

데이팅 앱 스타일 매칭 시스템: Metric Learning과 Vision-Language Model을 활용한 접근

초록 (Abstract)

본 연구는 기존 텍스트 기반 데이팅 매칭 시스템의 한계인 시각적 비언어 정보(Non-verbal Cues)의 부재를 해결하기 위해, Vision-Language Model (VLM)과 Metric Learning을 결합한 새로운 스타일 매칭 시스템을 제안한다. 실제 서비스에서 성사된 121건의 매칭 데이터를 기반으로 Qwen3-VL 모델을 통해 사용자 프로필의 패션 스타일과 분위기를 추출하고, Triplet Margin Loss를 적용하여 스타일 유사도를 학습했다. 제안된 시스템은 텍스트 정보만으로는 포착하기 어려운 사용자의 심미적 취향을 정량화하여 매칭 만족도를 제고하는 것을 목표로 하며, 실험 결과 Recall@1 60% 이상의 성능을 달성하여 실제 서비스 적용 가능성을 확인했다.

1. 프로젝트 개요 (Project Overview)

1.1 연구 배경 및 필요성

현재 운영 중인 데이팅 서비스(가입자 약 5,000명)는 가입 시 사용자가 입력한 텍스트 정보(이상형, 성격, MBTI 등)에 기반한 매칭 알고리즘을 사용하고 있다. 그러나 텍스트 기반 매칭은 **사용자의 외모 스타일과 분위기 (Vibe)를 반영하지 못한다는** 근본적인 한계가 있다.

실제 서비스 운영 데이터 분석 결과, 다음과 같은 문제점이 식별되었다:

- 외모/스타일 불일치로 인한 만족도 저하:** 텍스트 매칭 점수가 높아도, 실제 프로필 사진 확인 단계에서 스타일 불호로 인한 매칭 실패 빈번
- 낮은 상호 좋아요(Like) 비율:** 텍스트 정보만으로는 상대방의 시각적 매력을 예측하기 어려워 매칭 성공률 정체

본 프로젝트는 서비스 내 **121건의 실제 성사 데이터(실제 만남까지 이어진 상호 좋아요 케이스)**를 기반으로, **사용자의 프로필 이미지에서 패션 스타일과 분위기를 추출하여 매칭 알고리즘에 통합함으로써 매칭 만족도를 획기적으로 개선하고자 한다.**

1.2 연구 목표

기술적 목표

1. VLM 기반 스타일 특징 추출기(Feature Extractor) 구축

- Qwen3-VL-2B 모델을 활용하여 이미지의 시각적 정보와 텍스트 프롬프트("Describe this person's appearance.")를 결합한 고차원 스타일 특징 추출

2. Metric Learning 기반 임베딩 모델 개발

- Triplet Margin Loss와 PKSampler를 적용하여 동일 스타일 간의 거리는 좁히고, 다른 스타일 간의 거리는 넓히는 임베딩 공간 학습
- 5가지 핵심 스타일(Casual, Street, Sporty, Chic, Classy)에 대한 분류 및 임베딩 성능 최적화

3. 검증 지표 달성

- Recall@1: 60% 이상
- Recall@5: 80% 이상

1.3 연구의 의의 및 기대 효과

1) 학술적/기술적 의의 (Research Contributions)

- **멀티모달 데이터 매칭 방법론 제안:** 기존 텍스트(프로필) 중심의 매칭 시스템 한계를 극복하기 위해, VLM(Vision-Language Model)을 활용하여 이미지의 시각적 스타일 정보를 결합한 새로운 멀티모달 매칭 프레임워크를 제안한다.
- **실제 서비스 데이터 기반의 실증 연구:** 공개 데이터셋(Public Dataset)이 아닌, 실제 운영 중인 서비스의 성공 매칭 데이터(Ground Truth)를 활용하여 현실적인 매칭 성능을 검증한다.
- **VLM과 Metric Learning의 결합:** 범용 VLM(Qwen3-VL)의 강력한 표현력과 Metric Learning(Triplet Loss)의 정교한 거리 학습을 결합하여, 적은 데이터로도 패션 스타일의 미세한 차이를 효과적으로 학습할 수 있음을 보인다.

