Hybrid Style Siamese Network: Incorporating style loss in complimentary apparels retrieval

Mayukh Bhattacharyya
mayukhbhattacharyya18@gmail.com

Sayan Nag
nagsayan112358@gmail.com

Abstract

Image Retrieval grows to be an integral part of fashion e-commerce ecosystem as it keeps expanding in multitudes. Other than the retrieval of visually similar items, the retrieval of visually compatible or complimentary items is also an important aspect of it. Normal Siamese Networks tend to work well on complimentary items retrieval. But it fails to identify low level style features which make items compatible in human eyes. These low level style features are captured to a large extent in techniques used in neural style transfer. This paper proposes a mechanism of utilising those methods in this retrieval task and capturing the low level style features through a hybrid siamese network coupled with a hybrid loss. The experimental results indicate that the proposed method outperforms traditional siamese networks in retrieval tasks for complimentary items.

1. Introduction

Fashion e-commerce is a booming industry right now. According to studies, the industry is $500 billion strong. A big factor which has led to its growth and success is the large inventory it can offer to the customers. As an unwanted by-product of this advantage is the problem of plenty. How to effectively select and present the items which the customer will prefer to buy has been the problem. The solution to this has been recommendation systems. The prevalent recommendation systems has been mostly content based where we pick up their brands, colors, type and other metadata to recommend new unseen items, in this case, clothes. Latest works on this problem have been using image retrieval techniques to find out similar items from the pool of all items. So as an example, if an user looks at a pink striped shirt, it will most likely offer the user other variations of pink and striped apparels. That may be a very good solution for same apparels but this idea does not work well if the user is looking for complementary apparels. In the above case, a pink or striped trouser may not be the best choice of style to go with a pink striped shirt.

There have been several studies involving traditional image retrieval techniques but very few have approached the problem of finding the complementary clothing of a certain apparel. Finding a complementary match to a certain clothing is a subjective problem to some extent. The sense of style and fashion varies from person to person, so there is no single correct answer. But based on peoples choices of combinations of clothings, we can obtain a general sense of style of people. The closest works have been [1], [2], [3] and [4] where they have mostly used variations of Siamese Networks for this problem. This paper tries to form a general sense of this style from combinations of apparels worn by people in real life and proposes and evaluates techniques based on their retrieval of complimentary apparels.

The contribution of this paper is two-fold:
1. It proposes a Hybrid Siamese Network model which takes into account the style information of the input images to better aid in retrieval of complementary fashion items.
2. It introduces a style loss function suited to image retrieval applications which coupled with traditional triplet loss leads to better performance of models in complimentary items retrieval.

2. Related Work

Siamese Network was introduced in [5] and in the recent years, it is being extensively used for image retrieval. The basic advantage of the Siamese setup that it can convert an image similarity problem to a latent space of metric distances has been exploited with use cases from varied domains. Perhaps the domain that led to the fame of Siamese Network is facial recognition where Chopra et al. [6] and Hu et al. [7] first used it in this context. Although the main idea of Siamese Networks being that images with certain degree of similarity in certain context will lie closer in the latent space, it was shown in recent years that Siamese Networks can indeed be used in context of compatibility too. Bell and Bala [8] first explored the stylistic similarity of items in images through Siamese Networks. The idea was extended to the domain of clothing style compatibility by Veit et al. [1]. Veit et al. first tackled the problem of retrieval of clothes based on compatibility in style instead of retrieval of stylistically similar items. It is the closest work...
to our paper in this domain of research. Similar to our work, they aimed at learning a notion of style and retrieving cloths from different categories that will be compatible with each other. They base their experiments on data collected from Amazon.com and create positive and negative pairs based on user interests estimated from their activities. Although after [1], quite a few work has been done in this domain, such as [2], [4]; those are variations and modifications of [1] and mostly evaluate on the same dataset as [1]. Vasileva et al. in [2] used type-aware latent transformations to better model the Siamese networks representation of items of different types. In [4], color histograms information is introduced in the network and a single network is formed by connecting the Siamese network at the top to turn it into a classification problem of predicting positive or negative for an input pair of two clothes. It is evaluated on the basis of LIFT metric on the Polyvore dataset. In Gatys et al. [9], neural style transfer was introduced which aimed at creating artificial artistic images by imposing a style image over a content image. It served its way by optimizing a loss function comprised of a content loss and a style loss. The work in this paper follows the Siamese structure of [1] but incorporates secondary ideas from this work [9] in regards to estimating the style loss part. In [9], the style loss is calculated by means of a gram matrix, which discards the spatial or content information and only stores the style information from the images. A similar approach has been taken in constructing a style loss as a part of triplet loss described in this paper.

