강건한 이미지-텍스트 검색을 위한 소프트 레이블 앙상블과 피처 융합 기법

양태영[1], 신동현[2], 이승철[3], 박창현[4]
[1]영남대학교 컴퓨터공학과 학부생, [2]영남대학교 컴퓨터공학과 석사과정,
[3]영남대학교 컴퓨터공학과 박사수료, [4]영남대학교 컴퓨터공학과 교수
xodud120016@gmail.com, dkvk486967@gmail.com, fatalist316@gmail.com,
park@yu.ac.kr

Robust Image-Text Retrieval via Soft-Label Ensemble and Feature Fusion

Tae-Young Yang[1], Dong-Hyun Shin[2], Seung-Cheol Lee[3], Chang-Hyun Park[4],
[1234]Dept. of Computer Engineering, Yeungnam University

요 약

이미지-텍스트 검색(Image-Text Retrieval, ITR)은 이미지와 텍스트 간 의미적 대응 학습을 목표로 한다. 그러나 기존 contrastive learning 기반 ITR 모델은 false negative 문제와 intra-modal semantic loss라는 구조적 한계를 가진다. 이를 보완하기 위해 soft label을 활용하는 연구가 제안되었으나, 교사-학생 간 도메인 불일치로 인해 강건성이 저하되는 문제가 발생한다. 본 논문은 이를 해결하기 위해 교사 소프트 레이블 앙상블과 교사-학생 피처 융합 기반 표현 보강을 통합한 학습 방식을 제안하며, 도메인 불균형 상황에서도 강건한 성능 향상을 기대할 수 있다.

1. 서론

이미지-텍스트 검색(ITR)은 이미지와 텍스트의 의미적 대응을 학습하는 연구분야다. 대조학습 기반 ITR은 cross-modal 매칭 누락과 intra-modal 의미 붕괴가 잦고, CUSA가 soft label로 이를 완화했지만 교사-학생 도메인 불일치에서 편향이 발생한다 [1]. 본 연구는 이를 해결하기 위해 교사 soft label 앙상블과 교사-학생 특징 융합을 결합한 학습전략을 제안하여 ITR 모델의 도메인 강건성과 검색 성능을 향상시키고자 한다.

2. 관련 연구

CLIP은 대규모 이미지-텍스트 쌍을 활용해 공통 임베딩 공간을 구축 높은 성능을 보였으나, 동일 이미지를 설명할 수 있는 다양한 텍스트를 부정 예시로 처리하는 거짓 부정 문제가 발생한다 [2]. 이를 보완한 CUSA는 hard label의 이진적 한계를 넘어 cross-modal 및 uni-modal soft label을 도입했으나, 교사-학생 간 도메인 불일치 상황에서는 soft label이 오히려 잡음 신호로 작용할 수 있다. Rao et al.은 대조 학습과 지식 증류를 결합하였으나, 단일 교사 모델에 대한 의존을 완전히 해결하지 못했다 [3].

3. 제안 학습전략

본 논문은 이미지-텍스트 검색에서 발생하는 도메인 불일치와 false negative 문제를 해결하기 위해, Soft-label Ensemble 정렬과 의미적 표현 보강을 결합한 학습전략을 제안한다. 제안하는 학습전략은 그림 1과 같다.

3.1 Soft-label Ensemble

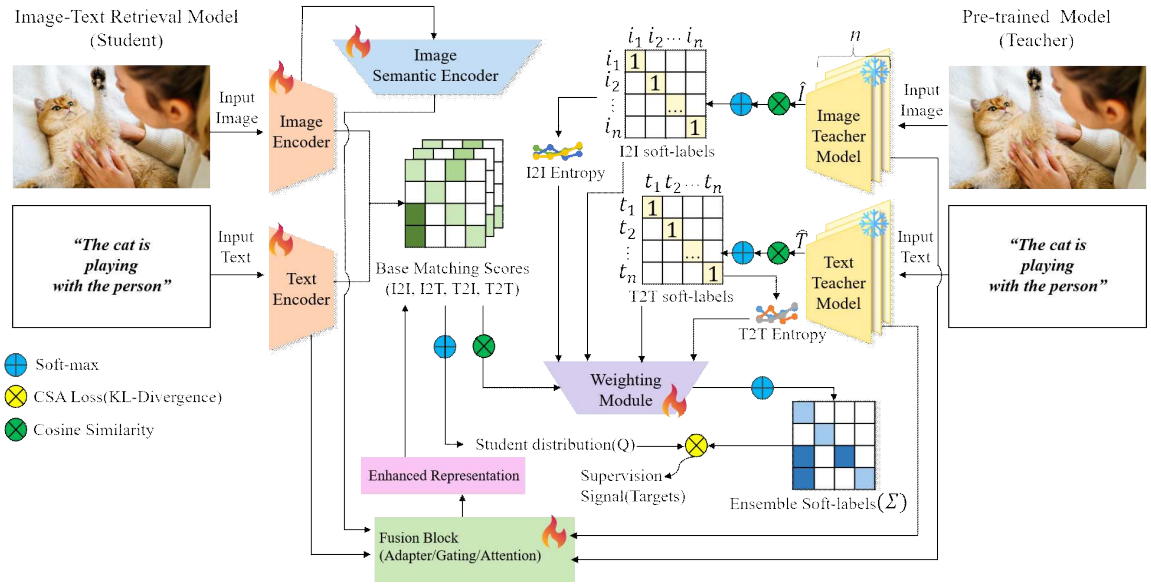
사전학습된 이미지 및 텍스트 교사 모델은 입력 샘플을 기반으로 각각 I2I, T2T soft-label 분포를 생성한다. 각 soft-label은 엔트로피와 교사 간 코사인 유사도를 이용해 가중치 모듈에서 가중치 $w_{i,k}^{I2I}$, $w_{t,l}^{T2T}$가 부여된다. 최종 앙상블 soft-label은 수식 (1)과 같이 정의된다:

