COSE474-2022F: Final Project Report
Fashion Item Recommendation Methodology with Natural Language

Jiwon Jeong

1. Introduction

For the classical fashion item recommendation services, Collaborative Filtering, Content-based Filtering, and a filtering technology in which the two are mixed or different technology is fused are used (An, Kwon, & Park, 2019)(1). These three technologies are based only on users' purchase history and user characteristics such as preferences. However, these technologies have problems in that the interaction with users is limited. These services can only recommend fashion items to users within given categories, such as the kinds and colors.

In recent years, companies have introduced AI chatbot to recommend fashion items. As the result, customers showed high confidence in the product recommendation service chatbot, which had a positive effect on consumer responses (Lee, Kim, & Park, 2022)(2). The researchers suggest that if the consumer's conversational input is further segmented, it will lead to a more positive effect on consumer purchase.

Therefore, these above studies propose that a methodology for recommending fashion items through greater interaction with users using natural language is important to user custom services.

Accordingly, this study intends to suggest a method of recommending fashion items using natural language arbitrarily written by the users, beyond recommendation within a limited category.

Then, Our contributions are threefold:

- We present a model for measuring similarity between reference image and target image and a caption explaining the differences between the images.
- We train the model by applying the difference in image features and the captions between the two images to CLIP model(3).
- Given an image by user, we present an efficient image retrieval method that does not compare with all images in the dataset.

2. Related Works

Radford et al. (2021)(4) have pointed out that the existing training method which predicts within predefined categories is deficient in the usability and generalization, and it is difficult to collect a dataset since each data needs labeling. They presented a new model which is called CLIP, Contrastive Language-Image Pre-training, that learns relationships between images and the sentences which describe the images.

Following the above research, Chia et al. (2022)(5) introduced a model called FashionCLIP that could be used in the fashion industry. FashionCLIP showed better performance in Multi-modal Retrieval, Zero-shot Classification and Compositionality than CLIP in single-cloth photos.

3. Methods

3.1. Challenges

The goal of this study is to suggest Retrieval System for recommending a new fashion item based on the user's purchase data when the user inputs a natural language description.

Vo et al. (2019)(6) have suggested a new way to combine image and text using Text Image Residual Gating (TIRG) function for the image retrieval task. The TIRG function encodes the reference image and text in the same dimension, then combines the two features. After that, training proceeds in the direction of maximally reducing the loss between the composition feature, which is a combination of image and text, and the target image feature.

However, this study focuses on the difference between images. Unlike previous studies, we try to match the feature difference between the reference image and the target image and the text which explains the images.

3.2. Approach

The overall training process is similar to CLIP, but the difference is that out model maximizes the similarity between the image feature difference and the captions of images.

Simplified Pseudocode

```
# tgt = target, ref = reference
```

```
# img = image, cap = caption
# embed = embedding, sim = similarity
# Getting Image and Text Features
tgt_img_f = image_encoder(batch[tgt_img])
ref_img_f = image_encoder(batch[ref_img])
cap_f = text_encoder(batch[cap])
# Getting Image and Text Embeddings
tgt_img_embed = image_projection(tgt_img_f)
ref_img_embed = image_projection(ref_img_f)
cap_embed = text_projection(cap_f)
# Calculating the Loss
img_embed = tgt_img_embed - ref_img_embed
logits = cap_embed  img_embed.T
img_sim = img_embed @ img_embed.T
cap_sim = cap_embed @ cap_embed.T
targets = softmax((img_sim+cap_sim)/2)
img_loss = cross_entrophy(logits, targets)
cap_loss = cross_entrophy(logits.T, targets.T)
loss = (img_loss + cap_loss) / 2
```

Figure 2. Psuedocode for the core of of our model(3)

Given a batch of N (target image, reference image, caption) pairs, our model encodes two images and text in the same multi-modal embedding space. Then, the model trains image encoders and text encoders in a direction that maximizes the cosine similarity of the difference between the two images and the text. The model optimizes a symmetric cross-entropy loss for these similarity values. In Figure 1, we summarize and illustrate this process. In Figure 2, We include the simplified pseudocode of the core part of our model.

