National College of Ireland

Text and Image Based Multi-model Fashion Image Retrieval system

MSc Research Project

Aafaq Iqbal Khan
Student ID: x20108851

School of Computing
National College of Ireland

Supervisor: Cristina Hava Muntean

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:……. ……Aafaq Iqbal Khan……

Student ID: ……x20108851……..……

Programme: ……MSc Data Analytics…… **Year:** ……2021……..

Module: ……MSc Thesis Project…….……..……

Supervisor: …… Cristina Hava Muntean…….……

Submission Due Date: ……31/01/2022…….……

Project Title: Text and Image Based Multi-model Fashion Image Retrieval system

Word Count: …… **Page Count**……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Aafaq Iqbal Khan……

Date: ……16/12/2021……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Text and Image Based Multi-model Fashion Image Retrieval system

Aafaq Iqbal Khan
X2010851

Abstract

Interactive image retrieval is an emerging research topic playing a significant role in the success of a wide variety of applications, especially in the fashion domain. However, as fashion product catalogues have grown in size and the number of features each product has increased, it has become more challenging for users to express their needs effectively. In traditional fashion e-stores, users may not be able to specify the details of the outfits more accurately by utilizing a text-based query. Therefore, we focus on a multi-model image retrieval system to integrate the query image with a text query that describes the visual differences between the query image and the search target image. To tackle this task, we investigate a similarity metric between a target image and a candidate image (query image) plus a text query. Both target images and query image are encoded with Efficient-Net and ResNet-50 (only one at a time) into feature vector representation, and encode caption text to a text feature vector using LSTM. Then we compose the query image vector and text feature vector into a single vector which is expected to be biased toward the target image vector with the help of state-of-the-art TIRG Vo et al. (2019). The compositional query-based TIRG achieved a higher average recall with 29.2, than other methods, text-only (21.93), image only Efficient-Net (8.74), and image only Resnet-50 (8.75). The TIRG outperforms text-only, image-only Efficient-Net, and image-only Resnet-50 methods in terms of batch-based classification training loss with values 0.192, 0.42 (65% more), 0.91 (79% more), and 0.52 (63% more) respectively.

1 Introduction

Fashion is a multibillion-dollar business having strong cultural and financial impacts for any society. Due to the digital revolution, the fashion industry has adopted a new paradigm. Especially with the recent pandemic, more people are shifting towards online shopping than ever, as in USA 32.4% more online retail sales are recorded in 2020[1]. Therefore, it is very critical for fashion e-commerce businesses to retrieve the relevant products for customers according to their requirements. It helps to increase the probability of selling of product and improve the sales of the company. Currently, almost all fashion e-commerce platforms employ a recommendation system to help their online visitors discover what they're looking for. The text-based image retrieval system based on natural language processing has been used as the fundamental search engine or recommender system by e-

[1] https://www.digitalcommerce360.com/article/coronavirus-impact-online-retail/

commerce platforms. However, the recommendation of similar items just based on textual characteristics such as name, category, description, might not be an efficient method. Zhang et al. (2019). Deep learning advancement has transformed retrieval systems and actually provided new possibilities for increasing the efficiency of retrieval systems. But there is a limitation exit that is the traditional information retrieval systems only permit a unimodal query that can be based on text or image. By permitting a multi-modal query i.e. including both text and image, smart image retrieval systems should empower users to express properly the product in their minds.

The major constraint in any fashion product retrieval system is that users have some "idea" in their mind and they want to find the product or images exactly related to the idea. However, they must communicate that "idea" to the retrieval system. There are multiple ways to convey the "idea" to the retrieval system by giving query image or query text or combination of both. In this study, we investigate enhancing our search by incorporating the query image and query text features as proposed in Wu et al. (2021). For instance, a person has a query image of a dress that will be purchased on an e-commerce website but the user wants to buy a similar dress with the white color as shown in figure. 1. In this scenario, the image retrieval system should be capable to formulate the query by combining the query image with a text query that defines an intended image modification. Due to the advancement of deep learning methods, image retrieval systems can have capabilities to integrate the image and text features. It means, the user can refine the retrieval output by providing a text query, which explains the difference between the reference image and the target image.

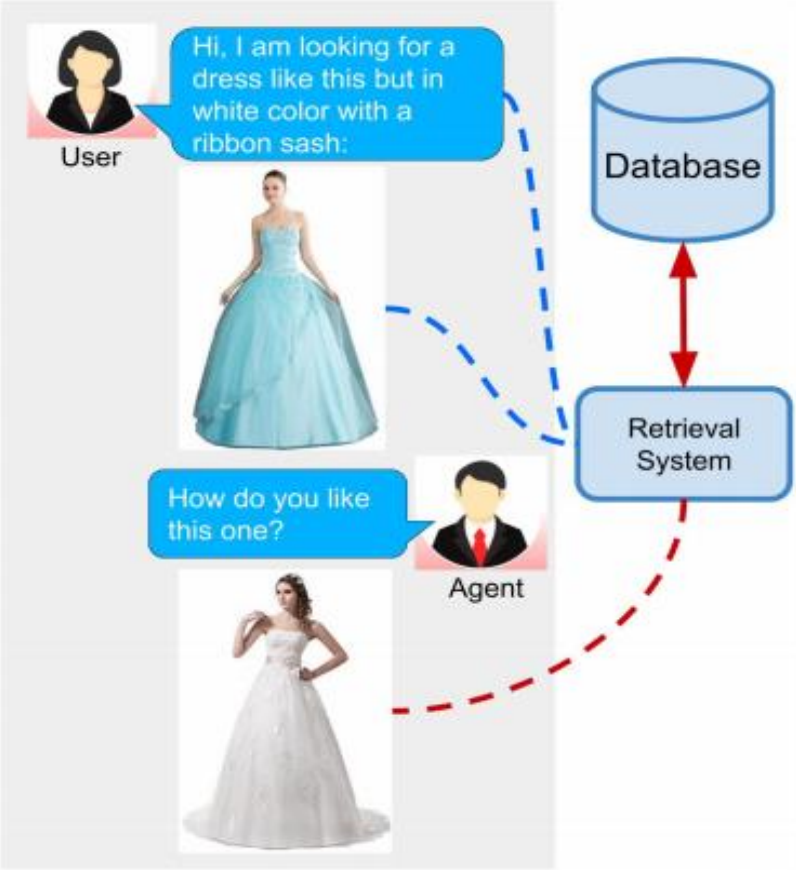

Figure 1. A potential application of this concept. Anwaar et al. (2020)

The objective of this study revolves around the research questions **RQ:** *To what extent, integrating both text query and image query help to improve the fashion image retrieval system by investigating a multi-model approach TIRG (Text Image Residual Gating*)?
The major objectives of the studies are:

- Compare the performance (recall as major evaluation metric) of the Only-Text based, Only-Image based, and both text and image multi-model based image retrieval system.
- Based on recall, training loss, and training time performance, evaluate the image retrieval system with different image encoders ResNet-50 and Efficient-Net.

Existing image retrieval systems are commonly optimized for the text-to-image or image-to-image retrieval task, which has its limitations due to the nature of fashion images that can be described more accurately by a combination of text and visual features Zhang et al. (2019). To eliminate the gap between the textual and imagery modalities, Faghri et al. (2017) proposed the state-of-the-art visual semantic embeddings concept by which it is possible to compare semantically text and image features in a shared common space. As mentioned earlier, we use TIRG for the composition layer, it employs a convolutional neural network to acquire image embedding, a recurrent neural network to generate text embedding, and a gated connection to fuse both embeddings.

