

Multi-modal LLM 을 활용한 다음 시즌 의류 예측 및 추천*

¹ 김설아*^{○1} 장승우* ¹ 남희정* ^{2,3} 이영완*

¹ 모두의 연구소 ²ETRI ³KAIST

jena121898@gmail.com, seungu1108@gmail.com, to-chris@hanmail.net, yw.lee@etri.re.kr

Predicting and Recommending Clothing for Next Season Based on Multi-modal LLM

¹Seol-A Kim*^{○1} ¹Seuung-Woo Jang* ¹Hee-Jung Nam* ^{2,3}Young-Wan Lee*

¹Modulabs ²ETRI ³KAIST

요 약

최근 전 세계적으로 대형언어모델 (Large Language model, LLM)이 비약적 발전을 이룸과 동시에 LLM 을 활용한 연구 및 비즈니스모델들이 큰 주목을 받고 있다. 본 논문은 LLM 과 시각 모델 (Vision Model)을 결합해 이미지와 텍스트를 동시에 처리할 수 있고 챗봇 기능을 구현할 수 있는 Foundation model 중에 하나인 FROMAGe 를 활용하여 다음 시즌 유행할 의류를 예측하고 추천해 주는 챗봇 시스템을 제안한다. 이를 위해 패션쇼 이미지 데이터를 수집하고 LLaMA-Adapter v2 multimodal LLM 모델을 활용하여 패션 데이터의 캡션을 생성하였다. 이렇게 구축된 패션 데이터 (이미지-캡션)을 이용하여 FROMAGe 모델을 Fine-tuning 하였다. 본 연구를 통해 소비자들이 시즌 정보 및 특정 브랜드의 특징을 가지고 있는 시장의 기출시된 제품들에 보다 빠르게 접근할 수 있도록 하며, 챗봇 형식의 AI 의류 추천 방식으로 소비자의 편의성 향상을 도모했다.

1. 서 론

Open AI 의 Chat GPT, Meta 의 LLaMA 등 Transformer 의 등장 이후 LLM 은 활발하게 발전하고 있으며[1] Vision Model 과 함께 연구되어 Multi-modal 모델의 발전에도 기여하고 있다.

본 연구에서는 2023 년 2~3 월경 발표된 2023 년도 F/W Ready-to-Wear(RTW) 컬렉션 이미지 데이터를 활용해 LLaMA Adapter v2[2]로 캡션을 생성하고, FROMAGe[3]를 Fine-tuning 한다. 이를 통해 실제 2023 년도 가을, 겨울 시즌의 판매 아이템 중 앞서 발표된 RTW 컬렉션의 제품과 유사한 제품을 추천한다.

LLaMA-Adapter v2[2]와 FROMAGe 는 Multi-modal LLM (MLM)으로 LLaMA-Adapter v2 는 기존의 이미지 캡셔닝(captioning) 모델들에 비해 더 상세한 캡션을 생성해 주었으며, 특히, FROMAGe 는 기존의 MLM 과 달리, 이미지와 텍스트 임/출력과 자유 형식의 텍스트 생성이 가능한 모델이다.

한편, RTW 는 높은 가격대로 맞춤 생산된 제품군(High Fashion)과 달리 다양한 사이즈로 대량 생산되어 판매되는 제품군[4]으로 RTW 컬렉션은 패션위크(Fashion-week)에 유명 패션 하우스에 의해 선보여진다. 이후 트렌드를 대중에게 소개하는 다양한 매체 혹은 인물을 통해 소비자에게 공유되고 유행에 영향을 미친다[5].

모델 학습 후 시즌(2023 F/W) 및 브랜드 정보 학습 여부와 시즌 정보가 실제 유행을 예측할 수 있는지에 대한 평가를 위해 모델의 출력 결과와 RTW 컬렉션, 2023 년 10 월 월간 인기도 랭킹 Top 100 제품을 각각 비교하였다.

