# 멀티모달 임베딩 유사도 학습과 교차어텐션 융합 구조를 활용한 의류 신제품 수요 예측 연구

이정환[1], 김현철[2]
[1]고려대학교 SW·AI 융합대학원 빅데이터융합학과 석사과정
[2]고려대학교 정보대학 컴퓨터학과
jhl114@korea.ac.kr, harrykim@korea.ac.kr

# A Multimodal New Apparel Demand Forecasting Model Based on Embedding Similarity Learning and Cross-Attention Fusion

JungHwan Lee[1], Hyeonchoel Kim[2]
[1]Dept. of Big Data Convergence, Graduate School of SW·AI Convergence, Korea University
[2]Dept. of Computer Science Engineering, Korea University

요 약

신제품 의류 수요 예측은 과거 판매 이력이 없어 예측 난이도가 높은 문제다. 본 연구는 제품 이미지, 의류 도메인 특화 BLIP 캡션, 시계열 판매 데이터를 결합한 Cross-Attention 기반 멀티모달 모델(CrossAttnRNN)을 제안한다. 각 모달리티는 CNN(이미지), BERT(캡션), GRU(시계열) 인코더로 임베딩되며, 이미지·텍스트는 Cross-Attention과 contrastive loss(임베딩 유사도 학습)로 의미 정렬 후 융합된다. 특히 캡션의 도메인 적합성이 예측 성능에 미치는 영향을 정량적으로 분석한 결과, 의류 도메인 특화 BLIP 캡션과 범용 BERT 조합이 MAE 0.94, WAPE 84.72%로 최적 성능을 기록하며 전통적 시계열 모델보다 우수했다. 실험 결과, 캡션의 품질과 도메인 적합성이 모델 성능의 핵심 요인임을 확인하였으며, 정적·동적 정보를 융합한 Cross-Attention 기반 모델이 패션 신제품 수요 예측에 효과적임을 확인하였다.

## 1. 서론

패션 산업은 변화 주기가 짧고 소비 트렌드 변화에 민감하여, 정확한 수요 예측이 재고 및 매출 관리에 직접적인 영향을 미친다. 그러나 신제품의 경우 판매 이력이 거의 없거나 부족하여, 전통적인 시계열 기반 방식만으로 수요 변화를 예측하기에는 한계가 있다.

기존 연구에서는 이미지·텍스트·시계열 데이터를 함께 활용하거나 어텐션 기반 융합 모델을 제안하는 시도가 있었으나, 캡션의 도메인 적합성이 성능에 미치는 영향을 세밀하게 분석한 사례는 드물다. 또한 정적 데이터(제품 속성)와 동적 데이터(판매 흐름)를 결합해 성능을 향상시키는 구조를 실증적으로 검증한 연구 역시 한정적이다. 이에 본 연구는 contrastive learning(유사도 학습)을 통한 이미지-텍스트 의미 정렬과 Cross-Attention 융합 구조를 결합한 멀티모달 수요 예측 모델(CrossAttnRNN)을 설계하였다. BLIP 모델로 생성한 의류 도메인 특화 캡션을 다양한 텍스트 임베더와 결합하고, 이를 유사 제품군의 시계열 판매 데이터와 함께 학습하여 성능을 개선하였다. 또한, 캡션의 도메인 적합성 여부가 예측 성능에 미치는 영향을 정량적으로 비교·분석하여, 모델 구조뿐만 아니라, 도메인에 최적화된 데이터 활용 전략이 예측 성능 향상에 중요한 변수임을 실험을 통해 확인하였다.

## 2. 관련 연구

최근 신제품 수요 예측 분야에서는 딥러닝 기반의 정적·동적 정보 통합 모델에 대한 연구가 점차 확대되고 있으며, 일부 연구에서는 이미지, 텍스트, 시계열 등 멀티모달 데이터를 통합하려는 시도도 나타나고 있다. [1] Ekambaram 등은 이미지, 텍스트, 시계열 데이터를 계층적 어텐션 메커니즘으로 통합하여 신제품 판매 예측 정확도를 향상시켰다. [2] Wu 등은 Dual Forecaster 모델을 통해 시계열 판매량과

제품 설명 텍스트 간 의미 정렬을 위해 contrastive loss 기반 멀티모달 구조를 제안하였으며, 이는 시계열 예측 정확도를 향상시켰다. [3] Lei 등은 어텐션 기반의 다중 예측 모델 융합 프레임워크를 통해 시계열 변동성과 제품 특성 간의 상관관계를 효과적으로 학습할 수 있음을 보였다.

