# 한국 패션 도메인 합성 데이터 구축

정헌영○, 신기영
디자이노블, 한양대학교
jhyoung@designovel.com, skyworldgo@hanyang.ac.kr

# Construct Synthetic Text-Image Dataset on Korean Fashion Domain

Hun-young Jung○, Ki-young Shin
Designovel, Hanyang University

**요 약**

본 연구는 한국 패션 도메인에서 텍스트-이미지 병렬 데이터를 구축하기 위한 합성 데이터셋(Synthetic dataset)을 제안한다. 이를 위해 한국 패션 이미지 데이터에 캡션 생성 모델을 적용하여 텍스트를 생성하고, 해외 패션 텍스트-이미지 데이터의 키워드 중심 설명 텍스트를 한국 패션 도메인 방식의 긴 문장형 설명으로 변환하는 방법을 사용하였다. CLIP 임베딩을 활용한 데이터 품질 평가 결과, 합성 데이터셋의 동일 상품 간의 유사도 및 비동일 상품간 유사도의 분포가 실제 데이터셋과 유사하게 나타났으며, 이는 구축된 합성 데이터셋이 실제 데이터와 유사한 특성을 가짐을 시사한다. 본 연구의 성과는 저자원 환경에서 패션 도메인의 텍스트-이미지 데이터를 확장하는 데 기여할 것 이다.

주제어: Synthetic Dataset, Image Captioning, Prompt Rewriting

## 1. 서론

패션 산업에서 전자상거래 플랫폼의 확산과 더불어 상품 기획, 추천, 광고 등에 디지털 데이터의 중요성이 점점 더 커지고 있다. 특히 이미지와 텍스트를 결합한 데이터셋은 이미지 생성 모델을 통해 패션 상품을 기획하거나, 캡션 생성 모델을 사용하여 아이템 설명 또는 광고 문구를 자동으로 생성하는 데 활용될 수 있다. 그러나 한국 패션 도메인에서는 패션 상품 이미지 수집이 비교적 용이한 반면, 텍스트 설명이 이미지 형태로 제공되는 경우가 많아 텍스트-이미지 병렬 데이터의 구축이 어려운 실정이다.

본 연구는 한국 패션 도메인의 합성 데이터셋(Synthetic dataset)을 구축하는 것을 목표로 한다. 이를 위해, 한국 패션 상품 이미지에 캡션 생성 모델을 적용하여 설명문을 생성하거나, 해외 패션 도메인의 이미지-키워드 병렬 데이터를 변환하여 한국 패션 도메인의 특성에 맞는 텍스트-이미지 데이터로 전환하는 방식을 사용한다. 이러한 합성 데이터셋은 향후 한국 패션 산업에서 다양한 응용 프로그램에 활용될 수 있는 중요한 자원이 될 것이다.

## 2. 관련 연구

Generative Adversarial Networks (GAN)[1] 및 Diffusion Model[2]의 발달로 인해 텍스트-이미지 합성 데이터셋 구축이 다양한 분야에서 활발히 활용되고 있다. 이러한 합성 데이터셋은 이미지 분류, 이미지 인식 등의 문제를 해결하는 데 기여하고 있으며, 특히 few-shot 학습과 같은 저자원 환경에서 모델 사전학습에 유용하다는 연구 결과가 있다[3][4]. 이러한 모델들은 텍스트를 기반으로 고해상도의 이미지를 생성함으로써 시각적 표현 학습에 중요한 역할을 하고 있다.

이미지 생성을 이외에, 이미지에서 텍스트를 생성하는 이미지 캡션 생성 기술을 활용한 합성 데이터셋 구축 방법도 존재한다. 이러한 데이터셋은 이미지 인식 및 분류 모델을 훈련하는 데 사용될 수 있으며[5], 특히 저자원(low resource) 환경에서 다국어 (multilingual) 멀티모달 (multi-modal) 모델을 학습시키는 데 있어 이미지 캡션 생성과 번역을 통한 합성 데이터 구축이 유효한 방법으로 연구되고 있다[6].

## 3. 제안 방법

본 연구에서는 소규모로 구축된 한국 패션 텍스트-이미지 데이터셋에 기반하여 비교적 대량으로 수집된 한국 패션 이미지 데이터셋과 해외 패션 텍스트-이미지 데이터셋을 변환하는 것으로 더 큰 규모의 패션 도메인 합성 텍스트-이미지 데이터셋을 구축한다.

한국 패션 이미지 데이터셋에 대해서는 텍스트-이미지 데이터셋을 사용한 이미지 캡션 생성 모델을 적용하는 방법을 사용하고, 해외 패션 텍스트-이미지 데이터셋에 대해서는 텍스트 데이터를 한국 패션 데이터 특징과 비슷하게 변환하는 방법을 적용한다. (그림 1.)