4. Experiments

4.1. Datasets & Resources

For a professional AI-based fashion model study, Jin, Piao, Gu, & Yoo (2019)(8) have investigated and analyzed image datasets used in fashion analysis research in terms of the purpose, research field, and characteristics. They have argued that Fashion IQ dataset is useful for research on search systems incorporating natural language feedback, since contain relative captions derived from product descriptions.

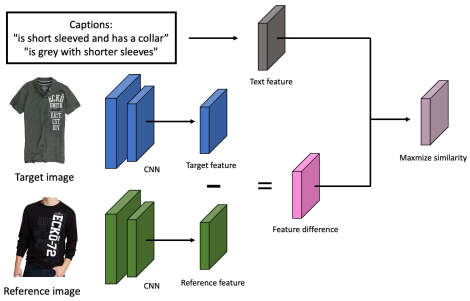

Figure 1. Summarized Our Approach for training. We encoded the reference image and the target image as features using CNN, and encoded the text using transformer. Then, we trained to maximize the similarity between the feature difference of images and feature of text.

	# Image	# With Attr.	# Relative Cap.
Dresses			
Train	11,452	7,741	11,970
Val	3,817	2,561	4,034
Test	3,818	2,653	4,048
Shirts			
Train	19,036	12,062	11,976
Val	6,346	4,014	4,076
Test	6,346	3,995	4,078
Tops&Tees			
Train	16,121	9,925	12,054
Val	5,374	3,303	3,924
Test	5,374	3,210	4,112

Table 1. Fashion IQ Dataset statistics (Wu et al., 2021)(7)

Fashion IQ dataset consists of 77,684 images including a total of 3 categories (dresses, shirts, tops&tees), of which 49,464 images include side information and 60,272 relative captions. The side information consists of about 1,000 labels grouped into five annotations: texture, fabric, shape, composition and style. And, the relative caption consists of a description expressing the difference between the reference image and the target image specified by human in natural language (Wu et al., 2021)(7).

In this study, the computer resources are as follows.

- Environment: Google Colab
- Platform: Linux-5.10.133+-x86_64-with-glibc2.27
- OS: Ubuntu 18.04.6 LTS
- CPU: Intel(R) Xeon(R) CPU @ 2.00GHz
- Total Memory: 13297228 kB
- GPU: NVIDIA Tesla T4

4.2. Design

Initially, we expected that training on the entire dataset would yield meaningful results regardless of the type of clothing. However, when we directly conducted training and testing, no meaningful results were obtained. Thus, we train the model focusing on 'Shirts' with the most data among

the three types (Dresses, Shirts, Tops&Tees) in Fashion IQ dataset.

We use ResNet50 model as the image encoder, and DistilBERT model from HuggingFace library as the text encoder. SANH, DEBUT, CHAUMOND, & WOLF (2019)(9) have proposed a smaller and more efficient general purpose language representation model, called DistilBERT. Since DistilBERT is 40% smaller and 60% faster than the BERT while retaining 97% of the capability, we use this model.

The overall training method of our model is the same as described above. More details like setting hyperparameters are below.

- Batch Size is 32.
- Epoch is 5.
- Image embedding size is 2048 and Text embedding size is 768.
- The number of projection layers is 1, Dimension of projection is 256, and Dropout is 0.1.
- Learning rate and Weight Decay is 0.001.
- Temperature (Softmax parameter) is 2.0.

During training, We update the model parameter values only when the validation loss became smaller.

After completing to train the model, we proceed with Image Retrival in the following methods. First, the model encode all fashion item images in Fashion IQ dataset in the embedding space. Given an image and query, the model encodes the input image and query in the embedding space, and calculates the difference between all the embedded images and input image. Then, the model compares the cosine similarity values to find the image most similar to the query in the dataset and shows the found image. This method enables efficient Image Retrival through cosine similarity calculation with the pre-stored embedding value of the images without comparing input image to all images in the dataset whenever given the image.