패션 도메인에 특화된 연구로는 [4] Chae와 Kim이 머신러닝 기반 예측 모델을 통해 의류 아우터 제품의 계절성을 반영할 경우 기존 ARIMA 모델 대비 예측 오차(RMSE)를 23% 개선할 수 있음을 실증하였다. 또한 [5]와 같이 적대적 학습(adversarial learning) 기법을 도입하여 신제품과 기존 제품 간 특징 공유와 도메인 불변 표현 학습을 통해 데이터 부족 문제를 극복하였다.

최근에는 크로스 어텐션(cross-attention) 메커니즘을 활용한 멀티모달 융합 연구도 주목받고 있다. [6] 송인권 등은 Cross Attention 기반의 GTM Transformer를 활용하여 이미지와 시계열 데이터를 융합하는 의류 제품 수요 예측 모델을 제안하였으며, 패션 도메인에서도 멀티모달 융합 구조의 가능성을 제시하였다. [7] Li와 Liu는 연합 학습(federated learning) 기반의 멀티모달 수요 예측 모델을 제안하여 데이터 프라이버시를 보장하면서도 높은 예측 정확도를 달성하였다.

이와 같이 선행 연구들은 멀티모달 데이터의 통합, 어텐션 메커니즘, 도메인 특화 모델링 등 다양한 접근법의 효과를 입증하였으나, 패션 신제품 수요 예측에서 정적 데이터(이미지, 텍스트)와 동적 데이터(판매 시계열)를 통합적으로 활용하는 연구는 상대적으로 부족하다. 이러한 한계를 극복하기 위해, 의류 신제품 수요 예측에 특화된 크로스 어텐션 기반 멀티모달 융합 모델을 제안한다.

## 3. 제안 모델

본 논문에서 제안하는 의류 신제품 수요 예측 모델은 이미지, 이미지로부터 생성된 텍스트 캡션, 그리고 시계열 판매 데이터 세 가지 입력을 통합적으로 활용한다. 모델 아키텍처는 모달리티별 인코딩 단계, 이미지-텍스트 융합 단계, 그리고 수요 예측 디코더 단계로 구성된다.

### 3.1. 모달리티별 인코딩

이미지 인코딩: 의류 제품 이미지는 사전 학습된 CNN 기반 인코더(ImageEncoder)를 통해 고정 차원의 벡터로 임베딩된다. ResNet 기반 CNN 인코더를 fine-tuning하여 도메인 특화 학습을 수행하였으며, 최종적으로 이미지는 embedding_dim 차원의 벡터로 표현하였다.

텍스트(캡션) 인코딩: 각 제품 이미지로부터 BLIP 모델이 생성한 캡션은 사전 학습된 BERT 모델을 이용해 임베딩된다. 이때 문장 전체 의미를 대표하는 [CLS] 토큰 출력을 선형 투영(caption_proj)하여 이미지 임베딩 차원과 일치시킨다. 또한 BLIP 기반 도메인 특화 캡션과 범용 캡션 모두를 실험에 활용하여, 캡션의 도메인 적합성이 예측 성능에 미치는 영향을 비교·분석하였다.

시계열 인코딩: 과거 유사 제품의 주별 판매량은 GRU 기반 시계열 인코더(ts_embedder)를 통해 처리하였다. GRU는 순차적 입력의 시간적 패턴을 효과적으로 학습할 수 있어, 시계열 판매 데이터의 변동성을 모델링하는 데 적합하다.