### 3.1. 이미지 캡션 생성 (Image captioning)

보다 큰 규모로 구축된 한국 패션 이미지 데이터셋에 대응되는 텍스트를 생성하기 위해, 본 연구에서는 이미지 캡션 생성 모델을 적용하였다. 한국 패션 도메인의 상품 설명은 일반적으로 색상, 재질, 착용 방식 등 다양한 속성을 구체적이고 자세히 설명하는 긴 문장 형태로

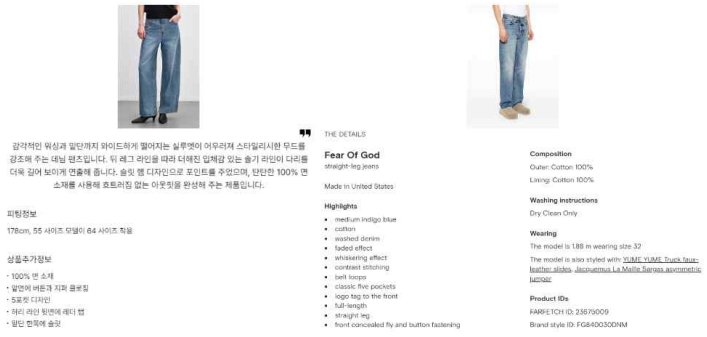


그림 2 패션 상품에 대한 설명 텍스트 한국(좌)/해외(우)

제공되므로, 생성되는 캡션 역시 이러한 특징을 반영해야 한다. 이를 위해 본 연구에서는 한국 패션 텍스트-이미지 데이터를 사용하여 이미지 캡션 생성 모델을 fine-tuning하여, 패션 텍스트의 특성을 더 잘 표현할 수 있도록 조정하였다.

### 3.2. 설명 텍스트 변환 (keywords to sentence)

해외 패션 데이터의 경우 대규모로 수집된 텍스트-이미지 병렬 데이터가 존재하나 텍스트 데이터의 구성이 주로 키워드 목록과 같은 형식으로 간결하게 구성되어 있어, 한국 패션 도메인의 긴 문장형 설명과는 차이가 있다. 예를 들어 해외 데이터의 'blue jeas, cotton, slim fit'과 같은 키워드 목록 방식의 설명문은 "These jeans are designed with a slim fit and made of cotton for a comfortable fit." 같은 방식의 문장형 텍스트로 변환되어야 한다. (그림 2)

따라서 본 연구에서는 이러한 키워드 목록을 문장형 설명으로 변환하여 합성 데이터를 구축하는데 사용하였다. 이를 위해 대규모 언어 모델(LLM)을 활용하여 문장을 생성하였다. LLM은 자체적으로 키워드를 사용하는 문장 생성이 가능하여 스타일 가이드를 위한 예시를 제공하는 in-context learning을 하는 것으로 별도의 추가 학습 없이 사용할 수 있다.

## 4. 데이터 구축

한국 패션 텍스트-이미지 데이터로 약 7600개 상품에 대한 약 87K개의 상품 이미지가 수집되었고 각 상품에 대한 설명 텍스트도 수집되어 구글 번역기[1]를 통해 영어로 번역되었다. 한국 패션 이미지 데이터로 약 23만개의 상품 이미지가 수집되었다. 해외 패션 텍스트-이미지 데이터로 약 7.2만개 상품에 대한 이미지 24.4만장이 수집되었고 각 상품에 대한 설명 텍스트도 수집되었다.

이미지 캡션 생성을 위해 사용한 멀티모달 LLM은 Blip-2[7] 로 한국 패션 텍스트-이미지 데이터로 미세조정되었다. 미세조정 단계의 학습에서 계산량을 줄이기 위해 Low-Rank Adaptation (Lora)[8] 기법을 사용하였다.

키워드 텍스트에서 문장생성을 위한 LLM은 llama [9] 모델로, 최신 버전인 Llama3.1-8B-Instruct[2] 모델을 사용하였다.

표 1. 이미지-텍스트 임베딩 유사도 통계

| 데이터셋 | 동일 상품 | | 비동일 상품 | |
|---|---|---|---|---|
| | 평균 | 분산 | 평균 | 분산 |
| 원천 | 0.204 | 0.0002 | 0.182 | 0.0002 |
| 캡션 생성 | 0.205 | 0.0004 | 0.182 | 0.0003 |
| 문장 생성 | 0.244 | 0.0007 | 0.172 | 0.0007 |

## 5. 데이터 평가

데이터 품질을 평가하기 위해 CLIP 임베딩 (Embedding) [10] 모델에 기반한 텍스트-이미지 연관성을 계산하여 이미지에 걸맞는 텍스트가 생성되었는지를 평가하였다. 각 이미지에 적절한 텍스트가 생성되었다면 동일 패션 상품에 대한 각 임베딩 사이의 cosine 값이 서로 다른 상품에서 나온 임베딩 사이의 값보다 클 것이라 상정하였다.

