



(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2024-0146801
(43) 공개일자 2024년10월08일

(51) 국제특허분류(Int. Cl.)
G06N 3/096 (2023.01) G06N 3/0455 (2023.01)
(52) CPC특허분류
G06N 3/096 (2023.01)
G06N 3/0455 (2023.01)
(21) 출원번호 10-2023-0041837
(22) 출원일자 2023년03월30일
심사청구일자 2023년03월30일

(71) 출원인
인하대학교 산학협력단
인천광역시 미추홀구 인하로 100(용현동, 인하대학교)
(72) 발명자
배승환
인천광역시 미추홀구 인하로 100(용현동)
박훈범
인천광역시 서구 완정로64번길 7, 102동 1001호
(마전동, 영남탑스빌아파트)
(74) 대리인
양성보

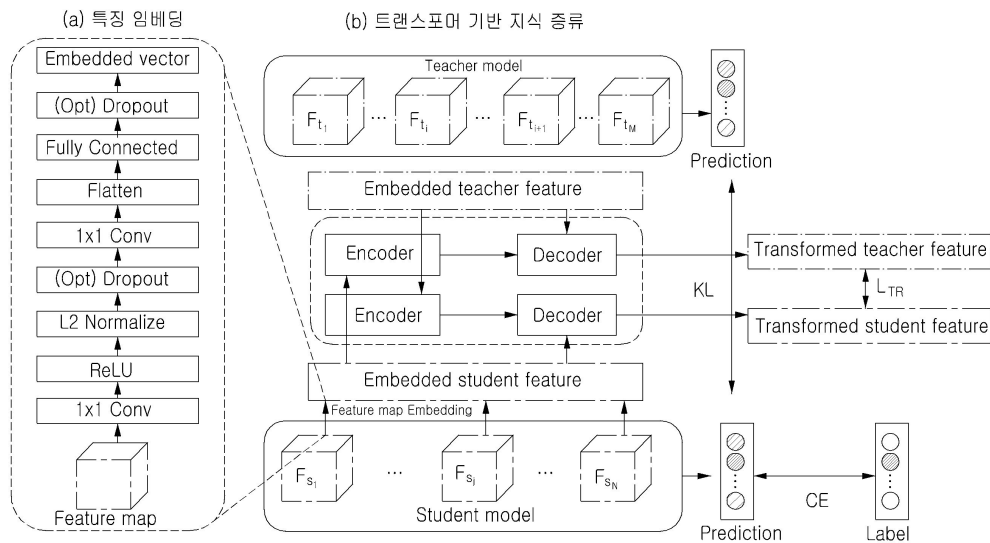
전체 청구항 수 : 총 10 항

(54) 발명의 명칭 특징 상관성 학습 기반 이중 모델 지식 증류 기술

(57) 요약

특징 상관성 학습 기반 이중 모델 지식 증류 기술이 개시된다. 일 실시예에 따른 지식 증류 시스템에 의해 수행되는 트랜스포머 기반 지식 증류 방법은, 교사 모델의 특징맵과 학생 모델의 특징맵을 각각 임베딩함에 따라 상기 교사 모델의 임베딩 벡터와 상기 학생 모델의 임베딩 벡터를 획득하는 단계; 및 상기 획득된 상기 교사 모델의 임베딩 벡터와 상기 학생 모델의 임베딩 벡터에 대해 트랜스포머 기반의 지식 증류를 수행하여 상관성 정보가 포함된 특징맵을 생성하는 단계를 포함할 수 있다.

대표도



이 발명을 지원한 국가연구개발사업

과제고유번호 1711190278
 과제번호 2022R1C1C1009208
 부처명 과학기술정보통신부
 과제관리(전문)기관명 한국연구재단
 연구사업명 우수신진연구
 연구과제명 [Ezbaro] 대규모 확장 가능한 On-Vision AI를 위한 E3-AI 네트워크와 협업 학습 연

구

기 여 율 60/100
 과제수행기관명 인하대학교
 연구기간 2023.03.01 ~ 2024.02.29

이 발명을 지원한 국가연구개발사업

과제고유번호 1345362086
 과제번호 2022R1A6A1A03051705
 부처명 교육부
 과제관리(전문)기관명 한국연구재단
 연구사업명 중점연구소
 연구과제명 [Ezbaro] 3D 나노융합소자 연구 센터

기 여 율 20/100
 과제수행기관명 인하대학교
 연구기간 2022.06.01 ~ 2023.02.28

이 발명을 지원한 국가연구개발사업

과제고유번호 1711160586
 과제번호 2022-0-00448-001
 부처명 과학기술정보통신부
 과제관리(전문)기관명 정보통신기획평가원
 연구사업명 방송통신산업기술개발사업
 연구과제명 [Ezbaro] 인간처럼 회상이 가능한 인공 신경망 지속학습 플랫폼 개발(2차년도)
 기 여 율 10/100
 과제수행기관명 인하대학교 산학협력단
 연구기간 2023.01.01 ~ 2023.12.31