The dataset used for analysis is the benchmark Fashion-IQ[2] dataset that is first time used by Wu et al. (2021). The dataset has 77,683 images that cover three product categories Dresses, Tops&Tees, and Shirts. Moreover, each training item is represented by a triplet (reference image, caption, target image), captions are annotated by human annotators. Training, validation, and test splits proposed by Wu et al. (2021) are used. Each split is about evenly distributed throughout the three categories and every experiment is repeated for each individual category.

Each training triplet (reference image, caption, target image) passes through three major parts: the image encoder, the text encoder, and the composition layers. Both reference and target images are encoded by image encoder using pre-trained Efficient-Net Tan & Le. (2019) and Resnet-50 one at a time, both output the same embedding dimensions 1024. Next, for the text encoder, standard LSTM is used with 1024 LSTM hidden dimensions and 0.1 dropout. We set our setting as both image encoder and text encoder output has the same dimension that is 1024. We carried out different combinations of image and text encoders to evaluate the performance of models. Recalls at different levels e.g. 5, 10, and 50 are major evaluation metrics for each two fashion categories (dresses, shirts). The average of Recall is used to measure overall performance at R@5, Recall@10, and Recall@50 on the test dataset. The training time performance and batch-based classification losses are also used as evaluation metrics.

In our study, we hypothesize that the image retrieval results can be refined by integrating the textual and visual queries in common space. Particularly in the fashion industry, where it is challenging to represent the sophisticated styles of fashion using visual

[2] https://github.com/XiaoxiaoGuo/fashion-iq

or keyword-based searches alone. Our detailed set of experiments show how our hypothesis improves retrieval performance on the fashion-IQ dataset.

In subsequent sections of this work, we will discuss the research study in further detail. In section 2, we critically assess prior researchers' work, which helps us enhance our approach to answer our research questions. Later in this paper, sections 3, 4, and 5 will discuss the research methodology, design specification, and implementation respectively. Finally, the paper will conclude with a conclusion followed by a short discussion in section 7.

2 Related Work

Despite the progress of conventional machine learning, new advances in deep neural networks have made substantial progress in the domain of image retrieval systems. Finding the appropriate method for feature extraction of the images has remained one of the primary issues of researchers in order to design an effective image similarity retrieval system. This section will highlight the previous work done in image retrieval systems. The section is divided further into subsections. The Content-Based Image Retrieval System section discusses the uni-model image retrieval system in the fashion domain. The Composition Interactive Retrieval section sheds light on a multi-model image retrieval system.

2.1 Content-Based Image Retrieval System (CBIR) in Fashion domain

Fashion items, such as dresses, have a variety of characteristics, for instance, they can have different attributes like size and stuff, e.g. two shirts that appear to be identical, may have different sizes and colour. Any retrieval system requires a comprehensive understanding of these attributes. The research community has put a lot of effort into developing deep neural network models that can distinguish the features of fashion photos.

In recent studies, Sidharth et al. (2020) performed image retrieval on different images categories with the help of the Convolutional Neural Network. The sequential CNN model was built using layered architecture and the softmax function, which yields output between 0 and 1 for all classes. In their study, 1889 images were used for training while 188 images were used for testing. The accuracy on training was 99.12% and 76.19% for testing. However, no cross-validation has been undertaken to ensure that the model is not over-fitted. Moreover, the amount of data that is used to train the CNN model did not seem sufficient that's why cross-validation was necessary to even check the under-fitting of the model. They prominently highlighted that they did not use any pre-trained CNN model, instead, they implement their own CNN model with three convolutional layers with "RELU" as activation function followed by two hidden layers.

Hsiao & Grauman (2018) used the Resnet-50 as the deep neural network to recognize the cross-domain features for capsule wardrobe problems. To differentiate the features of the dresses catalogue, their approach uses transfer learning with ResNet-50 that is pre-trained on the ImageNet dataset with 1.2 million images spanning 1000 classes. They use a capsule neural network to classify such features into top and bottom apparel sets. Finally, the network is tweaked to classify the outfit based on these attributes. The plus point is that they used the

real-world data that were collected from *polyvore.com* consisting of 7,478 images of outfits with meta-data.

Ak et al. (2018) worked on a novel approach named *FashionSearchNet* that utilizes input image and product attributes to retrieve the fashion products in a system. Their proposed network is similar to the famous *AlexNet*[3] Network, but the difference is that they used only seven convolutional layers and removed all fully connected layers from the network. At the end, their approach is to integrate all of these features into a unified representation that can be deployed to any image retrieval system. They highlight that their results are affected by the presence of the background in the image dataset. Their network is not able to handle the background in images. They can use object detection or other computer vision concepts to extract only ROIs and eliminate the background. So one important thing we understand from this paper is that our dataset should not have a background in fashion product images just like we usually see on e-commerce websites.

Jo et al. (2020) created a sketch-based outfit search method that uses a deep neural network to convert a given apparel sketch to the level of an image. Then, find the similarities of the given image with the rest of the images in the database and retrieve only similar images for users. They used a different concept of Sketch-Product fashion retrieval model based on image2vec feature extraction model to overcome the limitations of a text-based search method. A total of approx. 66k cases containing 33 amazon fashion products from amazon.com were used for the study and the best performing model showed 77.47% of "Precision at 5". For further improvement, they integrated their model with user profiling for personalized results but they did not consider any textual information while modeling.

In the most state-of-the-art work for image classification or image retrieval tasks, convolutional neural networks (CNNs) are more commonly used. However, as per KINLI & KIRAC (2020) Capsule Neural Networks which are an enhancement of CNN showed more promising results than conventional CNNs due to some of their limitations, Such as they are less resilient to affine transformations due to the pooling layers. They come up with a four-layer stacked convolutional Capsule Network and name it "FashionCapsNet". They did not use the transfer learning instead of it they trained their model on the publically available state-of-the-art dataset "Deep Fashion" which consists of 290k images of 46 different categories. It outperformed the existing CNN- based models trained on deep fashion datasets with an accuracy of 89.83%.

2.2 Composition Interactive Retrieval

The goal of Interactive Retrieval is to include user feedback or requirement into an image retrieval system in order to refine the image retrieval results that are customized to the users' expectations. The user feedback or requirement can be conveyed to the system in various ways such as modification text Vo, Jiang & Hays (2019), sketch, and feature attribute Sadeh et al. (2019). A study Yu et al. (2020) suggested that textual query is a natural way for users to convey their fine-grained ideas for interactive retrieval. It is an empirical results-

[3] https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

based study that shows that the probability of retrieving the targeted item is more if we include the textual feature information in the image retrieval system. This supports our hypothesis that is stated in the introduction section.

Verma et al. (2018) proposed a multi-stage deep neural network to learn input image features. In their architecture there are three levels, initially to extract the fashion product image's features, there is a three-layered CNN model with three max-pooling layers. Second, a Recurrent Neural Network (RNN) based module that concentrates only the key information inside the feature vector space learned in the first phase. In the third stage, a texture encoding layer captures the clothing texture of the attentions generated in the second phase. Because the similarity scores are calculated using multiple sections of the input image, this is highly beneficial for measuring visual similarity. They conducted the experiments on three publically available datasets Fashion144K, Fashion550k, and DeepFashion dataset. The limitation of the study is that their proposed model is unimodal that takes only images as its input to find similar images within the database. On the good side, they released their code publically for further improvement and research.

Tan et al. (2019) worked on an attribute-based product retrieval task and they use a challenging dataset "Visual Genome dataset". This dataset has images that contain multiple objects which increases its complexity. Because of it, they used an attribute-based retrieval system that helps to identify the specific object in the image. Every image in the dataset is annotated with multiple captions or descriptions. Dataset with a handsome amount of images 105, 414 and their code are publically available on GitHub. Their approach used Hierarchical Recurrent Encoder (HRE) for effective interactive retrieval of images with multiple objects. The shortcoming in their work is that query texts are restricted to a fixed set of relative attributes that need to be specified in input text which can only be one word only i.e. attributes such as colour, fabric type, etc.

