멀티모달 대조학습을 활용한 실세계 이상형 추천

김지훈, 최은기, 박준형, 박천음*

국립한밭대학교

{20227128, 20197123, 20221991}@edu.hanbat.ac.kr, parkce@hanbat.ac.kr

Multimodal Contrastive Learning for Real-World Ideal Type Recommendation

Jihun Kim, Eungi Choi, Junhyeong Park, Cheoneum Park*

Hanbat National University

요 약

본 논문에서는 실제 데이팅 앱에서 성사된 커플 데이터를 활용하여, 사용자 프로필 이미지 기반의 매칭 예측 시스템을 제안한다. 제안 방법은 Qwen3-VL을 백본으로 하여 시각적 특징을 추출하고, 경량화된 프로젝션 헤드를 결합하여 매칭 임베딩을 contrastive learning으로 학습한다. 본 논문에서는 실험을 위해 실제 매칭된 495쌍의 커플 데이터를 활용하여 남녀 이상형 쌍 데이터를 구축한다. 실험 결과, 제안 방법모델은 Recall@20에서 50.0%와 MRR 0.216으로 유의미한 성능을 달성하였다.

I. 서 론

온라인 데이팅 서비스에서 사용자 매칭은 주로 나이, 거주 지역, 관심사나 간단한 자기소개 문구 같은 텍스트 프로필 정보에 의존한다. 그러나 실제 사용자는 프로필 사진에서 드러나는 시각적 요인에 크게 좌우되며, 이러한 암묵적 선호를 정량화하는 것은 어려운 과제이다. 기존 연구 [1, 2]는 대부분 텍스트 프로필 유사도나 협업 필터링에 의존하기 때문에 사용자의 시각적 선호를 반영하지 못하는 한계가 있다. 사회심리 이론에는 사람은 자신과 유사한 수준의 신체적 매력도나 사회적 바람직성을 가진 상대를 선호하는 경향을 가진다는 매칭 가설(The Matching Hypothesis)이 연구되었다 [3]. 이 가설에 기반하여, 본 논문에서는 남녀 프로필 사진과 프로필 정보, 이미지 설명을 기반으로한 이상형 추천 문제를 정의하고 해결 방법을 제안한다.

본 논문에서 제안한 방법은 Qwen3-VL 모델 [4]을 백본으로 활용하고, 저차원 프로젝션 레이어를 추가하여 텍스트-이미지 기반 추천 문제를 학습한다. 학습을 위하여, contastive learning 방법 [5] 중 하나인 InfoNCE Loss [6]를 적용한다. 이는 서로 이상형인 커플 쌍을 임베딩 공간에서 가깝게, 비커플 쌍을 멀어지도록 학습한다. 본 논문의 주요 기여점은 다음과 같다: (1) 실제 커플 데이터 495쌍을 활용한 contrastive learning 기반 모델 제안, (2) VLM 백본을 기반으로 한 추천 모델 설계, (3) 사전 학습된 VLM 베이스라인 대비 Recall@5 2.23배, MRR 2.38배 성능 향상을 달성한다.

II. 본 론

사용자의 프로필 이미지를 x_i라 정의한다. 커플 쌍 (x_f, x_m)에 대해, 두 사용자의 임베딩(embeddings) e_f, e_m이 임베딩 공간에서 가깝도록 학습하며, 비커플 쌍은 멀어지도록 대조 학습을 수행한다. 실제 데이팅 앱에서 상호 좋아요를 통해 매칭된 커플 495쌍의 프로필 이미지를 수집하였으며, Train 346쌍(70%), Validation 74쌍(15%), Test 75쌍(15%)으로 분할한다.

본 논문에서 제안하는 방법은 [그림 1]과 같으며, 멀티모달(multimodal) 이해 능력을 활용하기 위해 이미지와 함께 구조화된 메타데이터(metadata)를 입력으로 사용한다. 메타데이터는 GPT-5-nano를 활용하여 각 프로필 이미지로부터 자동 생성하며, 외모 유형(appearance type), 스타일 분위기(style vibe), 추정 성격(personality impression), 단정함 수준(grooming level), 촬영 유형(photo style), 이미지 품질(image quality), 외형적 특징(physical features), 이미지 설명(caption)의 8가지 속성으로 구성된다. 메타데이터는 Vision Language Model (VLM)이 매칭 호환성 관련 시각적 특징을 효과적으로 추출하도록 유도한다.