이 발명을 지원한 국가연구개발사업

과제고유번호 1711179347
 과제번호 00155915
 부처명 과학기술정보통신부
 과제관리(전문)기관명 정보통신기획평가원
 연구사업명 정보통신인재양성(교육운영수익)
 연구과제명 [Ezbaro][정부+민간] 인공지능융합혁신인재양성(2차년도)
 기 여 율 10/100
 과제수행기관명 인하대학교 산학협력단
 연구기간 2023.01.01 ~ 2023.12.31

명세서

청구범위

청구항 1

지식 증류 시스템에 의해 수행되는 트랜스포머 기반의 지식 증류 방법에 있어서,

교사 모델의 특징맵과 학생 모델의 특징맵을 각각 임베딩함에 따라 상기 교사 모델의 임베딩 벡터와 상기 학생 모델의 임베딩 벡터를 획득하는 단계; 및

상기 획득된 상기 교사 모델의 임베딩 벡터와 상기 학생 모델의 임베딩 벡터에 대해 트랜스포머 기반의 지식 증류를 수행하여 상관성 정보가 포함된 특징맵을 생성하는 단계

를 포함하는 지식 증류 방법.

청구항 2

제1항에 있어서,

상기 지식 증류를 수행하는 단계는,

상기 획득된 교사 모델의 임베딩 벡터를 트랜스포머의 인코더에 입력하고, 상기 획득된 학생 모델의 임베딩 벡터를 트랜스포머의 디코더에 입력하는 단계

를 포함하는 지식 증류 방법.

청구항 3

제2항에 있어서,

상기 지식 증류를 수행하는 단계는,

상기 수행된 트랜스포머 기반의 지식 증류를 통해 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵과 내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵을 생성하는 단계

를 포함하는 지식 증류 방법.

청구항 4

제3항에 있어서,

상기 지식 증류를 수행하는 단계는,

상기 트랜스포머의 멀티 헤드 어텐션에 트랜스포머의 인코더 출력과 디코더 셀프 어텐션의 출력을 사용하여 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵을 생성하는 단계

를 포함하는 지식 증류 방법.

청구항 5

제3항에 있어서,

상기 지식 증류를 수행하는 단계는,

상기 트랜스포머의 멀티 헤드 어텐션에 트랜스포머의 인코더 출력과 디코더 셀프 어텐션의 출력을 사용하여 내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵을 생성하는 단계

를 포함하는 지식 증류 방법.

청구항 6

제3항에 있어서,

상기 지식 증류를 수행하는 단계는,

내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵, 상기 내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵의 전치 행렬 간 내적, 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵 및 상기 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵의 전치 행렬 간 내적의 평균 제곱 오차를 손실 함수로 사용하는 단계

를 포함하는 지식 증류 방법.

청구항 7

제6항에 있어서,

상기 지식 증류를 수행하는 단계는,

상기 내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵과 상기 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵을 사용하여 상기 교사 모델과 학생 모델 간의 외부적 상관성을 계산하는 단계

를 포함하는 지식 증류 방법.

청구항 8

제7항에 있어서,

상기 지식 증류를 수행하는 단계는,

상기 계산된 교사 모델과 학생 모델 간의 외부적 상관성과 상기 교사 모델의 특징맵과 학생 모델의 특징맵 간의 손실 함수를 스케일링함으로써 상관성이 높은 특징맵끼리 유사도가 높아지도록 매칭하는 단계

를 포함하는 지식 증류 방법.

청구항 9

제8항에 있어서,

상기 지식 증류를 수행하는 단계는,

상기 교사 모델의 특징맵과 상기 학생 모델의 특징맵 간의 채널과 공간적 크기를 일치시키기 위해 상기 교사 모델의 특징맵과 상기 학생 모델의 특징맵을 다층 퍼셉트론과 적응적 풀링으로 구성된 변환 함수에 각각 통과시키는 단계

를 포함하는 지식 증류 방법.

청구항 10

지식 증류 시스템에 있어서,

교사 모델의 특징맵과 학생 모델의 특징맵을 각각 임베딩함에 따라 상기 교사 모델의 임베딩 벡터와 상기 학생 모델의 임베딩 벡터를 획득하는 임베딩 벡터 획득부; 및

상기 획득된 상기 교사 모델의 임베딩 벡터와 상기 학생 모델의 임베딩 벡터에 대해 트랜스포머 기반의 지식 증류를 수행하여 상관성 정보가 포함된 특징맵을 생성하는 지식 증류 수행부

를 포함하는 지식 증류 시스템.

발명의 설명

기술 분야

[0001] 아래의 설명은 지식 증류 기술에 관한 것이다.