3. Methodology

3.1. Siamese Networks

Siamese Networks [5][6][7] involve two or three identical networks which share their weights between them. Those can be tied at the top by a dense layer to provide some predictions or can be used separately to obtain n-dimensional representations or embeddings of the images in a latent space. The later case is predominantly used in the applications of Siamese networks with a distance metric between the end layer embeddings of the networks used as the loss.

3.2. Hybrid Style Siamese Network

Our modified Siamese Network follows the same basic backbone of any Siamese Network. Over that, we add a style network to implement our style loss. As can be seen in fig[1], outputs from the 4 initial convolutional feature layers are taken out and each such output is passed through a style network to produce 4 auxiliary outputs which are then ingested by our style loss. The choice of using 4 initial layers for auxiliary outputs is found to perform best for VGG16. However it may be different for other backbone architectures. Apart from these 4 auxiliary outputs, the Hybrid Siamese Network still has the main output from the final embedding layer of the original network. It is only this output that is considered for any evaluation purposes while the auxiliary outputs are used only in the style loss.

3.3. Style Network

Each chosen initial layer feature maps are passed through a style network. The style network first applies a batch normalization layer over the incoming feature map. After which, a gram-matrix of the features is computed in-network. Finally the computed gram-matrix layer is flattened or unrolled into a dense fully connected layer, which is then provided as an auxiliary output to the model. The gram-matrix and formulations are explained in detail in section 3.5.

The batch normalizations are critical for the style features since it limits them to a feature space of mean 0 and variance 1, thus ensuring that the auxiliary outputs and style component of the hybrid triplet loss remains within bounds and more importantly the gradient doesn’t explode. In absence of the batch normalization layers the style component of the loss goes out of bounds very quickly and thus could not have been used as a loss function.

A slightly different style network structure is also possible where the batch normalization layer is applied after the gram-matrix layers, just before the formation of auxiliary outputs. However the previous structure is favoured because of better performances.

3.4. Triplet Loss

Triplet loss was introduced in [10] and involves calculating the loss among 3 images, namely anchor image A, positive image P and negative image N. (A,P) form a positive pair (conveying a sense of similarity in most use cases) while (A,N) form a negative pair (conveying a sense of dissimilarity in most cases). In general, triplet loss combines 2 distance metrics; one between A and P, d(A,P) and the other between A and N, d(A,N). In majority of the cases, the distance metric used is the Euclidean distance, where d(A,P) represent the Euclidean distance between the latent n-dimensional representations of A and P.

\[
d(x, y) = \left[ \sum_{i=1}^{N} (x_i - y_i)^2 \right]^{\frac{1}{2}}
\]  

(1)

Triplet loss strives to decrease the distance \(d(A, P)\) and increase \(d(A, N)\). A certain margin \(\alpha\) is set to separate the two distances with the positive pair distance being lower. For all cases where the loss becomes negative, it is considered as zero.

\[
L(A, P, N) = \max(\alpha + d(A, P) - d(A, N), 0)
\]  

(2)
In our case here, instead of similarity, (A,P) will form a compatible pair whereas (A,N) will be a non-compatible pair.

3.5. Style Loss

Style information from a layer is measured as the amount of correlation present between features maps of the channels in the given layer. This correlation information is computed through a gram-matrix calculated from the feature maps. Gatys et. al. [9] defines the gram-matrix $GM$ for layer $l$ of an image $x$ such that $GM(x)_{ij}$ is the inner product between the vectorized feature maps of channel $i$ and $j$ respectively.

$$GM(x)_{ij} = \sum_k F(x)_{ik} F(x)_{kj}$$  \hspace{1cm} (3)

The style loss is defined as the difference in the gram matrices for a layer evaluated between two images. In neural style transfer applications, these two images are usually the style image and the generated image. The style loss is then used to make the generated image visually similar to the style image by learning the required low-level features to represent the style image.

$$L^s_l = \frac{1}{4n_l^2m_l^2} \sum_{i,j} (GM(G)_{ij} - GM(S)_{ij})^2$$ \hspace{1cm} (4)

Here the style loss for layer $l$, $L^s_l$ is calculated as a difference between the gram matrices of the generated image $G$ and style image $S$. $n_l$ gives the number of channels in the layer $l$ and $m_l$ is the height\*width of the channels of the layer.

