Fashion Fusion: Deep Learning-Based Outfit Recommendation

Sharada HP[1], Dr. Sandeep Bhat[2]
[1]*Mtech Student, Computer Science and Engineering, Srinivas Institute of Technology, Valachil, Mangaluru -575001 Karnataka, India*
[2]*Associate Professor, Computer Science and Engineering, Srinivas Institute of Technology, Valachil, Mangaluru -575001 Karnataka, India*
doi.org/10.64643/IJIRTV12I12-206798-459

***Abstract*—The increased growth of online fashion platforms expands the desire of smart systems that will provide personalized styling recommendations. This paper gives a deep learning–based framework for generating fashion outfit suggestions leads to individual user preferences. The present system tries to understand the compatibility between clothing products while all past methods were trying to find the compatibility based on similarity of the cloths and customer purchase history. There are different types of CNN models but here we use pre-trained CNN Resnet50. Shared space embedding and metric learning are the main two techniques which takes major role in identifying matching and non-matching outfits. All specific parts of the image are identified by attention mechanism. These attention mechanisms are also used to focus on relevant style parts. Experiments shows that the given method effectively calculate difficult relationships between fashion items and gives more relevant and visually matching things compared to conventional methods. So, this generated system can be applied in e-commerce and fashion applications to improve user experience by providing fully automatic, smart, and personalized styling assistance.**

***Index Terms*—Complementary Product Recommendation, cross domain recommendation visual compatibility learning, feature extraction, Triplet's loss.**

I. INTRODUCTION

Outfit recommendation is taking a major role in the field of E-Commerce because it is helping the customers to find the products that match their style. All users use their own selfie or street photo to find recommendations in today's life. This real-world photo is also called scene image. Each scene image is different in clothing combination, body posture and background. This paper concentrates on the concept of "Complete the Look" (CTL), which helps to find visually compatible products based on a given photo. In old product-to-product recommendation systems, this method uses scene-to-product compatibility, where the system observes the overall context of an image and suggests suitable items that improve the user's appearance. In order to achieve this, deep learning models such as Resnet50 convolutional neural network is used to get visual features from both scene and product images. In addition to this, attention mechanisms are taking important role in identifying important parts within the photo and gives most applicable compatibility decisions. By mixing both general and specific feature representations, the system generates a unique style space where suitable items are highly related to each other.

II. LITERATURE SURVEY

Existence of new research in computer vision and deep learning are strongly influencing the fashion suggestion system. All traditional approaches use older versions of convolutional neural network to identify the clothing images. These CNN models are helpful in gaining visual features and tries to address the problem of clothes separation and suitable attribute finding. These all-old methods are only performing the task of finding similar items. But they are not performing the task of compatibility prediction.

So, the research on outfit recommendation takes a major role. Because it finds the relationship between each product. A big direction in outfit recommendation research contains learning the relationships between products. The studies on this

research revealed the new idea. The new idea is that the recommendation of alternative item by using extra features. For example, if we give our shirt photo as our input data, then model learns from this input image and suggest suitable set of shoes, pants that exactly matches to the given image. The new research uses different techniques such as siasme network and metric learning rather than using older CNNs.

III. SYSTEM ARCHITECTURE

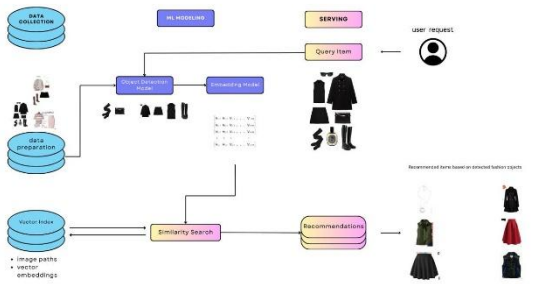

Fig 1: Block diagram

3.1 Input Acquisition

The is the starting step in the system architecture which involves collection of input data required for the recommendation process. A real-world photo for example a selfie or a product image is send to system The requested image from a system is a product image which indicates an individual fashion item.

3.2. Preprocessing of image

In preprocessing step, the noise present in the input image is removed by using different filter techniques. Normalization of image also takes place to increase the learning capacity. This will also help to improve the quality of image.

A Deep Learning Residual network model is used to extract all useful information from processed images. All features of input images are analyzed by neural network. After analyzing the model produces two types of features. The first type of feature represents the full image and second type of feature explains the different parts of the image.

3.3 Style Embed vector generation

Once the deep learning Resnet50 model extracted features from input image, these extracted features are converted into some unique description. We called this type of description are style embedding's After that the embedding with the same style are placed close to each other. And the embedding with the different style is placed far away. All these embedding are normalized to improve learning capacity and maintain consistency.