$$Soft-labels = \sum_{i,k}^{n} w_{i,k}^{I2I} p_{i,k}^{I2I} + \sum_{t,l} w_{t,l}^{T2T} p_{t,l}^{T2T} \quad (1)$$

여기서 $p_{i,k}^{I2I}, p_{t,l}^{T2T}$는 각각 교사 분포이며, 가중치는 엔트로피가 낮고 교사 간 일관성이 높은 분포에 더 크게 부여된다.

3.2 학생 분포 정합

학생 모델은 입력 이미지와 텍스트로부터 Base



(그림 1) 이미지-텍스트 검색 성능 향상을 위한 학습 전략 개요도

Matching Scores를 계산한다. 이를 Soft-max를 통해 학생 분포 Q로 정규화한 뒤, 앙상블 soft-label와의 차이를 수식 (2)와 같이 CSA Loss(KL Divergence)로 최소화한다.

$$L_{CSA} = KL(Soft-labels \,\|\, Q) \qquad (2)$$

3.3 이미지 시맨틱 인코더와 의미적 표현 보강

학생 이미지 임베딩을 Image Semantic Encoder로 고수준 의미 특징으로 변환하고, Fusion Block (어댑터/게이팅/어텐션)에서 학생과 교사 및 이미지와 텍스트 특징을 경량 결합한다. 융합 특징은 매칭 입력에 주입되어 도메인 불일치 환경에서도 의미 보존을 강화한다. Fusion Block은 학생 인코더의 중간 임베딩(z^I,z^T)과 교사 임베딩(t^I, t^T)을 입력으로 받아, 경량 어댑터 A_I, A_T로 교사 표현을 학생 공간에 정렬한뒤 결합하여 보강 표현 $z^{I'}$,$z^{T'}$를 생성한다.

L_{fusion}은 수식 (3)과 같이, 학생 표현이 보강된 입력에서 교사 표현의 정렬 방향을 따르도록 지식 증류를 적용하고, 보강 전후 표현의 과도한 편향을 완화하기 위해 잔차 규제를 적용한다. $\beta > 0$는 보강 강도를 조절한다. 또한 cross-modal 직접 가중(I2T)은 라벨 누수 및 FN 증폭을 야기하므로 제외하고, 해당 정보는 Base Matching-Fusion 경로에서 간접 반영된다.

$$L_{fusion} = \|z^{I'} - A_I t^I\|_2^2 + \|z^{T'} - A_T t^T\|_2^2 + \beta(\|z^{I'} - z^I\|_2^2 + \|z^{T'} - z^T\|_2^2) \qquad (3)$$

3.4 최종 목적함수

최종 손실은 (4)과 같으며, λ는 정렬 손실과 융합 규제 간 균형 계수다.

$$L_{total} = L_{CSA} + \lambda L_{fusion} \qquad (4)$$

4. 결론 및 향후 연구

본 연구는 엔트로피·합의도 기반 교사 소프트-레이블 앙상블과 피처 융합을 단일 목적함수로 결합해, 도메인 불일치와 false negative에 강건한 ITR 학습을 제안하였다. 향후 연구로는 MS-COCO (1K/5K)와 Flickr30K에서 R@K, mR, mAP (유의성)로 CLIP, CUSA, DCD, AMMKD와 cost 비교 및 ablation study (단일 교사, 엔트로피·합의 가중모듈 및 Fusion block)를 수행해 I2I 및 T2T 정렬만으로 R@K 개선과 도메인 편차 강건성을 입증할 계획이다.

참고문헌

[1] H. Huang, Z. Nie, Z. Wang, Z. Shang, "Cross-Modal and Uni-Modal Soft-Label Alignment for Image-Text Retrieval," AAAI Conf. Artif. Intell., Vancouver, 2024, pp. 18298 - 18306.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., "Learning Transferable Visual Models from Natural Language Supervision," ICML, Vienna, 2021, pp. 8748 - 8763.

[3] J. Rao, L. Ding, S. Qi, M. Fang, Y. Liu, L. Shen, D. Tao, "Dynamic Contrastive Distillation for Image-Text Retrieval," IEEE Trans. Multimedia, vol. 26, no. 3, pp. 832 - 844, 2023.