https://doi.org/10.5392/JKCA.2022.22.01.138

멀티모달 패션 추천 대화 시스템을 위한 개선된 트랜스포머 모델

Improved Transformer Model for Multimodal Fashion Recommendation Conversation System

박영준, 조병철, 이경욱, 김경선
엔에이치엔 다이퀘스트

Yeong Joon Park(yjpark@diquest.com), Byeong Cheol Jo(byeongcheol7879@diquest.com), Kyoung Uk Lee(arp1710@diquest.com), Kyung Sun Kim(kksun@diqueset.com)

요약

최근 챗봇이 다양한 분야에 적용되어 좋은 성과를 보이면서 쇼핑몰 상품 추천 서비스에도 챗봇을 활용하려는 시도가 많은 이커머스 플랫폼에서 진행되고 있다. 본 논문에서는 사용자와 시스템간의 대화와 패션 이미지 정보에 기반해 사용자가 원하는 패션을 추천하는 챗봇 대화시스템을 위해, 최근 자연어처리, 음성인식, 이미지 인식 등의 다양한 AI 분야에서 좋은 성능을 내고 있는 트랜스포머 모델에 대화 (텍스트) 와 패션 (이미지) 정보를 같이 사용하여 추천의 정확도를 높일 수 있도록 개선한 멀티모달 기반 개선된 트랜스포머 모델을 제안하며, 데이터 전처리(Data preprocessing) 및 학습 데이터 표현(Data Representation)에 대한 분석을 진행하여 데이터 개선을 통한 정확도 향상 방법도 제안한다. 제안 시스템은 추천 정확도는 0.6563 WKT(Weighted Kendall's tau)으로 기존 시스템의 0.3372 WKT를 0.3191 WKT 이상 크게 향상시켰다.

■ 중심어 : | 대화시스템 | 트랜스포머 | 멀티모달 | 자연어처리 | 인공지능 |

Abstract

Recently, chatbots have been applied in various fields and have shown good results, and many attempts to use chatbots in shopping mall product recommendation services are being conducted on e-commerce platforms. In this paper, for a conversation system that recommends a fashion that a user wants based on conversation between the user and the system and fashion image information, a transformer model that is currently performing well in various AI fields such as natural language processing, voice recognition, and image recognition. We propose a multimodal-based improved transformer model that is improved to increase the accuracy of recommendation by using dialogue (text) and fashion (image) information together for data preprocessing and data representation. We also propose a method to improve accuracy through data improvement by analyzing the data. The proposed system has a recommendation accuracy score of 0.6563 WKT (Weighted Kendall's tau), which significantly improved the existing system's 0.3372 WKT by 0.3191 WKT or more.

■ keyword : | Dialogue System | Transformer | MultiModal | NLP | AI |

I. 서론

자연어처리[1]는 인간의 언어를 시스템이 잘 알아듣고 이해할 수 있도록 분석하여 처리하는 분야이다.

* 이 논문은 '산업통상자원부 산업기술혁신사업'의 지원을 받아 수행된 연구 결과입니다. [과제번호 : 20008625, 과제명 : 패션속성기반 혼합현실 시각화 서비스 제공을 위한 패션 온라인채널용 딥태깅기술 및 2D기반 가상피팅 기술 개발]
접수일자 : 2021년 09월 13일
수정일자 : 2021년 10월 18일
심사완료일 : 2021년 11월 01일
교신저자 : 김경선, e-mail : kksun@diquest.com

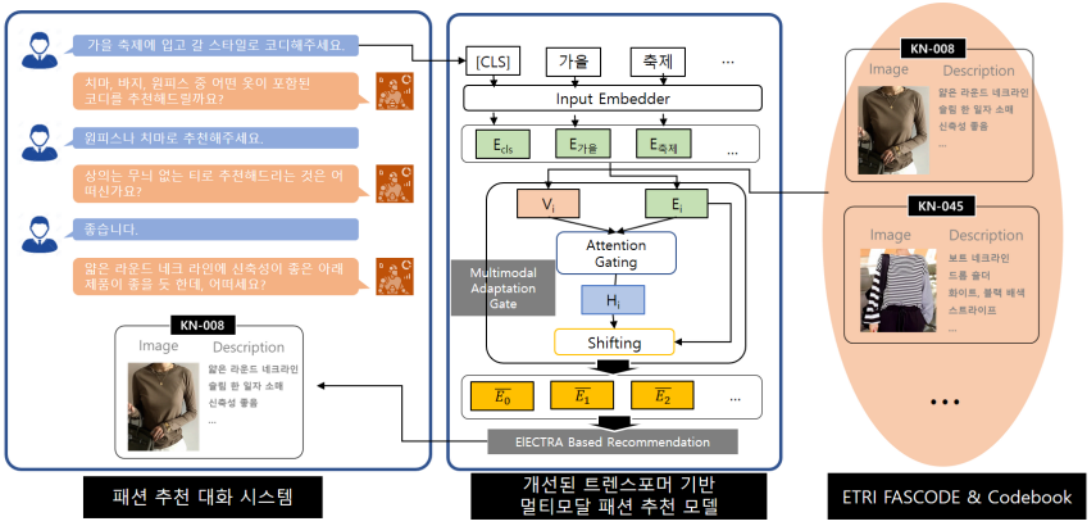

그림 1. 멀티 모달 패션 추천 대화 시스템 개요

대화 시스템[2][3]은 자연어처리의 주요 연구 중 하나로, 시스템이 사용자의 발화를 이해하고 적절한 응답 생성을 연구하는 분야로 최근 쇼핑몰, 헬스케어, AI 기반 콜센터 등 다양한 도메인에서 챗봇 서비스 형태로 도입되고 있다[4].