제안 방법의 백본(backbone) 모델은 Qwen3-VL-2B[4]를 사용한다. Qwen3-VL은 이미지와 텍스트를 함께 이해하는 VLM로, 다양한 시각적 태스크에서 뛰어난 일반화 성능을 보인다. 소규모 도메인 데이터에서의 과적합을 방지하기 위해, 백본 네트워크의 약 20억 개 파라미터는 모두 동결(freezing)한다.

백본을 통과한 고차원 특징 벡터 h_i는 [식 1]과 같으며, 여기서 d는 2048이다.

$$h_i = f_\theta(x_i) \in \mathbb{R}^d \tag{1}$$

매칭 호환성 학습에 특화된 저차원 임베딩을 얻기 위해, 학습 가능한 프로젝션 헤드(projection head) g_ϕ를 추가하며 식은 아래와 같다. 여기서 프로젝션 헤드는 2개의 선형 계층과 BatchNorm, ReLU 활성 함수로 구성된다. L2 정규화를 통해 모든 임베딩 벡터를 단위 초구(hypersphere) 상에 투영하여, 코사인 유사도와 유클리드 거리가 단조 관계를 가지게 한다.

$$z_i = g_\phi(h_i) = W_2 \cdot \mathrm{ReLU}(\mathrm{BN}(W_1 \cdot h_i)) \tag{2}$$

$$e_i = \frac{z_i}{\|z_i\|_2} \in \mathbb{R}^d \tag{3}$$

*교신 저자

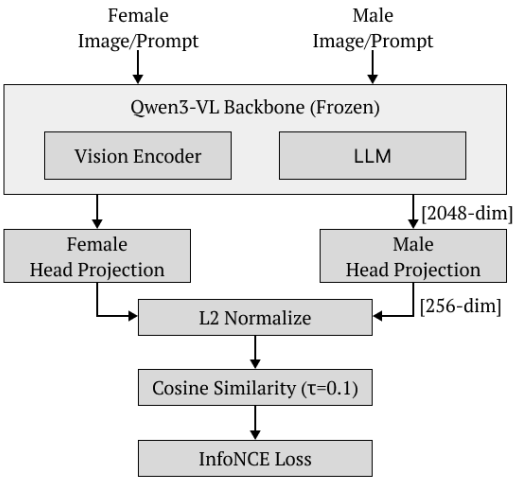

그림 1: 제안 모델의 전체 아키텍처

실제 커플 쌍을 Positive, 배치 내 다른 사용자들을 Negative로 하여 대조 학습을 수행한다. CLIP[7]에서 검증된 InfoNCE Loss를 사용하며, 손실 함수는 [식 4]와 같다.

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(e_f, e_m)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(e_f, e_{m_j})/\tau)} \quad (4)$$

여기서 τ는 temperature 하이퍼파라미터이며, sim은 코사인 유사도, N은 배치 크기를 나타낸다.

III. 실험 및 결과

본 논문에서 수행한 실험은 다음과 같다. 학습 인프라는 AWS SageMaker의 NVIDIA A10G GPU (24GB)를 사용한다. 학습률 5e-5, 배치 크기 48쌍, Temperature(τ) 0.1, Weight Decay 1e-3으로 설정하며, 30 에폭까지 학습하고 Early Stopping을 적용한다.

[표 1]은 제안 방법의 실험 결과를 확인하며, 베이스라인(Qwen3-VL)과 성능을 비교한다. 실험 결과, 제안 모델은 베이스라인 대비 R@5에서 2.59배, R@30에서 1.40배 향상된 성능을 달성한다. 이에 따라, 도메인 특화 프로젝션 헤드를 학습하는 것이 유의미함을 알 수 있다.