Chen et al. (2020) extract image features at many scales using outputs from several levels of a convolutional network, however, their model combines the text and image characteristics into a single feature vector. They validate their models on two datasets Fashion 200k and Fashion-IQ (the same we are using for this study) and use recall as their evaluation metrics. Their approach performs well on the fashion 200k dataset but comparatively does not good on the Fashion-IQ dataset with Recall@10 values 50.8 and 24.15 respectively. As per our understanding, maybe their approach is more suitable for simple or shorter text caption as the fashion-200k dataset uses single-word text modifications in contrast to the Fashion-IQ dataset which has an average caption length of 10.69.

Parekh et al. (2021) address the challenge of visual feature extraction for very difficult fashion data from Flipkart, a big e-commerce website in India, whose data is highly unbalanced since it is directly obtained from the website. In addition to dealing with data imbalance, they present an end-to-end system that merges multi-task learning with a transformer as an attention module. They used several CNN architectures such as VGG16, VGG19, ResNet, and Efficient-Net. But efficient-net outperforms the other model by giving 5 – 6% more accurate results that's why they used it as their base model as it has less

trainable parameters which make it more efficient, which motivated us to use it in our work as well.

2.3 Summary

After the critical review of the papers, we understood that choice of an appropriate method to combine the image features and textual features could be the important factor of the interactive image retrieval system. We discovered that CNN models, particularly ResNet, are the most often used methods among image-retrieval system researchers. So we select ResNet as one of our CNN models for feature extraction along with Efficient-Net. It also showed great results in image classification that's depicted well in the above-discussed papers but as per our best knowledge it has not been used for multi-model image retrieval systems. Therefore this research paper extends the use of Efficient-Net for image retrieval tasks on our Fashion IQ dataset which has not been implemented yet.

3 Research Methodology

One of the most significant sections of any research proposal is the methodology section, which examines the methodologies and approaches that will be used in the study. We will use Knowledge Discovery in Databases (KDD) for this research project. KDD was chosen for this research study. The justification for selecting KDD over CRISP-DM is because CRISP - DM normally concludes with project deployment because it is meant to work with business applications, however, with KDD, this is not a required step to complete, which is more suitable for our current research. Figure 2 depicts the KDD framework Overview of the KDD Process (1996) as it is updated for this project.

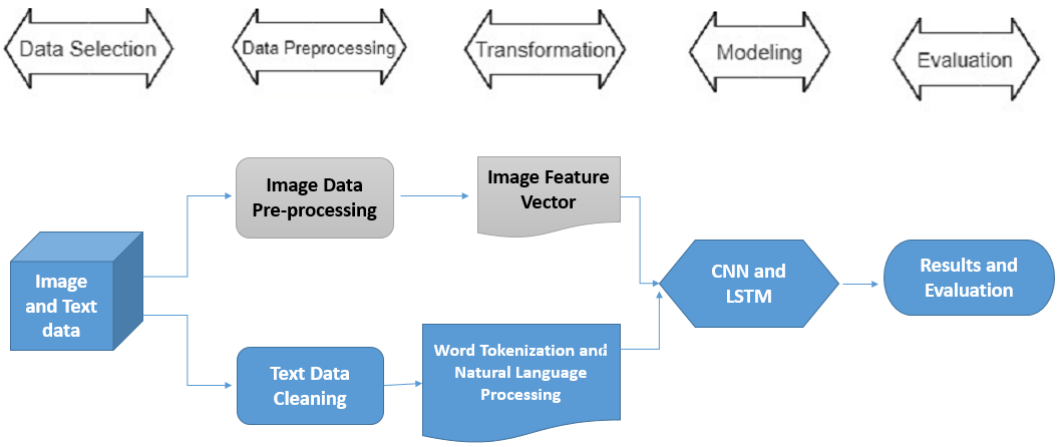

Figure 2: KDD Methodology

In terms of our research concerns, we have both image data and textual data, and the Data Pre-processing, Transformation, and Modeling stages are applied to both. It is critical to begin every project by planning out the flow, design, technologies, and algorithms while keeping various limitations in mind. This section offers an overview of the study's methodology.

3.1 Data Selection

Deep-Fashion and ImageNet were found to be the most commonly used datasets for the implementation and evaluation of the Content-Based Image Retrieval (CBIR) system which takes the only image as input query according to the above-mentioned published papers in the literature review section. We are analyzing an approach to embed the image-based and text-based recommendation system, so we need a comprehensive dataset that contains not only images of the product as well as has the textual information of the images like product name or their descriptions to address the research questions. For specifically these types of problems, Fashion 200k and Fashion 550k like datasets are used by the researchers Chen et al. (2020) and Verna, et al. (2019) but the limitation of these datasets is that these are attributes based datasets and each attribute can be contained only 1-2 words to specify the attribute. Due to this less textual information, we will have the very little vocabulary or corpus for the natural language processing model so these would not be good options.

To address the research question, we need a dataset with the following features.

- Dataset should be available publically and there should not be any ethical concerns associated with the usage of the dataset.
- The dataset should have fashion product images along with textual information in the form of captions.
- There should be ground truths available for validating the results on testing data. So, we will confident about whether the retrieved images are targeted images or relevant to the query or not.

A dataset Fashion-IQ[4] is found to have all these features which were first time used by Wu et al. (2021). The fig. 3 from the same paper clearly illustrates that the Fashion IQ dataset comprehensively covers the above-mentioned shortcomings.

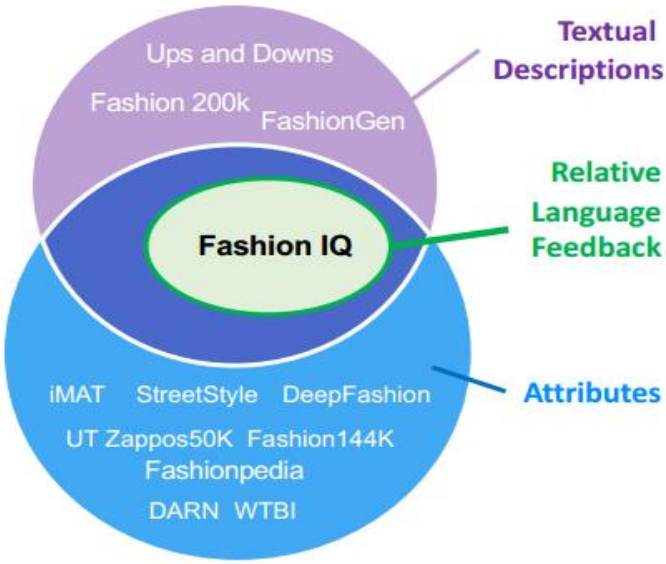

Figure 3. Fashion IQ dataset illustration. Wu et al. (2021)

[4] https://github.com/XiaoxiaoGuo/fashion-iq

The dataset has a total of 77,683 fashion product images that cover three product categories Dresses, Tops&Tees, and Shirts. There are 19,087 (24.5%) images of dress, 31,728 (40.8%) images of shirt, and 26,869 (34.7%) images of top & Tees. Moreover, images are divided into 30,134 pairs of reference/candidate and target images, annotated with relative captions by human annotators, which describe the difference between candidate and target images (e.g., "shorter more fancy"). So, each pair can be represented by a triplet (reference image, caption, target image) as in figure 4.