3.6. Hybrid Triplet Loss

For the proposed hybrid siamese network described earlier, a hybrid version of the triplet loss is employed which combines the traditional triplet loss explained above with a modified version of the style loss. Contrary to the use-case in neural style transfer, we use the style loss in a different way. We use a negative style loss between two images to help the network learn low-level features that can demarcate between different visual styles of two images even though they may have the same type of content. In essence, the loss forces the distance between the gram-matrices of the two images to be as large as possible. Since while evaluating compatibility of apparels, people use style as the main yardstick, learning to identify different styles in clothings will help the model to better learn the sense of compatibility.

We derive the style loss for our case as a modification of the traditional one shown in eq.(4)

$$L^s_l(P,N) = K - K_l \frac{1}{4n_l^2m_l^2} \sum_{i,j} (GM(P)_{ij} - GM(N)_{ij})^2$$ \hspace{1cm} (5)

Here, the style loss is computed between the positive and negative images since they are expected to be stylistically different. $K$ and $K_l$ are constants. Based on experiments, the value of $K$ is chosen to be 2. The value of $K_l$ is chosen to be $m_l$ for a certain layer $l$. We compute this loss over
the models initial layers output features gathered from the positive and negative images of the triplets. Combining the style loss with the traditional triplet loss, we thus obtain a hybrid triplet loss.

\[ L_h(A, P, N) = w_1 * L(A, P, N) + w_2 * \sum_l L^l_s(P, N) \]  

Hybrid triplet loss, \( L_h(A, P, N) \) is thus a weighted sum of the traditional triplet loss, \( L(A, P, N) \) and the sum of the style losses, \( L^l_s(P, N) \) over all the chosen initial layers. Here \( w_1 \) and \( w_2 \) are the corresponding weights.

It should be noted here that for an image \( x \), the style network provides a flattened version of \( GM(x) \) as output for each layer which serve as auxiliary outputs to the main network. But there is no any clear demarcation between between computing the gram-matrices inside the style network or inside the style loss, since there are no learnable parameters after the batch normalization layer. However, the style network version is chosen here since the batch normalization layer can also be applied after the gram-matrix computation. The possible variability in the location of the batch normalization layer makes the style network representation far more reasonable.

4. Experiments

4.1. Dataset

In this paper, the model is evaluated on the iMaterialist [11] dataset. It contains real life images of people wearing clothes and corresponding category and segmentation details of the clothes in the images. For the experiments, among the category combinations available, images having people wearing both skirts and t-shirts have been selected. The combination of skirts & t-shirts were chosen because it enables the model to capture a bigger extent of the variability in style. This would have been problematic in case of other categories like trousers and pants where there are only a limited few types of designs and styles prevalent. There are a total of 3595 images having the valid (skirt & t-shirt) combination in the iMaterialist dataset.

4.2. Preprocessing

Since the datasets comprise of people in real life, in order to create pairs of images (each containing an apparel/type of dress) from the main image, we need to segregate the image. The pixel level segmentation annotations are used for each apparel in an image to generate masks. The masks serve the purpose of not only capturing the portion of the image containing the apparel but also eliminating the surrounding scene information of the images so that it does not influence the model. In a sense, this serves the purpose of preventing the model from learning latent scene-based features for matching apparels. All the masked apparel images are then resized to 128x128 to generate the final training data.

4.3. Experimental Settings and Details

The Hybrid Style Siamese Network is compared with the normal Siamese network described in Veit et. al. [1]. To ensure a fair comparison, both the models were implemented with a VGG16 backbone. Pretrained weights trained on ImageNet were used as initialization for all the compared models. 4 initial layers were chosen for style loss computation from the VGG16 backbone, namely the 1st, 2nd, 3rd and 5th layers. Triplet loss with Adam optimizer is used for training purposes. To create the triplets, the anchor and the positive images are selected by pairing masked apparel images of skirt and t-shirt worn by the same person in an image. The negative is selected by choosing a different image and using the masked apparel image of the skirt in it if the anchor is a t-shirt or a t-shirt if the anchor is a skirt. Since, the number of negative combinations possible is very large, the triplets are selected at random. Due to the existing randomness, in order to make the experiment robust, the experiment is carried out 3 times with 3 different random seeds and each trial is comprised of a 5-Fold cross validation. Two different learning rate decay schedules were tested and the best performing results was chosen for each of the models for each run.