3.4. Calculation of global compatibility score

How efficiently the overall photo matches with the given product are calculated by the system. The distance between the corresponding vectors in the shared embedding space is calculated to compare the style embedding of scene image and product image. If the product is correctly match with the photo, then it indicates that the distance is minimum. If the product is not matching exactly with the image, then it indicates that the distance is maximum and it has less compatibility. So, this is the general style relationship between entire image and the product in terms of color and theme.

3.5. Local Compatibility with high performance Mechanism

The system gives more focus on detailed matching by analyzing different parts of the image instead of considering it as a whole. Each small part which is partitioned from whole image is used for comparison. This comparison takes place between product embedding and each part. The comparison finds out the matching. All the parts of the images are not so much important. In order to give importance attention mechanism is used. The attention mechanism assigns higher score to the most important regions for example clothing regions. This mechanism assigns lowest score to the part which are less important for example background.

All region-level comparisons and attention weights are combined. All combined results help to find the most suitable compatibility and it will improve model accuracy and performance.

3.6. Total compatibility calculation

Both the global compatibility and local compatibility score are combined by the system. The combined score is called as final compatibility score. This is the score which neither defines overall appearance not all small details.

3.7. Triplet loss training for model

A supervised training process which uses labelled data is used to train the model. After model training it learns compatible and incompatible items. A triplet structure is used for model learning. This structure contains three elements, first one is a scene image which is also called anchor. Second one is a compatible product which is also called positive and third one is incompatible product which is also called negative. Model Parameters are adjusted in such a way that the distance between compatible product and scene should be minimum. The distance between scene image and incompatible item should be maximum. A margin-based loss function is used for this purpose. Loss function clearly predicts positive and negative pairs. Slowly system performance is increasing by repeated learning of triplets. So, this helps to make better compatibility prediction.

3.8. Final Recommendation and ranking

Finally, a trained model is used to generate recommendations for a new input image. The system first extracts all features from input image. These features are converted into embedding vector. After that compatibility score is calculated for all candidate products. By using hybrid compatibility score, rankings will be given to all products. Highest ranking will be given to most compatible product and least rankings will be given to incompatible products. The system selects Top K products with highest ranking and gives recommendation.

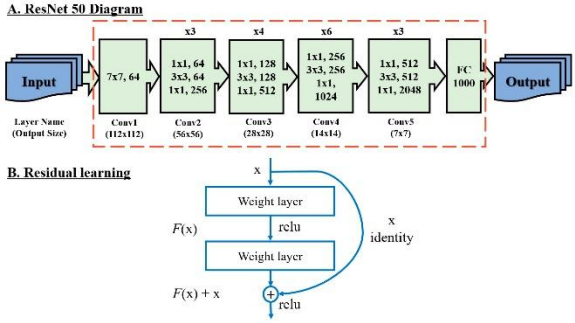

Fig 1: Working of Resnet

IV. RESEARCH METHODOLOGY

4.1 Style Embedding

Style Embedding are the numerical representation of images in which all visual information of images is converted into number vector. At the first step the given input image is passed to the pre-trained Resnet50 to get all features. From this given input image, Resnet50 will extract 2 types of features: one is global feature and another is local feature. The feature which we called as global attribute shows the overall looks of the image and local attribute is one which shows varies parts of the image. Both global and local attributes are converted into corresponding vector.

In this platform both types of features are represented in terms of constant length vectors. So, these vectors help to find compatibility. In addition to this the embedding that are generated for each local region of the image which helps to understand all minute details. Both types of embedding are normalized to make sure that it is stable and consistent. This is very much important that because it creates a representation which is unique for the similar items. This similar style items are placed each other so that it will be helpful for prediction of compatibility in future stages.

4.2 Measuring Compatibility

Matching products for a given input image is determined by computing compatibility in a shared embedding space.

Global compatibility is the learning of style embedding between scene image and product image. Embedding's which are placed nearby, have the high compatibility. Embedding's which are placed far away, have the low compatibility.

$d_{global}(s,p) = \| f_s - f_p \|^2$

where $\|\cdot\|$ is the d_2 distance, f_s and f_n are scene embedding and product embedding.

The distance d_2 gives the global compatibility.

To make the matching more detailed and accurate, we compare each part of the scene image with the product image. However, every region in the scene does not contribute equally, and its importance can vary depending on the category. So, the model first checks how well each small region of the scene matches the product.

After that, it applies an attention mechanism to focus more on the important areas, helping it makes better and more accurate predictions about complementarity.

