr

Explanation-focused Adaptive Multi-teacher Knowledge Distillation

Chih-Yun Li[1], Inwhee Joe[2]
[1]Dept. of Computer Science, Hanyang University
[2]Dept. of Computer Science, Hanyang University

요 약

엄청난 성능에도 불구하고, 심층 신경망은 예측결과에 대한 설명이 없는 블랙 박스로 작동한다는 비판을 받고 있다. 이러한 불투명한 표현은 신뢰성을 제한하고 모델의 대한 과학적 이해를 방해한다. 본 연구는 여러 개의 교사 신경망으로부터 설명 중심의 학생 신경망으로 지식 증류를 통해 해석 가능성을 향상시키는 것을 제안한다. 구체적으로, 인간이 정의한 개념 활성화 벡터 (CAV)를 통해 교사 모델의 개념 민감도를 방향성 도함수를 사용하여 계량화한다. 목표 개념에 대한 민감도 점수에 비례하여 교사 지식 융합을 가중치를 부여함으로써 증류된 학생 모델은 양호한 성능을 달성하면서 네트워크 논리를 해석으로 집중시킨다. 실험 결과, ResNet50, DenseNet201 및 EfficientNetV2-S 앙상블을 7 배 작은 아키텍처로 압축하여 정확도가 6% 향상되었다. 이 방법은 모델 용량, 예측 능력 및 해석 가능성 사이의 트레이드오프를 조화하고자 한다. 이는 모바일 플랫폼부터 안정성이 중요한 도메인에 걸쳐 믿을 수 있는 AI 의 미래를 여는 데 도움이 될 것이다.

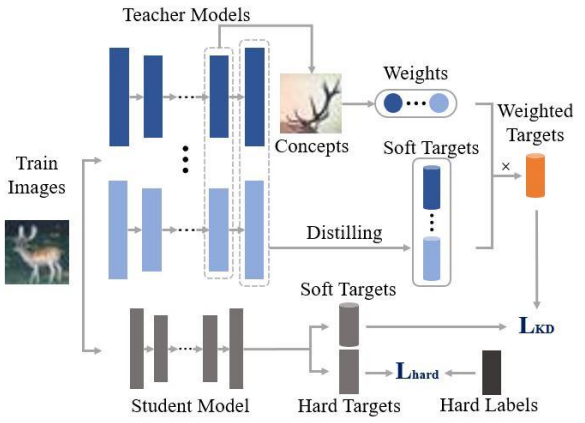
1. 서론

최근 몇 년 동안 인공 지능의 발전은 엄청난 진전을 보여왔으며, 신경망은 가장 능력이 있고 우수한 머신러닝 알고리즘 중 하나로 빛나고 있다. 모델 용량과 계산 효율성 사이의 트레이드오프를 탐색하는 것은 중요한 도전과 활발한 연구 분야로 발전해 왔다. 현재의 연구는 신경구조 탐색[1], 네트워크 pruning[2] 및 지식 증류[3]와 같은 기술을 탐구하며 정확도와 효율성을 최적으로 균형을 맞추려고 한다.

딥러닝의 신속한 발전은 산업 전반에 혁명적인 변화를 가져오지만, 신경망은 결정 과정에서 심각한 불투명성이 있다. 모델의 행동을 해석하고 설명할 수 없는 경우, 안전, 책임 및 신뢰가 최우선인 상황에서 신경망을 응용하는 것은 위험할 수 있다. 로컬 설명 맵[4]부터 대표적인 프로토타입 선택[5] 까지 다양한 기술들은 모델의 논리를 역공학하고 결론에 영향을 미치는 특성을 탐구한다. 또한 전역 설명 방식은 개별 노드나 레이어와 같은 구성 요소의 변화가 전체 출력에 어떻게 영향을 미치는지를 분석한다[6]. 전통적인 컨볼루션 및 다층 퍼셉트론 신경망의 대안으로 내재적으로 해석 가능한 모델 아키텍처를 제안하기도 했다[7].

