패션 추천에서 멀티모달 파운데이션 모델에 관한 연구

데레 로시다트 올루와부콜라[1], 김경백[1]
[1]전남대학교 인공지능융합학과
roshidatdere@gmail.com, Kyungbeakkim@jnu.co.kr

A Study of MultiModal Foundation Model in Fashion Recommendation

Dere Roshidat Oluwabukola[1], Kyungbeak Kim[1]
[1]Dept. of Artificial Intelligence Convergence, Chonnam National University

Abstract

Influenced by societal trends, cultural standards, and individual personalitiees, fashion is a potent means of self-expression. Many industries have benefited from the advancement of Artificial Intelligence(AI), with the fashion industry emerging as one of the most notable. AI has assisted the fashion industry in a number of areas, including product design and marketing. Online buying has proliferated as the fashion business has expanded into a multibillion-dollar industry, offering customers easy, stress-free shopping experiences. By advising customers on what to buy there could be potential increase in the sales of such and other products. The goal of this study is to investigate qualitatively mutimodal foundation models for fashion critics and advice. In this paper, we adapted a Gemini 1.5 flash on our dataset for compatibility prediction and complementary commentary on clothing. Qualitatively, the model provided very indepth review with varying images while also criticing fashion combination that are not compabible. The study alludes to the robotuness of mutimodal models with reommendation on quantitative evaluation in future studies.

1. Introduction

Fashion is a dynamic and multifaceted concept that encompasses the styles of clothing and accessories that are popular at any given time. It is not just about the clothes we wear; it reflects cultural, social, and economic influences, and it evolves continuously [1]. Fashion compatibility prediction and complementary item retrieval are two fundamental task in fashion [2].

Fashion Compatibility prediction involves assessing if a combination of fashion items in an ensemble complements one another, taking into account the connections between each piece [3]. Complimentary item retrieveal involves finding an appropriate item from a sizable database in order to finish a partially assembled outfit. It gives description of missing items that matches well with the ones that are already there[4].

In this study, we qualitatively investigate the ability of multimodal foundational model in compatibiity prediction.

2. Related Works

FashionViL was introduced in [5], a novel representation learning framework for Large-scale Vision-and-Language (V+L) tasks in the fashion domain. FashionViL includes two pre-training tasks tailored to the unique characteristics of fashion V+L data. The first task, Multi-View Contrastive Learning, addresses the presence of multiple images in the fashion domain by pulling closer the visual representation of one image to the compositional multimodal representation of another image+text. The second task, Pseudo-Attributes Classification, capitalizes on the rich fine-grained concepts in fashion text by encouraging the learned unimodal (visual/textual) representations of the same concept to be adjacent. A flexible, versatile V+L model architecture is also proposed for various downstream tasks. The paper reports that FashionViL achieves new state-of-the-art results across five downstream tasks.

Authors in [6] proposed a fashion compatibility modeling approach with a category-aware multimodal attention network (FCM-CMAN) which used contextual attention mechanism and dynamics representation of categories to augument and aggregate multimodal representation of fashion products focused. The researchers also developed a graph convolutional network to learn the semantic correlations between categories, taking into account the category correlations are always dynamic and variable for different fashion products.

3. Methodology

3.1 Dataset Description

This study, we utilized the Fashion Product Image Dataset [7] hosted on Kaggle and adopted in several peer-reviewws studies. The dataset includes 44,424 images with 11 categories has shown in Figure 1.

	id	gender	masterCategory	subCategory	articleType	baseColour	season	year	usage	productDisplayName
0	15970	Men	Apparel	Topwear	Shirts	Navy Blue	Fall	2011.0	Casual	Turtle Check Men Navy Blue Shirt
1	39386	Men	Apparel	Bottomwear	Jeans	Blue	Summer	2012.0	Casual	Peter England Men Party Blue Jeans
2	59263	Women	Accessories	Watches	Watches	Silver	Winter	2016.0	Casual	Titan Women Silver Watch
3	21379	Men	Apparel	Bottomwear	Track Pants	Black	Fall	2011.0	Casual	Manchester United Men Solid Black Track Pants
4	53759	Men	Apparel	Topwear	Tshirts	Grey	Summer	2012.0	Casual	Puma Men Grey T-shirt

(Figure 1) Dataset Categories

3.2 Preprocessing

The images in the dataset were resized into shape of (600,600) to enable effective model training and evaluation on an Nvidia GeForce 2060. OpenCV, Torch 2.2 and Google-generative AI are some of the tools used in this study.