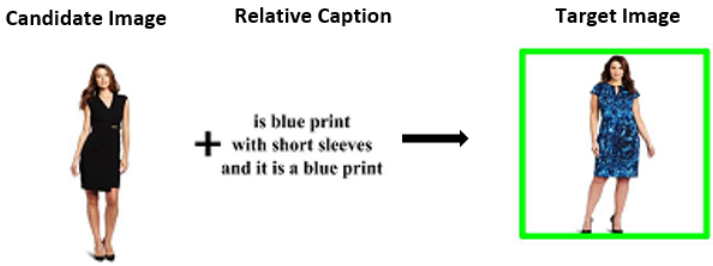

Figure 4. Example of Fashion IQ dataset


```
{
  "target": "B008BHCT58",
  "candidate": "B003FGW7MK",
  "captions": ["is blue print with short sleeves", "it is a blue print"]
}
```


Figure 5. Example of single entity of JSON file

This dataset has been published for the Fashion-IQ Challenge[5], the goal of which is to create an interactive image retrieval system from the given input, which includes both images and text in the fashion domain. To be more precise, the composition with text captions connects the input source image to the target image. For training and test data, the caption along with target and candidate image's id is provided in JSON format, a chunk of data is shown in fig. 5.

3.2 Data Pre-processing and Data Transformation

The stage of data preparation is critical for effective modeling. The data provided is not always in the proper format to be fed directly into the model. Text data, for example, might have mixed occurrences and other abnormalities. The model thus finds it difficult to generalize the data, resulting in the model over-fitting, which indicates the model isn't suitable for unknown data points. On this background, pre-processing is performed prior to modeling and training that is explained in the following subsection for image and text data separately. Once the data has been pre-processed, it needs to be transformed into the required format.

[5] https://sites.google.com/view/lingir/fashion-iq

3.2.1 Image Data

There is no information about the sizes of the images provided in the description of the dataset. Wu et al. (2021) resized the all fashion-IQ dataset image to 256x256 pixels and proves that with resizing the images recall can be improved, we did the same. Initially, images need to be converted into "RGB" format by making a triple 8-bit value. Image transformation seeks to enhance the original data by making some desired information more evident or expressive. These approaches include the random horizontal flip and cropping according to the given probabilities. The CNN models perform well with normalized images so images need to be normalized with the pre-defined sequence of means and standard deviations for each three-channel.

3.2.2 Text Data

For text pre-processing, each caption is further split into two captions or sentences that are associated with pair of candidate images and target images which describes the difference between both images. Each caption is written by human annotators as described in Wu et al. (2021). Firstly, concatenate these two captions into a single sentence, because these two captions are given in inverted commas separately. By further investigation, we found that the Fashion IQ dataset includes some misspelled words like (wihte instead of white) to correct this pyspellchecker1 python library is used to make sure the probability of misspelled words in the corpus is minimum.

We used splitting of training and testing data for our analysis that is described in Fashion-IQ Challenge and mentioned by previous researchers in their work e.g. Chen et al, (2020). The training, validation, and testing files are already provided in JSON format, and each split are about evenly distributed throughout the three categories, some statistics are shown in table 1.

		Split	# images	# Relative Caption
Dresses	Train	75%	11,452	11,970
	Test	25%	3,818	4,048
	Total		15,270	**16,018**
Shirts	Train	75%	19,036	11,976
	Test	25%	6,346	4,076
	Total		25,382	**16,052**
Tops & Tees	Train	75%	16,121	12,054
	Test	25%	5,374	4,112
	Total		21,495	**16,166**

Table 1: Fashion IQ dataset statistics

3.3 Modelling

CNN is made up of a series of neural networks that extract features. It concatenates all characteristics from preceding layers to circulate features at different levels. In the

architecture of the CNN, there are primarily five levels, as stated in Beevi et at. (2019). After the input layer, a sequence of deep convolution layers will be implemented, with RELU as the activation. The Rectified Linear Unit (ReLU) is used as an activation function in CNN. It speeds up CNN performance by using batch normalization and Max pooling layers to extract features from images. The batch normalization layer regulates the preceding layer's output and transfers it to the following activation layer. Then the use of the flatten layer converts all of the picture characteristics into a one-dimensional list.

3.3.1 Image Encoder

For image encoding, we used two CNN models EfficientNet and Resnet-50. EfficientNet is a Google-developed family of Convolutional Neural networks that not only increase accuracy but also model efficiency by lowering the number of parameters used during training in comparison to other state-of-the-art models. EfficientNet is available in a range of variants. We used EfficientNet-B0 because it is a simple small-size baseline model. The reason for selecting EfficientNet-B0 is due to the limited computing resources that will be discussed in detail in the implementation section. EfficientNet-B0 was trained on the ImageNet dataset, and we used the ImageNet weights to extract features from both the candidate and target images. We utilized the final layer before the classifier, with a dimensionality of 1024, to represent the image. The Resnet-50 is used as another CNN model which is also trained on the ImageNet dataset we use the pre-trained version without the last classification layer for image feature extraction.

3.3.2 Text Encoder

The words from concatenated captions are tokenized into text embeddings. Anwaar et al. (2020) use a pre-trained BERT model for a similar type of work but for our study we prefer to use the pre-trained LSTM model for textual feature because it is originally used in our compositional model TIRG Vo et al. (2019) we feed the text into LSTM with hidden layer 1024 and dropout layer with 0.1, followed by a linear projection layer. In summary, we begin by tokenizing text, then input the token sequence into the text encoder to produce the final text representation.

In the literature review, we see in Chen et al. (2020) that the performance of the model depends upon how much textual information we have in a dataset like fashion 200k[6] dataset contains a shorter or less descriptive caption that's why textual information is not much contributing in model performance. But in our dataset, the average caption length is 6.96 tokens and collectively they make 4,769 words in the vocabulary. Fig 6 shows sentence length distribution over all three product categories. We can see that the sentence length of captions is almost normally distributed. All three categories dresses, top & tees, and shirts have peaked around 5-6, which means that most of the captions in all categories have an average sentence length of 5-6 words.

[6] https://github.com/xthan/fashion-200k

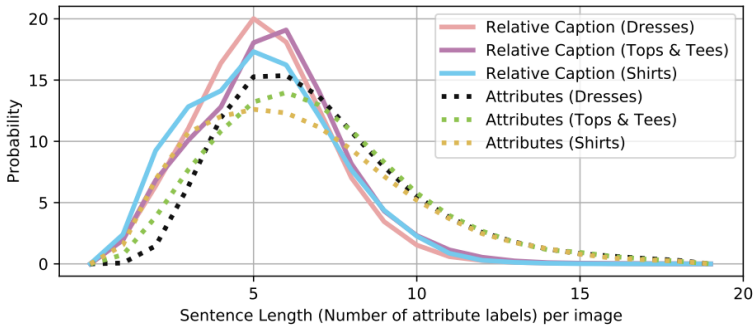

Fig. 6 Sentence length distribution. (Wu et al., 2019)

3.3.3 Composition Network

Composition Network selects the best-matched target image from a set of reference images and captions. We employ a gating module inspired by Text Image Residual Gating (TIRG) to modify the vector transition to combine image and text. We customize TIRG by using a different combination of models, TIRG uses a CNN (Efficient-NET, Resnet-50) to obtain image embedding, and LSTM to retrieve text embedding and then they combine the two using a gated connection.

TIRG learns the gating and residual features modules from images and captions, and the output is the weighted sum of the two features. Each module's projection layer is a fully-connected layer, followed by L2 normalization, which maps the features (image or text) onto a shared latent space. The residual features are learned by concatenating the candidate image and text with two 3x3 convolution layers with non-linearity. The gating function employs the same module followed by a sigmoid function. The assumption here is that matching image and text characteristics should be near together, while mismatched ones should be far apart. For model optimization, we use Adam with any suitable pre-defined learning rate e.g. 1e-4.