본 연구는 RTW 컬렉션과 소비자의 중간 단계를 생략하고, 소비자가 실제로 구매하는 플랫폼에서 유사한 상품을 추천해 본인의 스타일과 트렌드가 접목된 제품을 선택함으로써, 소비자가 유행을 이끄는 주도적인 역할로의 전환을 기대한다.

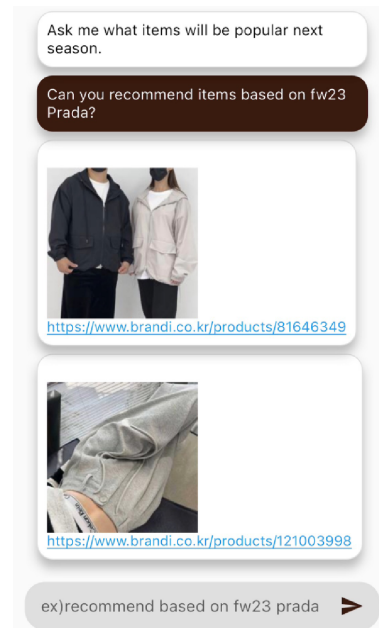


그림 1. 서비스 구현 데모

2. 제안 방법

FROMAGe[3]는 Vision Model 과 LLM 의 가중치는 동결시킨 채 Linear mapping 되는 레이어(layer)들의 가중치와 [RET] 토큰의 임베딩 값만 업데이트가 진행되며, Image Captioning 과 Image-

* 이 논문은 2023 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. RS-2022-00187238, 효율적 사전학습이 가능한 한국어 대형 언어모델 사전학습 기술 개발)

* 본 연구는 과학기술 발전을 돕는 비영리단체 브라이언임팩트의 지원을 받았습니다.

*: equal contribution

+: corresponding author

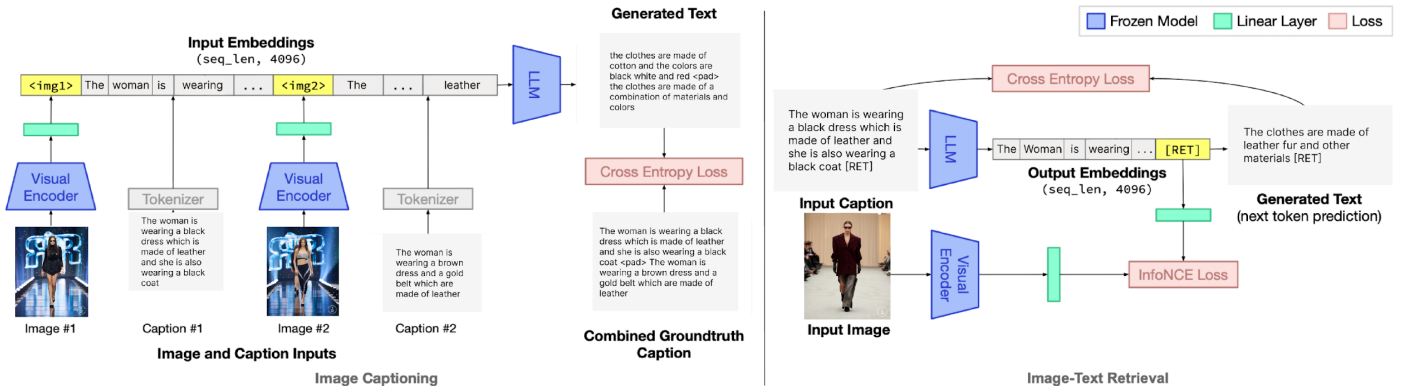


그림 2. FROMAGE fine-tuning 아키텍처 [3]

Text Retrieval 로 구성되어 있다. 이때 도출되는 InfoNCE Loss 와 CE Loss 값을 최소화하는 방식으로 Training 을 진행한다.