이미지 임베딩과 캡션 임베딩 간의 의미적 유사성을 정렬하고 멀티모달 표현의 품질을 향상시키기 위해, 학습 과정에서 contrastive loss(유사도 학습)를 함께 사용한다. 이는 이미지와 텍스트 쌍의 임베딩 거리를 최소화하고 비매칭 쌍의 거리를 최대화함으로써 Cross-Attention 이전의 표현 정렬을 유도한다.

### 3.2. Cross-Attention 기반 이미지-텍스트 융합

추출된 이미지 임베딩과 캡션 임베딩은 Cross-Attention 메커니즘을 통해 의미적으로 융합된다. 이 과정의 핵심 목표는 시각적 정보(이미지)와 언어적 정보(캡션) 간의 상호 보완적인 관계를 학습하는 것이다. 캡션 임베딩을 **쿼리(Query)**로, 이미지 임베딩을 키(Key) 및 **값(Value)**로 설정한 Multihead Attention 구조를 적용한다. 이를 통해 텍스트의 각 의미 단위가 이미지의 어떤 시각적 특징과 관련되는지를 어텐션 가중치를 통해 학습할 수 있도록 유도한다. 다중 어텐션(Multihead Attention)은 여러 개의 어텐션 헤드를 병렬적으로 활용함으로써, 이미지와 텍스트 간의 의미적 연관성을 다양한 관점에서 포착할 수 있는 장점을 제공한다.

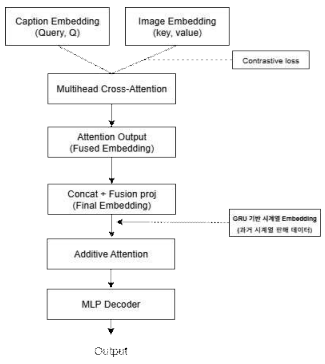
어텐션 연산의 결과로 생성된 attn_output은 원본 이미지 임베딩과 결합(concatenate)되어, 선형 계층(fusion_proj)을 통해 최종적인 이미지-캡션 통합 표현으로 변환된다. 구체적으로, 캡션 임베딩(caption_

emb)이 쿼리로 사용되고, 이미지 임베딩(img_encoding)이 키 및 값으로 사용되며, 이 연산을 수행하는 모듈은 코드에서 image_text_attn (Multihead Attention)으로 구현되어 있다.

본 연구에서는 캡션이 이미지로부터 생성된 2차적 정보라는 점에서, 이를 다시 이미지와 결합하는 방식의 정당성을 검토하였다. 본 모델은 이미지 임베딩과 캡션 임베딩이 동일한 원천을 갖더라도 BLIP이 생성한 캡션은 이미지의 전체적인 맥락과 핵심 속성을 언어적 기호로 요약 및 추상화한 정보이며, CNN 인코더가 추출한 시각적 특징은 저수준(low-level) 텍스처부터 고수준(high-level) 객체까지의 시각적 계층 정보를 담고 있다. 따라서 Cross-Attention은 언어적으로 요약된 핵심 속성(쿼리)을 바탕으로, 원본 이미지의 어떤 시각적 증거(키, 값)가 그 속성을 뒷받침하는지를 명시적으로 학습하는 과정이며, 이는 단순 정보 중복이 아닌 상호 보완적인 의미 강화(semantic enrichment) 효과를 가진다.

### 3.3. 수요 예측

융합된 이미지-텍스트 표현은 Additive Attention을 통해 GRU 기반 시계열 임베딩과 결합되어 최종 예측 모듈에 입력된다. 본 모델의 예측기는 선형 계층과 ReLU 활성화 함수, 드롭아웃(Dropout)으로 이루어진 MLP 구조로 설계되었으며, 이 통합된 특징 벡터를 바탕으로 미래의 수요량을 회귀 방식으로 산출한다. 학습 손실은 MAE(0.6)·MSE(0.4)의 가중 합에 SMAPE(0.01)를 소량 가중하여 절대/제곱/상대 오차의 균형을 도모하였다. 또한, 캡션 사용 시 이미지-텍스트 임베딩 정렬 강화를 위해 contrastive loss(λ=0.04)를 추가하였다. 해당 예측 모듈은 코드상에서 decoder 모듈로 구현되어 있다. 그림 1에 제안 모델의 구조를 제시하였다.