4.4. Evaluation Metric

The MAP (Mean Average Precision) metric is used for evaluating the retrieval performance. During evaluation, each typeA- typeB pair (skirts and t-shirts in this case) are fed to the network and scored. Then for each typeA apparel, we rank all the apparels of typeB. Similarly for each typeB apparel, the apparels of typeA type are ranked. Now for each apparel image of typeA type, there will be only a single correct apparel image of typeB type which was part of the same actual image. This means the precision score of each typeA apparel reduces to \( 1/(\text{rank of correct typeB}) \) and vice versa. The mean of the two MAP scores for each type of apparel in the apparel pair is considered as the MAP score for this pair.

\[ P_i^A = 1/R_i^B \]

Where \( P_i^A \) is the precision corresponding to the ith item of type A and \( R_i^B \) is the rank of ith item of type B among all the items of type B in terms of ascending order of distances.

\[ mAP = \frac{1}{2} \sum_i (P_i^A + P_i^B) \]

Since 5-fold cross validation was used, the MAP scores are computed on 20% of the dataset, i.e. on 719 samples.
4.5. Results

The results consolidating the MAP scores from the 3 trials with the 3 different random seeds are given in below table.

| Model                        | Seed 1  | Seed 2  | Seed 3  | Mean    |
|------------------------------|---------|---------|---------|---------|
| Siamese Network Veit et. al. | 0.1226  | 0.1323  | 0.1263  | 0.1271  |
| Hybrid Style Siamese Network | 0.1251  | 0.1343  | 0.1329  | 0.1308  |
| Improvement (%)              | 2.04    | 1.51    | 5.22    | 2.91    |

As can be seen, the Hybrid Style Siamese Network outperforms the traditional Siamese Network in all 3 of the trials. The extent of the improvement in performance varies due to the difference in randomness in the trials.

The mean validation curve of the 3 trials is shown in the figure below.

![Figure 2. Mean validation MAP scores of the experimental trials.](image)

Based on the results, it can be inferred that the style network and associated style loss provides the Siamese network extra feature information which makes a bigger difference in the later stages of training.

5. Conclusion and Future Work

This paper presented a new method of complementary clothes retrieval. The paper presented a novel way of encompassing the style loss of neural style transfer applications into the application of image retrieval. The presented method gave superior results to traditional method of siamese networks and thus showed how leveraging the style information is beneficial for the retrieval of visually complementary items.

Although the paper only explores the domain of complementary clothes retrieval, it can easily be inferred that this method will also be beneficial in normal non-complementary fashion retrieval scenarios where one finds items similar to a given one. Since the method emphasizes on patterns at the lower level, it can be expected that retrieval using this technique will be more uniformly dependant over all features and not just color combinations and broader patterns.

References

[1] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV ’15, pages 4642–4650, Washington, DC, USA, 2015. IEEE Computer Society.

[2] Mariya I. Vasileva, Bryan A. Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 405–421, Cham, 2018. Springer International Publishing.

[3] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *CoRR*, abs/1608.03016, 2016.

[4] Luisa F. Polanía and Satyajit Gupte. Learning fashion compatibility across apparel categories for outfit recommendation. *CoRR*, abs/1905.03703, 2019.

[5] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS ’93, pages 737–744, San Francisco, CA, USA, 1993. Morgan Kaufmann Publishers Inc.

[6] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05) - Volume 1 - Volume 01, CVPR ’05*, pages 539–546, Washington, DC, USA, 2005. IEEE Computer Society.
[7] J. Hu, J. Lu, and Y. Tan. Discriminative deep metric learning for face verification in the wild. In 2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 1875–1882, June 2014.

[8] Sean Bell and Kavita Bala. Learning visual similarity for product design with convolutional neural networks. ACM Trans. Graph., 34(4):98:1–98:10, July 2015.

[9] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[10] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. Large scale online learning of image similarity through ranking. J. Mach. Learn. Res., 11:1109–1135, March 2010.

[11] Sheng Guo, Weilin Huang, Xiao Zhang, Prasanna Srikhanta, Yin Cui, Yuan Li, Matthew R. Scott, Hartwig Adam, and Serge J. Belongie. The imaterialist fashion attribute dataset. CoRR, abs/1906.05750, 2019.