International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Enhancing Fashion Image Retrieval with Multi-Modal Query and Zero-Shot Learning for Cross-Domain

Swathy S

swathyecengg[at]gmail.com

Abstract: ***CBIR (Content-Based Image Retrieval) system has two main challenges in a) Generalizability and b) Retrieval on Cross-Domain data. Fashion Image Retrieval (FIR) encounters the challenge of retrieving images in cross- domain data due to the difference in user shot photograph and product photographs. This is due to the viewpoints, lighting conditions, and the presence of complex backgrounds a relevant query is crucial for retrieving the closest match. The scenario of inadequate relevant query, to search and retrieve images is a major cause for low generalizability in CBIR. This research targets both these challenges by implementing multi-modal queries for retrieval to handle the first challenge. And the second challenge is addressed by a zero-shot learning model for retrieval to enhance the retrieval accuracy on cross- domain data for FIR. DeepFashion (Liu et al., 2016) dataset with the cross-domain data will be used to propose a system that can retrieve based on user shot images and text queries. Evaluation metrics like Recall, Retrieval Accuracy, F1 score, and Mean Average Precision (mAP) will be used to evaluate the model. The evaluation metrics for each attribute type will be presented in this research.***

Keywords: CBIR, Image Retrieval, Cross-Domain, Multi-Modal Query, Zero-Shot Learning

1. Introduction

Advancements in data storage technologies lead to a continuous increase in the creation of large multimedia content databases. Specifically, image acquisition technology and digitization have led to an increase in many image datasets. The abundance of content generation introduces new challenges in data management and data retrieval. Content-based image retrieval (CBIR) is a technique to retrieve images from the database according to the user's requirement with a search query.

The capability of deep learning algorithms to learn the features from the data is leveraged in CBIR algorithms. The usage of an efficient information retrieval system from image data (DAVIES, 2005), with CBIR, is focused on in recent research (Chen et al., 2021). CBIR is the problem of searching visually similar images or semantically matched images from the image database. To enhance retrieval accuracy, one of the key research objectives is decreasing the gap between the features presented and the visual perception of humans. Some of the applications of CBIR are medical (Murala and Wu, 2014), remote sensing image retrieval (Walter et al., 2020), e-commerce websites (where users can search an item with an example image), and forensic applications. The basic technique adopted in CBIR is to search and rank the images (Faria et al., 2010) based on a visual semantic relationship with the search query given by the user. Search engines on the Internet provide search results based on the text query or image query (Google Image Search). The user retrieves the search results based on the input keywords and/or input image. The user's search query plays a crucial role in retrieval. The characteristics of a query are dependent on the dataset domain from which the image is to be retrieved and the user's knowledge to give an appropriate search query to retrieve intended results accurately. There are many types of search queries, popular query types among them are text-based, image-based, sketch-based, and key-point-based. The query can be flexible with different types of queries called multi-modal queries (Xu et al., 2012) used with the CBIR framework. For accurate required content retrieval, the example image query should be close enough to the user's requirement information.

The retrieval accuracy, similarity score, F1 score, MAP (Mean Average Precision), and ranking are some of the widely used performance metrics to evaluate the performance of the CBIR algorithm. The ranking strategy (Faria et al., 2010) used by CBIR systems is to order image content descriptors, so that returned images that are most like the query image are placed higher in the rank. The common challenges faced in CBIR are extracting discriminative features and selecting the best similarity score those results in high retrieval accuracy and best retrieval ranking orders. Based on the literature review conducted during this research, two main challenges were identified a.) Insufficient training instance for each class b.) Generalization in cross-domain retrieval. Image retrieval is a challenging task in the case of cross-domain, which entails pairing photographs from one domain to their equivalent in another. One of the real- time challenges occurs during the retrieval process, when retrieval is based on new (unseen style during training) product images or user images, large variations in the viewpoint and style, lighting conditions, complex background, shape deformations, and the occlusions result in undesired search results. This thesis focuses on the research in developing a CBIR methodology for fashion image retrieval with multi-modal (Image and Text) queries addressing the cross- domain retrieval challenge.

2. Literature Review

Introduction

From the literature survey for fashion image retrieval, the

challenges faced in real-time applications, proposed methodologies for improving the retrieval accuracy, image feature extraction techniques, image and text fusion techniques that bridge the semantic gap between low-level characteristic and global image features were studied. Benchmark results that were presented in the proposed methods by the authors, the advantages and disadvantages of proposed methods, and challenges that are still need to be addressed for improving the retrieval accuracy in the real-time application are focused and presented in this section. Cross-domain retrieval is the major challenge that was identified, and existing research works on addressing the same are studied. Leveraging zero-shot learning approach for a similarity-based retrieval system, implementing techniques that improve the feature extraction from the query image, and choosing the best training strategy and loss function is the aim of this research work.

Content-Based Image Retrieval

Retrieval of images is required in different domains like web search engines, biometrics, medical diagnosis, face finding, textiles industry, retail catalogues, remote sensing systems, etc. A retrieval system works on the principle of finding similar images from the database, according to the given query information. Text-query describing the texture, shape, colour, or keywords related to the image are being used popularly in image retrieval systems. But text- query based retrieval techniques require a huge number of annotated data with tags utilizing human efforts, which is a limitation. There are certain key challenges in CBIR like reducing the semantic gap, improvising the retrieval scalability, and balancing both retrieval accuracy and efficiency. (Lai et al., 2015) proposed Triplet Ranking Loss that effectively utilizes the inter-class and intra-class differences for image retrieval. Intermediate image features are generated by mid-level convolutional layers from the deep CNN architecture. These image features are used to generate the hashing bits for images. (Zhang et al., 2015) proposed a deep regularized similarity comparison hashing (DRSCH) to address the problem of generating hash functions from pre-defined feature space leading to ignoring the significance level of the different bits and ignoring the generalizability.

Multi-Modal Query

Multi-modal query introduces flexibility in searching content and retrieving it. In search engines, when queries with a different modality like text-based, image-based, or sketch-based are used, the design imitates on how the human mind works to create and process information to search. This increases user experience. Some existing multi-modal search enabled web applications are the Google Images search engine, Bing image search engine, and MMRetrieval (Ciaccia et al., 2010). Multi-modal query in CBIR system for a clothing-fashion domain helps the user to retrieve appropriate clothes. Let $\phi_{i,t}$ be the combined feature encoding for image and text with the function defined as $f_{combined} = (\phi_i , \phi_t)$. From the existing methods for combining text and image features, two of them are:

1) TIRG (Text Image Residual Gating)

(Vo et al., 2019a) proposed TIRG to combine image and text features and it is defined by equation □□□□□□□□ = □□,□ = □□□□□□□(□□, □□) + □□□□□□(□□, □□), where □□□□□,□□□□ are gating and residual features. □□, □□ these are the weights learned during the training. Gating connection features □□□□□ and residual connection features □□□□ are used to create a new feature fusion method in which the input image features and the output image features are in the same image feature space with the input image as the reference. The residual connection adds a modification in this feature space based on the input text.