배경 기술

[0003] 신경망이 발전함에 따라 컨볼루션 네트워크 기반의 깊은 신경망이 다양한 컴퓨터 비전 분야에서 좋은 결과를 보여주고 있다. 하지만 신경망 모델을 한정된 자원의 임베디드 기기에서 사용하기 위해서는 가지치기, 양자화, 지식 증류 등 모델 압축(Model compression) 기술을 통해 경량화된 신경망 모델을 만들어야 한다.

[0004] 최근에는 모델 압축 기술 중에서 높은 성능의 깊은 신경망 모델인 교사 모델이 학습한 지식을, 경량화된 모델인 학생 모델의 학습에 사용하는 지식 증류(Knowledge distillation)가 많이 주목받고 있다. 지식 증류는 교사 모델과 학생 모델의 소프트맥스 출력인 소프트 라벨 간의 손실을 최소화하여, 경량화된 학생 모델이 깊은 교사 모델과 비슷한 성능을 갖도록 학습하는 방법이다. 하지만 지식 증류 학습 방법은 교사 모델과 학생 모델의 깊이 차이가 커질수록, 모델의 capacity 차이가 커지게 되어 학생 모델의 지식 증류 학습 최적화가 어려워진다. 이러한 문제점을 해결하기 위해 교사 모델과 학생 모델의 특징맵 간 거리를 최소화하는 사용하는 지식 증류 기술이 제안되었다. 하지만 이러한 기술 역시 모델의 구조가 다르면 교사 모델과 학생 모델 간에 매칭할 특징맵을 임의로 선택할 경우 성능이 하락하는 문제점이 있다. 따라서 매칭할 특징맵을 선택하는 것이 중요한 이슈가 되었으며, 교사 모델과 학생 모델 간 이중 모델의 상관성인 외부적 상관성을 가중치로 사용하여 모델 간 특징맵의 유사도를 최대화할 수 있는 또 다른 지식 증류 기술이 제안되었지만, 이중 모델 간 외부적 상관성을 이용하는 기존 방법은 동일 모델의 특징맵 간 상관성인 내부적 상관성을 무시하고 있다. 이에, 교사 모델의 내부적 상관성과 학생 모델의 내부적 상관성을 유사하게 만들어 두 모델 간 특징맵이 더욱 유사해지도록 하는 기술이 요구된다.

발명의 내용

해결하려는 과제

[0006] 교사 및 학생 모델의 내부적/외부적(Internal/External) 상관성을 고려한 트랜스포머 기반의 이중 모델 지식 증류 방법 및 시스템을 제공할 수 있다.

과제의 해결 수단

[0008] 지식 증류 시스템에 의해 수행되는 트랜스포머 기반의 지식 증류 방법은, 교사 모델의 특징맵과 학생 모델의 특징맵을 각각 임베딩함에 따라 상기 교사 모델의 임베딩 벡터와 상기 학생 모델의 임베딩 벡터를 획득하는 단계; 및 상기 획득된 상기 교사 모델의 임베딩 벡터와 상기 학생 모델의 임베딩 벡터에 대해 트랜스포머 기반의 지식 증류를 수행하여 상관성 정보가 포함된 특징맵을 생성하는 단계를 포함할 수 있다.

[0009] 상기 지식 증류를 수행하는 단계는, 상기 획득된 교사 모델의 임베딩 벡터를 트랜스포머의 인코더에 입력하고, 상기 획득된 학생 모델의 임베딩 벡터를 트랜스포머의 디코더에 입력하는 단계를 포함할 수 있다.

[0010] 상기 지식 증류를 수행하는 단계는, 상기 수행된 트랜스포머 기반의 지식 증류를 통해 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵과 내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵을 생성하는 단계를 포함할 수 있다.

[0011] 상기 지식 증류를 수행하는 단계는, 상기 트랜스포머의 멀티 헤드 어텐션에 트랜스포머의 인코더 출력과 디코더 셀프 어텐션의 출력을 사용하여 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵을 생성하는 단계를 포함할 수 있다.

[0012] 상기 지식 증류를 수행하는 단계는, 상기 트랜스포머의 멀티 헤드 어텐션에 트랜스포머의 인코더 출력과 디코더 셀프 어텐션의 출력을 사용하여 내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵을 생성하는 단계를 포함할 수 있다.

[0013] 상기 지식 증류를 수행하는 단계는, 내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵, 상기

내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵의 전치 행렬 간 내적, 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵 및 상기 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵의 전치 행렬 간 내적의 평균 제곱 오차를 손실 함수로 사용하는 단계를 포함할 수 있다.

[0014] 상기 지식 증류를 수행하는 단계는, 상기 내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵과 상기 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵을 사용하여 상기 교사 모델과 학생 모델 간의 외부적 상관성을 계산하는 단계를 포함할 수 있다.

[0015] 상기 지식 증류를 수행하는 단계는, 상기 계산된 교사 모델과 학생 모델 간의 외부적 상관성과 상기 교사 모델의 특징맵과 학생 모델의 특징맵 간의 손실 함수를 스케일링함으로써 상관성이 높은 특징맵끼리 유사도가 높아지도록 매칭하는 단계를 포함할 수 있다.