3.3.4 Evaluation:

Evaluation is an essential component of every research approach. Our primary retrieval evaluation metric is recall at rank k (R@K), where K can be 1, 3, 5, 10, 20, or 50, and is determined as the proportion of test queries where the target image is among the top K recovered photos. The average of recall at different levels is used to measure overall performance. Along with it, each model will also be evaluated based on batch-based classification training loss Zhu et al. (2020) and time performance, calculated as the time taken by each model for training.

4 Design Specification

The proposed methodology follows the KDD is shown in figure 2 represents the outline of strategy with little modified KDD process. Figure 7 depicts the overview of our proposed design framework, which consists of five modules, namely, image transformation module, image encoder module, text embedding, text encoder, and multimodal composition module. The multimodal composition module intends to extract and combine candidate image and

caption text characteristics in order to generate a composite feature representation. It then associates the composite characteristics with the target image representation. The image representation is retrieved specifically utilizing the efficient-net and resnet-50 CNN models. The LSTM model is used to embed the natural language caption text of each image.

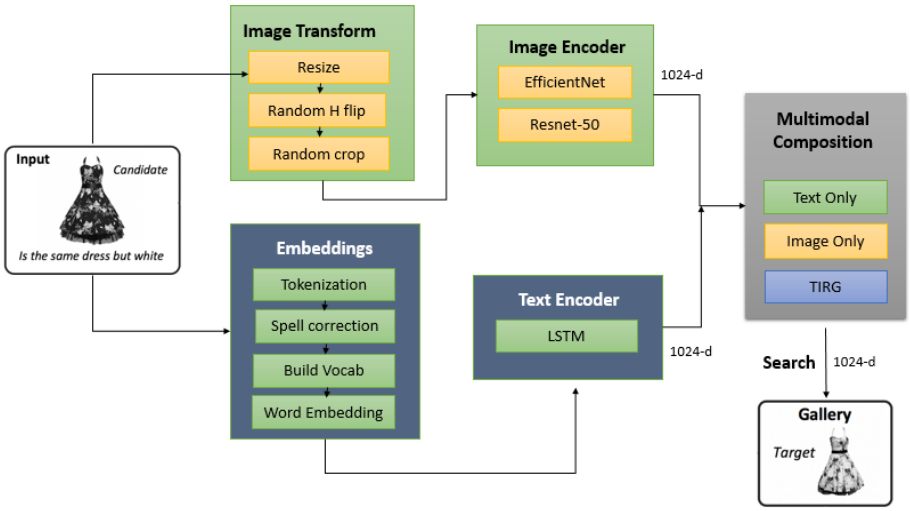

Fig. 7 Design Architecture

The following are the specifications used for training of each triplet candidate, target image, and caption, we proceed with the following steps:

Image Transformation: We used different image transformation techniques like image resize to a fixed dimension 256x256, random horizontal flip, and random crop according to the given probabilities.

Image Encoder: The image encoder represents the images into linear space, which has 1024 dimensions.

Text Encoder: Convert the concatenated captions into tokens and then into text embeddings and feed those to the LSTM which gives a 1024 dimension feature vector.

Composition layer: Using TIRG composition, create candidate vectors and caption representations in the same space. When we do this, we assume that the candidate image vector is biased toward the target image vector, so that the similarity between the candidate and target image vectors is greater than the similarity between the candidate and non-target image vectors.

Batch Normalisation: allowing our neural network to use normalized inputs across all the layers, the technique can ensure that models converge faster and hence require less computational resources to be trained. We use one-dimensional Batch Normalization "*BatchNorm1d*" provided by PyTorch

Dropout layer: a dropout layer with a probability 0.1 is used with the LSTM model. The dropout layer helps in reducing the overfitting of the model

Weight: both Efficient-Net and ResNet-50 are trained on ImageNet datasets so we use "ImageNet" pre-trained weight for training

Optimizer: Adam optimizer is used here with a starting learning rate of 1e-4.

5 Implementation

This section demonstrates how models are implemented to retrieve the most relevant images from the dataset. CNN and LSTM models need more time to process the images and text, training the model and dataset also have more than 77k images. Initially, we tried to execute the code in Jupyter Notebook on the local system but it takes so much time then Collab (Google Collaboratory) is used to carry out the experiments. It comes with roughly 12 GB of RAM and nearly 35 GB of storage and requires no setup. It also incorporates a shared GPU and TPU for quick deep neural network training processing. PyTorch, is used in the implementation of the project. On google colaboratory notebook Python version 3. The data is then uploaded to Google Drive and accessible using Google Drive mounting. The numpy, PIL, and Keras libraries from Python are used to convert and normalize images.

The recall at rank K (R@K), defined as the proportion of test queries where the target image is among the top-K retrieval samples, is used to quantify the performance of the image retrieval system. In our experiments. The Efficient-Net and ResNet-50 image encoders are pre-trained on ImageNet, and the outputs are loaded to PyTorch tensors. For image enhancement, we utilize a random crop and a random horizontal flip. For text embedding, the LSTM is employed, followed by a ReLU activation layer in each FC layer and a hidden dimension of 1024. We employ batch normalization and dropout with a regularization probability of 0.2. Each training batch contains B = 32 triplets of (reference image, query text, target image) for 15 epochs using Adam as the optimizer and a learning rate of 1e-4 as the starting point.

6 Evaluation

In this section all the models built will be evaluated in order to determine whether the results obtained are good enough for a multi-model image retrieval system or not, and to see which method is better for retrieving the target fashion image. Although evaluation of the retrieval systems is not as simple as it seems since subjectivity has a key role in the evaluation; usually, better visual similarities consider better results. Moreover, it will be more difficult if there is no ground truth, but our comprehensive dataset has ground truth in the form of target images labels which will be used for evaluating the models. Our main evaluation metric for the retrieval system is recall at rank k (R@K) as used by Shin et al. (2020) where K can be 1, 3, 5, 10, 20 and 50. Each pair in our dataset is associated with one true positive image which we call target image. So we can calculate the recall rate among K retrieved images which is the percentage of target images found in the top-K retrieval images. Along with it, each model will also be evaluated based on-time performance for training the models. Moreover, we compare batch-based classification triplet loss Zhu et al. (2020) against epoch during training of each model.

We conduct different experiments to evaluate models that are mentioned below. All models are evaluated on given categories (dresses, shirts) of Fashion-IQ dataset, using the

same data split as given in Table 1. Due to the limited computational resources, we are not able to train our models in all categories (dress, shirts, and Top&Tees). So, we randomly select the top&Tee category to drop, more details are mentioned in section 7. The dress category is divided into 75 % (11,452 images) train and 25 % (3,818 images) test, similarly, 75 % (19036 images) train and 25 % (6,346 images) test for shirt category.

Text-Only: This model does not consider candidate images and only fetches the target image based on the user text query. A LSTM network encodes the text into textual features, while Efficient-Net encodes the target images.

Image-Only Efficient-Net: In this model, only the candidate image is used to retrieve the target image and does not utilize textual information. Both candidate image and target image are encoded by Efficient-Net. It measures visual similarity between the candidate and the target image.

Image-Only Resnet-50: Just like the same Image-Only Efficient-Net, but instead of Efficient-Net, Resnet-50 is used as an image encoder.

TIRG: A novel technique that combines visual and textual information with an extra gating connection to transfer image features straight to the learned joint feature space. TIRG uses Efficient-Net to encode both candidate and target images, and LSTM network to encode Text.