2) Attention Fusion

(Zhang et al., 2021) presents four different methods for combining text and image features. These four different methods have experimented with different few-shot learning classification models and backbones. From the experiment result presented, the Attention Fusion method has shown the highest accuracy of 78.40 ± 0.81 for image and text multi-modal query with ProtoNet (Snell et al., 2017a). The attention mechanism is utilized to capture the correlation within a sequence of features. From the text query, each sentence is encoded separately and stacked upon each other. The corresponding image feature, the image channel is reshaped and obtained feature vector by applying 1 X 1 convolutional layer. A single head attention module consisting of queries and keys is used. Using this attention mechanism is achieved to calculate text-guided image features.

Image Segmentation and key point detection

Locating and recognizing clothing attributes is a challenging step in fashion image retrieval. Key points for clothing attributes describe the structure of a clothing item which can be used to determine the bounding box and its category. Detection and key point estimation are done together parallelly in Mask R-CNN (He et al., 2017a) in which the feature extracted from ROI- Align is used for both instance segmentation and object detection. Match R-CNN built upon Mask R-CNN, proposed by (Ge et al., 2019) solves four tasks including clothes detection, pose estimation, segmentation, and retrieval as a multi-task learning framework. Semantic segmentation is a type of segmentation where it is a process of classifying each pixel belonging to a particular label. There are many deep learning architectures for semantic segmentation like AlexNet (Krizhevsky et al., n.d.), GoogleNet (Szegedy et al., 2014), VGGNet (Simonyan and Zisserman, 2014), ResNet (He et al., 2015), etc. (Martinsson and Mogren, n.d.) proposed a sematic segmentation approach using feature pyramid network (FPN). Clothing detection is a fundamental task in fashion analysis, an improved detection method by using clothing key points is proposed by (Qian et al., 2021a). This is a Key Point-Guided (KGDet) clothing detection approach that utilizes two main aspects i) Local features are integrated around the key points identified to enhance both classification and regression. ii) Accurate bounding box is generated from key points. KGDet takes in an input image and outputs a bounding box, class, and key points of each clothing item. KGDet consists of three stages, one initial stage, and two final refined stages. Each stage has a classification branch that gives a class based on the score map from prediction and a regression branch that regresses key points. KGDet uses ResNet (He et al., 2015) as a backbone along with the FPN (Lin et al., 2017). Since CNN's are limited to geometrics transformations, viewpoint,

multi-scale objects, and part deformation. CNN's can be modeled only using data augmentations that are chosen by the user limited to their knowledge. To tackle this scenario, deformable convolutions proposed by (Dai et al., 2017) are used at the refinement stage with the informative key points taken as an offset for deformable convolutions. For classification, anchor-free detectors are used like in (Zhang et al., 2017). Key points are detected with the offset vector calculated from the center of the object and regression is performed directly on offset values instead of using heatmaps. These key points detected are like RepPoints used by (Yang et al., 2019).

Zero-shot Learning

Zero-shot learning is an approach where the model predicts unseen classes by combining auxiliary information that contains any specific attributes and seen objects during training. Zero-shot learning is implemented in two stages. Training is stage 1 where the attributes information is captured, stage 2 is Inference where the information is then used to recognize unseen object categories during training. Siamese networks proposed by (Koch et al., n.d.) is a learning methodology in which the model learns to the rank similarity between the input data. With this learning, the methodology model gains a powerful discriminating ability to generalize the network to new data and new classes from unknown distributions. (Ong et al., 2017) addressed the challenge of image retrieval from a large-scale database. A Siamese network consisting of two computational strands. These strands consisted of a CNN component and a Fisher Vector (Csurka and Perronnin, 2011) component to produce a final global descriptor. And achieved 77.3% accuracy on the Oxford Building dataset. For improving the retrieval accuracy (Deng et al., 2017) presented a unique embedding method named as "focus ranking unit" that can be added into a CNN for joint learning the image representations and metrics for fabric image retrieval. This is achieved by training the model with images as a focus ranking unit. Each of these units contains a probe image, one image of the same fabric as the probe image (a positive image example), and several images of different fabrics (negative image examples). These image units are then fed to the network to compute "probe to negative" and "probe to positive". Model learning is targeted to decrease the distance between probe-to-positive than the probe-to-negative distance. Thus, the ranking disorders in all units are penalized. Cross-entropy loss with regularization term is used. Triplet networks proposed by (Hoffer and Ailon, 2014), contain three networks that are identical to each other and are trained as triplets $\{x_+, x_a, x_-\}$ which is of the form (positive, anchor, negative). The anchor and the positive samples are from the same class. The negative sample is from a different class. The loss calculated during this learning process is called triplet loss which targets the anchor to be closer to the positive sample than the negative sample in the embedding space. This approach is used in many applications (Wei et al., n.d.; Schroff et al., 2015; Elezi et al., 2019; Huang et al., 2021; Khaertdinov et al., 2021).

Deep metric learning loss functions

Deep metric learning aims to develop a similarity metric that computes the similarity or dissimilarity of two or more objects through informed samples. With this approach, learning a non-linear transformation of the feature space leads to the possibility of capturing a non-linear feature structure. Deep metric learning can be leveraged for the task of retrieval by designing appropriate loss functions and training strategies. The most widely used loss functions are contrastive loss as used in Siamese networks and triplet loss for triplet networks.

1) Triplet loss

Triplet loss (Yuan et al., 2019) (Wu et al., 2019) with the underlying idea of comparing the anchor input image to a positive sample belonging to the same class and a negative sample belonging to a different class.(Xiao et al., 2017) proposed a Margin Sample Mining Loss based on triplets. (Xiao et al., 2017) stated that a triplet contains three different images I_A, $I_{A'}$ and I_B, where I_A and $I_{A'}$ are similar images while I_B is an image of a different identity This makes sure that all points in the same class form a single cluster but are not required to be present at the same point. Triplet loss suffers from poor generalization since it shares the same image for both negative and positive pairs. Quadruplet loss (Chen et al., 2017), extends the triplet loss by adding a different negative pair. $\{I_A, I_{A'}, I_B, I_C\}$, where I_A and $I_{A'}$ are similar images or from the same class and I_B and I_C are images dissimilar to I_A and $I_{A'}$ or from a different class.