4.3. Quantitative results

	Loss
Train	3.48
Validation	3.47

Table 2. Train & Validation Loss

Table 2 shows the test loss and validation loss of our model.

Table 3 shows R@10 results compared to SOTA (Kim, Yu, Lee, & Kim, 2021)(10). However, for R@10 measurements, the results may vary depending on the person measuring and the input image set.

SOTA(10)	R@10
Validation	0.3229
Test	0.3376

Ours	R@10
Validation	0.38
Test	0.27

Table 3. Comparison R@10 with SOTA

4.4. Qualitative results

Figure 3. Qualitative result example in validation set.

Figure 4. Qualitative result example in test set.

Figure 3 and 4 show the example result in validation and test set. In figure 3, It can be seen that about 5 out of 10 results were relevant. In figure 4, It can be seen that about only 2 out of 10 results were relevant.

5. Limitation and Future Work

This study has several limitations. First, we obtained poor results for the test dataset since we were not really good at fine tuning hyperparameters. We had a hard time tuning the hyperparameters directly since it took a lot of time to train and test the model. Second, As Shin, Cho, & Hong(11) have argued that there were problems with the Fashion IQ dataset, we also found some problems in Fashion IQ dataset. For example, the image in the dataset contains a lot of human bodies, not just clothes. These problems will act as a major distraction to learning. Following this study, if hyperparameter tuning is done well and we use high-quality

datasets, we would be obtain better results.

6. Appendix

https://github.com/stop1one/CLIP

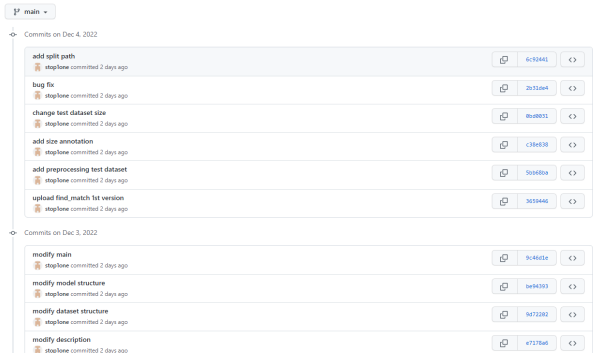

Figure 5. Our Commit History

References

[1] Hyosun An, Suehee Kwon, and Minjung Park. A case study on the recommendation services for customized fashion styles based on artificial intelligence. *Journal of the Korean Society of Clothing and Textiles*, 43(3):349–360, 2019.

[2] Y Lee, H Kim, and M Park. The effects of perceived quality of fashion chatbot's product recommendation service on perceived usefulness, trust and consumer response. *Journal of the Korean Society of Clothing and Textiles*, 46(1):80, 2022.

[3] M. Moein Shariatnia. Simple CLIP, 4 2021.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[5] Patrick John Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Magalhães, Diogo Goncalves, Ciro Greco, and Jacopo Tagliabue. Fashionclip: Connecting language and images for product representations. *arXiv preprint arXiv:2204.03972*, 2022.

[6] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6439–6448, 2019.

[7] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11307–11317, 2021.

[8] Hailin Jin, Zhegao Piao, Yeong Hyeon Gu, and Seong Joon Yoo. A survey of fashion datasets for ai training. In *Proceedings of the Korean Society of Broadcast Engineers Conference*, pages 637–642. The Korean Institute of Broadcast and Media Engineers, 2020.

[9] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[10] Jongseok Kim, Youngjae Yu, Seunghwan Lee, et al. Cycled compositional learning between images and text. *arXiv preprint arXiv:2107.11509*, 2021.

[11] Minchul Shin, Yoonjae Cho, and Seongwuk Hong. Fashion-iq 2020 challenge 2nd place team's solution. *arXiv preprint arXiv:2007.06404*, 2020.