The evaluation section is divided into the following two broader parts that will be discussed in more detail in the respective subsection.

- Quantitative Results
- Qualitative Results

6.1 Quantitative Results

Each experiment is conducted for both dress and shirt categories separately and results are recorded. Each training batch contains N=32 triplets of candidate image, target image, and caption. During training, batch shuffling is used for each epoch. The similarity between target and query image is calculated by normalized dot products determined by each model. For each experiment, we set the Initial learning rate as lr = 1e-4. For stable performance evaluation, each experiment is repeated 2 times and results are reported as the mean of them. The reason for only 2 times repetition is that it takes too much time for execution or training even on Colab with provided GPUs. But it is necessary to run the experiments multiple times because, in Colab, GPUs and resources are provided dynamically, so performance can vary for each time.

6.1.1 Epoch vs Training Loss

The training loss of every model is calculated as batch-based classification loss suggested by Zhu et al. (2020). The loss is gradually decreases with increasing epoch. At the initial stage, all models have almost the same loss but in the end, the multi-model TIRG has the lowest loss as compared to other models. Both category dress and shirt depict the same trend as shown in fig. 7. The image-only method with efficient-net image encoder performs worst in terms of training loss on both shirt and dress categories. Even at 15 epoch, its batch-based classification loss value is around ~1.0 as compared to TIRG which has around ~0.25 for

both categories. For image-only Resnet-50 and text-only method, the trend is different in dress and shirt categories. The text-only method has a lower loss value than Resnet-50 method for the dress category, and opposite is true for shirt category.

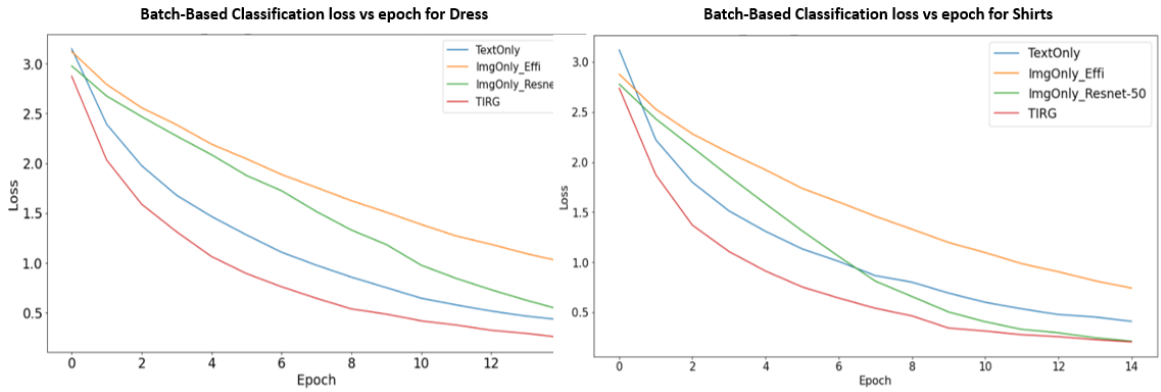

Figure. 7 Training loss vs Epoch for dress and shirt category.

6.1.2 Recall Mean vs Training epoch

Figure. 8 shows the mean of recalls vs epoch for the dress and shirt category. The mean of Recalls is calculated by taking the average of the R@1, R@3, R@5, R@10, R@20, R@50. These recalls are measured on training data against each epoch during training. We can see that Text-Only and TIRG perform well and have higher recall mean, while both Image-only models (Resnet-50 and Efficient-Net) do not show a high recall mean when compare to TIRG and Text-only method. The justification of this point is that, as the captions of the images are annotated by the human experts, so dataset has a very accurate and real description of the difference between query and target images. The image-only based models do not utilize the textual information while retrieving the images and do not show promising results. This shows that the combination of the textual and visual features gives better performance of fashion image retrieval system, which supports our research motive.

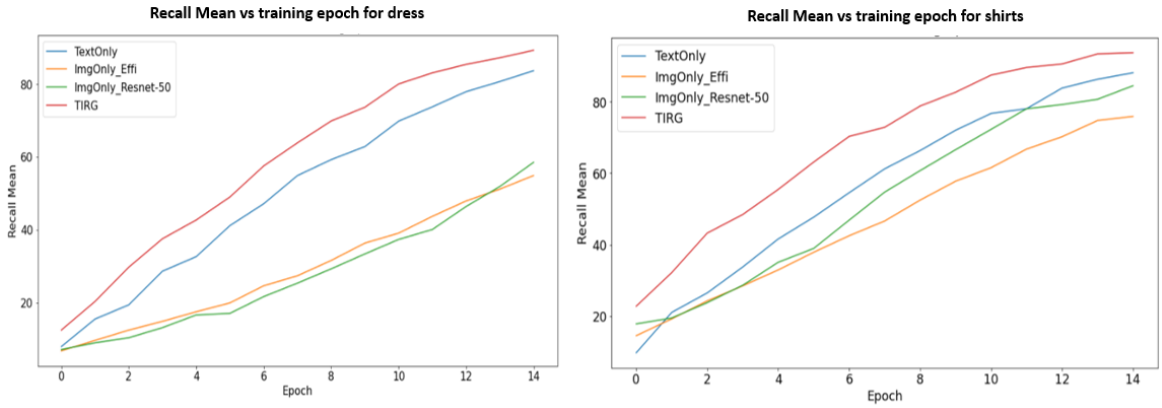

Figure. 8 Recall Mean vs Epoch on Training Dataset for dress and shirt category.

6.1.3 Evaluation on Test Data:

After evaluating the model on training data, evaluation is also done on the test dataset. The evaluation metrics are kept the same. Figures 9, 10 graphically represent the results on test data for dress and shirt category respectively. The recall values are plotted against

different ranks (Top K) for each model. The TIRG model outperforms the other models with having the highest recall values for all ranks (Top K). Similar to training data, only image-based models do not perform well on test data in terms of recall. We can conclude that in terms of recall, our models perform well when the sample size of retrieve images is large. It means, the higher the rank (Top K), the higher will be recall. Figure 9, 10 also plots the recall mean for all models on the test dataset. Here recall mean is calculate as the average of R@1, R@3, R@5, R@10, R@20, and R@50. For both dress and shirt categories, TIRG has the highest recall mean (Approx. 24) on test data followed by text-only method (Approx. 17).

