International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.152

Fashion Products Retrieval and Recommendation Pipeline Using Multi-Modal Representations from Contrastive Learning

Harsh Sharma
dept. of Data Science and Business Systems SRM Institute Of Science and Technology
Chennai, India
hs7685@srmist.edu.in

Dr. A. Shanthini
dept. Of Data Science and Business Systems SRM Institute Of Science and Technology
Chennai, India
shanthia@srmist.edu.in

AarshChaube
dept. Of Data Science and Business Systems SRM Institute Of Science and Technology
Chennai, India
ac5973@srmist.edu.in

***Abstract*—This research explores the use of multi-modal representations and contrastive learning in the retrieval and recommendation of fashion products. By leveraging multiple modalities, such as images and textual descriptions, the study aims to improve the accuracy and diversity of product recommendations. The proposed method utilizes contrastive learning to learn representations that capture both the similarities and differences between products, enabling effective retrieval and recommendation. The study presents promising results, demonstrating the potential of multi-modal representations and contrastive learning in fashion product retrieval and recommendation systems.**

***Index Terms*—Deep Learning, Generative Modelling, Recommendation System, Transformers**

I. Introduction

Due to the growth of e-commerce and online buying, the fashion industry has seen a substantial transition in recent years. The difficulty of efficiently retrieving and recommending products has grown more important than ever due to the expansion of the availability of fashion products online. To address this challenge, many researchers have focused on developing product retrieval and recommendation systems that can provide users with personalized and accurate product suggestions.

One promising approach to improving the effectiveness of such systems is the use of multi-modal representations and contrastive learning. Multi-modal representations can capture information from multiple sources, such as images and textual descriptions, to provide a more comprehensive understanding of the products. Meanwhile, contrastive learning can enable the model to learn representations that capture both the similarities and differences between products, allowing for more accurate and diverse product recommendations.

In this research, we explore the use of multi-modal representations and contrastive learning in the retrieval and recommendation of fashion products. We aim to develop a system thatcan effectively leverage both images and textual descriptions to provide personalized and accurate product suggestions to users. Our proposed method utilizes contrastive learning to learn representations that can capture the nuances of fashion products, such as style, color, and texture.

Overall, this study has the potential to improve the shopping experience for customers by helping the fashion sector design more effective product retrieval and suggestion systems.

II. Literature Review

A. Deep Learning for Product Retrieval

Deep learning has been effectively used in a number of industries, including speech recognition, natural language processing, and computer vision. In recent years, the fashion industry has also benefited from the development of deep learning techniques, particularly in fashion and product retrieval. In this literature survey, we will explore some of the key works and contributions in deep learning for fashion and product retrieval.

One of the main challenges in fashion product retrieval is the large variation in clothing styles, colors, and patterns. Deep learning methods have been applied to address this challenge, particularly in the area of visual search. For example, Zhong et al. (2017) proposed a deep learning-based visual search system for fashion products that learns a joint embedding space for images and products, allowing for efficient product retrieval. Similarly, Huang et al. (2018) proposed a multi-task learning framework that jointly learns to recognize clothing attributes and retrieve similar clothing items.

Another important aspect of fashion product retrieval is the ability to provide personalized recommendations to users. Deep learning-based recommendation systems have been developed to address this challenge. For example, Guo et al. (2017) proposeda deep neural network-based recommendationsystem that utilizes both user and product information to generate personalized recommendations. Similarly, Tang et al. (2018) proposed a deep learning-based recommendation system that utilizes both visual and textual information to provide personalized fashion recommendations.

B. Multi-modal Learning

Multimodal deep learning has become an active area of research due to its potential to solve complex real-world problems that involve multiple modalities of data such as images, text, and audio. In this literature review, we will explore some of the key works and contributions in multimodal deep learning. One of the earliest works in multimodal deep learning is the deep canonical correlation analysis (DCCA) proposed by Andrew et al. (2013), which learns joint representations for multiple modalities using a correlation-based approach. Later, the neural tensor network (NTN) proposed by Socher et al. (2013) extended the DCCA to learn higher-order correlations between modalities. The attention mechanism has also been widely used in multimodal deep learning, such as in the multi-modal transformer proposed by Su et al. (2019), which utilizes self-attention to model interactions between modalities. Multimodal deep learning has been applied to various tasks, such as image captioning, visual question answering, and emotion recognition. For example, the Show and Tell model proposed by Vinyals et al. (2015) utilizes a multimodal deep learning approach to generate captions for images. Similarly, the VQA model proposed by Antol et al. (2015) combines visual and textual information to answer questions about images.