2) TriHard Loss

TriHard (Triplet loss with Hard sampling) loss (Chen et al., 2019)(Hermans et al., 2017) is a method where simple samples are filtered by hard sample mining, which increases the model's resilience. Tri hard loss can be calculated on a group of samples. There are multiple identities in each batch, each with the same number of samples. It creates a triplet for each sample by selecting the most dissimilar sample with the same identification and the most identical sample with a different identity.

3) Margin sample mining loss

(Xiao et al., 2017) proposed a loss function for metric learning in which, the most dissimilar positive pairs and the most similar negative pairs are as follows:

$$Loss = (\max_{A,A'}(\|f_A - f_{A'}\|_2) - \max_{C,B}(\|f_C - f_B\|_2) + \alpha)$$

where C and B are the hardest negative pair and A and A' are the hardest positive pair.

3. Proposed Solution

Two types of approaches are proposed for a) Enhancing Multi-Modal query based image retrieval b) Enhancing image retrieval accuracy on cross-domain with few-shot learning. The proposed pipeline was implemented with pytorch modules.

Dataset

DeepFashion Database, a large-scale clothes dataset is utilized for this research. The dataset is very huge consisting of 800,000 diverse fashion images with product images from the shop and consumer images. To conduct the experiments

with available limited resources, a subset of the DeepFashion image dataset is created that resembles the dataset distribution of the original dataset. The fashion images are annotated with textual and attribute information, such as the texture of cloth, style of cloth, part of the cloth, category, or label stating whether the cloth is bottom wear, top, or a dress. The DeepFashion image database consists of four different types of image datasets. From which two different sub-datasets are used in this research. The first dataset is called the "Attribute Prediction Dataset". And the second sub-dataset is called the "Consumer-to-shop Clothes Retrieval Benchmark" consisting of cross-domain image pairs used for the research for cross-domain retrieval.Data subset is created by acquiring 300 random images for each of the 50 categories, totalling 15000 images from the" Attribute and Prediction Dataset". From the "Consumer-top-shop Clothes Retrieval Benchmark" consisting of 2,13,000 images, under the broad categories of Clothing, Dresses, Tops, and Trousers. The subset is created by randomly choosing 100 datasets from each sub-category under the categories of clothing, dresses, tops, and trousers.

Two types of dataset a)"Attribute and Prediction Dataset" is used for the research on Fashion Image retrieval with multi-modal query and b) "Consumer-top-shop Clothes Retrieval Benchmark" is used for research on cross-domain retrieval. A sample datapoint or image with the available information is presented in Figure 1 from "Attribute and Prediction Dataset".

A sample datapoint or image with cross-domain pair is presented in figure 2. And a snapshot of the dataset structure is presented in figure 3.

Figure 2: Cross-Domain Pair of a Shop product image and a Consumer Image

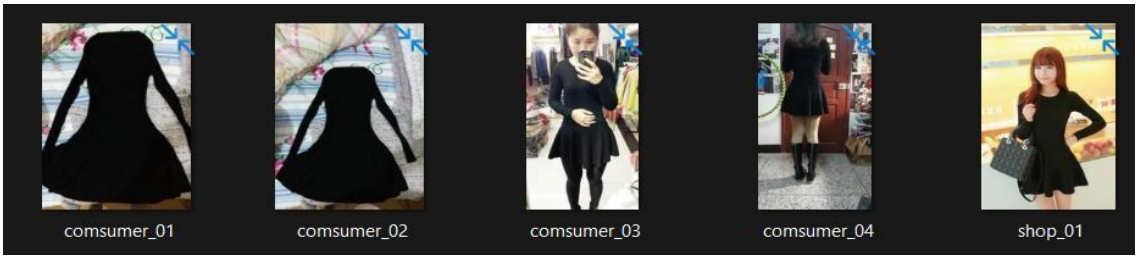


Figure 3: Snapshot previews the dataset structure for an image with 4 consumer images and matching 1 product image

For training a learning algorithm for multi-modal retrieval, an image with equivalent text information is required. The dataset does not consist of image captions that can be readily used. The annotation data consists of 5 different descriptive cloth features encoded as labels and one-hot encoded annotations. To solve this problem, the textual information is constructed by gathering all the descriptive features and thus forming caption-like textual information for every image. An example datapoint loaded into data frame can be referred from figure 4.

ID	image	text_info	label_group
0	img/Single-Button_Blazer/img_00000143.jpg	Blouse button part	Single-Button_Blazer
1	img/Tribal_Print_Linen_Blazer/img_00000088.jpg	Blouse print texture trench style tribal texture	Tribal_Print_Linen_Blazer
2	img/Single-Button_Blazer/img_00000049.jpg	Blouse button part fitted shape	Single-Button_Blazer
3	img/Textured_Collarless_Jacket/img_00000061.jpg	Jersey collarless part dark style light style ...	Textured_Collarless_Jacket
4	img/On_The_Range_T-Shirt_Dress/img_00000026.jpg	Jumpsuit basic style stretch fabric	On_The_Range_T-Shirt_Dress

Figure 4: Snapshot of a loaded data frame with image path and respective textual information

"label_group" is the feature specifying the exact style of the image represented with a snapshot of the dataset in figure 5. The "text_info" feature provides the information on category, attribute, style, texture, and part concatenated as a string. The string is tokenized further for the experiments.

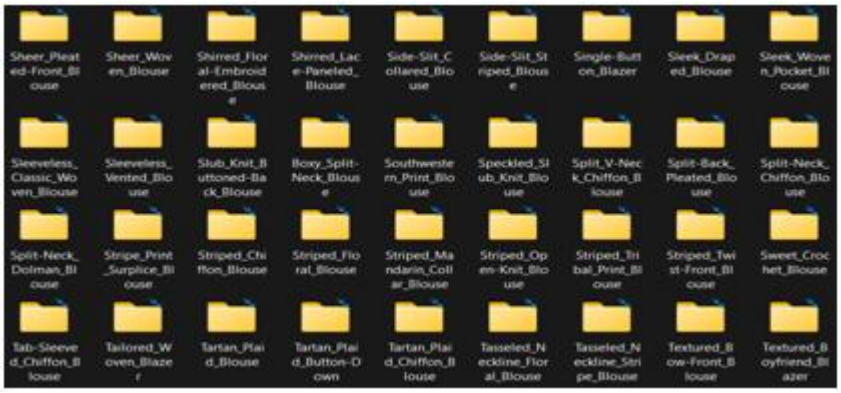


Figure 5: Snapshot of image segregation based on the style. Images within a folder can have a few common attributes and some different attributes due to the texture or type of fabric