[0016] 상기 지식 증류를 수행하는 단계는, 상기 교사 모델의 특징맵과 상기 학생 모델의 특징맵 간의 채널과 공간적 크기를 일치시키기 위해 상기 교사 모델의 특징맵과 상기 학생 모델의 특징맵을 다층 퍼셉트론과 적응적 풀링으로 구성된 변환 함수에 각각 통과시키는 단계를 포함할 수 있다.

[0017] 지식 증류 시스템은, 교사 모델의 특징맵과 학생 모델의 특징맵을 각각 임베딩함에 따라 상기 교사 모델의 임베딩 벡터와 상기 학생 모델의 임베딩 벡터를 획득하는 임베딩 벡터 획득부; 및 상기 획득된 상기 교사 모델의 임베딩 벡터와 상기 학생 모델의 임베딩 벡터에 대해 트랜스포머 기반의 지식 증류를 수행하여 상관성 정보가 포함된 특징맵을 생성하는 지식 증류 수행부를 포함할 수 있다.

발명의 효과

[0019] 내부적 상관성과 외부적 상관성을 활용한 트랜스포머 기반 지식 증류 방법을 통해 학생 모델의 성능을 향상시킬 수 있다.

도면의 간단한 설명

[0021] 도 1은 일 실시예에 있어서, 트랜스포머 기반의 지식 증류 동작을 설명하기 위한 도면이다.

도 2는 일 실시예에 있어서, 지식 증류 시스템의 구성을 설명하기 위한 블록도이다.

도 3은 일 실시예에 있어서, 트랜스포머 기반의 지식 증류 방법을 설명하기 위한 흐름도이다.

도 4는 일 실시예에 있어서, 트랜스포머 기반의 지식 증류 방법의 기술적 효과를 설명하기 위한 성능 분석 표이다.

발명을 실시하기 위한 구체적인 내용

[0022] 이하, 실시예를 첨부한 도면을 참조하여 상세히 설명한다.

[0024] 실시예에서는 특징 상관성(내부적 상관성과 외부 상관성)을 활용할 수 있는 트랜스포머 기반 이중 레이어(이중 모델) 간 지식 증류 동작에 대하여 설명하기로 한다.

[0025] 실시예에서 내부적 상관성이란 동종 모델 내부의 특징맵 간 상관관계(correlation)를 나타내며, 이는 동종 모델 내부 레이어 간 관계성을 나타낸다. 지식 증류 과정에서 내부적 상관성을 활용하여, 학생 모델이 교사 모델의 레이어 간 관계를 모방하도록 학습한다.

[0026] 실시예에서 외부적 상관성이란 이중 모델의 특징맵 간 상관관계를 나타낸다. 교사와 학생 모델의 특징맵 사이의 관계성을 나타내며, 지식 증류 과정에서 외부 상관성을 활용하여, 학생 모델의 특징맵이 자기 자신과 유사한 교사 모델의 특징맵을 모방하도록 학습한다.

[0027] 도 1은 일 실시예에 있어서, 트랜스포머 기반의 지식 증류 동작을 설명하기 위한 도면이다.

[0028] 지식 증류는 교사 모델과 학생 모델의 소프트 맥스 출력 s_t , s_s 사이의 손실을 측정한다. 일반적으로 소프트 라벨 사이의 손실 함수는 쿨백-라이블러 발산(Kullback-Leibler divergence; KL)을 사용한다. 또한, 학생 모델

의 분류 정확도를 향상시킬 수 있도록 s_s 와 GT y_{gt} 간의 크로스-엔트로피(Cross-Entropy; CE)를 손실 함수로 추가한다. 따라서 지식 증류 손실 함수는 수학적 식 1과 같이 정의할 수 있다.

수학적 식 1:

$$L_{KD} = KL(s_s, s_t) + CE(s_s, y_{gt})$$

도 1(b)를 참고하면, 지식 증류 시스템은 트랜스포머를 사용하여, 내부적 상관성과 외부적 상관성을 활용할 수 있는 트랜스포머 기반 지식 증류를 수행할 수 있다. 지식 증류 시스템은 특징맵을 트랜스포머에 입력시키기 전에, 교사 모델의 특징맵 F_T 과 학생 모델의 특징맵 F_S 을 시퀀스 형태로 임베딩할 수 있다. 임베딩 함수 $\psi(\cdot)$ 는 도 1(a)처럼 구성될 수 있다.