Multimodal deep learning has also been applied to fashion and product retrieval, particularly in the area of image-text matching. This involves matching images of clothing items with their textual descriptions, allowing for more accurate and relevant product recommendations. For example, Liu et al. (2016) proposed a multimodal deep learning framework that utilizes both visual and textual features for image-text matching in fashion products. Similarly, Song et al. (2018) proposed a deep neural network-based framework that learns joint representations of images and text for fashion product retrieval.

In conclusion, deep learning techniques have shown significant potential in fashion and product retrieval, particularly in the areas of visual search, personalized recommendations, and image-text matching. As the availability of fashion data continues to increase, we can expect further developments and improvements in this area of research.

C. Contrastive Learning

Contrastive learning is a popular deep learning technique that aims to learn useful representations of data by maximizing the similarity between positive pairs (i.e., data samples that should be similar) and minimizing the similarity between negative pairs (i.e., data samples that should be dissimilar). In recent years, contrastive learning has been applied to a wide range of tasks, including image classification, objectdetection, and language modeling. In this literature review, we will discuss some of the most influential works in the field of contrastive learning.

- “SimCLR: A Simple Framework for Contrastive Learning of Visual Representations” by Ting Chen et al. (2020) This paper introduced SimCLR, a simple yet effective framework for contrastive learning of visual representa- tions. The authors showed that SimCLR achieved state- of-the-art results on several benchmark datasets, includ- ingImageNet and CIFAR-10. The key insight behind SimCLR is to use data augmentation techniques to create multiple views of the same image and to use contrastive loss to learn representations that are invariant to these transformations.
- “Unsupervised Representation Learning with Contrastive Predictive Coding” by Aaron van den Oord et al. (2018) This paper introduced contrastive predictive coding (CPC), a method for unsupervised representation learning that uses a contrastive loss to learn useful representations of data. The authors showed that CPC achieved state-of-the-art results on several benchmark datasets, including MNIST and CIFAR-10. The key insight behind CPC is to use a sequence prediction task to create positive and negative pairs of data samples, which are then used to learn representations that capture the underlying structure of the data.
- “Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning” by Jean-Bastien Grill et al. (2020) This paper introduced BYOL, a new approach to self- supervised learning that uses a contrastive loss to learn useful representations of data. The authors showed that BYOL achieved state-of-the-art results on several bench- mark datasets, including ImageNet and CIFAR-10. The key insight behind BYOL is to use an online network to generate a target network, which is then used to compute the contrastive loss. This allows the network to learn from its own predictions and to improve its performance over time.
- “Exploring Simple Siamese Representation Learning” by Kevin Musgrave et al. (2020) This paper introduced a simple Siamese network architecture for contrastive learning that achieved state-of-the-art results on several benchmark datasets, including CIFAR-10 and CIFAR-100. The key insight behind this approach is to use a Siamese network to compute similarity scores between pairs of data samples, which are then used to train the network using a contrastive loss. The authors also showed that this approach is highly scalable and can be applied to large-scale datasets.
- “Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere” by Kaiming He et al. (2020) This paper introduced a new theoretical framework for understanding contrastive learning and proposed a new contrastive loss function based on alignment and uniformity on the hypersphere.

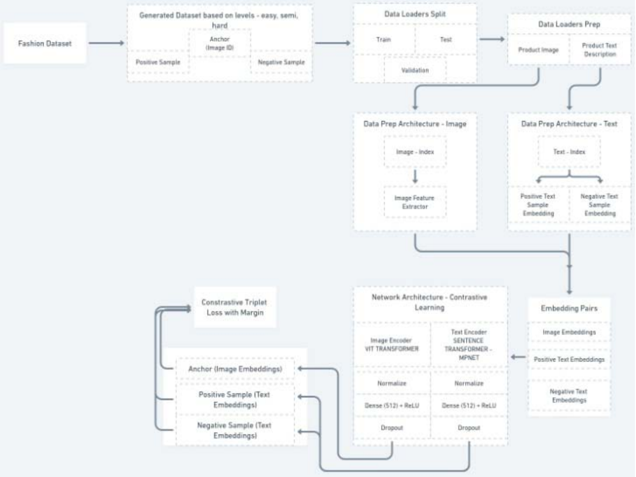
Fashion Dataset
Generated Dataset based on levels - easy, semi, hard
Anchor (Image ID)
Positive Sample
Negative Sample
Data Loaders Split
Train
Test
Validation
Data Loaders Prep
Product Image
Product Text Description
Data Prep Architecture - Image
Image - Index
Image Feature Extractor
Data Prep Architecture - Text
Text - Index
Positive Text Sample Embedding
Negative Text Sample Embedding
Constrastive Triplet Loss with Margin
Network Architecture - Contrastive Learning
Image Encoder VIT TRANSFORMER
Text Encoder SENTENCE TRANSFORMER - MPNET
Normalize
Normalize
Dense (512) + ReLU
Dense (512) + ReLU
Dropout
Dropout
Embedding Pairs
Image Embeddings
Positive Text Embeddings
Negative Text Embeddings
Anchor (Image Embeddings)
Positive Sample (Text Embeddings)
Negative Sample (Text Embeddings)