Baseline Model

A non-parametric algorithm KNN (K-Nearest Neighbor) was used for the retrieval model based on image and text was used as a baseline model. DeepFashion dataset was used and tested with K = 10 to 100, where K number of retrieved top-k images with KNN model. Specific models were selectively chosen for generating Image and text feature embeddings. Resnet101, Mobilenet and Distil BERT. Combinations of with and without image and text feature embeddings were experimented with to analyze the impact of query (Image and Text) embeddings. The text and image embeddings are combined by simple feature embeddings concatenation. The selection of models was based on the best MMAP evaluated. Mean MAP (Mean Average Precision) is calculated for retrieved images. MMAP is derived by calculating the average of precision at each hit for each of the categories or styles. A correct image in top-k is called a hit. The experiment is carried out for different k values and is presented. All the experiment results with baseline models are compared in terms of MMAP. The flow of the baseline model is depicted in figure 6. With this baseline model, the reference MMAP score observed is 0.23.

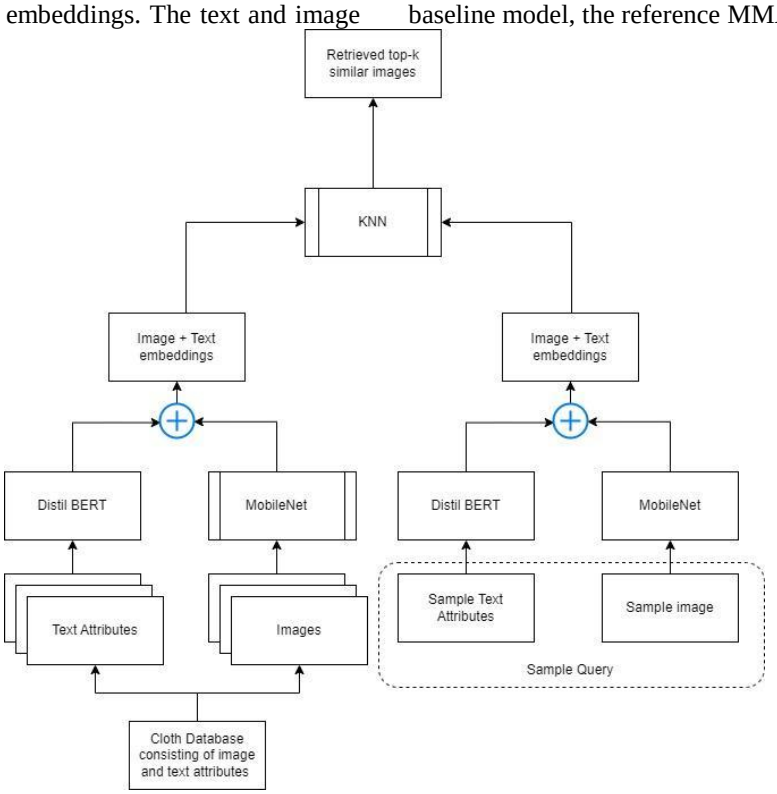


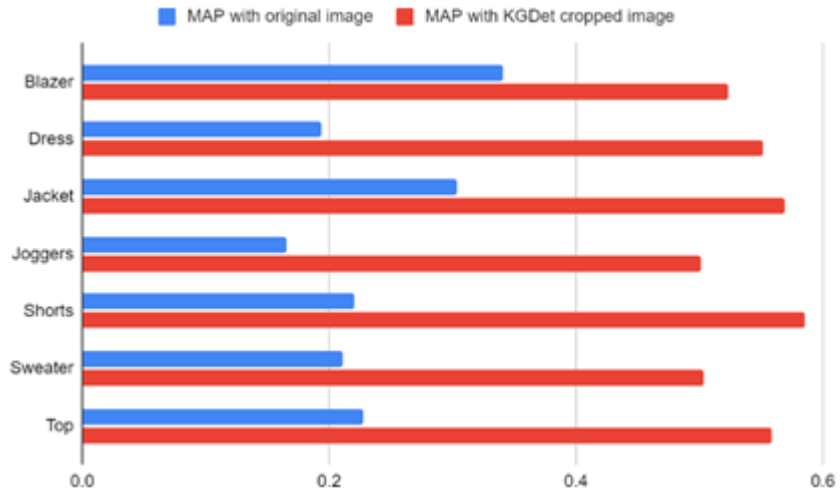
Figure 6: Baseline model Flow Diagram

Impact of segmenting cloth area in baseline model pipeline

KGDet is used to detect the cloth area in the image and is cropped to get rid of the background. The performance of KGDet was evaluated with the DeepFashion dataset and resulted in 0.65 mAP for DeepFashion dataset images for cloth detection (object detection). From the bounding box predicted for an image, the bounding box with the highest confidence score greater than 0.4 is selectively chosen refer to figure 7. This bounding box is then used to crop the cloth region.

Figure 7: KGDet bounding box predictions

The impact of the cropping cloth area from the image on MMAP score is evaluated. For this evaluation, with the same baseline model, the image embeddings are generated from the cloth area segmented image using KGDet. Then the KNN model is used to predict similar images and MAP is generated for each attribute. The MAP score is compared with the original full image and cloth area segmented image and is presented in figure 8. Significant improvement of MMAP score of 0.304 is achieved. This step is used as image preprocessing step for both the two proposed methods in this research.



4. Proposed Methods

1) Multi-Modal query based retrieval using CLIP (Contrastive Language-Image Pre-training) with Vision Transformer (ViT)

In this proposed system, KGDet is used for identifying clothing area and the region is cropped for further downstream tasks. For generating embeddings into a space where a similar style of clothes, having similar attributes will be closer in distance, CLIP is used in this architecture. Image embeddings are generated by Resnet 50 with architectural modifications proposed by (Radford et al., 2021) as follows (a) The global average pooling layer is replaced with an attention pooling mechanism. (b) For text encoder, a Transformer (Vaswani et al., 2017) with the architecture modified as described in Radford et al. (2019) is used. As a base size, a 63M-parameter 12-layer 512-wide model with 8 attention heads is used. The transformer operates on a lower-cased byte pair encoding (BPE) representation of the text with a 49,152-vocab size as stated in (Sennrich et al., 2015). The max sequence length is squeezed to 77. The text sequence is padded with a [SOS] and [EOS]. For generating embeddings from a multi-modal space, the model is trained with concatenated image and text features with triplet loss. (c) CLIP pre-trained model ViT-L/14, which is fine-tuned at a 336 Pixels higher resolution with an additional 1 epoch as stated by (Radford et al., 2021). Hence, the ViT-L/14 model is used to train image and text concatenated feature embeddings with Triplet loss for similarity learning (refer figure 11 for a sample triplet). Illustration of embeddings in similarity matching space is presented in figure 10.