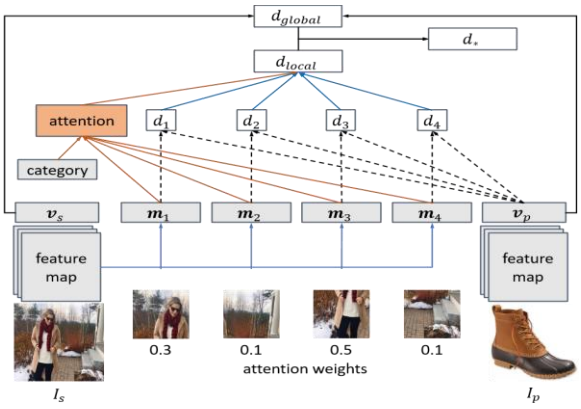

Fig 4: Hybrid Compatibility

The above figure shows how the model breaks the scene image into smaller regions and compares each one with the product image to understand how well they match. It then focuses more on the important regions using attention and combines both detailed and overall information to makes an accurate prediction.

$d_{local}(s,p) = \sum a_i \| f_i - f_p \|^2$

$1 \le i \le w*h$

$\hat{a}_i = -\|\hat{f}_i - \hat{e}_c\|^2$, $a = \text{softmax}(\hat{a})$,

Where c is the category of the product and $\hat{e}_c$ is the l_2 normalized category for product c. Softmax is the activation function which gives final output in the form of probability.

Finally, we measure compatibility by using a hybrid distance that combines both global and local features, helping the model capture overall similarity as well as fine details for better matching.

$d_*(s,p) = \frac{1}{2}[d_{global}(s,p) + d_{local}(s,p)]$

4.3 Recommendation Performance

Fig 5: Prediction of compatible products

As shown in Figure 5, the model is provided with a scene image, a category, and two products-one that is compatible (positive) and another that is less compatible (negative), both chosen from the same category. The model's goal is to identify which product better matches the scene. Its performance is evaluated based on how often it makes the correct choice, and this accuracy is equivalent to AUC, as both measure how well the model ranks products according to their compatibility.

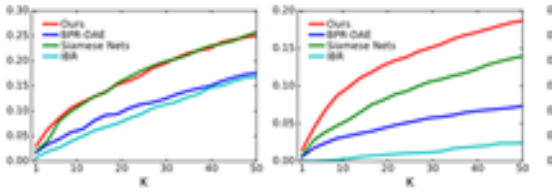

Fig 6: Top 2 accuracy

Along with overall ranking performance, we also consider Top- k accuracy, since it better represents real-world recommendation situations. It measures how often the correct item appears within the top K results. As shown in figure 6, our method performs much better than the baseline models on the last two datasets and slightly better than the strongest baseline on the first dataset.

4.4 Analysis of Attended Regions

Fig 7: Visualization of attention maps

To better understand what the model is focusing on, we visualize the attention region in the scene image and check whether they correspond to meaningful areas. Figure 7 presents cropped test images along with the attention maps from our model and the saliency maps generated by Deep Saliency. While Deep Saliency highlights visually noticeable objects, our model learns to focus on regions that are important for matching the scene with the product. For the first two fashion datasets, both methods correctly identify

the main subject across different backgrounds. However, our model tends to ignore faces and concentrate more on clothing, suggesting that it has learned clothing details are more important than the person's appearance when predicting compatibility.

Fig 8: Results

V. CONCLUSION

This paper will present an effective approach for recommending complementary products by analyzing real-world images. The older methods that depend only on product-to-product relationships, the proposed system gives scene context to better understand style and compatibility. Deep learning techniques with attention mechanisms makes the model to capture both overall appearance and fine-grained details within an image.

The use of a shared embedding space will allow the system to compare scenes and products efficiently, but the combination of global and local compatibility guarantees more accurate recommendations. This high performing attentive mechanism will increase performance by focusing on particular parts such as clothing areas. Experiments show the result that the proposed method will vanish existing approaches and produces recommendations that comes close to the human judgment. Totally the system is providing a more practical and realistic solution for fashion recommendation tasks. It will be helpful for real-world applications in e-commerce and personalized styling, where every client can get suitable suggestions based on their images.

REFERENCES

[1] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis, "Learning Fashion Compatibility with Bidirectional LSTMs," in *Proc. ACM Multimedia (MM)*, 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[3] W.-L. Hsiao and K. Grauman, "Creating Capsule Wardrobes from Fashion Images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.

[4] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial Transformer Networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2015.

[5] Y. Jing, D. C. Liu, D. Kislyuk, A. Zhai, J. Xu, J. Donahue, and S. Tavel, "Visual Search at Pinterest," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2015.

[6] W.-C. Kang, C. Fang, Z. Wang, and J. McAuley, "Visually-Aware Fashion Recommendation and Design with Generative Image Models," in *Proc. IEEE International Conference on Data Mining (ICDM)*, 2017.

[7] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to Buy It: Matching Street Clothing Photos in Online Shops," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2015.

[8] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Hipster Wars: Discovering Elements of Fashion Styles," in *Proc. European Conference on Computer Vision (ECCV)*, 2014.

[9] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[10] S. Kornblith, J. Shlens, and Q. V. Le, "Do Better ImageNet Models Transfer Better?" *arXiv preprint* arXiv:1805.08974, 2018.