본 논문에서는 쇼핑몰 영역에서 사용자가 원하는 의상을 사용자와 대화를 통해 추천하는 패션 추천 대화 시스템에 사용되어 추천 정확성과 신뢰도를 높일 수 있도록 개선한 트랜스포머 기반 멀티모달 패션 추천 대화 모델을 제안한다.

[그림 1]은 본 논문의 목표 서비스인 멀티모달 패션 추천 대화 시스템의 서비스 흐름으로 사용자와의 대화를 통해 여러 개의 의상 중 사용자의 요구사항에 가장 적합한 의상을 사용자와의 대화(텍스트)와 패션(이미지) 분석을 기반으로 딥러닝 추천 기술을 이용하여 추천한다.

기존 대화 시스템에서는 전통적인 TF-IDF(Term frequency-Inverse Document Frequency)[5] 등의 유사도 기반 추천 방식을 사용하여 사용자의 이전 대화를 완벽하게 이해하지 못하고 단편적인 정보만을 이용하여 응답을 생성해 추천 정확성이 낮은 단점이 존재한다. 최근 CNN(Convolutional Neural Network)[6], LSTM(Long-Short Term Memory)[7], 트랜스포머(TRANSFORMER)[8] 등의 딥러닝 모델들은 어텐션(Attention) 등을 반영하여 이전 대화를 이해하여 적절한 응답을 생성하여 추천 정확성을 높이고 있으며, 특히 셀프 어텐션 등이 반영된 트랜스포머 기반 모델 들이 좋은 성능을 보이고 있다.

트랜스포머 모델은 어텐션(attention)을 이용하여 거리가 먼 단어 간의 관계를 잘 잡아내지 못하는 CNN의 단점을 극복하였으며, 문장이 길어질 경우에 RNN(Recurrent Neural Network)에서 발생하는 의존성 문제 또한 극복했다.

본 논문의 목표 서비스인 패션 추천 대화 시스템에서는 사용자의 긴 발화와 이전 발화 내용에 대한 정확한 이해를 통한 정확한 추천이 필요하므로 이 분야에서 가장 좋은 성능을 보이는 트랜스포머를 기본 모델로 사용하여 대화(텍스트) 기반의 기존 트랜스포머 모델을 패션(이미지) 정보를 융합 적용한 멀티모달 기반 트랜스포머 모델로 개선하고 데이터 전처리 및 학습데이터 표현에 대한 분석을 진행하여 이를 학습데이터 개선에 반영한 개선된 멀티모달 기반 트랜스포머 모델을 제안한다. 본 논문 제안한 시스템은 기존 시스템의 정확도 0.3372 WKT를 0.6563 WKT로 크게 향상 시켰다.

II. 관련 연구

1. 기존 대화시스템

전통적인 대화 시스템은 통계적 모델을 기반으로 하는 경우가 많았으나 최근에는 다양한 딥러닝 모델이 적용되고 있다. 대화 시스템은 일반적인 대화를 다루는 대화지향(Non-task-oriented) 대화 시스템[9]과 특정 task를 해결하기 위한 목적지향(task-oriented) 대화 시스템[10][12]으로 나뉜다. 대화지향 대화 시스템은 일반적으로 사람과의 오픈 도메인에서의 의사소통을 목표로 하며, 목적지향 대화 시스템은 비교적 작은 도메인에서 대화를 진행하여 특정 목적을 수행하는 것을 목표로 한다. 일반적으로 목적지향 대화 시스템은 여러 단계를 거쳐 응답하는 파이프라인(pipeline) 방법과 종단간(end-to-end) 방법으로 나뉘며, 파이프라인 방법은 각 모듈을 디자인을 해야 하기 때문에 모델 디자인이 복잡하고 개발에 많은 노력이 들어 최근에는 어플리케이션에 적용하기 쉽고 재학습/사용이 가능한 종단간 방법에 대한 연구가 활발히 진행 중이다.

2. 트랜스포머

자연어 처리를 위해서 CNN[6], RNN[13], [14] 등의 다양한 딥러닝 모델들이 사용되고 있다. CNN은 지역 정보(local information)를 잘 이해하기 때문에 다양한 자연어처리 태스크에서 많이 사용되고 있다. 그러나 CNN은 입력 문장 또는 문단의 길이가 길어질 경우에 이전 발화 또는 이전 대화의 내용을 완전히 활용할 수 없다는 구조적인 단점이 존재한다. 긴 대화 속에서 발화자의 의도를 정확히 찾아야 하는 대화 시스템에서는 CNN의 단점이 치명적이다. RNN은 시계열 데이터(sequence data)를 잘 처리하기 위해 고안된 모델로서 자연어처리에 특성에 잘 맞아 다양한 분야에 사용되고 있다. 그러나 RNN도 아주 긴 대화 등의 데이터에서는 장기 의존성(long-term dependency) 문제가 발생한다. 트랜스포머[8]는 기계번역을 위해 고안된 모델로서 기존에 좋은 성능을 보였던 CNN, RNN 등의 네트워크를 사용하지 않고 MLP와 어텐션 방법만을 사용한 모델이다. 어텐션 방법을 이용하기 때문에 CNN이 가지고 있던 지역 정보 문제와 RNN이 가지고 있던 장기 의존성 문제를 둘 다 해결이 가능하다. 이후 트랜스포머의 인코더 레이어 또는 디코더 레이어를 이용하고, 대용량 코퍼스에 대해 사전학습을 한 모델들[15][16]이 고안됐다. 이러한 모델들은 대부분의 자연어 처리에서 SOTA (State-of-the-art) 성능을 기록하고 있다. 대화의 길이가 길고 대화 내용을 정확히 이해해 적절한 응답을 해야 하는 대화 시스템의 경우 대용량의 코퍼스를 이용해 사전학습 된 트랜스포머 기반의 모델들[17][18]을 사용했을 때 좋은 성능을 보인다.