수학적 식 2:

$$V_T = \psi(F_T)$$

수학적 식 3:

$$V_S = \psi(F_S);$$

여기서, $F_T \in \mathbb{R}^{B \times C_{t_i} \times H_{t_i} \times W_{t_i}}$ 는 교사 모델의 특징맵이며, $F_S \in \mathbb{R}^{B \times C_{s_j} \times H_{s_j} \times W_{s_j}}$ 는 학생 모델의 특징맵이다. i 와 j 는 교사 모델과 학생 모델 레이어의 인덱스를 나타낸다. 교사 모델과 학생 모델의 임베딩 벡터는 $V_T \in \mathbb{R}^{B \times M \times E}$ 와 $V_S \in \mathbb{R}^{B \times N \times E}$ 이며, M 과 N 은 교사 모델과 학생 모델의 레이어 수를 나타낸다. E 는 임베딩 벡터의 차원을 나타낸다.

지식 증류 시스템은 내부적 상관성을 활용하기 위해 학생 모델의 임베딩 벡터 V_S 를 트랜스포머 인코더 Enc에 입력하고, 교사 모델의 임베딩 벡터 V_T 를 트랜스포머 디코더 Dec의 셀프 어텐션에 입력할 수 있다. 지식 증류 시스템은 트랜스포머의 멀티 헤드 어텐션에 트랜스포머 인코더 Enc 출력과 디코더 Dec 셀프 어텐션 출력을 사용하여, 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵 P_T 를 획득할 수 있다.

수학적 식 4:

$$P_T = Dec(Enc(V_S), V_T)$$

동일한 방법으로, 지식 증류 시스템은 트랜스포머의 멀티 헤드 어텐션에 트랜스포머 인코더 Enc 출력과 디코더 Dec 셀프 어텐션 출력을 사용하여, 내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵 P_S 를 획득할 수 있다.

수학적 식 5:

$$P_S = Dec(Enc(V_T), V_S)$$

이를 통해 지식 증류 시스템은 내부적 상관성과 외부적 상관성이 모두 고려된 특징맵 P_T 와 P_S 를 생성할 수 있다. 특징맵 P_S 와 P_S 의 전치 행렬간 내적과 P_T 와 P_T 의 전치 행렬간 내적의 평균 제곱 오차(Mean Square Error, MSE)를 손실 함수 L_{TR} 로 수학적 식 6처럼 사용할 수 있다.

수학적 식 6:

$$L_{TR} = MSE(P_T \cdot P_T^T, P_S \cdot P_S^T)$$

[0046] 지식 증류 시스템은 교사 모델과 학생 모델 간 특징맵을 비교하여, 외부적 상관성이 높은 특징맵 간의 유사도가 최대화할 수 있는 SemCKD를 사용할 수 있다. SemCKD 는 이중 모델의 특징맵 간의 외부적 상관성을 어텐션 기법을 사용하여 계산한다. 지식 증류 시스템은 외부적 상관성을 특징맵 간 손실 함수의 가중치로 사용하여 상관성이 높은 특징맵끼리 유사해지도록 한다. 지식 증류 시스템은 교사 모델과 학생 모델 사이의 외부적 상관성을 구하기 위해 앞선 P_t 와 P_s 를 사용할 수 있다. 외부적 상관성 α 를 계산하기 위해 수학적 7을 사용한다.

[0047] 수학적 7:

$$\alpha = \frac{\exp(P_T \cdot P_S^T)}{\sum \exp(P_T \cdot P_S^T)}$$

[0049] 수학적 8처럼 α 를 두 특징맵 F_T 와 F_S 간의 손실 함수 MSE 을 스케일링함으로써, 상관성이 높은 특징맵끼리 유사도가 높아지도록 한다.

[0050] 수학적 8:

$$L_{Local} = \alpha \cdot \text{MSE}(\varphi(F_T), \varphi(F_S))$$

[0052] 지식 증류 시스템은 특징맵 F_T 와 F_S 의 채널과 공간적(Spatial) 크기를 일치시키기 위해 특징맵 F_T 와 F_S 을 다층 퍼셉트론과 적응적 풀링(Adaptive Pooling)으로 구성된 변환 함수 $\varphi(\cdot)$ 을 통과시킬 수 있다. 전체적인 손실 함수 $L_{Distill}$ 은 수학적 9와 같다.

[0053] 수학적 9:

$$L_{Distill} = L_{KD} + \beta L_{Local} + \zeta L_{TR}$$

[0055] 수학적 9에서 β 와 ζ 는 각 손실 함수를 스케일링하는 하이퍼 파라미터이다.

[0056] 도 2는 일 실시예에 있어서, 지식 증류 시스템의 구성을 설명하기 위한 블록도이고, 도 3은 일 실시예에 있어서, 트랜스포머 기반의 지식 증류 방법을 설명하기 위한 흐름도이다.

[0057] 지식 증류 시스템(100)의 프로세서는 임베딩 벡터 획득부(210) 및 지식 증류 수행부(220)를 포함할 수 있다. 이러한 프로세서의 구성요소들은 지식 증류 시스템에 저장된 프로그램 코드가 제공하는 제어 명령에 따라 프로세서에 의해 수행되는 서로 다른 기능들(different functions)의 표현들일 수 있다. 프로세서 및 프로세서의 구성요소들은 도 3의 트랜스포머 기반의 지식 증류 방법이 포함하는 단계들(310 내지 320)을 수행하도록 지식 증류 시스템을 제어할 수 있다. 이때, 프로세서 및 프로세서의 구성요소들은 메모리가 포함하는 운영체제의 코드와 적어도 하나의 프로그램의 코드에 따른 명령(instruction)을 실행하도록 구현될 수 있다.

