의미적 정보를 보존하는 지식 증류에 대한 연구

박성현[1], 이상근[2]
[1]고려대학교 사이버국방학과 학부생
[2]고려대학교 정보보호대학원 교수
shyun00124@korea.ac.kr, sangkyun@korea.ac.kr

A study on knowledge distillation to preserve semantic information

Seong-hyun Park[1], Sangkyun Lee[2]
[1]Dept of cyber defense, Korea university
[2]School of Cybersecurity, Korea university

요 약

의미적 정보까지 학생 모델에게 학습시키기 위한 지식 증류 기법은 많이 논의되어 왔다. 그러나 학생 모델의 용량이 교사 모델의 용량에 비해 부족함에서 발생하는 의미적 정보 손실에 대한 논의는 아직 진행되지 않았다. 본 논문에서는 의미적 정보의 최소 단위를 교사 모델의 레이어로 설정하여 학생 모델이 지식 증류를 시작하기 전 최적의 지식 증류 대상을 설정하는 최적 은닉층 선정 알고리즘을 제시한다.

1. 서론

기존의 지식 증류 기법은 오직 교사 모델의 출력만을 교사 모델에서의 학습 대상으로 삼기 때문에 학생 모델이 교사 모델의 의미적 정보를 학습하였음은 보장하지 않는다. 이를 해결하기 위해 교사 모델의 학습하고자 하는 계층의 Feature map을 손실함수에 적용하고자 하는 연구[1], 멀티모달 모델에서의 모듈별 기여도를 적용하고자 하는 연구[2]는 있었으나 교사 모델과 학생 모델의 계층 수의 차이로 인해 매핑 함수를 임의로 정의해야 하는 문제점이 있다. 이를 해결하기 위해 의미적 정보를 보존하는 매핑 함수를 찾아내는 연구를 진행하고자 한다.

2. 실험 방법

두 모델이 동일한 의미적 정보를 학습하였다는 것은 같은 입력에 대해 동일한 특징정보를 통해 결론을 도출했다는 것을 의미한다. 이때 특징정보는 모델의 각각 다른 은닉층에 포함되어 있다[3]. 그러므로 임의의 입력에 대해 교사 모델의 출력과 학생 모델의 매핑된 레이어의 출력이 동일한 출력을 보인다면 두 모델이 동일한 의미적 정보를 학습하였다고 말할 수 있다.

또한 교사 모델이 의미적으로 많은 정보를 담고 있음은 상대적으로 적은 용량을 가지는 학생 모델이 학습하였을 때 모든 정보를 학습할 수 없어 상대적으로 높은 손실에서 수렴하게 된다. 이를 이용해 은닉층 L개의 교사 모델, 은닉층 K개의 학생 모델에 대해 Layer-Wise Distilation을 진행할 교사 모델의 은닉층 K개에 대한 최적 은닉층 선정 알고리즘을 구현하면 그림 1과 같다.

Algorithm 1: 최적 은닉층 선정 알고리즘

Input: Teacher model T, Student model S, teacher model layer number L, student model layer number K, training dataset D
Output: mapping list $\{M_1, M_2, ..., M_K\}$

for $1 \le i \le K$ **do**
  $M_i = (L//K) \cdot i$
  $L_i^M = 0$

while *S converges* **do**
  $total\_loss = 0$
  for *d in D* **do**
    for $1 \le j \le K$ **do**
      $total\_loss \mathrel{+}= MSE(S_i(d), T_{M_j}(d))$
      $L_j^M \mathrel{+}= MSE(S_i(d), T_{M_j}(d))$
  $maxIndex = index(max(L^M))$
  $minIndex = index(min(L^M))$
  if *maxIndex != 1* **then**
    if $M_{maxIndex} - 1 > M_{maxIndex-1}$ **then**
      $M_{maxIndex} = M_{maxIndex} - 1$
    else
      if $M_{maxIndex} - 1 \ge 1$ **then**
        $M_{maxIndex} = M_{maxIndex} - 1$
    if *minIndex != K* **then**
      if $M_{minIndex} + 1 < M_{minIndex+1}$ **then**
        $M_{minIndex} = M_{minIndex} + 1$
      else
        if $M_{minIndex} + 1 \le L$ **then**
          $M_{minIndex} = M_{minIndex} + 1$
    return *M*

(그림 1) 최적 은닉층 선정 알고리즘

그러나 이는 모델이 수렴할 때까지 학습을 진행해야 한다는 현실적인 문제가 있다. 현재 알고리즘상 매핑 리스트가 진동하지 않고 유의미한 매핑 리스트의 변화가 발생하려면 교사 모델의 하나의 레이어에 담긴 의미적 정보보다 추가로 학습한 의미적 정보가 손실함수에 큰 영향을 주어야 하는데, 이 점에 착안하여 추가로 학습한 의미적 정보가 안정 상태를 깨뜨리지 않을 때까지 알고리즘을 진행하도록 알고리즘의 종료 조건을 수정하였다. 수정 사항은 표 1과 같다.

수정 전	**6 while** S *converges* **do**
수정 후	**6 while** M *is unstable* **do**

<표 1> 알고리즘 수정 사항

3. 실험 결과

학습 및 평가 데이터로 MNIST 데이터셋을 분할해서 사용했고, 교사 모델은 7 layer CNN, 학생 모델은 4 layer CNN을 사용해 실험하였다. 이 중 처음과 마지막 CNN layer의 경우 입력과 출력의 형태가 다른 layer와 달라 학습 대상에서 제외하였다. 하이퍼파라미터로 학습률 0.001, epoch 15, batch size 100, 손실함수 cross entropy, 최적화 함수 Adam을 학생 모델과 교사 모델 모두에 사용하였다. 교사 모델의 정확도는 0.9888을 기록하였다.

실험 결과는 표 2와 같다. 최적 은닉층 선정 알고리즘을 통해 지식 증류 대상을 탐색한 결과 기존 교사 모델의 3, 5번째 CNN layer에서 3, 4 번째 CNN layer로 지식 증류 대상이 변경되었다. 변경 후의 지식 증류 성능이 매핑 최적화 전보다 약 0.05 상승해 유의미한 차이를 보임을 확인할 수 있었다.

	정확도	증류 대상
학생 모델 (매핑 함수 미최적화)	0.9177	교사 모델의 layer 3, 5
학생 모델 (매핑 함수 최적화)	0.9642	교사 모델의 layer 3, 4

<표 2> 매핑 함수 최적화에 따른 지식 증류 성능 차이

4. 결론

지식 증류에서 의미적 정보를 최대한 추출하기 위해서는 학습시킬 대상의 용량에 맞추어 학습할 대상을 선정해야 한다. 본 논문에서는 학습시킬 대상의 용량을 고정한 채 학습할 대상의 최소 단위를 하나의 레이어로 삼고 이를 최적화하였다. 이는 교사 모델의 주요한 의미적 정보를 학생 모델의 각 레이어에 분배하는 효과를 보였다.

참고문헌

[1] Liang, Chen, et al. "Less is more: Task-aware layer-wise distillation for language model compression." Proceedings of the 40th International Conference on Machine Learning, Honolulu Hawaii USA, 2023, 20852-20867.

[2] Liang, Chen, et al. "Module-wise Adaptive Distillation for Multimodality Foundation Models." Advances in Neural Information Processing Systems, New Orleans USA, 2023, 69719-69735.

[3] Pasad, Chou, et al. "Layer-wise analysis of a self-supervised speech representation model." 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 2021, 914-921