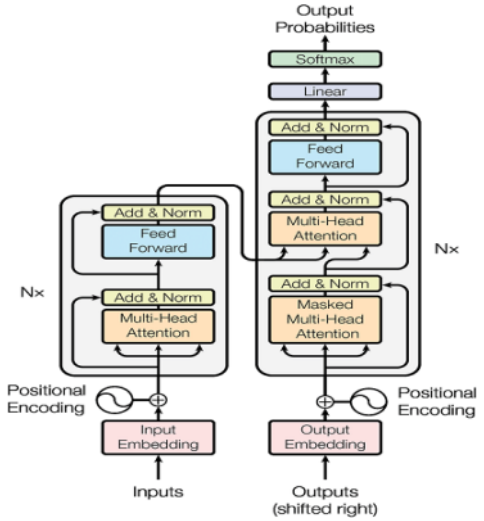

그림 2. 트랜스포머 모델

트랜스포머 인코더는 여러 개의 동일한 레이어로 구성된다. [그림 2]과 같이 각 레이어의 출력은 다음 레이어의 입력으로 사용된다. 각 레이어는 각 단어 사이의 관계를 계산하는 멀티헤드 셀프 어텐션(Multi-head self-attention) 부분과 MLP를 사용하는 순방향 신경망(feed-forward neural network) 부분으로 구성된다. 셀프 어텐션은 주어진 query 벡터에 대한 모든 key 벡터와의 점수를 계산하고, 이후 이 점수를 각각의 value 벡터와 함께 가중 합(weighted sum) 연산을 통해 자기 자신, 즉 인코더에 입력된 벡터들의 최종 표현을 계산한다. 멀티헤드 셀프 어텐션은 어텐션 연산을 여러 개의 가중치 쌍에 대해 계산한 후 여러 개의 결과 벡터를 이어 붙여 사용하는 방법이다. 이러한 과정을 통해 벡터 간의 다양한 관계에 대한 학습이 가능하다.

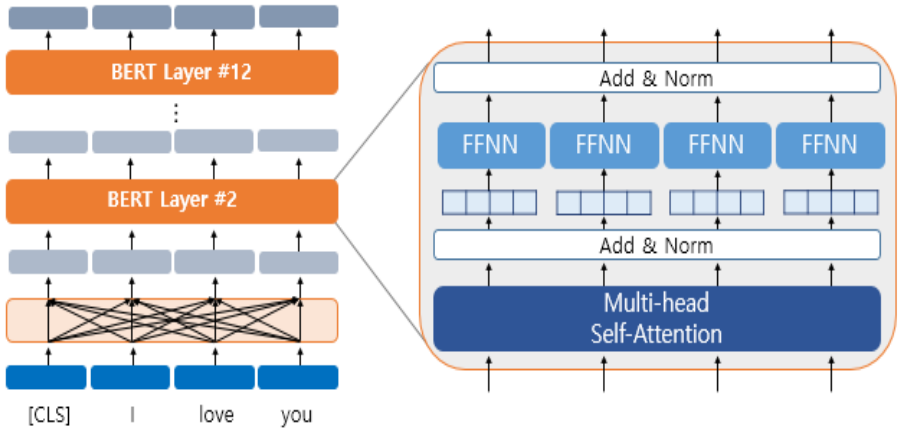

그림 3. BERT 모델[24]

멀티헤드 셀프 어텐션을 지난 후에 정규화(normalize) 과정을 거치고, 순방향 신경망을 거친 다음 다시 한 번 정규화 단계를 거친 벡터를 최종 출력으로 사용한다.

트랜스포머 디코더는 입력 문장이 트랜스포머의 인코더의 모든 레이어를 거친 후에 계산된다. 트랜스포머의 디코더 또한 여러 개의 동일한 레이어로 구성된다. 인코더와 달리 인코더 결과에 대해 인코더-디코더 어텐션을 수행하는 부분이 추가된다. 이 때 디코더의 입력 문장의 각 단어 토큰의 query 벡터들과 이미 계산된 인코더의 입력 문장들의 key 벡터, value 벡터들 간의 어텐션 계산을 통해 벡터값을 계산한다.