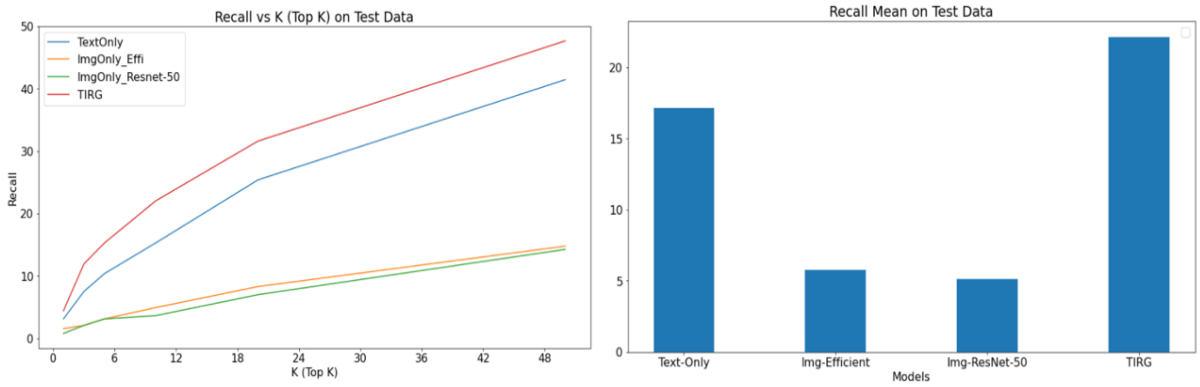

Fig. 9 Recall on test data for Dress category

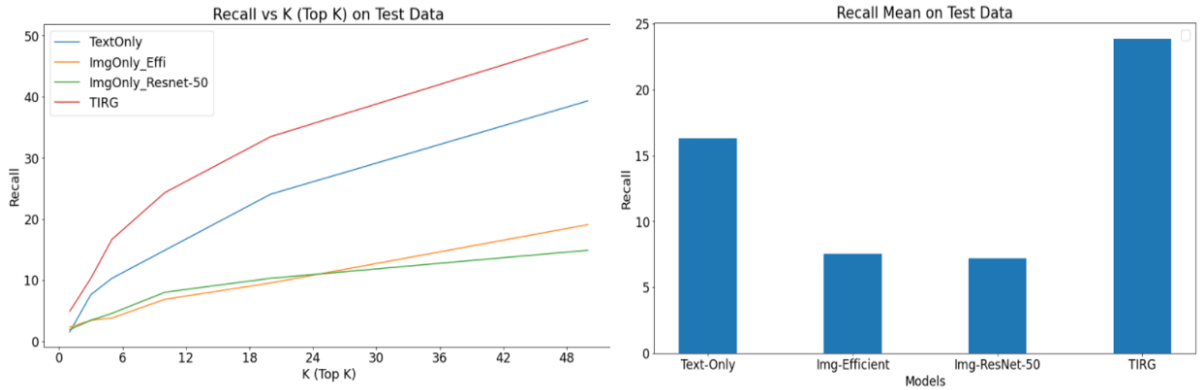

Figure 10. Recall on test data for Shirt category.

Table. 2 shows some evaluation statistics on test data. Here for convenience, only R@5, R@10, and R@50 are represented in table similar to Shin et al. (2020). Table 2 also shows the overall summary of the methods used in our research project. Here in table 2, the recall mean depicts the collectively representation of the recall results on both dress and shirt category. The mean of recall is calculated by adding R@5, R@10, and R@50 of both categories and divided by 6. The TIRG model has the highest average recall (29.24) followed by the text-only method having a value of 21.93 (25% less than TIRG). The recall mean of the Image-Only resnet-50 method (8.75) is as same as the image-Only efficient-net (8.74). Although, the time taken by the image-Only resnet-50 method for training is higher than image-Only efficient-net as shown in figure 11. So, we can safely conclude that Image-Only efficient-net is a better method than Image-Only resnet-50 method in terms of time performance. All methods show the high value of recall at R@50. It means that the chances of getting more relevant images will be high if more images are retrieved by the retrieval

system. Taken TIRG and Text-Only method as they perform well in terms of recall. By using TIRG instead of Text-only method, the performance is increased by 46% in R@5, 44% in R@10, and 21% in R@50 in recall on average.

Method	Text Encoder	Image Encoder		Dress			Shirt			Recall Mean
		Query	Target	R@5	R@10	R@50	R@5	R@10	R@50	
Text only	LSTM	-	Efficient-Net	10.36	15.28	41.45	10.31	14.89	39.31	21.93
Image only (Resnet-50)	-	Efficient-Net	Efficient-Net	3.11	4.92	14.77	3.82	6.78	19.08	8.74
Image only (Efficient-Net)	-	Resnet-50	Resnet-50	3.63	7.1	14.25	4.58	8.02	14.89	8.75
TIRG	LSTM	Efficient-Net	Efficient-Net	**15.28**	**22.02**	**47.67**	**16.68**	**24.32**	**49.47**	**29.24**

Table 2. Statistics on test data

6.1.4 Training Time performance

For large datasets, training time constrain is also very important. The fashion-IQ dataset has around 77K images so the training time of the models is also measured during each method. The time taken by the model is depends upon multiple factors like a number of epoch, learning rate etc. due to the limited computation resources we set the number of epochs to 15. Figure. 11 shows the combined time performance of both dress and shirt categories for each method. We can see that dress category takes more time to train comparatively than the shirt category data. The reason for this can be, as shirts are visually simple or have less style versatility than dresses, which can have a more sophisticated design or style. So it will be more easier and faster for deep learning models to train on shirt dataset as compared to dress data The image only Resnet-50 method takes the highest time to train 5385 sec (dress) and 4241 sec (shirt) followed by the TIRG method 4022 sec (26% less) and 3865 sec (9% less) for dress and shirt categories respectively.

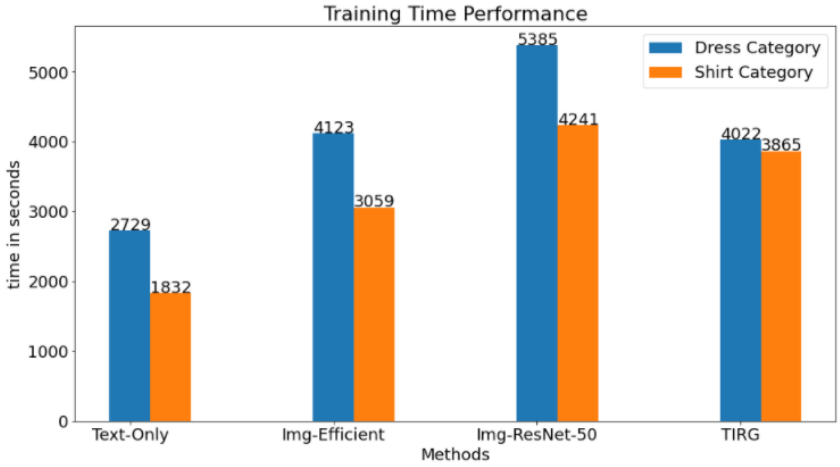

Fig. 11 Training time performance

6.2 Qualitative Results

Some qualitative results are reported in figure. 12 that shows that the top 5 retrieved similar images with the help of the TIRG model. For simplicity, only the TIRG results are shown, other method results are provided in the code notebook. The candidate image and text queries along with its ground truth target image are provided at left and the top 5 highest-scored retrieved images are on the right side. Our methods assign the higher value of dot product similarity to the images that best meet the requirement of the user query while keeping the style of the original query image. Figure. 12 shows three examples (a), (b) are from the shirt category, and example (c) from the dress category.

Fig. 12 qualitative results

7 Discussion

This section goes through the findings from the prior experiments on image retrieval systems. In section 6, all four methods Only-Text, Only-Image Resnet-50, Only-Image Efficient-net, TIRG are evaluated on the basis of Recall, training time performance, and batch-based classification loss. In all evaluations, we can conclude that TIRG outperforms the other methods significantly. It supports our research motive as the performance of the image retrieval system can be improved by utilizing the textual and visual features simultaneously. The conventional only-text based retrieval system also shows good results in terms of recall and losses. But the multi-model based TIRG approach outperforms the only-text based image retrieval system as well. The methods that are only based on image query i.e. image-only Resnet-50 and image-only efficient-net, do not show betters results in terms of recall because they do not utilize the textual information while retrieving the images. In our data case, the captions of the images are written by human experts, so the quality of the textual information is very accurate clearly defines the difference between query and target images. That is also a reason for better recall values of TIRG and text-only method. This shows that the combination of the textual and visual features gives better performance of the fashion image retrieval system, which supports our research motive.

First, the time and effort invested in pre-processing and transformation of the dataset are crucial since it will serve as the groundwork for our implementation. This project used the

state-of–the-art LSTM model for the extraction of textual features. However, some other alternative natural language processing techniques like BERT might be used for this purpose as well. The BERT-based text encoder results can be compared with the LSTM text encoder. But due to the time limitation, we had to drop the idea to implement the BERT model.