(그림 1) 예측 모델 구조

## 4. 실험 및 결과

### 4.1. 데이터셋 및 실험 환경

실험에는 실제 이탈리아 여성 의류 브랜드 쇼핑몰의 제품 이미지, BLIP 모델을 활용한 도메인 특화 이미지 캡션, 그리고 제품군의 2016년부터 2019년까지의 주별 과거 시계열 판매 데이터를 활용하여 제안 모델의 수요 예측 성능을 평가하였다. 이미지 캡션은 다음 두 가지 유형으로 비교 실험을 수행하였다:

(1) 범용 캡션 생성 모델이 생성한 일반 캡션
(2) 의류 도메인에 특화된 BLIP 모델이 생성한 캡션

성능 평가는 MAE, MSE, WAPE(%) 등의 수요 예측 지표를 기준으로 이루어졌다.

### 4.2. 실험 결과 및 분석

실험 결과, 제안하는 Cross-Attention 기반 멀티모달 수요 예측 모델은 단일 모달리티(이미지 또는 텍스트만 사용) 모델이나 단순 결합 방식보다 전반적으로 우수한 성능을 보였다.

<표 1> 의류 신제품 수요 예측 성능 비교

| 모델 유형 | 모델 상세 | MAE | MSE | WAPE(%) |
|---|---|---|---|---|
| 전통적 시계열 모델 | Naïve | 1.31 | 3.14 | 101.93 |
| | DES (Double Exponential Smoothing) | 1.74 | 6.51 | 156.73 |
| 베이스라인 모델 | Cross-Attention (이미지 Only) | 0.97 | 2.16 | 87.34 |
| 제안 모델 (이미지+캡션) | 의류 특화 BLIP Caption + 패션 BERT | 0.95 | 2.24 | 85.85 |
| | 범용 BLIP Caption + 범용 BERT | 0.96 | 2.26 | 86.57 |
| | 범용 BLIP Caption + 패션 BERT | 0.97 | 2.23 | 87.59 |
| | **의류 특화 BLIP Caption + 범용 BERT (최적 모델)** | **0.94** | **2.15** | **84.72** |

전통적인 시계열 모델(Naïve, DES)과 비교할 때, 이미지와 시계열 데이터를 함께 활용한 모델(Cross-Attn (이미지))은 MAE 0.97로 상대적으로 낮은 오차를 기록하였다. 여기에 이미지 캡션을 통합한 모델들은 대부분 더 낮은 오차를 보였으며, 본 실험 조건에서는 '의류 특화 BLIP 기반 캡션 + 범용 BERT' 조합이 MAE 0.94, WAPE 84.72%로 가장 낮은 예측 오차를 기록하였다. 또한, 성능 비교에 더해 캡션의 도메인 적합성이 수요 예측 성능에 미치는 영향을 정량적으로 분석하였다. 비교 결과, 범용 BLIP 캡션 생성 모델보다 의류 도메인 특화 BLIP 모델이 생성한 캡션을 활용했을 때 대부분의 경우 오차가 더 낮았다. 이러한 결과는 두 가지 요인으로

해석될 수 있다. 첫째, BLIP 모델이 생성한 캡션 자체가 의류 도메인 특성을 충분히 반영하고 있어, 범용 BERT 임베딩과 결합 시 불필요한 특화 효과보다는 보완적 효과가 두드러졌을 가능성이 있다. 둘째, 도메인 특화 BERT의 경우 BLIP 캡션과 학습된 도메인 정보가 중첩되면서 표현 다양성이 제한될 수 있으며, 이로 인해 일반 BERT 대비 성능 개선 효과가 크지 않았던 것으로 해석된다. 한편, 캡션 임베딩 모델의 차이를 비교한 실험에서는 BLIP 캡션에 대해 일반 BERT 임베딩(MAE 0.94)이 패션 도메인 특화 BERT(MAE 0.95)보다 소폭 우수한 결과를 보였다. 이는 BLIP 모델 자체가 도메인에 특화되어 있어, 임베딩 과정에서의 추가적인 특화 효과가 제한적이었을 것으로 해석된다. 이 결과는 캡션 자체의 품질과 도메인 적합성이 임베딩 방식보다 더 중요한 변수로 작용할 수 있음을 보여주었다.