3. BERT

BERT(Bidirectional Encoder Representations from Transformers)[15]는 여러 개의 트랜스포머 인코더 블록을 쌓아 만든 언어 모델이다. 대용량의 말뭉치로 사전학습한 후 다운스트림 태스크(down-stream task)에서 미세조정해서 사용하는 것이 일반적이다. 사전학습 단계에서는 마스크드 언어 모델링(masked language modeling)과 다음 문장 예측(next sentence prediction) 2개의 목적함수(objective)를 수행한다. 마스크드 언어 모델링은 입력 문장의 임의의 몇 개의 단어 토큰에 대해 마스킹(masking)을 진행한 후, 원래 그 단어 토큰이 어떤 토큰이었는지를 맞추도록 학습된다. 다음 문장 예측에서는 입력으로 들어오는 2개의 문장이 실제로 연속된 문장인지, 아니면 관련 없는 2문장인지를 모델이 분류하도록 학습한다. 미세조정 방법은 다운스트림 태스크에 따라 달라진다. 분류 문제를 해결할 때는 입력 문장의 맨 앞 토큰인 [CLS] 토큰에 MLP를 연결한 후 softmax 함수를 이용해 해결하고, 개체명 인식 등의 시퀀스 레이블링(sequence labeling) 문제를 해결할 때는 BERT를 통과한 각 단어 토큰 위에 MLP를 연결한 후 마찬가지로 softmax 함수를 사용해 해결한다.

4. 멀티모달

기존의 머신러닝은 많은 부분 동형 데이터를 모델에 입력으로 사용되는 제한이 있었다. 딥러닝의 발전으로 비정형 데이터 처리가 가능해짐에 따라, 이종 데이터로부터 추출된 특성을 하나의 딥러닝 모델에 사용할 수 있다. 이미지 편집을 위해 음성과 제스처를 결합한 멀티모달 시스템[19]은 더 많은 이미지 편집을 위한 작업 속도를 증가시키는데 도움을 준다. 또한, 언어 모델링과 시각적 특징(visual features)들을 결합한 멀티모달에서는 이미지 질문에 대한 자연어 답변[20], 이미지 캡션[21]에서 우수한 성능을 보여 주고 있다.

Ⅲ. 제안 방법

1. 데이터 전처리

본 논문에서는 FASCODE & Codebook(FAShion

COordination DatasEt / FAShion CODE)[23] 패션 대화 데이터(토큰 분리 미적용 버전)을 사용한다. [그림 4][그림 5]는 본 논문에서 사용한 데이터 샘플이고, [표 1]는 발화자 태그 및 의도 태그 설명이다. 본 논문에서 사용한 데이터는 시스템과 사용자가 나눈 대화에 대해 발화자 정보, 대화 내용, 대화의 의도 태그, 추천받은 옷에 대한 텍스트 정보로 구성 되어 있다.

상품	설명
OP-147	무릎을 덮는 롱길이
OP-147	바스트 라인 아래로 자연스럽게 퍼지는 전체 A라인
OP-147	옆 라운드 네크라인, 뒤 V 네크라인
SE-071	샌들
SE-071	스트랩 샌들

그림 4. 추천 받은 옷에 대한 텍스트 정보 예시

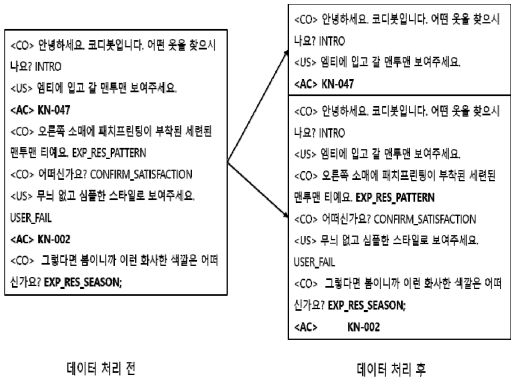

그림 5. 데이터 전처리 전후 예시

True Data	False Data Sampling
가을 축제에 입고 갈 옷 추천해주세요 KN-001 1	가을 축제에 입고 갈 옷 추천해주세요 KN-004 0
	가을 축제에 입고 갈 옷 추천해주세요 KN-011 0
	가을 축제에 입고 갈 옷 추천해주세요 KN-082 0

그림 6. 데이터 증강 예시 (True:1, False: 0)

[그림 5]은 본 논문에서 사용한 대화 데이터를 전처리하는 과정이고, 방법은 다음과 같다: 먼저, 발화자 tag 〈AC〉가 나왔을 때까지 하나의 대화 데이터로 사용한다. 그리고 〈AC〉 태그 이후에 나온 〈CO〉태그의 의도가 〈EXP_*〉 태그일 경우 두 인스턴스의 위치 순서를 변경해준다. 의미가 없는 의도 태그들[Wait, Closing, Confirm_show]에 대해 대화 데이터에서 제외한다. 노이즈 데이터를 줄이기 위해 이전 대화 안에 포함된 〈AC〉 태그 발화 내용 데이터는 대화 데이터에서 제외한다. 대화 의도 태그들은 대화 데이터를 구축을 위해 사용하지만, 실제 학습 데이터에서는 대화 의도 태그들은 제외한다. 최종적으로는 유효한 대화 내용 데이터를 순차적으로 이어 붙여 하나의 대화 데이터를 생성하게 된다. 높은 퍼포먼스를 위해서는 보통 양질의 데이터에 의존하는 경향이 있다. 때문에 본 논문에서는 소규모의 데이터를 보완하기 위해 [그림 6]과 같이 데이터 증강(data augmentation)을 한다. 데이터 전처리 과정을 통해 나온 트루 데이터(True Data)에 대해 1:3, 1:4, 1:5의 다양한 비율로 거짓 데이터 견본 추출(False Data sampling)을 한다. 거짓 데이터 견본 추출을 할 때, 기존의 데이터에 동일한 의상 타입(Type)으로 추출(sampling) 한다.