모델 성능, 효율성 및 해석 가능성 사이의 트레이드오프를 조화시키기 위해, 본 연구는 새로운 지식 증류 모델을 소개한다. 지식 증류는 복잡한 교사 모델 또는 모델 앙상블에서 학습된 특징을 더 작고 더 단순화된 학생 네트워크로 전달하는 것을 목표로 한다[3]. 지식을 통합함으로써, 증류된 학생은 교사와 비슷한 정확도를 달성하면서 계산적 요구를 크게 줄이려고 한다. 학생 모델의 단순화는 모델의 투명성과 설명 가능성을 본질적으로 향상시킨다. 그러나 대부분의 증류 기술은 교사를 동등한 기여자로 취급하고 특징을 선택적으로 강조하는 메커니즘의 대한 연구가 부족하다. 본 연구는 인간이 정의한 개념 민감도 분석[8]을 기반으로 교사 지식을 융합하는 것을 제안한다. 구체적으로, 방향성 도함수는 모델 매개변수에 대한 주어진 개념의 활성화 변화율을 계량화한다. 민감도가 높은 교사들은 관련 개념에 대한 주요 지식 출처로 우선 순위를 매길 수 있다.

2. 설명 중심 적응형 다중 교사 지식 증류 프레임워크



(그림 1) 제안된 모델 아키텍처.

그림 1 에 설명된 대로, 설명 중심 적응형 다중 교사 지식 증류 프레임워크는 복잡한 교사 모델 앙상블과 단일 단순화된 학생 모델로 구성된다. 학생이 여러 분야에서 나온 선생들한테 전문적인 지식을 배우는 것과 유사하게, 주요 목표는 지식 전달을 교사의 특화를 보완하는 데 맞춤화하는 것이다. 각 CNN 교사 모델의 인간이 해석 가능한 개념에 대한 분류 민감도를 개념 활성화 벡터(CAV)를 통해 평가한다[8]. CAV 는 의미론적 개념에 해당하는 표현 공간에서 해석 방향을 식별한다. 그런 다음 방향성 도함수는 역전파 중 모델 매개변수 변경에 대해 이러한 개념 채널이 얼마나 반응하는지를 계량화한다.

[8]에서 정의된 대로, 인간이 해석할 수 있는 개념 C에 대한 클래스 k 의 개념적 민감도 $S_{C,k,l}(x)$ 는 v_C^l 에 대한 방향성 도함수로, Concept Activation Vectors (CAV)로도 불립니다. 수학적으로, CAV 는 특정 모델 레이어 l 에서 C 와 관련된 입력 샘플의 활성화와 일반적인 샘플을 분리하는 데 훈련된 선형 이진 분류기의 법선 벡터입니다. 직관적으로, CAV 는 C의 시맨틱 의미와 강하게 상관된 고차원 표현 공간에서의 방향을 식별한다. 그런 다음 방향성 도함수 $S_{C,k,l}(x)$ 는 v_C^l을 따라 변화하는 것이 모델 출력에 어떤 영향을 미치는지를 보여주며, 이로써 클래스 k 가 개념 C 에 대한 개념적 민감도를 계량화한다. [8]에서 소개된 측정법 TCAV 점수는 개념 C 에 의해 긍정적으로 영향을 받는 클래스 k 의 샘플 비율에 의해 계산된다.

$$S_{C,k,l}(x) = \nabla h_{l,k}\big(f_l(x)\big) \cdot v_C^l \tag{1}$$

$$TCAV_{C,k,l} = \frac{\left|\{x \in X_k : S_{C,k,l}(x) > 0\}\right|}{|X_k|} \tag{2}$$

식 (2)에서 Iverson bracket 지시 함수 $\{x \in X_k : S_{C,k,l}(x) > 0\}$는 샘플 x 가 CAV 를 따라 양의 민감도를 갖는지를 이산화한다. 따라서 TCAV 점수는 C 가 클래스 k를 예측하는 데 긍정적으로 활성화되는 X_k 샘플의 전체 비율을 요약한다.

각 목표 출력 클래스 k에 대해, 주요 개념 C_k에 대해 가장 설명력이 높다고 판단되는 것에 대한 교사 가중치 요소를 그들의 TCAV 점수에 비례하여 할당한다. 이로써, 높은 관련성을 부여한 개념에 대해 감도가 강한 교사 모델에 더 큰 영향력이 전달된다.

교사 모델의 i번째 샘플의 로짓을 소프트 타겟 $\tilde{y}_i$로 변환하는 과정은 온도 조절된 로짓을 사용하여 계산된다:

$$\tilde{y}_i = \frac{\exp(z_i/T)}{\sum_i \exp(z_i/T)} \tag{3}$$

여기서 z_i는 i번째 샘플의 로짓(소프트맥스 전의 화성화)을 나타내며, T 는 분포의 날카로움을 조절하는 온도 초매개변수다. 이를 통해 지식 표현은 뾰족한 소프트맥스 출력보다 더 많은 정보를 보존한다.