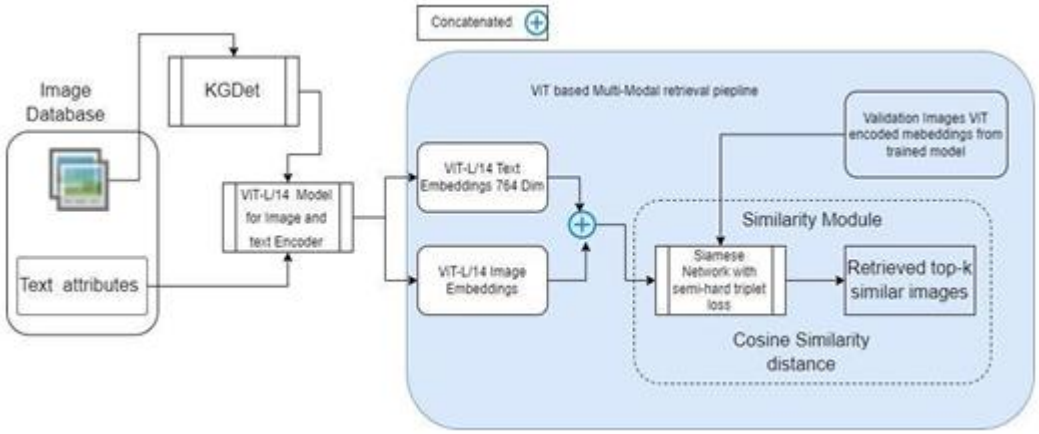


Figure 9: ViT based multi-modal query retrieval pipeline

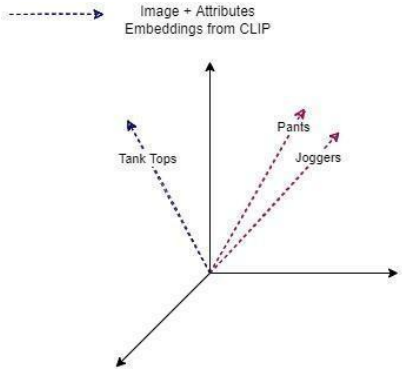


Figure 10: Illustration of feature vectors plotted in CLIP embedding space

The following are the best hyperparameters selected for training and pytorch was used for implementing the architecture:

1) Pre-trained text Encoder:
   a) Max sequence or context length: 77
   b) Encoded text is padded with "<|startoftext|>" and "<|endoftext|>"
2) Pre-trained Image Encoder:
   a) Resnet50 with architectural modifications specified by(Radford et al., 2021c)
3) Dataloader: Data is organized with the image path, text_info (attributes information) and its style group.
   a) Dataset is converted to the torch data loader object.
   b) The batch size is 8
   c) The number of workers is 2
   d) The sampler used is called "Same Group Sampler" for picking the sample for triplet semi-hard loss.
   e) Embedding size for ViT-L/14: 768
4) Model hyper parameters:
   a) Epochs: 20
   b) Optimizer : SGD (Stochastic Gradient Descent) , LR = 1e-2, momentum = 0.2
   c) Learning Rate Scheduler : "OneCyleLR" for optimizer , steps = Epochs * (2*len(train_data)-1) momentum=0.0, max_momentum=0.5, pct_start=0.1, div_factor=1e2, final_div_factor=1e4.
   d) Loss Criterion: Triplet-Semi-Hard Loss, the triplets consist of a negative sample that is more distant from the anchor than the positive sample, yet the result is still a positive loss.
   e) Similarity Module : Top-k, k = 10 is calculated between l2 normalized feature embeddings, and top matches are retrieved and evaluated for MMAP, F1- score.

Figure 11: Triplet Loss calculated for anchor with Positive and negative image

Zero-Shot Learning for image retrieval with CLIP

In this proposed system, the ZSL pipeline enable cross-domain retrieval through text query for unseen data. The design of this proposed text query based CDR (cross-domain retrieval) pipeline is mentioned in this section. As mentioned earlier, KGDet is used for image segmentation as the image preprocessing step. From the preprocessed image, embeddings are generated by ResNet50 image encoder. Each image is encoded into a fixed vector of size 2048, since ResNet50 is used. Global average pooling 2D layer is added for the output layer. Text embeddings are generated by DistilBERT text encoder. The text embedding vector size is 768 in the CLS token representation. The image and text encoder are set to trainable in the CLIP model. log_softmax is loss used for calculating cross entropy loss for image and text embeddings. To bring the 2048 dimensional image vectors and 768 dimensional text vectors, a projection head network is required. Projection network with output vector of size 256 is used.This projection head creates a spatial embedding space wherein the text and its closest matching images will be closer in the embedding vector space. The pipeline with Zero shot learning model for retrieving image with unseen attribute data is represented in the below figure 12.

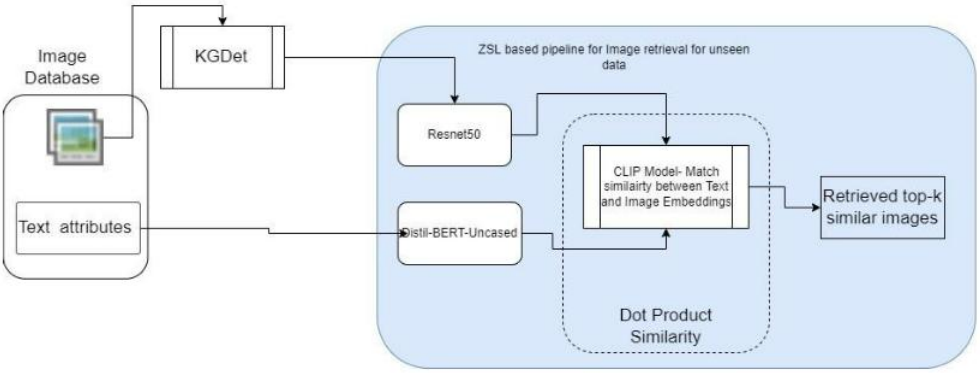


Figure 12: Zero-shot learning pipeline for Image retrieval with unseen text data

Following are the hyper parameter selected for the model (CLIP):

a) Batch size - 4
b) Head_lr - 1e-3
c) Image_encoder_lr - 1e-4
d) Text_encoder_lr - 1e-5
e) weight_decay - 1e-3
f) Factor - 0.8

g) Epochs – 20 with early stopping to avoid overfitting
h) image_embedding_size - 2048
i) text_encoder_model - "distilbert-base-uncased"
j) image_encoder_model - “resnet50”
k) text_embedding - 768
l) text_tokenizer - "distilbert-base-uncased"
m) max_length - 200
n) image size – 224
o) Dot product similarity is used for loss calculation and finding matches.
p) For projection layers used in both text and image encoders, num_projection_layers - 1 , projection_dim - 256, dropout - 0.1