표 1: 실세계 이상형 추천 성능 비교

모델	R@5	R@10	R@20	R@30	R@50
Random	6.7%	13.3%	26.7%	40.0%	66.7%
Qwen3-VL	11.3%	18.7%	30.0%	43.3%	72.7%
Ours	**29.3%**	**37.3%**	**50.0%**	**60.7%**	**82.0%**
향상률	2.59배	1.99배	1.67배	1.40배	1.13배

본 논문에서 제안한 방법의 모델과 하이퍼파라미터 최적화는 다음과 같다. 먼저, VLM의 hidden states에서 단일 임베딩 벡터를 추출하기 위해 EOS Token Pooling과 Mean Pooling 두 가지 전략을 비교한다. [표 2]은 각 Pooling 전략의 베이스라인 성능을 나타낸다. 실험 결과, Mean Pooling이 EOS Token 대비 R@5에서 1.41배, R@50에서 1.05배 높은 성능을 보여, 이후 모든 실험에서 Mean Pooling을 기본 전략으로 사용한다.

표 2: Pooling 전략별 베이스라인 성능 비교

Pooling	R@5	R@10	R@20	R@30	R@50
EOS Token	8.0%	14.0%	29.3%	**46.0%**	69.3%
Mean Pooling	**11.3%**	**18.7%**	**30.0%**	43.3%	**72.7%**
향상률	1.41배	1.34배	1.02배	0.94배	1.05배

표 3: 하이퍼파라미터 최적화에 따른 성능 변화

설정	R@5	R@10	R@20	R@30	R@50
기본 (dim=256)	25.3%	36.7%	43.3%	54.7%	74.7%
dim=512	**29.3%**	36.7%	49.3%	57.3%	78.0%
τ=0.2	29.3%	**37.3%**	**50.0%**	**60.7%**	**82.0%**

최적의 모델 성능을 확보하기 위해 주요 하이퍼파라미터에 대한 실험을 수행한다. [표 3]은 프로젝션 차원과 Temperature 변화에 따른 성능 변화를 보이며, 각각 베이스라인은 256과 0.1로 설정한다. 하이퍼파라미터는 프로젝션 차원을 512로 할 경우, 베이스라인에 비해 약 4%p 가량 향상된 R@5 29.3%을 보인다. Temperature를 0.2로 설정할 경우, 베이스라인인 0.1에 비하여 전반적으로 성능이 향상됨을 알 수 있다.

IV. 결 론

본 논문에서는 실제 데이팅 앱의 매칭 데이터를 활용하여 시각적 매칭 호환성을 예측하는 VLM 기반 방법을 제안하였다. Qwen3-VL을 기반으로 VLM 파라미터를 동결하고, 프로젝션 헤드만 학습하는 효율적인 구조를 설계하였으며, InfoNCE Loss 기반 대조 학습을 통해 실제 커플 쌍에 대한 학습을 최적화하였다. 실험 결과, 제안 모델은 사전 학습된 VLM 베이스라인 대비 R@5에서 2.59배, MRR에서 2.63배 향상된 성능을 달성하였다.

향후 연구로는 텍스트 임베딩(성격, 가치관)과 이미지 임베딩을 결합한 하이브리드 모델 개발, 그리고 더 많은 커플 데이터 확보를 통한 일반화 성능을 고도화할 계획한다.

참 고 문 헌

[1] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proc. CVPR*, pp. 815–823, 2015.

[2] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[3] E. Walster, V. Aronson, D. Abrahams, and L. Rottman, "Importance of physical attractiveness in dating behavior," *Journal of Personality and Social Psychology*, vol. 4, no. 5, pp. 508–516, 1966.

[4] J. Bai *et al.*, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.

[5] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, JMLR.org, 2020.

[6] A. Parulekar, L. Collins, K. Shanmugam, A. Mokhtari, and S. Shakkottai, "Infonce loss provably learns cluster-preserving representations," in *Proc. COLT*, pp. 1914–1961, 2023.

[7] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021.