표 1. 발화자 태그 및 대화 의도 태그 설명

태그	설명
〈CO〉	코디(시스템)발화
〈US〉	유저 발화
〈AC〉	추천 의상 ID
EXP_RES_*	추천 의상 설명
USER_SUCCESS	추천 의상 성공
USER_SUCCESS_PART	일부 추천 의상 성공
USER_FAIL	추천 의상 실패
FAIL	의상 추천 실패
ASK_*	사용자가 원하는 의상 에 대한 질문
INTRO	대화 도입부
CONFIRM_*	확인 질문
SUCCESS	의상 추천 성공
CLOSING	대화 종료
WAIT	대기 요청
SUGGEST_*	제안 발화
NONE	의상 없음
HELP	사용자 지원

2. 제안 모델

2.1 ELECTRA

ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)[22]는 BERT와 동일한 구조를 가지고 있는 모델이다. BERT보다 학습 속도가 빠르고, 비교적 적은 파라미터(parameter)를 가지고 BERT 이상의 성능을 기록했다.

본 논문에서는 한국어로 사전 학습된 KoELECTRA 모델을 사용한다[25]. 사전 학습된 KoELECTRA 모델

을 FASCODE & Codebook에 미세조정 학습 후 패션 추천에 사용한다. [그림 7]은 KoELECTRA 이용한 한국어 패션 추천 시스템의 구조다. 대화 데이터를 입력으로 받은 후 KoELECTRA 모델이 출력하는 [CLS] 토큰벡터에 MLP 연결 후 해당 의상 코디의 적절성 여부를 참/거짓으로 판단한다.

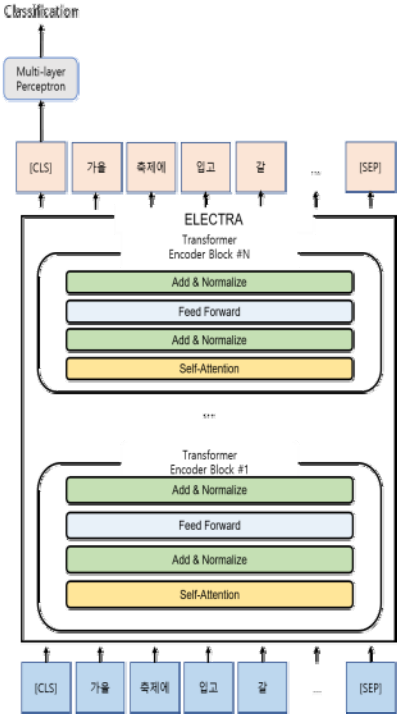

그림 7. ELECTRA를 이용한 패션 추천 시스템

2.2 Multimodal ELECTRA

멀티모달(Multimodal) ELECTRA 모델은 이전 대화 내용에 의상 코디의 설명 데이터뿐만 아니라 이미지 데이터도 이용해 의상 코디의 적절성 여부를 판단한다. 이미지 데이터는 먼저 합성곱 층(convolution layer)과 최대 풀링(max pooling)으로 이루어져 있는 자질 추출(feature extraction) 단계를 거친다. 이 과정을 통해 ELECTRA 모델의 입력 벡터와 같은 크기의 시각 자질 시계열 데이터 (visual feature sequence data)를 생성한다. 이후 Attention Gating 모듈과 Shifting 모듈이 포함되어 있는 Multimodal Adaptation Gate를 거쳐 이미지 데이터와 설명 데이터의 내용을 결합한다. 이렇게 생성된 멀티모달 입력 시계열 데이터를 ELECTRA 모델의 입력으로 사용하고 최종적으로 해당 의상 코디의 적절성 여부를 참/거짓으로 판단하게 된다. Multimodal Adaptation Gate의 과정은 아래와 같다.

$$g_i = ReLU(W_g([E_i; V_i])) \quad (1)$$

$$H_i = g_i \cdot (W_v V_i) \quad (2)$$

$$(E_i) = E_{i+\alpha H_i} \quad (3)$$

$$\alpha = \min \quad (\|E_i\|_2 / \|H_i\|_2 \beta, 1) \quad (4)$$

위와 같은 방법으로 본 논문에서는 이미지데이터와 텍스트 데이터를 결합한 방법인 초기 융합 멀티모달(early fusion multimodal) 패션 추천 시스템을 제안한다.

3. 학습 및 평가 데이터 모델 적용

본 논문에서는 사용자와 시스템이 나누었던 이전 대화를 활용하여, 주어진 의상 코디 중 사용자에게 적절한 코디를 추천하는 시스템을 제안한다. 모델을 미세조정 시에는 사용자와 시스템의 대화에 대해 해당 의상 코디가 적절한지의 여부를 분류하는 문제로 학습한다. 의상 코디 안의 각각의 의상 아이템이 대화 내용에 비추어 볼 때 적절하다면 참, 그렇지 않다면 거짓을 출력하도록 학습한다. ELECTRA 모델의 입력은 대화와 의상 아이템의 설명 데이터의 특징 단어들을 하나로 이어 붙여 긴 입력 시계열 데이터를 생성하여 사용한다. ELECTRA 모델을 통과해 출력된 [CLS] 토큰의 벡터를 MLP를 통과시킨 후 참, 거짓 여부를 출력할 수 있도록 Cross Entropy Loss를 사용해 학습했다.