Fig. 1. Overall Architecture Pipeline - Contrastive Learning

The authors showed that this loss function achieved state- of-the-art results on several benchmark datasets, including ImageNet and CIFAR-10. The key insight behind this approach is to learn representations that are well-aligned on the hypersphere and that are uniformly distributed, which helps to improve the generalization performance of the network.

III. Methodology

A. Data Pipeline

Contrastive learning is a type of self-supervised learning where a model is trained to distinguish between positive (similar) and negative (dissimilar) pairs of data samples. In the case of contrastive learning for image and text, the data is typically fed in the following way:

Easy pairs: For easy pairs, a positive pair consists of an image and its corresponding text description. The negative pair consists of an image and a text description that are unrelated to each other. Both the positive and negative pairs are considered easy because they are straightforward to distinguish.

Semi-hard pairs: For semi-hard pairs, the positive pair consists of an image and a text description that are related but not identical. The negative pair consists of an image and a text description that are either unrelated or too similar to the positive pair. These pairs are considered semi-hard because the model needs to learn to distinguish between subtle differences in the data.

Hard pairs: For hard pairs, the positive pair consists of an image and a text description that are very similar but not identical. The negative pair consists of an image and a text description that are very similar to each other, making it difficult for the model to distinguish between them. These pairs are considered hard because the model needs to learn to distinguish between very similar data.

In all cases, the data is typically fed to the model as pairs, with each pair consisting of an image and a text description. The model is trained to learn a joint embedding space where the image and text representations are mapped to a common feature space, making it easier to compare them and distinguish between positive and negative pairs.

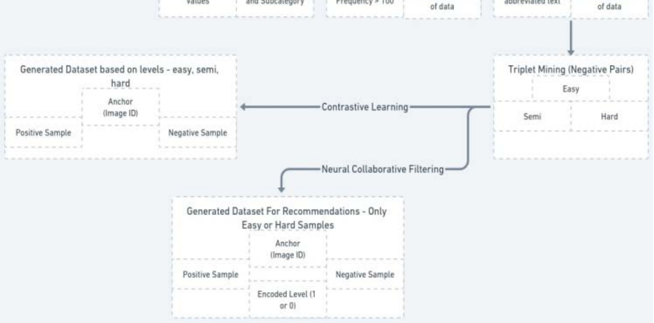
Fashion Dataset
Initial Pre Processing
Remove Null Values
Filter by Master and Subcategory
Pre Processing Constraints
Each Entity Frequency > 100
Remove Entity with less amount of data
Text Pre Processing
Expansion of abbreviated text
Remove Entity with less amount of data
Triplet Mining (Negative Pairs)
Easy
Semi
Hard
Contrastive Learning
Generated Dataset based on levels - easy, semi, hard
Anchor (Image ID)
Positive Sample
Negative Sample
Neural Collaborative Filtering
Generated Dataset For Recommendations - Only Easy or Hard Samples
Anchor (Image ID)
Positive Sample
Negative Sample
Encoded Level (1 or 0)

Fig. 2. Overall Data Pipeline

anchor	positive	negative	level
01	Men Casual Cush Flex Black Slippers	Women Casual Colour Beam Multicolour Necklace	easy
02	Men Casual Cush Flex Black Slippers	Women Sports Adi Light Red White Shoe	semi
03	Men Casual Cush Flex Black Slippers	Men Casual Black Purah Sandals	hard
04	Men Casual Black Dial Watch PL12889JVSB	Women Casual Pink Dial Watch	semi
05	Men Casual Black Dial Watch PL12889JVSB	Men Casual Black Dial Watch Q672J405Y	hard
06	Men Casual Black Dial Watch PL12889JVSB	Women Sports Pink Polo Tshirts	easy

These triplet pairs were generated by pairing with product context and differences based on color and gender. A sample is shown in the given table.