모든 k 클래스에 대한 각 교사 모델의 TCAV 점수를 고려할 때, 가중치는 다음과 같이 생성된다:

$$w_{t,k} = \frac{\exp\left(TCAV_{t,k}\right)}{\sum_{j=1}^{n} \exp\left(TCAV_{j,k}\right)} \tag{4}$$

여기서 $t \in \{1, \cdots, n\}$ 이고, 가중치는 n 개의 교사 모델에 걸친 TCAV 점수의 합으로 정규화되어 특화된 지식을 보존하는 유효한 압축된 확률 분포를 생성한다. 이는 다음과 같이 작성될 수 있다:

$$\tilde{y}_i^T = \sum_{t=1}^{n} w_{t,k} \tilde{y}_{t,i}^T \tag{5}$$

여기서 k는 i번째 샘플에 해당하는 클래스입니다.

학생 모델은 (5)에서 교사 모델의 통합된 소프트 타겟을 사용하여 (3)에서 자체 소프트 타겟을 학습하여 개념 인식 지식 증류를 구현한다. 이를 L_{KD}라고 한다. 동시에, 전통적인 분류 손실을 유지하기 위해 학생 예측과 실제 원핫 레이블을 맞추기도 한다. 이것은 L_{hard}로 표시된다. 따라서 설명 중심 적응형 다중 교사 지식 증류 모델의 손실 함수는 다음과 같이 표시될 수 있다:

$$\begin{aligned} L &= L_{hard} + \lambda L_{KD} \\ &= \sum_{i}^{m} H(y_i, y_i^S) + \lambda \sum_{i}^{m} H(\tilde{y}_i^T, \tilde{y}_i^S) \end{aligned} \tag{6}$$

여기서 y_i는 실제 레이블을 나타내고, y_i^S는 학생 모델에 의한 출력 타겟, $\tilde{y}_i^T$는 교사 모델의 통합된 소프트 타겟이며, $\tilde{y}_i^S$는 학생 모델에 의한 소프트 타겟이다. 이것은 총 m 개의 이미지 중 i 번째 이미지에 해당한다. λ 는 손실 함수에서 해당 부분의 영향을 제어하기 위한 하이퍼파라미터입니다. 이 설계로, 학생 모델은 가중 평균 소프트 분포와 동시에 하드 타겟을 맞추도록 학습해야 한다. 따라서 민감도가 높은 교사 전문가가 가장 강력하게 모델링한 목표 개념으로 특화된 지식의 적응적 융합이 이루어진다. 이는 복잡한 앙상블 내에서 해석 가능한 개념으로의 학습을 맞춤화한다.

<표 1> 학생, 교사 모델 및 우리 모델의 훈련 결과

Model	Parameters	Compression Ratio	CIFAR-10 Accuracy
RensNet50	~24M	x1.0	82.8%
DenseNet201	~18M	x1.3	85.7%
EfficientNetV2-S	~20M	x1.2	84.7%
Student-3	~3.4M	x7.1	67.38%
EAMT-KD(Ours)	-	-	**73.81%**

<표 2> 세 가지 크기의 학생 모델에 훈련 결과

Model	Student-1	Student-2	Student-3
Parameters	~0.4M	~2.6M	~3.4M
Accuracy	51.56%	71.92%	67.38%
EAMT-KD(Ours)	49.58%	70.07%	**73.81%**

3. 실험 및 결과

CIFAR-10 데이터셋[9]에서 제안된 지식 증류 방식을 평가한다. 이 데이터셋에는 10 가지 클래스에 걸쳐 50,000 개의 훈련 이미지와 10,000 개의 테스트 이미지가 포함되어 있다. 각 대상 클래스에는 인식을 위한 독특한 속성을 포착하는 대표적인 해석 가능한 개념이 할당된다. Airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck 를 포함한 클래스는 각각 plane wing, tire, beak, cat face, antler, dog face, frog feet, horsehair, ship bow, truck container 와 대응된다. 각 개념에 해당하는 시각적 패턴을 프로파일링하기 위해 검색 엔진을 통해 수집된 약 10 개의 샘플 이미지를 사용한다.

교사 앙상블은 ResNet50[10], DenseNet201[11], EfficientNetV2-S[12] 세 가지 구조적으로 다른 CNN 아키텍처로 구성된다. 각 모델에 대해 CIFAR-10 훈련 세트에서 클래스 당 이미지 100 개를 사용하여 방향성 도함수 기반 TCAV 민감도 점수를 계산한다. 점수 계산은 분류 전 글로벌 평균 풀링 이후 dense 레이어에서의 활성화를 활용한다. CAV 회귀 페널티를 0.1 로 정규화하면서 10 회 실행에 걸친 결과를 평균화한다.