5. Results

1) Improved retrieval accuracy for multi-modal query based retrieval system with vision transformer

During the literature survey, existing research works on Fashion Image Retrieval system on cross domain methodologies were reviewed. To the best of our knowledge, research on impact of cloth area segmentation in cross-domain retrieval was not available. Based on this research clothing area segmentation can be used as a image preprocessing step in existing retrieval pipeline that will further boost the retrieval accuracy. This is mainly due to the attentive feature generation of cloth area from the image. Achieved significant improvement of the ↑0.304 MMAP score across all 7 attributes after one step of clothing area segmentation as image preprocessing step before the downstream processes in the proposed multi-modal query based retrieval pipeline. This result is visualized be calculating the Precision @ K, K in 10 to 100 in steps of 10 plotted in figure 13. This plot shows the improvement of MMAP score, when the cloth area is cropped with KGDet in the baseline model retrieval pipeline. With this result, the cloth area cropping was added in the multi-modal query based retrieval system with vision transformer and the MMAP score across all the attributes are shown in the table 1

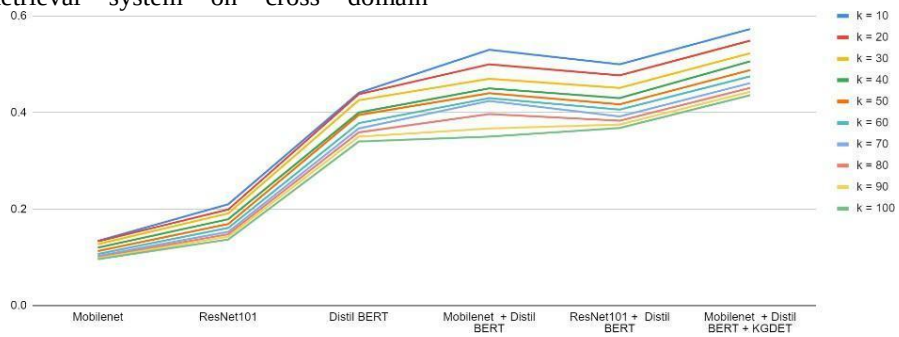


Figure 13: Comparison of P@K with feature embeddings

Table 1: Comparision of baseline model Vs ViT model for multi-modal query

Categories	MAP with KGDet cropped image	MAP with KGDet cropped image and Multi-Modal query generated using Vision Transformer Model
Blazer	0.523	**0.671**
Dress	0.551	**0.773**
Jacket	0.569	**0.813**
Joggers	0.501	**0.722**
Shorts	0.585	**0.634**
Sweater	0.504	**0.571**
Top	0.558	**0.639**

3) Effectiveness of Zero shot learning algorithm for Fashion Image Retrieval system in Cross-Domain for unseen dataset

The proposed ZSL based image retrieval pipeline enabled the retrieval of unseen data point based on custom query. Custom query can describe a dress based on multiple attributes like texture, style, sleeve type or collar type etc. This helps user to retrieve required dress from the shop database even in the absence of sample image from the product database or from customer. ZSL also leverages the advantage of meta-learning by reducing the requirement of huge training dataset. This is a major advantage for handling a data shift due to addition of new clothing styles periodically to the clothing database. The significance of ZSL with given text query for image retrieval is evaluated with MMAP score on unseen dataset across 7 attributes and presented in Table 1. MMAP score is based on top 5 retrieval results and one of the retrieval result for a given text query with multiple attribute is shown in figure 14.

Table 2: MMAP for KGDet

Categories	MMAP for KGDet ZSL
Blazer	0.746
Dress	0.725
Jacket	0.870
Joggers	0.944
Shorts	0.728
Sweater	0.812
Top	0.741

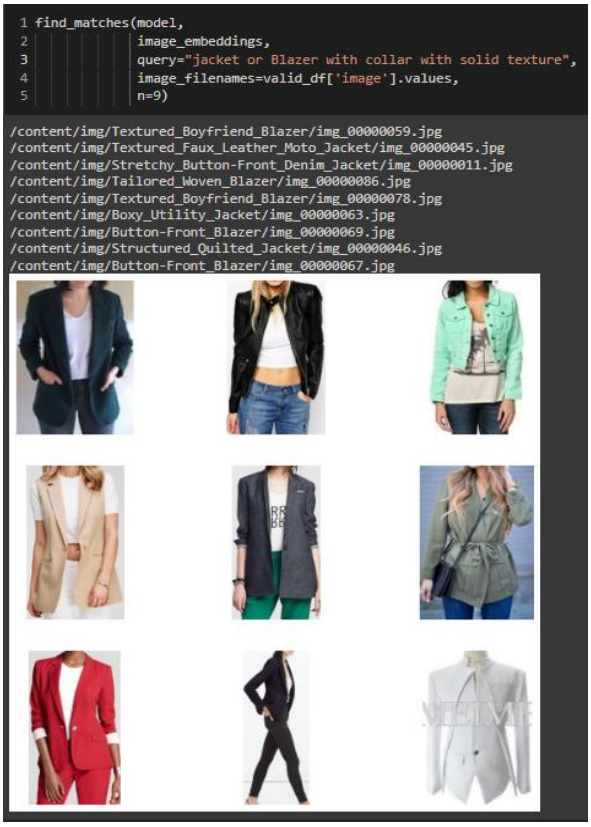


Figure 14: Output with Example query "jacket or Blazer with collar with solid texture"

6. Conclusion

In this research, solutions for the two major challenges of improving retrieval accuracy for multi-modal query for generalisability and on unseen data in cross-domain are proposed. These two proposed solution approach can be applied independently or combined depending upon the business use case.

7. Future Scope

The hyper parameters for ZSL model can be explored with MMAP score as evaluation criteria during hyperparameter optimization. Few-Shot learning with prototypical network can be researched for cross-domain retrieval. Fine-Tuning BERT with text attributes will improve the significance of text embeddings.