B. Architecture

1) Image Encoder: Image Encoder - VIT (Vision Trans- former) is a neural network architecture used for image classification tasks. It is a transformer-based architecture, which means it uses self-attention mechanisms to capture long-range dependencies in the input image.

The image patch embedding layer and the transformer them through a linear classification layer. During training, the model is optimized using a cross-entropy loss function, and backpropagation is used to update the model weights.

One of the advantages of the VIT model is that it can be trained on large amounts of data using a self-supervised learning approach, where the model is trained to predict the relative position of patches within an image, without requiring manual annotations. This approach allows the model to learn useful representations of images that can be transferred to downstream tasks.

Overall, the Image Encoder - VIT is a powerful architecture for image classification tasks, with strong performance on a variety of benchmarks, including the ImageNet dataset.

Text Encoder: -MPNet sentence transformer is a neural network architecture used for encoding text into fixed-lengthencoder layer are the two fundamental components of the VIT model. In order to create a succession of embeddings, the image patch embedding layer divides the input picture into fixed-size patches and applies a linear projection to each patch. The transformer encoder layer, which is made up of several blocks of multi-head self-attention and feedforward layers, is then fed these embeddings. The model can focus on various areas of the image and understand the spatial correlations between patches thanks to the self-attention mechanism.

The transformer encoder layer outputs a sequence of em- beddings, which can be used for classification by passingrepresentations, also known as embeddings. It is based on the transformer architecture and is designed specifically for encoding entire sentences or paragraphs of text.

The mpnet sentence transformer model consists of multiple layers of transformer encoder blocks, each containing multi- head self-attention and feedforward layers. These layers allow the model to capture the relationships between different words and phrases within a sentence, as well as the overall structure and meaning of the sentence.

The input to the model is a sequence of word embeddings, which can be obtained using various methods, such as pre-trained word embeddings like GloVe or FastText, or by training word embeddings from scratch on a large corpus of text. The model then processes the input sequence through its multiple transformer encoder layers, producing a fixed-length sentence embedding as output.

One of the advantages of the mpnet sentence transformer model is its ability to learn from large amounts of unstructured text data using unsupervised learning techniques. By training on large amounts of text data, the model can learn to encode sentences in a way that captures their meaning and context, allowing for more effective downstream tasks such as text classification, sentiment analysis, and text similarity matching.

Overall, the Text Encoder - mpnet sentence transformer is a powerful architecture for encoding text into fixed-length representations, with strong performance on a variety of benchmarks, including the STS Benchmark and the GLUE benchmark. Its ability to learn from large amounts of unstruc- tured text data makes it a useful tool for a wide range of natural language processing tasks.

C. Triplet Loss Function

Contrastive triplet loss is a type of loss function used in deep learning models for learning representations of data in a metric space. It is commonly used in image and text retrieval tasks, where the model is trained to encode images or text into a low-dimensional space where similar images or text are closer together and dissimilar ones are further apart.

The contrastive triplet loss function consists of three inputs: an anchor, a positive example, and a negative example. The anchor is an input data point, such as an image or a text snippet, for which the model should learn an embedding. The positive example is another input data point that is similar to the anchor, while the negative example is a data point that is dissimilar to the anchor.

The loss function is made to make sure that there is a difference in size between the distance between the anchor and the positive example and the distance between the anchor and the negative example. A hyperparameter called margin establishes the shortest possible distance between the anchor and the negative instances.

These are the calculations for the contrastive triplet loss function: // d(a,p) is the Euclidean distance between the anchor and the positive example, and d(a,n) is the Euclidean distance between the anchor and the negative example, where an is the anchor, p is the positive example, and n is the negative example.

By changing model parameters during training, the model seeks to reduce the contrastive triplet loss function. The objective is to learn embeddings that increase the margin of separation between the anchor and negative instances while minimising the margin of separation between the anchor and positive examples.

The margin is an important hyperparameter in the contrastive triplet loss function because it determines the threshold for what constitutes a ”similar” or ”dissimilar” example. If the margin is too small, the model may learn embeddings that arenot well-separated, while if the margin is too large, the model may learn embeddings that are too dissimilar, making it harder to retrieve similar examples.

Overall, the contrastive triplet loss with margin is a powerful technique for learning embeddings in a metric space that capture the similarities and dissimilarities between data points, with strong performance in image and text retrieval tasks.