[0058] 프로세서는 트랜스포머 기반의 지식 증류 방법을 위한 프로그램의 파일에 저장된 프로그램 코드를 메모리에 로딩할 수 있다. 예를 들면, 지식 증류 시스템에서 프로그램이 실행되면, 프로세서는 운영체제의 제어에 따라 프로그램의 파일로부터 프로그램 코드를 메모리에 로딩하도록 지식 증류 시스템을 제어할 수 있다. 이때, 임베딩 벡터 획득부(210) 및 지식 증류 수행부(220) 각각은 메모리에 로딩된 프로그램 코드 중 대응하는 부분의 명령을 실행하여 이후 단계들(310 내지 320)을 실행하기 위한 프로세서의 서로 다른 기능적 표현들일 수 있다.

[0059] 단계(310)에서 임베딩 벡터 획득부(210)는 교사 모델의 특징맵과 학생 모델의 특징맵을 각각 임베딩함에 따라 교사 모델의 임베딩 벡터와 학생 모델의 임베딩 벡터를 획득할 수 있다.

[0060] 단계(320)에서 지식 증류 수행부(220)는 획득된 교사 모델의 임베딩 벡터와 학생 모델의 임베딩 벡터에 대해 트랜스포머 기반의 지식증류를 수행하여 상관성 정보가 포함된 특징맵을 생성할 수 있다. 지식 증류 수행부(220)는 획득된 교사 모델의 임베딩 벡터를 트랜스포머의 인코더에 입력하고, 획득된 학생 모델의 임베딩 벡터를 트랜스포머의 디코더에 입력할 수 있다. 지식 증류 수행부(220)는 수행된 트랜스포머 기반의 지식증류를 통해 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵과 내부적 상관성과 외부적 상관성이 모두 고

려된 학생 모델의 특징맵을 생성할 수 있다. 지식 증류 수행부(220)는 트랜스포머의 멀티 헤드 어텐션에 트랜스포머의 인코더 출력과 디코더 셀프 어텐션의 출력을 사용하여 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵을 생성할 수 있다. 지식 증류 수행부(220)는 트랜스포머의 멀티 헤드 어텐션에 트랜스포머의 인코더 출력과 디코더 셀프 어텐션의 출력을 사용하여 내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵을 생성할 수 있다. 지식 증류 수행부(220)는 내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵, 상기 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵 및 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵의 전치 행렬 간 내적, 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵 및 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵의 전치 행렬 간 내적의 평균 제곱 오차를 손실 함수로 사용할 수 있다. 지식 증류 수행부(220)는 내부적 상관성과 외부적 상관성이 모두 고려된 학생 모델의 특징맵과 내부적 상관성과 외부적 상관성이 모두 고려된 교사 모델의 특징맵을 사용하여 교사 모델과 학생 모델 간의 외부적 상관성을 계산할 수 있다. 지식 증류 수행부(220)는 계산된 교사 모델과 학생 모델 간의 외부적 상관성과 교사 모델의 특징맵과 학생 모델의 특징맵 간의 손실 함수를 스케일링함으로써 상관성이 높은 특징맵끼리 유사도가 높아지도록 매칭할 수 있다. 지식 증류 수행부(220)는 교사 모델의 특징맵과 학생 모델의 특징맵 간의 채널과 공간적 크기를 일치시키기 위해 교사 모델의 특징맵과 학생 모델의 특징맵을 다층 퍼셉트론과 적응적 풀링으로 구성된 변환 함수에 각각 통과시킬 수 있다.

[0061] 도 4는 일 실시예에 있어서, 트랜스포머 기반의 지식 증류 방법의 기술적 효과를 설명하기 위한 성능 분석 표이다.

[0062] 실시예에서 제안된 내부적 상관성과 외부적 상관성을 활용한 트랜스포머 기반 지식 증류 방법을 통해 학생 모델의 성능 개선에 기여할 수 있다. 실시예에서 제안된 트랜스포머 기반 지식 증류 방법은 모델의 내부적 상관성을 활용하기 위해 트랜스포머의 인코더를 사용하고, 외부적 상관성을 활용하기 위해 트랜스포머의 디코더를 사용할 수 있다.