개념 민감도를 측정하는 이러한 클래스별 모델별 TCAV 점수는 가중 지식 증류 패러다임에 흡수된다. 증류 온도 T 는 2 로 설정되고, 모델 압축을 지배하는 손실 항의 하이퍼파라미터 λ 는 0.1 로 구성된다.

표 1 은 설명 중심 증류 방식 대 비교 대조군에 대한 효율성 및 정확도 지표의 양적 결과를 보고한다.

단일 학생 모델인 Student-3 모델은 3.4 백만 개의 매개변수로만 구성되어 있으며, 이는 가장 큰 교사 ResNet50 의 2,400 만 개의 매개변수 대비 무려 7.1 배의 매개변수 감소를 가져온다. 교사 앙상블로부터 설명을 전달함으로써, EAMT-KD 방식은 학생의 정확도를 67.38%에서 73.81%로 높이면서 7 배의 매개변수 절감을 유지한다. 이는 기준선에 비해 상당한 증가로, 맞춤형 지식 통합이 경량 모델의 성능을 향상시키는 능력을 입증한다.

추가 실험에서는 학생 모델 용량이 전달된 지식 및 독립적 정확도에 미치는 영향을 탐구한다. 표 2 는 0.4 백만 개, 2.6 백만 개 및 3.4 백만 개의 매개변수로 구성된 세 가지 CNN 아키텍처(Student 1-3)에 대한 개념 기반 EAMT 증류의 성능을 비교한다.

직관적으로, 더 넓은 모델은 복잡한 교사 앙상블로부터 풍부한 정보를 흡수하기 위한 더 큰 표현력을 갖는다. Student 1 및 2 의 더 얕은 디자인은 복잡한 관계 지식을 완전히 활용하기에 충분한 특징 변환 레이어를 갖고 있지 않다. 또한, 교사와의 상대적인 표현 차이는 크게 더 크기 때문에, 복잡한 결정 경계를 모방하는 것을 방해한다. 그러나 Student-3 이 달성한 합리적인 점수는 충분한 모델 표현력이 전문 지식을 흡수하기에 충분한 지점을 넘었음을 나타낸다.

4. 결론

본 연구는 신경망 증류에서 지식 전달을 인도하기 위해 양적인 모델 해석성 통찰력을 활용하는 가능성을 입증한다. 교사 모델 앙상블에 대한 개념 활성화 민감도를 평가함으로써 학생 학습 중 보충적으로 해석 가능한 설명에 집중하는 통합을 맞춤화한다. 이를 통해 유의미한 입력-출력 특징 매핑을 통해 증류된 지식을 해석할 수 있다. 더욱이, 이 방법은 대형이지만 불투명한 교사로부터 다양한 정보를 가벼운 학생 아키텍처로 통합함으로써 정확도와 효율성 목표를 조화한다. 실험은 다양한 모델 용량에서 전통적인 증류 및 독립적인 학생 교육보다 상당한 테스트 정확도 개선을 검증한다.

미래의 방향은 최종 레이어 분석을 넘어 여러 중간 표현으로 확장하고, 샘플당 개념을 동적으로 가중하며, 각 클래스마다 개인화된 교사 후보로 구성하는 것이다. 더 정교한 융합 방법이나 커리큘럼 일정도 앙상블의 깊이와 학생의 단순함 사이의 표현적 간극을 극복하는 데 도움이 될 수 있다.

본 연구는 설명으로 이끄는 지식 전달을 통해 기계와 인간의 이해를 연결하며, 신뢰할 수 있고 능숙한 인공 지능을 위한 길을 열어준다. 중요한 해석을

성능적인 특징과 일치시키기 위해 지식을 맞춤화하는 것은 여전히 개방적이고 유망한 도전이다.

참고문헌

[1] T. Elsken, J. H. Metzen, and F. Hutter, "Efficient multi-objective neural architecture search via lamarckian evolution," *arXiv preprint arXiv:1804.09081,* 2018.

[2] T. Gale, E. Elsen, and S. Hooker, "The state of sparsity in deep neural networks," *arXiv preprint arXiv:1902.09574,* 2019.

[3] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531,* 2015.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618-626.

[5] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6541-6549.

[6] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for deep neural networks," *arXiv preprint arXiv:1711.06104,* 2017.

[7] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8827-8836.

[8] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, and F. Viegas, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*, 2018: PMLR, pp. 2668-2677.

[9] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

[11] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700-4708.

[12] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *International conference on machine learning*, 2021: PMLR, pp. 10096-10106.