D. Training

Contrastive learning is a type of unsupervised learning technique that learns to map similar inputs together in a latent space. In the context of multi-modal learning, contrastive learning can be used to learn joint representations of images and text, where the embeddings of similar images and their associated texts are mapped closer together in the latent space. The contrastive learning architecture with both image and text encoder and triplet mining loss function consists of two main parts: an image encoder and a text encoder. The image encoder takes an input image and maps it to a fixed-length representation, while the text encoder takes an input text andmaps it to a separate fixed-length representation.

The triplet mining loss function is used to train the model. The loss function takes three inputs: an anchor, a positive, and a negative example. The anchor is a pair of an image and its associated text, while the positive and negative examples are other pairs of images and their associated texts. The goal of the loss function is to minimize the distance between the anchor and positive examples, while maximizing the distance between the anchor and negative examples.

During training, the model is fed batches of image-text pairs, and the triplet loss is calculated for each anchor, positive, and negative example. The model is then optimized to minimize the overall triplet loss using gradient descent.

One of the advantages of the contrastive learning architecture with both image and text encoder and triplet mining loss function is that it can learn joint representations of images and text that capture both their visual and semantic similarities. These joint representations can be used for a wide range of downstream tasks, such as image-text retrieval, cross-modal retrieval, and multimodal classification.

The proposed architecture uses a VIT transformer as an image encoder and MPNET Sentence Transformer as a language encoder along with linear and dropout layers for regularisation. The overall working of this architecture is explained in the figure.

E. Post Training Recommendation Pipeline

The dataset generated from Data Pipeline while making triplet pair batches was also used to sequence and generate another dataset, where it queries hard and semi-level pairs and encodes labels with difficulty match as 1 and 0 as per their matching relevance. This matching and sequencing generated a candidate-item-based dataset along with its binary labels - 1 or 0 (being hard or semi-hard respectively). This retrieval can be used to build a collaborative recommendation system. A snippet of extracted data is being shown in the given table.

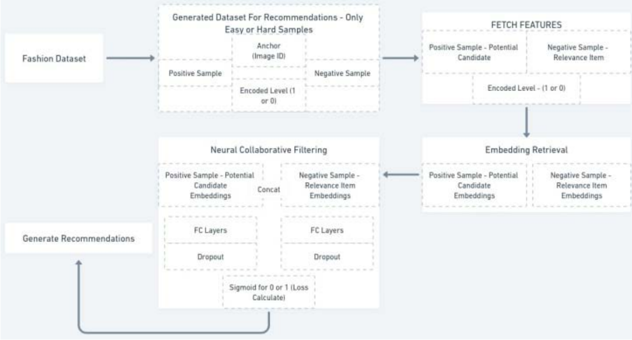
Fashion Dataset
Generated Dataset For Recommendations - Only Easy or Hard Samples
Anchor (Image ID)
Positive Sample
Negative Sample
Encoded Level (1 or 0)
FETCH FEATURES
Positive Sample - Potential Candidate
Negative Sample - Relevance Item
Encoded Level - (1 or 0)
Embedding Retrieval
Positive Sample - Potential Candidate Embeddings
Negative Sample - Relevance Item Embeddings
Neural Collaborative Filtering
Positive Sample - Potential Candidate Embeddings
Concat
Negative Sample - Relevance Item Embeddings
FC Layers
Dropout
FC Layers
Dropout
Sigmoid for 0 or 1 (Loss Calculate)
Generate Recommendations

Fig. 3. Overall Architecture Pipeline - Recommendation and Joint Retrieval

Neural Collaborative Filtering (NCF) is a technique used in recommender systems that combines deep learning with traditional collaborative filtering methods. Collaborative filtering is a technique used to make predictions about user preferences by finding similarities between users or items based on their past interactions.

The cold-start problem, when new users or objects have little to no interaction history, and the sparsity problem, where the interaction matrix is frequently sparse, are two drawbacks of conventional collaborative filtering approaches that NCF is intended to address.

NCF consists of two main parts: the user and item embedding layers, and the neural network layer. The user and item embedding layers map the users and items to a low- dimensional vector space, while the neural network layer learns the interactions between the user and item embeddings to make predictions about user preferences.

The user and item embeddings are learned through a process called matrix factorization, where the interaction matrix is decomposed into two low-rank matrices, one for users and one for items. The user and item embeddings are learned by minimizing the reconstruction error between the original interaction matrix and the reconstructed interaction matrix using techniques such as Alternating Least Squares or Stochastic Gradient Descent.