2.1 Image Captioning

Vision Model 로 추출한 이미지 데이터의 임베딩 값을 Linear mapping 한 후 텍스트 데이터 앞에 [RET] 토큰을 위치시켜 전체를 BPE Tokenizer 처리한다. 이때 Tokenizer 한 값에 대해 캡션(caption)을 추출한 뒤 Ground Truth 와 비교하여 CE Loss 를 최소화 한다.

2.2 Image-Text Retrieval

Vision Model 로 추출한 이미지 데이터의 임베딩 값을 Linear mapping 한 값과 텍스트 데이터를 Linear mapping 한 값의 Cosine similarity 를 계산하고, 이를 기반으로 InfoNCE Loss 값을 추출한다. 이후 텍스트 데이터 뒤에 [RET] 토큰으로 위치시켜 전체 Tokenizer 한 값에 대한 캡션을 추출한 뒤 Ground Truth 와 비교하여 CE Loss 를 추출한다.

3. 실험 방법

3.1 데이터 구축

Multi-modal 학습에 필요한 이미지, 캡션의 Paired-Dataset 을 구축하기 위해 Tagwalk¹ 에서 2023 년도 F/W Women's Wear 이미지에 대한 캡션을 생성하였다. 이를 위해 이미지 캡셔닝 모델 BLIP_{large}(BLIP)[6]과 MLM 인 LLaMA-Adapter v2 를 비교하였다. LLaMA-Adapter v2 에 사용된 프롬프트(prompt)는 'Summarize the object facts on the materials and colors of the clothes into one sentence'이다. 동일한 사진(그림 3)에 대해 LLaMA-Adapter v2 가 BLIP 에 비해 월등히 좋은 캡션을 도출하는 것을 확인하여 LLaMA-Adapter v2 를 캡션 생성 모델로 채택하였다(표 1). 생성된 Paired-Dataset 중 10,000 개를 Train 데이터로, 2,975 개를 Validation 데이터로 사용하였다.

표 1. BLIP 과 LLaMA-Adapter v2 의 캡션 생성 비교


| BLIP | woman in a hat and coat posing for a picture |
|--|---|
|  | The image shows a woman wearing a brown coat, which is made of a material that appears to be leather. The coat is long, reaching down to her ankles. The woman ¹ is also wearing a black shirts, which complements the brown coat. The overall appearance of the outfit is casual and comfortable, suitable for everyday wear or a cool. |
| LLaMA-Adapter v2 | |

그림 3².

3.2 학습 방법

사전 학습된 Fromage 에 가지치기 기법(pruning)이 적용된 모델 가중치를 활용하여 학습하였다. A100 GPU 환경에서 약 5 시간 동안 batch size 6 으로 진행되었으며, 대부분의 하이퍼 파라미터는 [3]에 따라 10 epoch, 0.0003 learning rate 등을 유지하였고, 생성된 캡션의 길이에 맞춰 max length 를 177 로, 메모리를 고려하여 precision type 을 bf16 으로 각각 설정하였다. 모델의 추천 결과 데이터 셋은 패션 플랫폼 앱인 브랜드³에서 2023 년 9 월 판매 중인 제품으로 설정하였다.

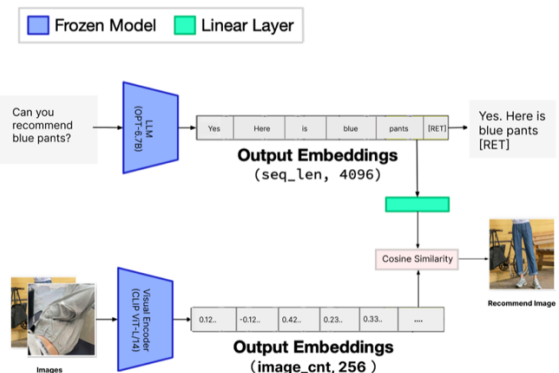


그림 4. 모델 출력 과정

4. 실험 결과

4.1 정량적 결과

시즌 및 브랜드 정보 학습 여부와 실제 유행과의 유사도를 평가하기 위해 프롬프트 작성하고 각 프롬프트에 따라 5 개 이미지 출력해 비교하였다. 각각 5 점 만점으로 평가가 진행되었으며, 핏, 색상, 분위기 등을 유사도의 평가 기준으로 정의하였고, 모델의 출력과 비교 대상 중 동일한 제품이 있다고 판단되면 1 점, 없으면 0 점, 유사한 제품이 있다고 판단되면 0.5 점으로 하였다.