References

[1] Chen, D., Chen, P., Yu, X., Cao, M. and Jia, T., (2019) Deeply-Learned Spatial Alignment for Person Re-Identification. IEEE Access, 7, pp.143684–143692.

[2] Chen, W., Chen, X., Zhang, J. and Huang, K., (2017) Beyond triplet loss: a deep quadruplet network for person re-identification. [online] Available at: http://arxiv.org/abs/1704.01719.

[3] Chen, W., Liu, Y., Wang, W., Bakker, E., Georgiou, T., Fieguth, P., Liu, L. and Lew, M.S., (2021) Deep Image Retrieval: A Survey. [online] Available at: http://arxiv.org/abs/2101.11282.

[4] Ciaccia, Paolo., Patella, Marco. and Association for Computing Machinery., (2010) SISAP 2010 : proceedings, Third International Conference on Similarity Search and Applications : 18- 19 September 2010, Istanbul, Turkey. Association for Computing Machinery.

[5] Csurka, G. and Perronnin, F., (2011) Fisher vectors: Beyond bag-of-visual-words image representations. In: Communications in Computer and Information Science. pp.28–42.

[6] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H. and Wei, Y., (2017) Deformable Convolutional Networks. [online] Available at: http://arxiv.org/abs/1703.06211.

[7] DAVIES, E.R., (2005) Image Acquisition. Machine Vision, [online] pp.781–803. Available at: https://linkinghub.elsevier.com/retrieve/pii/B9780122060939500307 [Accessed 25 Oct. 2021]. Deng, D., Wang, R., Wu, H., He, H., Li, Q. and Luo, X., (2017) Learning Deep Similarity Models with Focus Ranking for Fabric Image Retrieval. [online] Available at: http://arxiv.org/abs/1712.10211.

[8] Elezi, I., Vascon, S., Torcinovich, A., Pelillo, M. and Leal-Taixe, L., (2019) The Group Loss for Deep Metric Learning. [online] Available at: http://arxiv.org/abs/1912.00385.

[9] Faria, F.F., Veloso, A., Almeida, H.M., Valle, E., Torres, R. da S., Gonçalves, M.A. and Meira, W., (2010) Learning to Rank for Content-Based Image Retrieval. In: Proceedings of the International Conference on Multimedia Information Retrieval, MIR '10. [online] New York, NY, USA: Association for Computing Machinery, pp.285–294. Available at: https://doi.org/10.1145/1743384.1743434.

[10] He, K., Gkioxari, G., Dollár, P. and Girshick, R., (2017a) Mask R-CNN. [online] Available at: http://arxiv.org/abs/1703.06870.

[11] He, K., Gkioxari, G., Dollár, P. and Girshick, R., (2017b) Mask R-CNN. [online] Available at: http://arxiv.org/abs/1703.06870.

[12] He, K., Zhang, X., Ren, S. and Sun, J., (2015) Deep Residual Learning for Image Recognition. [online] Available at: http://arxiv.org/abs/1512.03385.

[13] Hermans, A., Beyer, L. and Leibe, B., (2017) In Defense of the Triplet Loss for Person Re-Identification. [online] Available at: http://arxiv.org/abs/1703.07737.

[14] Hoffer, E. and Ailon, N., (2014) Deep metric learning using Triplet network. [online] Available at: http://arxiv.org/abs/1412.6622.

[15] Hu, Y., Yi, X. and Davis, L.S., (2015) Collaborative fashion recommendation: A functional tensor factorization approach. In: MM 2015 - Proceedings of the 2015 ACM Multimedia Conference. Association for Computing Machinery, Inc, pp.129–138.

[16] Huang, F., Wang, Z., Wu, J., Shen, Y. and Chen, L., (2021) Residual triplet attention network for single-image super-resolution. Electronics (Switzerland), 1017.

[17] Khaertdinov, B., Ghaleb, E. and Asteriadis, S., (2021) Deep Triplet Networks with Attention for Sensor-based Human Activity Recognition. In: 2021 IEEE International Conference on Pervasive Computing and Communications, PerCom 2021. Institute of Electrical and Electronics Engineers Inc.

[18] Koch, G., Zemel, R. and Salakhutdinov, R., (n.d.) Siamese Neural Networks for One-shot Image Recognition.
[19] Li, A., Liu, L., Wang, K., Liu, S. and Yan, S., (2015) Clothing Attributes Assisted Person Reidentification. IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, [online] 255. Available at: http://www.ieee.org/publications_standards/publications/rights/index.html [Accessed 31 Oct. 2021].
[20] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S., (2016) Feature Pyramid Networks for Object Detection. [online] Available at: http://arxiv.org/abs/1612.03144. Lin, T.-Y., Goyal, P., Girshick, R., He, K. and Dollár, P., (2017) Focal Loss for Dense Object Detection. [online] Available at: http://arxiv.org/abs/1708.02002.
[21] Liu, Z., Luo, P., Qiu, S., Wang, X. and Tang, X., (2016a) DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, pp.1096–1104.
[22] Liu, Z., Luo, P., Qiu, S., Wang, X. and Tang, X., (2016b) DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp.1096–1104.
[23] Martinsson, J. and Mogren, O., (n.d.) Semantic Segmentation of Fashion Images Using Feature Pyramid Networks.
[24] Murala, S. and Wu, Q.M.J., (2014) Local mesh patterns versus local binary patterns: Biomedical image indexing and retrieval. IEEE Journal of Biomedical and Health Informatics, 183, pp.929– 938.
[25] Ong, E.-J., Husain, S. and Bober, M., (2017) Siamese Network of Deep Fisher-Vector Descriptors for Image Retrieval. [online] Available at: http://arxiv.org/abs/1702.00338.
[26] Qian, S., Lian, D., Zhao, B., Liu, T., Zhu, B., Li, H. and Gao, S., (2021a) KGDet: Keypoint- Guided Fashion Detection. [online] Available at: www.aaai.org.
[27] Qian, S., Lian, D., Zhao, B., Liu, T., Zhu, B., Li, H. and Gao, S., (2021b) KGDet: Keypoint- Guided Fashion Detection. [online] Available at: www.aaai.org.
[28] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I., (2021a) Learning Transferable Visual Models From Natural Language Supervision. [online] Available at: http://arxiv.org/abs/2103.00020.
[29] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I., (2021b) Learning Transferable Visual Models From Natural Language Supervision. [online] Available at: http://arxiv.org/abs/2103.00020.
[30] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I., (2021c) Learning Transferable Visual
[31] Models From Natural Language Supervision. [online] Available at: http://arxiv.org/abs/2103.00020.
[32] Razavian, A.S., Sullivan, J., Carlsson, S. and Maki, A., (2014) Visual Instance Retrieval with Deep Convolutional Networks. [online] Available at: http://arxiv.org/abs/1412.6574.
[33] le Roux, N., Lecun, Y., Reprint, S., Behnam, H., Hadsell, R. and Chopra, S., (n.d.) Dimensionality Reduction by Learning an Invariant Mapping Related papers Learning t he 2-D Topology of Images Scaling learning algorit hms t owards AI Dimensionality Reduction by Learning an Invariant Mapping. [online] Available at: http://www.cs.nyu.edu/∼yann.
[34] Sabokrou, M., Fayyaz, M., Fathy, M., Moayed, Zahra. and Klette, R., (2018) Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. Computer Vision and Image Understanding, [online] 172, pp.88–97. Available at: https://www.sciencedirect.com/science/article/pii/S1077314218300249.
[35] Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D. and Lillicrap, T., (2016) Meta-Learning with Memory-Augmented Neural Networks. In: Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, ICML'16. JMLR.org, pp.1842– 1850.
[36] Schroff, F., Kalenichenko, D. and Philbin, J., (2015) FaceNet: A Unified Embedding for Face Recognition and Clustering. [online] Available at: http://arxiv.org/abs/1503.03832.
[37] Simonyan, K. and Zisserman, A., (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. [online] Available at: http://arxiv.org/abs/1409.1556.
[38] Simo-Serra, E., Fidler, S., Moreno-Noguer, F. and Urtasun, R., (n.d.) Neuroaesthetics in Fashion: Modeling the Perception of Fashionability. [online] Available at: http://www.iri.upc.edu/people/esimo/research/.
[39] Simo-Serra, E. and Ishikawa, H., (2016) Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, pp.298–307.
[40] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., (2014) Going Deeper with Convolutions. [online] Available at: http://arxiv.org/abs/1409.4842.
[41] Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.-J., Fei-Fei, L. and Hays, J., (2019a) Composing Text and Image for Image Retrieval - an Empirical Odyssey. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp.6432–6441.
[42] Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L. and Hays, J., (2019b) Composing text and image for image retrieval-An empirical odyssey. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, pp.6432–6441.
[43] Walter, K., Gibson, M.J. and Sowmya, A., (2020) Self-