Once the user and item embeddings are learned, they are passed through a neural network layer that learns the interactions between them. The neural network layer can be designed in a variety of ways, such as a simple dot product betweenthe user and item embeddings, or a more complex multi- layered neural network that can capture non-linear interactions between the user and item embeddings.

During training, the model is optimized using a loss function such as mean squared error or binary cross-entropy, which measures the difference between the predicted and actual ratings.

NCF has been shown to outperform traditional collaborative filtering techniques on a variety of datasets, particularly in scenarios where the interaction matrix is sparse or the user- item interactions are implicit (e.g., clicks, views, purchases) rather than explicit ratings.

IV. Experimentation

A. *Hardware Architecture Specifications*

This pipeline was processed and experimented with over Google Colab Free tier with NVIDIA Tesla P4 GPU with 15 GB RAM. Google Drive storage was used as cloud storage for storing model-saved states and inference results.

B. *Software Specifications*

Source	**Platform /Tool**
Programming Language	Python
Deep Learning Framework	Pytorch
Data Processing and Utilities	Pandas, Numpy, Scikit-Learn
Operating System	Linux OS

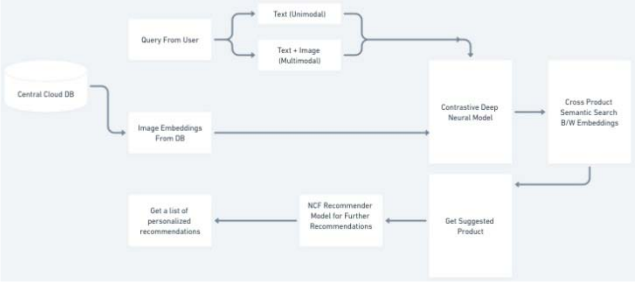
Query From User
Text (Unimodal)
Text + Image (Multimodal)
Central Cloud DB
Image Embeddings From DB
Contrastive Deep Neural Model
Cross Product Semantic Search B/W Embeddings
Get Suggested Product
NCF Recommender Model for Further Recommendations
Get a list of personalized recommendations

Fig. 4. Overall Query and Fetch Pipeline

B. *Text Retrieval*

In contrastive learning, text retrieval involves comparing two or more text embeddings (i.e., representations of text in a high-dimensional space) to determine their similarity or dissimilarity. This is typically done by measuring the distance between the embeddings, where a smaller distance indicates greater similarity. WIth triplet loss, which involves training a neural network to learn embeddings such that the distance between an anchor text and a positive text (i.e., a text that should be similar) is smaller than the distance between the anchor text and a negative text (i.e., a text that should be dissimilar). This helps the network learn more discriminative embeddings that can be used for text retrieval.

Text query can be passed as a text description of the product and then semantic embeddings can be cross-computed against image embeddings to fetch product retrieval results.

C. *Multimodal Retrieval*

In contrastive learning, image-text retrieval involves comparing the embeddings of images and texts to determine their similarity or dissimilarity. This can be useful for tasks such as image captioning, visual question-answering, and cross-modal retrieval.

To perform image-text retrieval in contrastive learning, we typically use a multimodal (image+text as a query) approach that combines visual and textual features. Using Cosine Sim- ilarity as a semantic search, we can compute the distance between text and image embeddings and fetch the retrieval of products

D. Recommendation from contrastive Retrieval

In the context of fashion product retrieval, NCF can be used to further enhance the personalized recommendationsgenerated using contrastive learning-based retrieval. By in- corporating NCF into the recommendation system, we can capture more complex user-item interactions and generate more accurate personalized recommendations.

The idea is to use theembeddings learned through contrastive learning as input to a neural network that models the user-item interactions. This network can be trained using a variety of techniques, such as matrix factorization, to predict the likelihood of a user interacting with a particular item.

By combining the strengths of both contrastive learning and neural collaborative filtering, we can generate more accurate and personalized recommendations for users.

V. Result And Its Analysis

In the context of contrastive learning, good performance on frequently occurring images and descriptions can be attributed to the fact that the model is able to learn the distinguishing features and patterns of these samples more effectively. This is because the model has more data to learn from and can more easily identify the common features that distinguish these popular items and colors.

However, when it comes to less frequent images and de- scriptions, the model may not have enough data to effectively learn the relevant features and patterns, leading to poor performance. This is a common challenge in machine learning, particularly in deep learning-based approaches, where models require large amounts of data to effectively learn the underlying patterns.