2) 사회적/경제적 기대 효과

- **매칭 만족도 및 성공률 제고:** 정량화하기 어려웠던 '스타일'과 '분위기'를 매칭에 반영함으로써, 사용자 만족도를 높이고 플랫폼의 신뢰도를 강화한다.
- **Cold-Start 문제 완화:** 생성형 AI를 활용한 데이터 증강 파이프라인을 구축하여, 초기 데이터가 부족한 신생 서비스에서도 고성능 매칭 모델을 도입할 수 있는 가능성을 제시한다.

2. 관련 연구 및 기술 분석 (Related Work)

2.1 Metric Learning & Triplet Loss

Metric Learning은 데이터 간의 유사도를 벡터 공간 상의 거리로 학습하는 방법론이다. 본 연구에서는 **Triplet Margin Loss**를 채택하여, 동일한 스타일 클래스(Positive) 간의 거리는 최소화하고, 다른 스타일 클래스(Negative) 간의 거리는 최대화하는 임베딩 공간을 구축했다.

- **Triplet Loss:** Anchor(\$A\$), Positive(\$P\$), Negative(\$N\$) 샘플 간의 거리 관계를 학습하는 순실 함수로, 수식은 다음과 같다.
$$L = \max(d(f(A), f(P)) - d(f(A), f(N)) + \alpha, 0)$$
 여기서 \$d\$는 거리 함수(Cosine Distance), \$\alpha\$는 마진(Margin)을 의미한다.
- **Online Semi-hard Mining:** 학습 효율성과 안정성을 동시에 확보하기 위해, 배치(Batch) 내에서 \$d(A, P) < d(A, N) < d(A, P) + \alpha\$ 조건을 만족하는 **Semi-hard Negative** 샘플을 동적으로 선별하여 학습에 반영했다. 이는 Hard Negative의 불안정성을 완화하면서도 충분한 학습 난이도를 제공한다.

2.2 Vision-Language Models (VLM)

기존의 CNN 기반 이미지 분류 모델(ResNet 등)과 달리, VLM은 이미지와 텍스트의 의미적 관계를 동시에 이해 할 수 있다.

- **Qwen3-VL:** 본 연구에서는 한국어 처리에 강점이 있고, 텍스트 프롬프트를 통해 이미지의 특정 속성(외모, 스타일)에 집중할 수 있는 Qwen3-VL 모델을 백본으로 선정했다. 이를 통해 단순한 객체 인식을 넘어, 이미지 전반의 '분위기(Vibe)'와 '스타일'을 포착하는 고차원 특징 추출이 가능하다.

3. 데이터셋 구축 (Data Preparation)

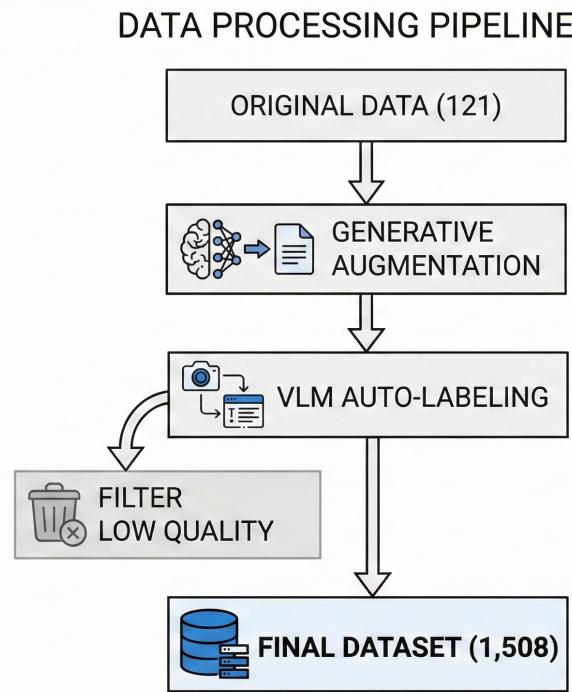
3.1 데이터셋 구성 (Dataset Construction)

- **Ground Truth Data:** 실제 서비스에서 성사된(상호 좋아요 및 만남) 고품질 데이터 121건을 확보하여, 스타일 매칭의 기준점(Anchor)으로 활용했다.