4.1.1 시즌 및 브랜드 정보 학습

브랜드 정보와 시즌 정보를 학습하였는지를 평가하기 위해 브랜드 가치 분석기관 LUXE Digital 에서 2023 년 9 월 21 일 발표한 온라인 인기도 상위 15 개 하이엔드(High-end) 브랜드 중 패션 카테고리의 상위 10 개 브랜드를 선정하였다. 사용된 프롬프트는 'Can you recommend clothes based on {브랜드} fw23?'로, 이에 대한 모델의 출력 결과를 해당 브랜드의 2023 년도 F/W RTW 컬렉션과 비교하였다. 그 결과 각 브랜드에

¹ <https://www.tag-walk.com/en/>

² <https://www.brandi.co.kr/>

³ <https://www.agnesb.eu/en/fashion-shows/women-s-summer-2022/>

대한 데이터의 수가 충분하지 않아 브랜드 정보에 대한 학습이 부족함을 알 수 있었다.

4.1.2 실제 유행과 비교

시즌 정보를 학습한 모델이 추천하는 제품이 실제 유행과 유사하게 적용되는지를 알기 위해 옷의 종류, 분위기, 기장 등에 기반해 분류를 나누었다. 사용된 프롬프트는 ‘Can you recommend {분류} based on fw23?’로, 이에 대한 모델의 출력 결과를 10 월 12 일 기준 브랜드 월간 인기도 랭킹 top 100 제품들과 비교하였다. 그 결과 절반 이상의 점수를 나타내는 분류를 볼 수 있어, 브랜드 정보에 대한 학습보다 시즌 정보에 대한 학습의 성능이 더 좋았음을 유추할 수 있었다.

표 2. 시즌 및 브랜드 정보 학습 여부와 실제 유행과의 유사도 평가

| 브랜드 (학습 데이터 수) | 점수 | 분류 | 점수 |
|-------------------|-----|-------------|-----|
| DIOR(96) | 1 | JACKET | 1 |
| GUCCI(53) | 1.5 | COAT | 1 |
| CHANEL(66) | 2 | CASUAL TOP | 1.5 |
| LOUIS VUITTON(44) | 1 | FORMAL TOP | 1 |
| HERMÈS(61) | 1 | SHORTS | 2.5 |
| PRADA(54) | 2 | PANTS | 2 |
| VERSACE(78) | 1.5 | SHORT DRESS | 2.5 |
| BURBERRY(55) | 0 | LONG DRESS | 2 |
| BALENCIAGA(54) | 0.5 | SHORT SKIRT | 4 |
| SAINT LAURENT(50) | 0.5 | LONG SKIRT | 0.5 |

4.2 정성적 결과

4.2.1 시즌 및 브랜드 정보 학습

학습하지 않은 시즌 정보인 2021 년도 F/W 에 대한 프롬프트 (‘Do you know fashion show on fw21?’)에 대해서는 ‘ I don’t know. I don’t know.’라고 답변하는 반면, 학습된 2023 년도 F/W 에 대한 프롬프트 (Do you know fashion show on fw23?)에 대해서는 ‘Yes. We have a lot of new products, and we will show them in the show.’라고 답변하여 훈련한 정보들이 학습되었음을 확인하였다. 또한, 모델의 출력 결과를 RTW 컬렉션 사진과 비교한 결과 완전히 같은 의상은 거의 없었으나, 전체적인 브랜드의 분위기 혹은 자주 사용되는 색상이 반영되었음을 확인하였다(그림 5). 특히 RTW 컬렉션의 두 제품의 특징을 합치면 모델의 출력 결과와 유사할 것으로 추정되는 결과를 볼 수 있었다.