<표 2> Contrastive loss 효과 비교

| 적용 여부 | MAE | MSE | WAPE (%) |
|---|---|---|---|
| 적용 전 | 0.942 | 2.150 | 84.72 |
| 적용 후 | 0.938 | 2.138 | 84.25 |
| 성능 변화 (향상) | -0.004 | -0.012 | -0.47%p |

표 2의 ablation 결과에서, contrastive loss를 적용한 결과 MAE가 0.942에서 0.938로(−0.004), MSE가 2.150에서 2.138로(−0.012), WAPE가 84.72%에서 84.25%로(−0.47%p) 감소하며 일관된 성능 향상이 관찰되었다. 이는 이미지-텍스트 임베딩 간 의미적 근접성을 강화함으로써 Cross-Attention 기반 멀티모달 융합 과정의 정렬 안정성이 높아진 결과로 해석된다. 개선 폭은 크지 않지만, 동일한 학습 조건에서 재현되어 모델의 예측 안정성 향상에 유의미하게 기여한 것으로 판단된다.

## 5. 결론

본 논문에서는 제품 이미지, 캡션, 판매 시계열 데이터를 결합한 Cross-Attention 기반 멀티모달 수요 예측 모델을 제시하였다. 각 모달리티별 인코더로 추출한 특징을 contrastive loss와 Cross-Attention으로 정렬·융합한 후, 시계열 디코더와 통합하여 미래 수요를 회귀 방식으로 예측하였다.

실험 결과, BLIP 기반 의류 특화 캡션과 범용 BERT 조합이 가장 낮은 MAE와 WAPE를 기록하였다. 이는 기존 시계열 모델이나 단일 모달리티 모델보다 일관되게 우수한 성능을 보였으며, 캡션의 도메인 적합성과 정적·동적 정보의 통합이 성능 향상에 핵심적인 요인으로 작용했음을 보여주었다.

향후 연구에서는 다양한 어텐션 변형 구조, contrastive 학습 전략 고도화, 모델 해석 가능성 제고, 다른 도메인(전자제품, 식품, 가구 등)에서의 적용 검증, 그리고 실시간 예측 시스템 개발을 통해 모델의 실용성과 확장성을 높이고자 한다.

## 참고문헌

[1] Ekambaram, V., Manglik, K., Mukherjee, S., Sajja, S. S. K., Dwivedi, S., & Raykar, V. Attention-based Multi Modal New Product Sales Time series Forecasting. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 3110 - 3118). ACM, 2020.

[2] Wu, W., Zhang, G., Tan, Z., Wang, Y., & Qi, H. Dual Forecaster: A Multimodal Time Series Model Integrating Descriptive and Predictive Texts. arXiv preprint arXiv:2505.01135, 2025.

[3] Lei, C., Zhang, H., Wang, Z., & Miao, Q. Multi Model Fusion Demand Forecasting Framework Based on Attention Mechanism. Processes, 12(11), 2612, 2024.

[4] Chae, J. M., & Kim, E. H. Sales Forecasting Model for Apparel Products Using Machine Learning Technique: A Case Study on Forecasting Outerwear Items. Fashion & Textile Research Journal, 23(4), 480-490, 2021.

[5] Chu, Z., Wang, C., Chen, C., Cheng, D., Liang, Y., & Qian, W. Learning invariant representations for new product sales forecasting via multi-granularity adversarial learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (CIKM '23) (pp. 3828 - 3832). ACM, 2023.

[6] 송인권, 김우주. Cross-Attention 기반의 GTM-Transformer를 활용한 의류 제품 수요 예측. 한국지능정보시스템학회 춘계학술대회 논문집, Vol. 2023(no. 5), 103. 한국지능정보시스템학회, 2023.

[7] Li, C., & Liu, W. Multimodal Transport Demand Forecasting via Federated Learning. IEEE Transactions on Intelligent Transportation Systems, 25(5), 4009 - 4020, 2024.