- **Augmented Data:** 딥러닝 모델 학습에 필요한 데이터 양을 확보하기 위해, 생성형 AI 기반 증강 기법을 적용하여 총 1,508장의 학습 데이터를 구축했다.
- **Validation Data:** 학습 데이터와 겹치지 않는 별도의 98장 이미지를 검증셋으로 구성하여, 모델의 일반화 성능을 평가했다.

3.2 데이터 증강 및 라벨링 파이프라인

데이터 부족 문제를 해결하기 위해 **Silver Standard** 방식의 증강 및 자동 라벨링 파이프라인을 구축했다.



1. Generative Data Augmentation:

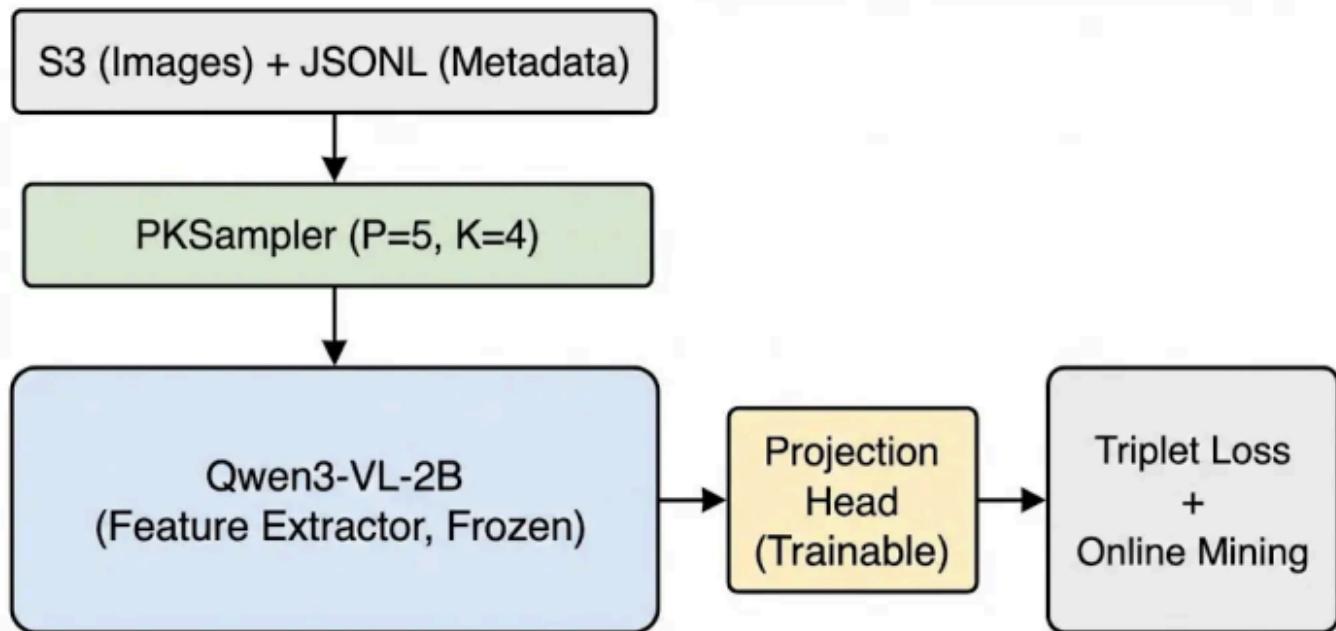
- **Method:** 원본 이미지의 핵심적인 스타일(Vibe)과 분위기는 유지하면서, 배경, 조명, 포즈 등 비핵심 요소를 변형하는 **Silver Standard** 증강 기법을 적용했다.
- **Tools:** Seedream 4 (이미지 생성)