학습이 완료된 이후에는 주어진 의상 코디 중 가장 적절한 의상 코디를 정해야 한다. 이때 각각의 의상 코디 내에 있는 아이템들에 대해 모델의 출력 값을 각 의상 아이템의 점수로서 사용한다. 한 의상 코디의 점수는 해당 의상 코디 내의 모든 아이템의 점수의 합으로 계산한다. 이후 점수가 높은 순서대로 대화에 적절한 의상 코디라고 판단한다.

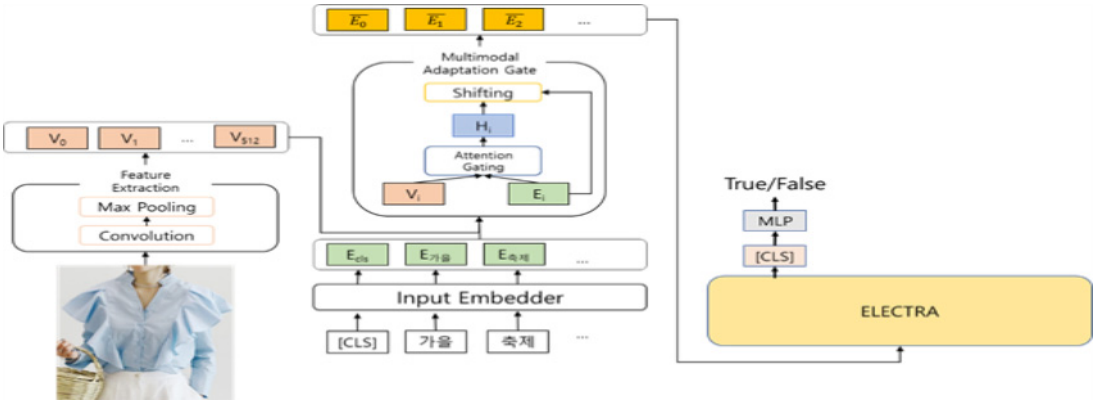

그림 8. Multimodal ELECTRA를 이용한 패션 추천 시스템

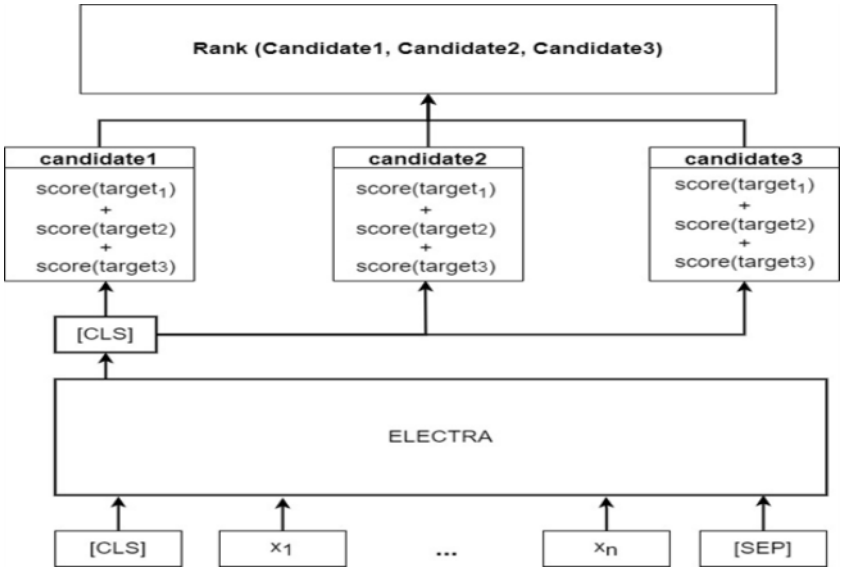

그림 9. 한 의상 코디가 3가지 종류로 이루어져 있을 때의 최종 의상 코디 선택 방법 예시

Ⅳ. 실험 결과

본 논문에서는 FASCODE DATA (토큰 미분류 데이터)를 이용하고, 다양한 하이퍼 파라미터(hyperparameter)와 데이터 증강을 적용하여 WKT Score를 구한다. WKT는 순위 상관계수(rank correlation coefficient)의 한 종류이며 두 변수들 간의 순위를 비교하여 연관성을 계산하고, 가중치가 고려된 상관계수를 측정하는 방법이다. 가중치는 각 요소에 음이 아닌 순위를 할당하는 순위 배열을 통해 정의된다. WKT 수식은 다음과 같다.

$$\frac{(Concordantpair - Disconcordantpair)}{(Concordantpair + Disconcordantpair)} \quad (5)$$

본 논문에서는 KoELECTRA 모델에 [표 2]와 같은

하이퍼 파라미터를 이용하여 실험을 진행한다.

표 2. 실험에서 사용한 하이퍼 파라미터

하이퍼 파라미터	값
Batch size	16
Sequence length	512
Optimizer	Adam
Learning rate	{ 3e-5; 4e-5 }
Epoch	{ 3; 4; 5; 10 }

[표 3]은 하이퍼 파라미터, 데이터 증강비율에 따른 실험 결과이다.