Supervised Remote Sensing Image Retrieval. In: International Geoscience and Remote Sensing Symposium (IGARSS). Institute of Electrical and Electronics Engineers Inc., pp.1683–1686.
[44] Wei, J., Huang, C., Vosoughi, S., Cheng, Y. and Xu, S., (n.d.) Few-Shot Text Classification with Triplet Networks, Data Augmentation, and Curriculum Learning. [online] Available at: https://github.
[45] Wu, C.-Y., Manmatha, R., Smola, A.J. and Krähenbühl, P., (2017) Sampling Matters in Deep Embedding Learning. [online] Available at: http://arxiv.org/abs/1706.07567.
[46] Wu, D., Zheng, S.J., Bao, W.Z., Zhang, X.P., Yuan, C.A. and Huang, D.S., (2019) A novel deep model with multi-loss and efficient training for person re-identification. Neurocomputing, 324, pp.69–75.
[47] Xiao, Q., Luo, H. and Zhang, C., (2017) Margin Sample Mining Loss: A Deep Learning Based Method for Person Re-identification. [online] Available at: http://arxiv.org/abs/1710.00478.
[48] Xiao, T., Xia, T., Yang, Y., Huang, C. and Wang, X., (n.d.) Learning from Massive Noisy Labeled Data for Image Classification.
[49] Xie, S., Girshick, R., Dollár, P., Tu, Z. and He, K., (2016) Aggregated Residual Transformations for Deep Neural Networks. [online] Available at: http://arxiv.org/abs/1611.05431.
[50] Xu, B., Luo, S. and Sun, K., (2012) Towards multimodal query in web service search. In: Proceedings - 2012 IEEE 19th International Conference on Web Services, ICWS 2012. pp.272–279.
[51] Xue, N., Wang, Y., Fan, X. and Min, M., (n.d.) INCREMENTAL ZERO-SHOT LEARNING BASED ON ATTRIBUTES FOR IMAGE CLASSIFICATION.
[52] Yan, C., Ding, A., Zhang, Y. and Wang, Z., (2021a) Learning Fashion Similarity Based on Hierarchical Attribute Embedding. In: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA). pp.1–8.
[53] Yan, C., Ding, A., Zhang, Y. and Wang, Z., (2021b) Learning Fashion Similarity Based on Hierarchical Attribute Embedding. Institute of Electrical and Electronics Engineers (IEEE), pp.1– 8.
[54] Yang, Z., Liu, S., Hu, H., Wang, L. and Lin, S., (2019) RepPoints: Point Set Representation for Object Detection. [online] Available at: http://arxiv.org/abs/1904.11490.
[55] Yuan, C., Guo, J., Feng, P., Zhao, Z., Xu, C., Wang, T., Choe, G. and Duan, K., (2019) A jointly learned deep embedding for person re-identification. Neurocomputing, 330, pp.127–137.
[56] Zhang, C., Liu, W., Ma, H. and Fu, H., (n.d.) SIAMESE NEURAL NETWORK BASED GAIT RECOGNITION FOR HUMAN IDENTIFICATION. [online] Available at:
[57] http://caffe.berkeleyvision.org/model.
[58] Zhang, R., Lin, L., Zhang, R., Zuo, W. and Zhang, L., (2015) Bit-Scalable Deep Hashing with Regularized Similarity Learning for Image Retrieval and Person Re-identification. [online] Available at: http://arxiv.org/abs/1508.04535.
[59] Zhang, S., Wen, L., Bian, X., Lei, Z. and Li, S.Z., (2017) Single-Shot Refinement Neural Network for Object Detection. [online] Available at: http://arxiv.org/abs/1711.06897.
[60] Zhang, Z., Ma, S. and Zhang, Y., (2021) Will Multi-modal Data Improves Few-shot Learning? [online] Available at: http://arxiv.org/abs/2107.11853.
[61] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang and Xiaoou Tang, (2021) Large-scale Fashion (DeepFashion) Database. [online] Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Available at: http://mmlab.ie.cuhk.edu.hk/projects/DeepFashion.html [Accessed 14 Nov. 2021].

Author Profile

Swathy S, experienced Data Scientist providing AI solutions across different domains. Currently working at Intellect Design Arena as Data Scientist and holding a master's degree from Liverpool John Moores University with specialization in Artificial Intelligence