One size does not fit all: Improving Fashion Recommendation relevance via fit-aware Neural Re-ranking

Vishal G* Sai Sreenivas Chintha* Harshitha Kolukuluru*

UMass Amherst UMass Amherst UMass Amherst

vishalg@umass.edu saisreenivas@umass.edu hkolukuluru@umass.edu

1. Motivation

Fashion recommendation systems are crucial for enhancing user experience and boosting sales through personalized suggestions. In our project, we aim to explore methods for improving the relevance of image retrieval systems given a textual description of custom needs for fashion recommendation. We plan to train neural networks in a learning to rank framework with custom features extracted from the images and the given query, to first retrieve set of all relevant images with high recall and further enhance the precision of the top-k retrieved items through re-ranking .

2. Contributions

Vishal would be responsible for setting up the retrieval-pipeline and CLIP fine-tuning with custom image features. Harshitha will investigate what custom extracted features are useful for improving relevance (sleeve bounding box, shirt bounding box etc.) and Srinivas will be responsible for training a custom neural network for re-ranking.

3. Literature Review

Learning joint image-text representations can be used to retrieve text given an image and vice versa (via maximum inner product search). [10] uses contrastive learning to learn joint representations of images and text. Positive pairs are taken from aligned text-image pairs while negative pairs are sampled randomly, with the InfoNCE loss [12] being optimized to align these representations via contrastive learning. A top-k item list is retrieved to maximize recall followed by a re-ranking step to optimize precision [11], where a neural network is trained to predict a relevance score $score_\theta(q, d)$ given a query q and document d, where parameters θ are learnable.

[2] show learning to rank can help improve the fashion recommendation systems by training models on various feature representations. The works of [13] further highlight how integrating custom extracted features like bounding boxes, colors, textures and other style elements can help improve the overall retrieval precision. Furthermore, recent work [4] showed modeling joint probabilities can improve retrieval relevance and better handle OOD data compared to CLIP.

4. Data

For the task of retrieving items (images) given text, there are many datasets with aligned text and images with the most popular one being MSCOCO [6]. Other datasets include Flickr30K [8], Deep fashion [7] and INRIA holidays [9]. For our project on fashion recommendation we plan to use Fashion 30K [1] and Deep Fashion [7] to benchmark our pipeline. While these datasets provide aligned text-image pairs, they often lack detailed sub-image and text-level annotations. [13] for example provides a dataset of annotation at the level of pixel and bounding box. [5] analyzes the effect of different features for text-visual alignment. We wish to investigate if we can use the above works to create a sub-image level annotation and if that's not sufficient annotate a small dataset of our own.

5. Approach

Previous works use CLIP embeddings of texts and images to compute maximum inner product search for retrieving images given text and vice-versa. We aim to extend this pipeline by re-ranking the retrieved images using the following methods:

1. **CLIP (Baseline):** We employ CLIP to generate embeddings for images and textual queries. By measuring the similarity between these embeddings, we effectively retrieve the top-k images that are most relevant to the query.
2. **Fine-tune CLIP with custom extracted features:** We will extract custom features from the top-k images, such as bounding box dimensions using image segmentation algorithms, dominant colors using color histograms, tex-

[1] all authors contributed equally

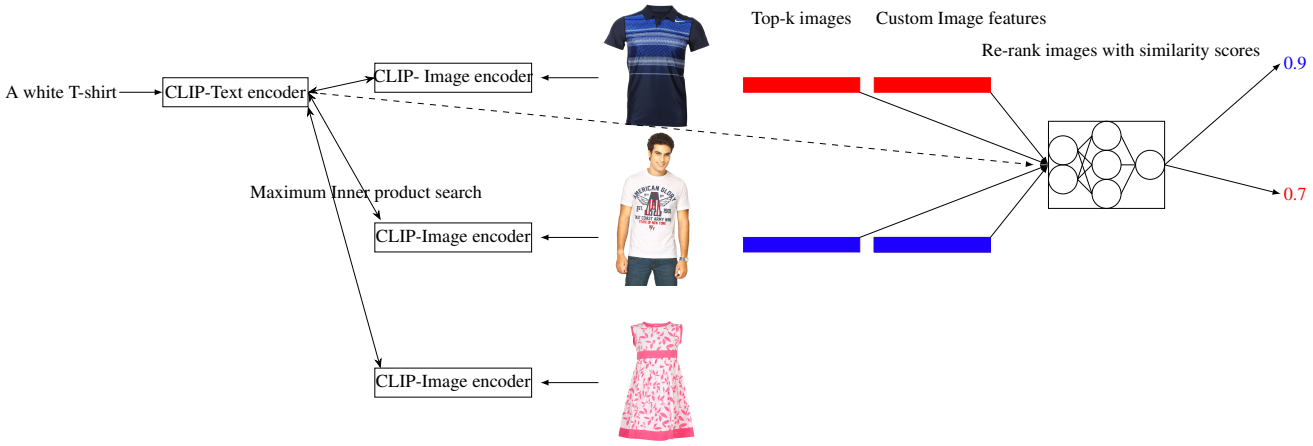

Figure 1. Diagram of our pipeline to first retrieve images with high recall and use re-ranking with custom image features such as bounding box, sleeve length etc. along with the embeddings of original image and query to achieve better precision

ture features and item attributes like sleeve length, neckline type etc. We will conduct ablation studies to assess the contribution of these features to re-ranking. We will fine-tune CLIP with these additional features using contrastive learning with **Low-Rank Adaptation (LoRA)** [3].

3. **Build a neural network for re-ranking:** Instead of fine-tuning CLIP with custom features, we plan to train a neural network to predict relevance scores given image embedding, query embedding and custom image features extracted in the previous step.

6. Evaluation Metric

1. **Qualitative Evaluation:** For each query, we will present examples of retrieved images to visually analyze the relevance and accuracy of the retrieval system.
2. **Quantitative Evaluation:** We will use performance metrics such as Recall at k (R@K), Precision at K (P@K), mean average precision (mAP) & nDCG.

References

[1] Param Aggarwal. Fashion product images dataset, 2019.

[2] Samit Chakraborty, Md. Saiful Hoque, Naimur Rahman Jeem, Manik Chandra Biswas, Deepayan Bardhan, and Edgar Lobaton. Fashion recommendation systems, models and methods: A review. *Informatics*, 8(3), 2021.

[3] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[4] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. Diffusionret: Generative text-video retrieval with diffusion model, 2023.

[5] Katrien Laenen and Marie-Francine Moens. Learning explainable disentangled representations of e-commerce data by aligning their visual and textual attributes. *Computers*, 11 (12), 2022.

[6] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[7] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[8] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870, 2015.

[9] Adrian Popescu, Etienne Gadeski, and Hervé Le Borgne. Scalable domain adaptation of convolutional neural networks. *CoRR*, abs/1512.02013, 2015.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[11] Xi SHEN, Yang Xiao, Shell Xu Hu, Othman Sbai, and Mathieu Aubry. Re-ranking for image retrieval and transductive few-shot classification. In *Advances in Neural Information Processing Systems*, pages 25932–25943. Curran Associates, Inc., 2021.

[12] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[13] Shuai Zheng, Fan Yang, M. Hadi Kiapour, and Robinson Piramuthu. Modanet: A large-scale street fashion dataset with polygon annotations, 2019.