2. AI-Assisted Labeling:

- **Method:** 대규모 VLM(Gemini-2.5-flash)을 활용하여 증강된 이미지에서 구조화된 메타데이터를 추출하는 자동 라벨링 파이프라인을 구축했다.
- **Label Schema:**
 - **fashion_style** (5 classes): Casual_Basic, Street_Hip, Sporty_Athleisure, Chic_Modern, Classy_Elegant
 - **visual_quality**: 학습 데이터의 품질 관리를 위해 Low quality 이미지는 필터링했다.

4. 제안하는 방법론 (Proposed Method)

4.1 모델 아키텍처 (Model Architecture)



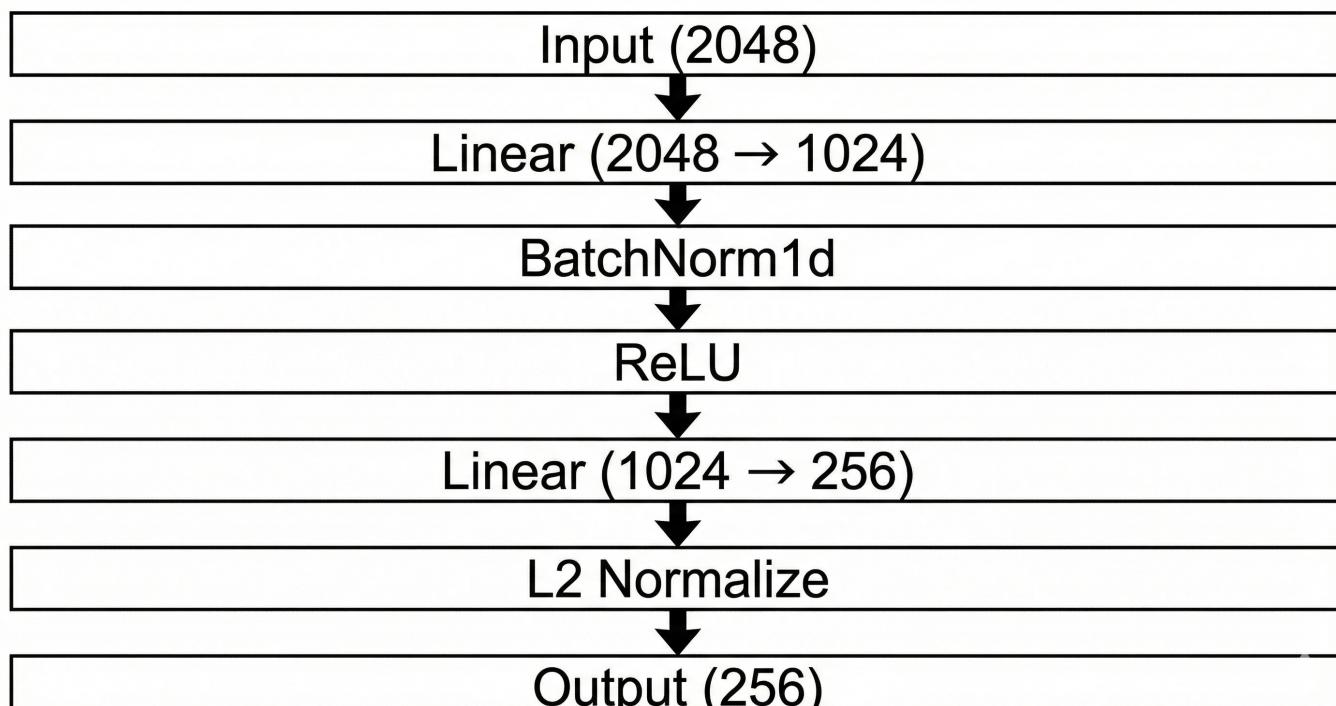
본 시스템은 **Qwen3-VL** 백본과 학습 가능한 **Projection Head**로 구성된다.