[0063] 실시예에서 제안된 트랜스포머 기반 지식 증류 방법은 CIFAR-100 데이터 셋을 통해 성능 평가가 진행될 수 있다. CIFAR-100 데이터 셋은 100 개 클래스의 50,000 장의 학습 샘플과 10,000 장의 평가 샘플로 구성된 이미지 분류 데이터 셋이다. 각 샘플은 32X32 의 해상도를 가진다. 교사 모델로는 ResNet-32 X 4 을 사용하였고, 학생 모델은 ResNet-8 X 4 을 사용하여 지식 증류가 수행될 수 있다. 여기서 "X 4"는 기존 모델의 채널 수의 4 배임을 의미한다.

[0064] 실시예에서는 SemCKD를 기반으로 지식 증류 방법을 구현할 수 있다. 트랜스포머의 인코더와 디코더 레이어는 6 개로 설정하였으며, 디코더의 멀티 헤드 어텐션의 헤드 수는 8 로 설정할 수 있다. 임베딩 벡터의 특징의 수 이 16 일 때, 가장 높은 성능임을 실험적으로 확인할 수 있다. 지식 증류 성능을 확인하기 위해 학생 모델의 이미지 분류 Top-1 정확도를 측정할 수 있다. 초기 학습률은 0.05로 240 에포크(epoch) 동안 Stochastic Gradient Descent 를 사용하여 학습을 진행한다. 그리고 150, 180, 210 에포크 마다 학습률을 0.1 배로 감소시킨다. 수학적 8에서 손실 함수 스케일링을 위한 하이퍼 파라미터 β 와 ζ 는 각각 400 과 0.05 로 사용할 수 있다.

[0065] 최근 지식 증류 방법을 포함하여 이미지 분류 태스크에서 Top-1 정확도를 비교한 결과는 도 4에 도시되어 있다.

[0066] 실시예에서 제안된 트랜스포머 기반 지식 증류 방법은 SemCKD 보다 높은 성능을 보여주는 것을 확인할 수 있다. 이는 모델의 내부적 상관성과 외부적 상관성을 모두 지식 증류에 사용하는 것이 학생 모델 학습에 도움이 되는 것을 보여준다. 최근 지식 증류 방법 중 하나인 ICKD는 소프트 맥스 직전 마지막 레이어의 특징맵의 내부 채널 상관 정보를 전달하는 방법으로 내부적 상관성만을 학습에 이용한다. TaT는 특징맵의 픽셀 단위 매칭을 위해 학생 모델과 교사 모델의 특징맵에 어텐션 매커니즘을 적용하여 외부적 상관성만 사용한다. 반면, 실시예에서 제안된 트랜스포머 기반 지식 증류 방법(IEKD)은 내부적 상관성과 외부적 상관성을 모두 활용하기 때문에 더 좋은 성능을 보여준다.

[0067] 이상에서 설명된 장치는 하드웨어 구성요소, 소프트웨어 구성요소, 및/또는 하드웨어 구성요소 및 소프트웨어 구성요소의 조합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성요소는, 예를 들어, 프로세서, 컨트롤러, ALU(arithmetic logic unit), 디지털 신호 프로세서(digital signal processor), 마이크로컴퓨터, FPGA(field programmable gate array), PLU(programmable logic unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(OS) 및 상기 운영 체제 상에서 수행되는 하나 이상

의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(processing element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 컨트롤러를 포함할 수 있다. 또한, 병렬 프로세서(parallel processor)와 같은, 다른 처리 구성(processing configuration)도 가능하다.

[0068] 소프트웨어는 컴퓨터 프로그램(computer program), 코드(code), 명령(instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성요소(component), 물리적 장치, 가상 장치(virtual equipment), 컴퓨터 저장 매체 또는 장치에 구체화(embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매체에 저장될 수 있다.