구축된 데이터의 텍스트 길이가 긴 편임을 감안하여 기본 CLIP 모델이 아닌 최근 공개된 Long-clip [11]모델을 사용하였다. 적절한 계산량으로 평가하기 위해 전체 임베딩 사이의 값을 계산하지 않고 상품 1000개를 임의로 선정하여 상품당 이미지 2개와 설명에 대한 임베딩을 가지고 유사도를 계산하였다.

계산한 대상 데이터셋은 다음 세가지이다. 기존 한국 패션 텍스트-이미지 데이터셋 (이하 원천 데이터셋), 한국 패션 이미지 데이터에 이미지 캡션 생성을 적용하여 만든 합성 데이터셋 (이하 캡션 생성 데이터셋), 해외 패션 텍스트-이미지 데이터에 텍스트 변환을 적용한 합성 데이터셋 (이하 문장 생성 데이터셋)

측정된 유사도의 평균과 분산은 세 데이터셋에서 모두 동일 상품의 임베딩간 유사도가 비동일 상품의 임베딩간 유사도보다 크게 계산되었고 유사도의 분산은 모든 경우에서 0.001 이하로 나타났다. (표1)

## 5. 결론

본 연구에서는 한국 패션 도메인에 맞춘 대규모 합성

1) https://translate.google.com/

2) https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

텍스트-이미지 데이터셋을 구축하고 이를 평가하였다. 소규모의 기존 한국 패션 텍스트-이미지 데이터셋을 바탕으로, 이미지 캡션 생성 및 텍스트 변환을 통해 더 큰 규모의 합성 데이터셋을 생성하는 방법론을 제안하였다. 한국 패션 데이터의 특성상 긴 문장형 설명이 필요하다는 점을 반영하여, 이미지 캡션 생성 모델을 미세조정하였고, 해외 패션 데이터의 간단한 텍스트 설명을 한국 패션 도메인의 스타일에 맞게 변환하였다.

데이터셋의 품질을 평가하기 위해 CLIP 임베딩 모델을 기반으로 이미지와 텍스트 간의 연관성을 분석하였으며, 실제 한국 패션 텍스트-이미지 데이터셋과 유사하게 동일 상품 간의 유사도가 비동일 상품 간의 유사도보다 높게 나타나는 것을 확인하였다. 이는 본 연구에서 구축한 합성 데이터셋이 실제 데이터와 유사한 특징을 보이며, 패션 도메인에서 활용 가능한 고품질의 데이터셋임을 시사한다.

본 연구의 성과는 한국 패션 도메인에서 텍스트-이미지 데이터를 확장하는 데 중요한 기여를 할 것으로 기대된다. 특히, 저자원 환경에서 멀티모달 학습을 지원할 수 있는 데이터셋을 제공함으로써, 패션 관련 애플리케이션에서 이미지 캡션 생성, 추천 시스템, 이미지 분류 등 다양한 연구와 산업적 응용에 활용될 수 있을 것이다.

향후 연구에서는 생성된 합성 데이터셋을 기반으로 보다 다양한 패션 아이템에 대해 세밀한 분석이 가능하도록 데이터의 다양성을 확장하고, 최신의 이미지-텍스트 모델을 추가적으로 활용하여 데이터 품질을 더욱 향상시키는 연구를 진행할 예정이다.

## 참고문헌

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. Generative adversarial nets. Advances in neural information processing systems, 27. 2014.

[2] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684-10695. 2022.

[3] Tian, Y., Fan, L., Isola, P., Chang, H., & Krishnan, D. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. Advances in Neural Information Processing Systems, 36. 2024.

[4] He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., ... & Qi, X. Is synthetic data from generative models ready for image recognition?. arXiv preprint arXiv:2210.07574. 2022.

[5] Nguyen, T., Gadre, S. Y., Ilharco, G., Oh, S., & Schmidt, L. Improving multimodal datasets with image captioning. Advances in Neural Information Processing Systems, 36. 2024.

[6] Santos, G. O. D., Moreia, D. A., Ferreira, A. I., Silva, J., Pereira, L., Bueno, P., ... & Avila, S. CAPIVARA: Cost-Efficient Approach for Improving Multilingual CLIP Performance on Low-Resource Languages. arXiv preprint arXiv:2310.13683. 2023.

[7] Li, J., Li, D., Savarese, S., & Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning. pp. 19730-19742. PMLR. 2023, July.

[8] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685. 2021.

[9] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... & Lample, G. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. 2023.

[10] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. Learning transferable visual models from natural language supervision. In International conference on machine learning. pp. 8748-8763. PMLR. 2021, July.

[11] Zhang, B., Zhang, P., Dong, X., Zang, Y., & Wang, J. Long-clip: Unlocking the long-text capability of clip. arXiv preprint arXiv:2403.15378. 2024. I. Mani and T. Maybury, Advances in Automatic Text, The MIT Press, 1999.