3.3 Models

The essensce of the study is to qualitatively investigate the performance of multimodal model in fashion recommendation.

To achvieve the aim, Gemini 1.5 flash [8] from Google was the multi-modal foundation model used in this study. Gemini 1.5 Flash is a large language model (LLM) developed by Google AI. Gemini is a trasnformer-based model with Mixture-of-Experts (MoE) routing. Gemini reported to excel in a wide range of tasks. The Gemini Flash model is a lightweight as compared to the Pro. We made the choice to adapt Gemini Flash because of Hardware constraint and its 39.5% reasoning performance on GPQA dataset benchmark. Gemini was used in a virtual enviroment alongside OpenCV and Torch for preprocessing.

Using a chain-of-though approach the Multimodal langauge model is give a combination of fashion wearables. Therefter, it LLM is prompted as follows:

1. “Given this combination of fashion items”.
2. “Based on the color, and current fashion trends,”
3. “Does it look great for the X season?”
4. “Can it be combined to look great?”
5. “Give the response with reason:”

The model reponds with two answers. The first answer is a determination if the combination is possible. Whereas the second respond is the reason why the combination is not possible.

4. Results and Conclusion

4.1 Result

(Figure 2) Reponse for a combination in winter

Gemini was prompter for the winter season given the combination of products in Figure 2. The response indicate that Gemini is able to reason. This is apparent because the top seems like an apparel for a woman while the jean trouers is a man. Therefore the combination would be incompatiible. Additionaly the top of the woman is short making it unsutiable for the winter season. Looking at the dataset, Product 26960 was calssified as a summer cloth.

Hence, we can allude to the effectivness of the model.

(Figure 3) Reponse for Fall season

The model was given products as shown in Figure 3. It gave a reponse as follows “The grey t-shirt with the bus graphic is a bit too casual for the fall season, especially when paired with dress pants. While sandals can be worn in fall, they are not typically paired with dress pants. The overall look is a bit mismatched and doesn't reflect a cohesive fall style.”. As it is appartent, that sandals are typicall not paried with Tourser during the Fall season. A good indicator of a good fashion sense of the model.

Figure 4, has a response of yes with reason “The blue plaid shirt with denim jeans and dark blue socks is a classic and timeless combination that looks great in the fall. The colors are appropriate for the season, and the overall look is both stylish and comfortable.”

(Figure 4) Reponse for Fall season

4.2 Conclusion

Large Multimodal models that take images and text are very powerful models with significant impact in not only recursive text generation but image understading with concept. In this study, we have been able to show the concept of fashion recommendation and reasoning of multimodal models. While the result are quite decent qualitatively, we plan to fine-tune such models in the future and quantitatively measure performance across academic and industy benchmarks.

Acknowledgement

This work was supported by Innovative Human Resource Development for Local Intellectualization program through the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(IITP-2024-RS-2022-00156287, 50%). This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629, 50%) grant funded by the Korea government (MSIT).

References

[1] Girard, A. (2024). History and Evolution of Fashion and Design in Different Regions and Periods in France. International Journal of Fashion and Design, 3(1), 49-59.

[2] Song, X., Feng, F., Liu, J., Li, Z., Nie, L., & Ma, J. (2017, October). Neurostylist: Neural compatibility modeling for clothing matching. In Proceedings of the 25th ACM international conference on Multimedia (pp. 753-761).

[3]Shirkhani, S., Mokayed, H., Saini, R., & Chai, H. Y. (2023). Study of AI-Driven Fashion Recommender Systems. SN Computer Science, 4(5), 514.

[4]Lin, Y. L., Tran, S., & Davis, L. S. (2020). Fashion outfit complementary item retrieval. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3311-3319).

[5]Han, X., Yu, L., Zhu, X., Zhang, L., Song, Y. Z., & Xiang, T. (2022, October). Fashionvil: Fashion-focused vision-and-language representation learning. In European conference on computer vision (pp. 634-651). Cham: Springer Nature Switzerland.

[6]Jing, P., Cui, K., Guan, W., Nie, L., & Su, Y. (2023). Category-aware multimodal attention network for fashion compatibility modeling. IEEE Transactions on Multimedia, 25, 9120-9131.

[7]Aggarwal, P. (2019). Fashion product images dataset. Retrieved from kaggle: https://www. kaggle.

com/paramaggarwal/fashion-product-images-dataset.

[8]Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J. B., ... & Mustafa, B. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.