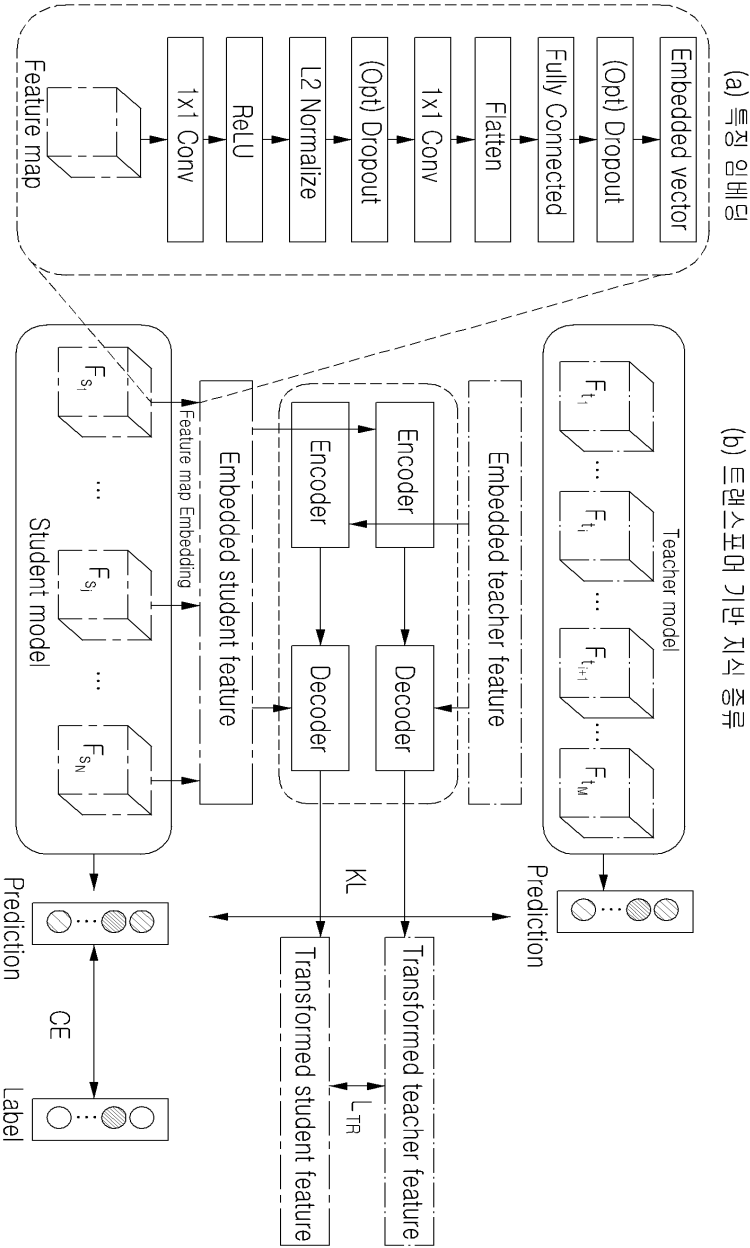
[0069] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(magnetic media), CD-ROM, DVD와 같은 광기록 매체(optical media), 플롭티컬 디스크(floptical disk)와 같은 자기-광 매체(magneto-optical media), 및 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다.

[0070] 이상과 같이 실시예들이 비록 한정된 실시예와 도면에 의해 설명되었으나, 해당 기술분야에서 통상의 지식을 가진 자라면 상기의 기재로부터 다양한 수정 및 변형이 가능하다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다.

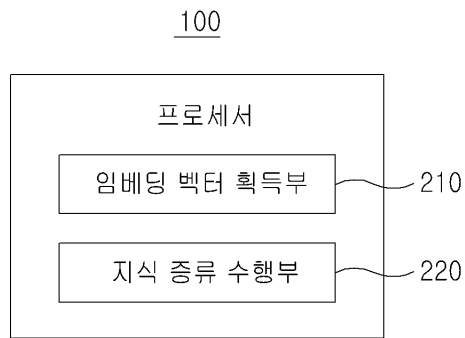
[0071] 그러므로, 다른 구현들, 다른 실시예들 및 특허청구범위와 균등한 것들도 후술하는 특허청구범위의 범위에 속한다.

도면

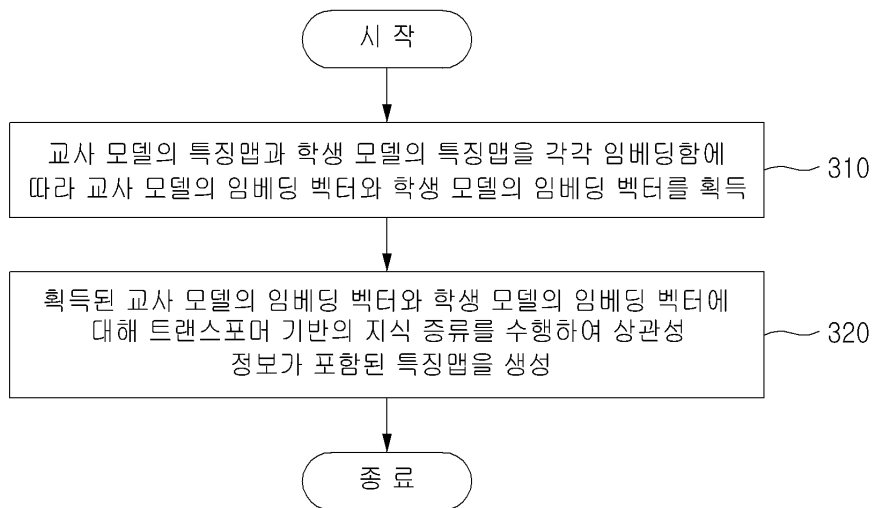
도면1



도면2



도면3



도면4

방법	Top-1 정확도
교사 모델 (ResNet-32×4)	79.42
학생 모델 (ResNet-8×4)	73.09 ± 0.30
KD	74.42 ± 0.05
FitNet	74.32 ± 0.08
AT	75.07 ± 0.03
SP	74.29 ± 0.07
VID	74.55 ± 0.10
HKD	74.86 ± 0.21
SemCKD	76.23 ± 0.04
ICKD	75.48
TaT	75.54
IEKD (Ours)	76.52