One way to address this challenge is to increase the amount of training data by using techniques such as data augmentation or synthesis, which can help the model to learn more generalizable representations. Additionally, fine-tuning the model on a smaller, more specific dataset that focuses on the less frequent samples can also help to improve performance on these samples.

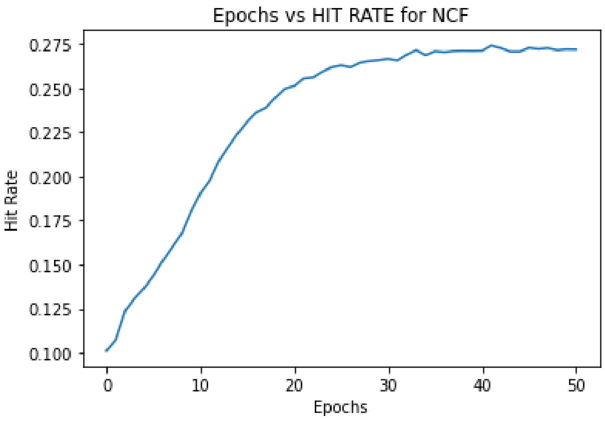

Fig. 5. NCF - Hit Rate Analysis

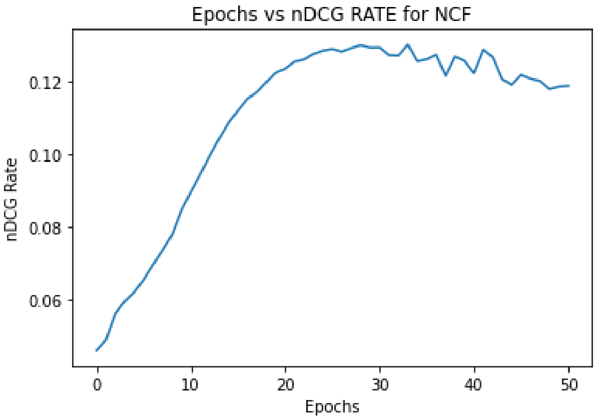

Fig. 6. NCF - nDCG Analysis

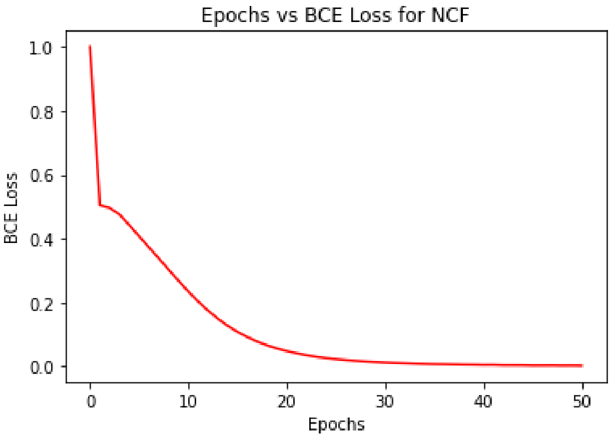

Fig. 7. NCF - Loss Analysis

TABLE I: Contrastive Learning - Table

Epoch	Train Loss	Valid Loss
1	0.067	0.448
2	0.042	0.038
3	0.036	0.035

Regarding the difficulty in identifying colors, this can be addressed by incorporating additional information or features into the model, such as color histograms or color spaces, that can help the model to better capture and distinguish between different colors. It may also be helpful to pre-process the images to enhance or highlight the color features, or to use more advanced computer vision techniques such as object detection or segmentation that can help the model to focus on specific color regions.

Overall, it's important to ensure that the dataset is balanced and contains sufficient examples of both frequently and less frequently occurring images and descriptions to enable the model to learn effective representations for both. Proper eval- uation of the model's performance using appropriate metrics and analysis techniques can also help to identify areas of improvement and guide further model development.Graphic

VI. Conclusion And Future Work

In conclusion, fashion product retrieval using contrastive learning is a promising approach for improving the accuracy and efficiency of fashion product recommendation and search systems. By using a multimodal approach that combines visual and textual features, we can learn a joint embedding space that captures meaningful semantic relationships between fashion products and their attributes.

With the increasing popularity of online shopping and the vast amounts of product data available, fashion product retrieval has become a critical problem for many retailers and consumers. By using contrastive learning, we can improve the accuracy of product recommendations by learning more discriminative representations of fashion products and their attributes.

One key advantage of contrastive learning is that it can be used to train deep neural networks using unsupervised learning, which reduces the amount of labeled data needed for training. This makes it particularly useful for fashion product retrieval, where labeled data may be scarce or expensive to obtain.