표 3. 다양한 모델에 따른 실험 결과

모델	WKT Score
KoELECTRA(1)	0.3372
KoELECTRA(2)	0.6218
KoELECTRA(3)	**0.6422**
KoELECTRA(4)	0.375
Multimodal KoELECTRA(1)	0.2999
Multimodal KoELECTRA(2)	**0.5781**
Multimodal KoELECTRA(3)	0.5304

KoELECTRA(1) 모델은 데이터 증강 없이 원본 데이터, learning rate 3e-5, epoch 5를 사용했을 때 0.3372 WKT score를 보였다. KoELECTRA(2) 모델은 learning rate 4e-5, epoch 5, 1:3 비율로 데이터 증강을 적용했고, KoELECTRA(3) 모델은 learning rate 4e-5, epoch 5, 1:4 비율로 데이터 증강을 했다. KoELECTRA(4) 모델은 learning rate 3e-5, epoch 5, 1:5 비율로 데이터 증강을 했다. 이때 각각 0.6218, 0.6422, 0.375 WKT score를 보였다.

Multimodal KoELECTRA(1)는 원본 데이터를 사용하고, 하이퍼 파라미터는 learning rate 3e-5, epoch 5를 했을 때 0.2999 WKT score를 보였다. Multimodal KoELECTRA(2)는 1:3비율, Multimodal KoELECTRA(3)는 1:4 비율로 데이터 증강을 했고, 동일하게 learning rate 3e-5, epoch 5를 하이퍼 파라미터로 사용했을 때 각각 0.5781, 0.5304 WKT score를 보였다.

KoELECTRA 모델의 경우 데이터 증강을 했을 때 최대 0.30의 WKT score 성능 향상이 있었고, Multimodal KoELECTRA의 경우 최대 0.27의 WKT score 성능 향상이 있었다. 모델의 종류와 무관하게 데이터 증강을 사용할 경우 성능 향상을 기록했다.

표 4. 최종모델에 대한 결과

모델	WKT Score
KoELECTRA(3)	0.6422
Multimodal KoELECTRA(2)	0.5781
Ensemble	**0.6563**

[표 4]은 본 논문에서의 최종 모델에 대한 결과이다. ensemble 모델은 KoELECTRA(3)모델과 Multimodal KoELECTRA(2)를 선형 앙상블(linear ensemble)한 모델이다. 최종적으로 ensemble모델이 0.6563 WKT score로 가장 높은 성능을 기록했다.

V. 결론

본 논문에서는 멀티모달 패션 추천 시스템을 위해 개선된 트랜스포머 모델을 제안한다. 본 논문에서 제안하는 모델은 자연어 데이터만 사용하는 기존 트랜스포머 모델과 달리, 자연어 데이터와 이미지 데이터를 동시에 사용해 의상 코디를 추천한다. 제안하는 모델은 대용량의 말뭉치에 사전 학습된 언어 모델을 사용함으로써 비교적 정확하게 대화 내용을 이해했으며, 본 논문에서는 모델의 성능 향상을 위해 데이터 증강 등의 데이터 전처리를 진행했다. 향후 연구에서는 멀티모달 모델의 결합 부분을 개선하여 좀 더 이미지 데이터를 효과적으로 사용할 수 있는 방법을 고안할 것이며, ELECTRA 모델이 아닌 다른 사전 학습 언어 모델들에 대해서도 추가적인 연구를 진행할 계획이다.

참 고 문 헌

[1] T. Young, D. Hazarika, S. Poria, and E. Cambria, “Recent Trends in Deep Learning Based Natural Language Processing,” IEEE Computational Intelligence Magazine, Vol.13, pp.55-75, 2018.

[2] Y. N. Chen, A. Celikyilmaz, and D. H. Tür, "Deep Learning for Dialogue Systems," in Proceedings of the 27th International Conference on Computational Linguistics: Tutorial Abstracts, pp.25-31, 2018.

[3] S. Koo, H. Yu, and G. G. Lee, "Adversarial approach to domain adaptation for reinforcement learning on dialog systems," Pattern Recognit Lett, Vol.128, pp.67-473, 2019.

[4] H. Chen, X. Liu, D. Yin, and J. Tang, "A Survey on Dialogue Systems," ACM SIGKDD Explor. Newsl, Vol.19, No.2, pp.25-35, 2017.

[5] B. H. Su, T. W. Kuan, S. P. Tseng, J. F. Wang, and P. H. Su, "Improved TF-IDF weight method based on sentence similarity for spoken dialogue system," 2016 International Conference on Orange Technologies, pp.36-39, 2016.

[6] Y. Kim, "Convolutional Neural Networks for Sentence Classification," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp.1746-1751, 2014.

[7] H. Palangi, L. Deng, Y. Shen, J. Gao, J. Chen, and R. Ward, "Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval," IEEE/ACM Trans. Audio, Speech and Lang, Vol.24, No.4, pp.694-707, 2016.

[8] A. Vaswani, N. Shazzer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in Advances in Neural Information Processing Systems 30 Curran Associates, Inc, pp.5998-6008. 2017.

[9] Z. Yu, Z. Xu, A. W. Black, and A. Rudnicky, "Strategy and Policy Learning for Non-Task-Oriented Conversational Systems," in Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp.404-412, 2016.