1. Backbone: Qwen3-VL-2B-Instruct (Frozen)

- **입력:** 이미지 + 텍스트 프롬프트 ("Describe this person's appearance.")
- **처리:**
 - 이미지와 텍스트를 멀티모달 프로세서로 처리
 - 마지막 Hidden State 추출 후 **Mean Pooling**을 통해 시퀀스 차원 축소
 - 출력 차원: 1536 (Qwen3-VL hidden size)
- **특징:** Vision Tower만 사용하는 것이 아니라, 텍스트 프롬프트를 통해 '외모'에 집중된 특징을 추출하도록 유도

2. Projection Head (Trainable)

백본에서 추출된 일반적인 특징을 스타일 매칭에 특화된 저차원 임베딩으로 변환한다.



- **구조:** Linear(1536 -> 2048) -> LayerNorm -> GELU -> Dropout(0.1) -> Linear(2048 -> 256)
- **역할:** 고차원 특징을 압축하고 비선형성을 추가하여 스타일 구분력 강화

3. Loss Function

- **Online Triplet Margin Loss:** Margin = 0.3
- **Distance Metric:** Cosine Similarity

4.2 학습 전략 (Training Strategy)

- **PKSampler:** 클래스 불균형 해소 및 Triplet 구성률 보장하기 위해 배치 당 P개의 클래스에서 K개의 샘플을 무작위 추출 (P=5, K=4, Batch Size=20)
- **Mixed Precision:** bfloat16 적용으로 메모리 효율성 및 학습 속도 증대
- **Optimizer:** AdamW (Learning Rate: 1e-4)

5. 실험 및 평가 (Experiments)

5.1 실험 환경

- **Infrastructure:** AWS SageMaker (g5.2xlarge)
- **GPU:** NVIDIA A10G (24GB VRAM)
- **Framework:** PyTorch 2.0+, Transformers 4.57+

5.2 평가 지표 (Metrics)

모델의 성능은 다음 지표를 통해 정량적으로 평가된다.

1. **Recall@K (R@1, R@5):** 질의 이미지와 가장 유사한 상위 K개 결과 중 동일한 스타일이 포함될 확률. 검색/매칭 시스템의 핵심 지표.
2. **MAP@R (Mean Average Precision at R):** 검색 결과의 순위까지 고려한 정밀도.
3. **Validation Loss:** 학습 과정(Overfitting) 여부 모니터링.

5.3 정성적 평가 (Qualitative Analysis)

- **t-SNE 시각화:** 256차원 임베딩 공간을 2차원으로 축소하여 시각화. 5가지 스타일 클래스가 명확하게 군집(Cluster)을 형성하는지 확인.

6. 결론 및 향후 계획 (Conclusion & Future Work)

6.1 결론

본 프로젝트는 텍스트 중심의 데이터 앱 매칭 시스템에 VLM 기반의 시각적 스타일 매칭을 도입하는 시도이다. Qwen3-VL의 강력한 표현력과 Metric Learning의 정교한 거리 학습을 결합하여, 사용자의 취향을 보다 입체적으로 반영할 수 있는 기술적 토대를 마련했다.

6.2 향후 계획

1. **하이브리드 매칭 엔진 구현:** 텍스트 임베딩(성격/가치관)과 이미지 임베딩(스타일)을 가중 합(Weighted Sum)하여 최종 매칭 점수 산출
 2. **실시간 서빙 최적화:** 임베딩 벡터를 Vector DB(Milvus, Pinecone 등)에 인덱싱하여 실시간 검색 속도 확보
 3. **A/B 테스트:** 실제 서비스 일부 유저군에 적용하여 매칭 성과율 및 유저 만족도 변화 측정
-

References

1. F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815-823.
2. A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8748-8763.
3. J. Bai et al., "Qwen-VL: A Versatile Vision-Language Model," *arXiv preprint arXiv:2308.12966*, 2023.
4. Z. Liu et al., "Visual Instruction Tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.