그림 5 . ‘Can you recommend clothes based on hermes/saint-laurent fw23?’에 대한 모델 출력 비교

4.2.2 실제 유행과 비교

모델의 출력 결과를 브랜드 월간 인기도 랭킹 상위 100 개의 제품과 비교한 결과 같은 종류의 의상(cargo pants, training pants), 동일하다고 판단되는 의상, 옷의 분류는 다르지만 동일한 디자인으로 판단되는 의상 등을 추천하는 것을 볼 수 있었다 (그림 6).

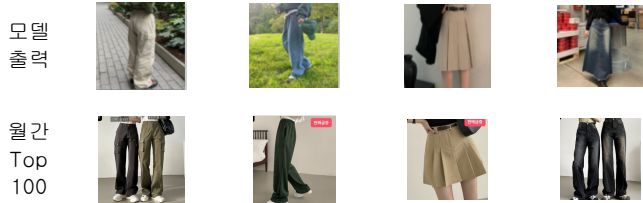


그림 6 . ‘Can you recommend shorts/pants/short skirt/long skirt based on fw23?’에 대한 모델 출력 비교

5. 결 론

본 논문의 목적은 사용자가 시즌 트렌드와 유행 제품을 각각 찾아야 했던 기존 방식의 번거로움을 해소하여 소비자가 유행을 주도한다는 경험과 만족감을 경험하도록 하는 데 있다.

이를 위해 MLM 인 LLaMA-Adapter v2 와 FROMAGe 를 활용하여 Fine-tuning 을 진행하였다. 챗봇형태로 구성된 결과물에서는 실제 학습된 시즌 트렌드 및 브랜드와 유사한 제품을 일부 추천하였다. 그리고 브랜드 정보에 대한 학습에 비하여 시즌 정보에 대한 학습의 성능이 더 좋음을 확인하였다. 반면 브랜드 혹은 특정 색상(purple, blue 등)에 대한 낮은 정확도를 보이므로 추가적인 학습이 필요할 것으로 보인다. 또한 캡션 생성 시 전문적인 패션 용어가 학습된 모델을 활용해 볼 수 있을 것이다. 학습 데이터는 배경을 삭제한 의류 사진 활용, SNS 데이터 등으로의 범위 확대, 매출 데이터를 기반으로 실제 유행에 영향력이 있는 브랜드의 데이터 활용 등의 개선 사항이 고려된다. 이외에도 한국어 지원, 경량화 적용, LLM 의 일반적 문제 개선 등 다양한 연구 과제들이 남아 있으며, 향후 실험을 통해 모델의 성능을 더 향상시킬 수 있을 것으로 기대된다.

참 고 문 헌

- [1] 주하영, 오현택, and 양진홍. "오픈 소스 기반의 거대 언어 모델 연구 동향: 서베이." 한국정보전자통신기술학회 논문지 16.4 : 193–202, 2023.
- [2] Gao, Peng, et al. "Llama-adapter v2: Parameter-efficient visual instruction model." arXiv preprint arXiv:2304.15010, 2023.
- [3] Koh, Jing Yu, Ruslan Salakhutdinov, and Daniel Fried. "Grounding Language Models to Images for Multimodal Inputs and Outputs." ICML, 2023.
- [4] 한지연. "돌체앤가바나 컬렉션에 나타난 맥시멀리즘 디자인 연구." 기초조형학연구 19.4 : 509–522, 2018.
- [5] 임성민, and 박민여. "패션에서 유행을 따르게 되는 내적 에너지에 대한 연구-Masochism 을 중심으로." 한국의류학회지 32.3 : 362–372, 2008.
- [6] Li, Junnan, et al. "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation." ICML, 2022.