[10] T. H. Wen, D. Vandyke, N. Mrksic, M. Casic, L. M. Rojas Barahona, P. H. Su, S. Ultes, and S. Young, "A Network-based End-to-End Trainable Task-oriented Dialogue System," CoRR, Vol.1604, p.1236, 2016.

[11] C. Gunasekara, J. K. Kummerfeld, L. Polymenakos, and W. Lasecki, "7 Task 1: Noetic End-to-End Response Selection," in Proceedings of the First Workshop on NLP for Conversational AI, 2019, pp.60-67, doi: 10.18653/v1/W19-4107.

[12] A. Bordes, Y. L. Boureau, and J. Weston, "Learning End-to-End Goal-Oriented Dialog," 2017, [Online]. Available: https://openreview.net/forum?id=S1Bb3D5gg.

[13] A. Sherstinsky, "Fundamentals of Recurrent Neural Network and Long Short-Term Memory Network," CoRR, Vol.abs, p.1808, 2018, [Online]. Available: http://arxiv.org/abs/1808.03314.

[14] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using Encoder-Decoder for Statistical Machine Translation," CoRR, Vol.abs, p.1406, 2014.

[15] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "Pre-training of Deep Bidirectional Transformers for Language Understanding," Human Language Technologies, Vol.1, pp.4171-4186, 2019. doi: 10.18653/v1/N19-1423.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Sotyanov, "RoBERTa: Robustly Optimized Pretraining Approach," CoRR, Vol.1907.1, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692.

[17] W. Rahman, M. K. Hasan, S. Lee, A. Zadeh, C. Mao, L. P. Morency, and E. Hoque, "Integrating Multimodal Information in Large Pretrained Transformers," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp.2359-2369, 2020. doi: 10.18653/v1/2020.acl-main.214.

[18] X. Zhou, L. Li, D. Dong, Y. Liu, Y. Chen, W. X. Zaho, D. Yu, and H. Wu, "Multi-Turn Response Selection for Chatbots with Deep Attention Matching Network," Association for Computational Linguistics, Vol.1, pp.1118-

1127, 2018. doi: 10.18653/v1/P18-1103.
[19] G. Laput, M. Dontcheva, G. Wilensky, W. Chang, A. Agarwala, J. Linder, and E. Adar, "PixelTone: a multimodal interface for image editing," in 2013 Conference on Human Factors in Computing Systems, 13, Paris, France, pp.2185-2194, 2013. doi: 10.1145/2470654.2481301.
[20] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual Question Answering," In Proceedings of the IEEE international conference on computer vision, 2015.
[21] J. Devlin, H. Cheong, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell, "Language Models for Image Captioning: The Quirks and What Works," Association for Computational Linguistics, Vol.2, pp.100-105, 2015. doi: 10.3115/v1/P15-2017.
[22] K. Clark, M. T. Luong, Q. V. Le, and C. D. Mannin g, "{ELECTRA:} Pre-training Text Encoders as Discr iminators Rather Than Generators," 2020, [Online] Available: https://openreview.net/forum?id=r1xMH 1BtvB.
[23] 정의석, 김현우, 오효정, 송화전, "인터렉션 기반 추천 시스템 개발을 위한 데이터셋 연구," 한글 및 한국어 정보처리 학술 대회, pp.1-5, 2020.
[24] https://wikidocs.net/115055, 2021.10.07.
[25] https://github.com/monologg/KoELECTRA, 2020

저 자 소 개

박 영 준(YeongJoon Park) 정회원

- 2018년 2월 : 서강대학교 컴퓨터공학과(공학사)
- 2020년 2월 : 서강대학교 컴퓨터공학과(공학석사)
- 2020년 4월 ~ 현재 : 엔에이치엔 다이퀘스트 연구원

〈관심분야〉 : 자연어처리, AI

조 병 철(Byeong Cheol Jo) 정회원

- 2016년 2월 : 한림대학교 융합소프트웨어(공학사)
- 2018년 2월 : 한림대학교 융합소프트웨어(공학석사)
- 2020년 1월 ~ 현재 : 엔에이치엔 다이퀘스트 연구원

〈관심분야〉 : 자연어처리, AI

이 경 욱(Kyoung Uk Lee) 정회원

- 2003년 2월 : 단국대학교 전자계산학과(공학사)
- 2011년 8월 : 성균관대학교 경영대학원 MBA(경영학석사)
- 2003년 8월 ~ 2010년 5월 : 엔에이치엔 다이퀘스트 연구원
- 2009년 1월 ~ 2010년 5월 : sk Planet 검색파트 매니저
- 2011년 7월 ~ 현재 : 엔에이치엔 다이퀘스트 수석연구원

〈관심분야〉 : 자연어처리, AI, 챗봇, 정보검색

김 경 선(Kyung Sun Kim) 정회원

- 1995년 2월 : 서강대학교 전자계산학과(공학사)
- 1997년 2월 : 서강대학교 컴퓨터공학과(공학석사)
- 2006년 2월 : 서강대학교 컴퓨터공학과(공학박사)
- 2006년 9월 ~ 2009년 3월 : 삼성전자 통신연구소 책임연구원
- 2009년 3월 ~ 현재 : 엔에이치엔 다이퀘스트 CTO

〈관심분야〉 : 자연어처리, 챗봇, 정보검색