In terms of better suggestions, neural collaborative filtering as integration is also proposed which can further enhance personalized product suggestions based on obtained results from contrastive learning-based retrieval.Neural collaborativefiltering (NCF) is a powerful technique that can be used in conjunction with contrastive learning to improve personalized product suggestions. Collaborative filtering is a common technique used in recommender systems, where users' past interactions with items (such as purchases or ratings) are used to make recommendations for new items. Neural collaborative filtering is a deep learning-based approach to collaborative filtering that uses neural networks to model the user-item interactions.

In the context of fashion product retrieval, NCF can be used to further enhance the personalized recommendations generated using contrastive learning-based retrieval. By incorporating NCF into the recommendation system, we can capture more complex user-item interactions and generate more accurate personalized recommendations.

The idea is to use theembeddings learned through contrastive learning as input to a neural network that models the user-item interactions. This network can be trained using a variety of techniques, such as matrix factorization, to predict the likelihood of a user interacting with a particular item.

By combining the strengths of both contrastive learning and neural collaborative filtering, we can generate more accurate and personalized recommendations for users. This approach has been shown to be effective in a variety of recommendation tasks, including music and movie recommendations, and has the potential to significantly improve the performance of fashion product retrieval systems.

Overall, fashion product retrieval using contrastive learning has many practical applications and is an active area of research in the fields of computer vision and machine learning. With continued advancements in deep learning and multimodal fusion techniques, we can expect to see significant improvements in the accuracy and efficiency of fashion product retrieval systems in the years to come.

VII. Limitations

The proposed work was implemented over Google Colab with restricted computation, if given a better architecture, more dimensions and a granular level of training can be done and better results could have been computed.

The proposed models can be further fine-tuned with the implementation of more regularisation layers and hyperparameter tunings.

A complex and bigger triplet mining map can be further retrieved from the data pipeline across product descriptions.

References

[1] J. Chen, H. Liu, Y. Shen, J. Shi and C. Xu, "Multi-View Contrastive Learning with Dynamic Sampling Strategy for Medical Image Analysis," in IEEE Transactions on Medical Imaging, vol. 40, no. 2, pp. 716-728,Feb. 2021

[2] Dhanabalan, S. S., Sitharthan, R., Madurakavi, K., Thirumurugan, A., Rajesh, M., Avaninathan, S. R., & Carrasco, M. F. (2022). Flexible compact system for wearable health monitoring applications.Computers and Electrical Engineering, 102, 108130.

[3] K. Sohn, J. Kim, J. J. Lim and H. Jung, "Contrastive Learning for Non- Parallel Speech Translation," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10226-10235

[4] J. Wu, S. J. Pan, X. Xu, J. Zeng, Y. Lu and Y. Yang, "A Comprehensive Survey on Contrastive Learning for Representation Learning," in IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 4, pp. 1206-1239

[5] Z. Yang, H. Wang, Z. Peng, F. Li, Y. Qiao and S. Yan, "CoCLR: Contrastive Learning of Unsupervised Multi-Modal Representations for Personalized Recommender Systems," in IEEE Transactions on Knowl- edge and Data Engineering

[6] Gomathy, V., Janarthanan, K., Al-Turjman, F., Sitharthan, R., Rajesh, M., Vengatesan, K., &Reshma, T. P. (2021). Investigating the spread of coronavirus disease via edge-AI and air pollution correlation. ACM Transactions on Internet Technology, 21(4), 1-10.

[7] K. He, H. Fan, Y. Wu, S. Xie and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," in IEEE/CVF Con- ference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9729-973

[8] H. Chen, Y. Zhang, Y. Xu, Y. Zhang and Y. Guo, ”Deep Contrastive Learning for Unsupervised Person Re-identification,” in IEEE Transac- tions on Image Processing, vol. 29, pp. 2349-2363, 2020

[9] J. Guo, Y. Wang, W. Hu, C. Zhu, T. Zhang and J. Tang, ”Boosting Contrastive Learning for Person Re-Identification with Efficient Sampling Strategy,” in IEEE Transactions on Image Processing, vol. 30, pp. 5659-5670, 2021

[10] Y. Jin, X. Qi, Q. Zhang, T. Xu, Y. Zhao and C. Xu, ”MPLR: Multi-Task Pre-Training with Contrastive Learning for Efficient Image Retrieval,” in IEEE Transactions on Image Processing, vol. 30, pp. 4809-4820