7.1.1 Resource limitation

For implementation, we used Colab, as the local machine is not a feasible solution to train the fashion-IQ dataset that has around 74k images. The Colab provides 12 GB of RAM in free mode which is not enough to train the model on the Fashion-IQ dataset. When we try to train our models on just a single category e.g. dress, all of the RAM is utilized and the session got crashes. To tackle this issue, we configure our code to make it compatible with colab provided GPUs that is NVidia Cuda. But still, with the restriction of a maximum of 12 hours processing in Colab and RAM limitation, we are managed to train our model on only two categories dress and shirt data individually. With more resources, we can be able to implement all categories of the Fashion-IQ dataset and come up with more comprehensive results.

7.1.2 Data Limitation:

The performance of any approach or model highly depends upon the quality of the used dataset. The Fashion-IQ dataset has better quality than other similar datasets like Fashion 200K[7], FashionGen[8] because captions are written by human annotators that more clearly describe the visual difference between candidate image and target image. However, there is a limitation of the Fashion-IQ dataset that is highlighted by Shin et al. (2020), who showed that there are some wrong data pairs. As the dataset is given in pairs and each pair contains candidate image, target image, and caption. The caption should explain the difference between the candidate image and target image but some annotations or captions are found to be wrong, some examples are shown in figure 13. In case 1, the target image is not a fashion image, the target image should be a fashion product image according to described caption. In case 2, the target image is not exactly matched with the prescribed caption and candidate image. In case 3, wrong candidate image or no noticeable relation between two images.

Fig. 13 wrong data pairs

[7] https://github.com/xthan/fashion-200k

[8] Fashion-Gen: The Generative Fashion Dataset and Challenge

Another limitation is low vocabulary. For the fashion-IQ dataset, each pair of candidate and target images is associated with two small captions. The average length of the caption is 6.96 tokens and collectively they make only 4,769 words in the vocabulary. The LSTM model performs well when there is large vocabulary available for training, it depicts that results can be improved if the dataset has more detailed or lengthy captions of images.

8 Conclusion and Future Work

Fashion e-commerce is expanding at an incredible rate. Moreover, as the quantity of fashion products available in online stores grows tremendously, so retrieving the desired product in online stores becomes a challenging task. As a result, online retailers and businesses recognize the need of sophisticated image retrieval systems. This paper compares the uni-model and multi-model image retrieval system and shows how results can be improved by using both text and image query simultaneously. For implementation, Resnet-50 and Efficient-Net are used as image encoders and LSTM as text encoder. We experimentally evaluate different methods such as text-only, image-only, and combination of both text and image on the Fashion-IQ dataset by using recall as a main evaluation metric.

Possible short-term future directions of this research work can be, to evaluate all our methods on the remaining top & tee category of Fashion–IQ dataset. This can be done with the availability of more computing resources. The long-term future goal is to integrate our methods with the knowledge base concept beheshti et al. (2020). Many features of fashion dress are connected to one another and have a significant positive or negative correlation, such as formal shirts has a proper collars and embroidered dress are usually wedding dress. These data may be efficiently used to produce more accurate results. Another future work is to extend the applicability of the multi-model image retrieval system because it can be applied to any image and text pair data. This approach is not limited to the fashion domain, in fact, it may apply to other domains such as food, shoes, and so on

References

Ak, K. E., Kassim, A. A., Lim, J. H. & Tham, J. Y. (2018), Learning attribute representations with localization for flexible fashion search, in ‘Proceedings of the IEEE conference on computer vision and pattern recognition’, pp. 7708–7717.

Beevi, K. S., Nair, M. S. & Bindu, G. (2019), ‘Automatic mitosis detection in breast histopathology images using convolutional neural network based deep transfer learning’, Biocybernetics and Biomedical Engineering 39(1), 214–223.

Beheshti, A., Yakhchi, S., Mousaeirad, S., Ghafari, S. M., Goluguri, S. R. & Edrisi, M. A. (2020), ‘Towards cognitive recommender systems’, Algorithms 13(8), 176.

Chen, Y., Gong, S. & Bazzani, L. (2020), Image search with text feedback by visiolinguistic attention learning, in ‘Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition’, pp. 3001–3011

Faghri, F., Fleet, D., Kiros, J. & Fidler, S. V. (2017), 'Improving visual-semantic embeddings with hard negatives', arXiv preprint arXiv:1707.05612 .

Hsiao, W.-L. & Grauman, K. (2018), Creating capsule wardrobes from fashion images, in 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 7161–7170.

Jo, J., Lee, S., Lee, C., Lee, D. & Lim, H. (2020), 'Development of fashion product retrieval and recommendations model based on deep learning', Electronics 9(3), 508.

KINLI, F. O. & KIRAC¸ , F. M. (2020), 'Fashioncapsnet: clothing classification with capsule networks', Bilis¸im Teknolojileri Dergisi 13(1), 87–96.

Overview of the KDD Process (1996.), http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_ kdd.html. Accessed: 2021-03-26

Parekh, V., Shaik, K., Biswas, S. & Chelliah, M. (2021), Fine-grained visual attribute extraction from fashion wear, in 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 3973–3977.

Sadeh, G., Fritz, L., Shalev, G. & Oks, E. (2019), 'Joint visual-textual embedding for multimodal style search', arXiv preprint arXiv:1906.06620 .

Sidharth, R., Rohit, P., Vishagan, S., Karthika, R. & Ganesan, M. (2020), Deep learning based smart garbage classifier for effective waste management, in '2020 5th International Conference on Communication and Electronics Systems (ICCES)', IEEE, pp. 1086–1089.

Shin, M., Cho, Y. and Hong, S. (2020) "Fashion-IQ 2020 challenge 2nd place team's solution," arXiv [cs.CV]. Available at: http://arxiv.org/abs/2007.06404.

Tan, F., Cascante-Bonilla, P., Guo, X., Wu, H., Feng, S. & Ordonez, V. (2019), 'Drill-down: Interactive retrieval of complex scenes using natural language queries', Advances in Neural Information Processing Systems 32, 2651–2661.

Tan, M. & Le, Q. (2019), Efficientnet: Rethinking model scaling for convolutional neural networks, in 'International Conference on Machine Learning', PMLR, pp. 6105–6114.

Umer Anwaar, M., Labintcev, E. & Kleinsteuber, M. (2020), 'Compositional learning of image-text query for image retrieval', arXiv e-prints pp. arXiv–2006.

Verma, S., Anand, S., Arora, C. & Rai, A. (2018), Diversity in fashion recommendation using semantic parsing, in '2018 25th IEEE International Conference on Image Processing (ICIP)', IEEE, pp. 500– 504.

Vo, N., Jiang, L. & Hays, J. (2019), 'Let's transfer transformations of shared semantic representations', arXiv preprint arXiv:1903.00793 .

Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.-J., Fei-Fei, L. & Hays, J. (2019), Composing text and image for image retrieval-an empirical odyssey, in 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 6439–6448.

Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K. & Feris, R. (2021), 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', pp. 11307–11317.

Yu, T., Shen, Y. & Jin, H. (2020), Towards hands-free visual dialog interactive recommendation, in 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 34, pp. 1137–1144.

Zhang, S., Yao, L., Sun, A. & Tay, Y. (2019), 'Deep learning based recommender system: A survey and new perspectives', ACM Computing Surveys (CSUR) 52(1), 1–38.

Zhu, Q., He, Z., Zhang, T. & Cui, W. (2020), 'Improving classification performance of softmax loss function based on scalable batch-normalization', Applied Sciences 10(8), 2950.