Image-based fashion recommendation with attention to users' interests

Image-based fashion recommendation with attention to users' interests

Yao WenXin
Graduate School of Computer and Information Sciences, Hosei University
Tokyo, Japan
wenxin.yao.9e@stu.hosei.ac.jp

Abstract—Building effective fashion recommendation systems is challenging due to the high level of subjectivity and the semantic complexity of the features involved. Users' decision depends largely on their own interest and the appearance of the product, such information is often hidden in implicit feedback from users' purchase histories and product images. Most of the interest based recommendation systems like Deep Interest Network (DIN) and Deep Interest Evolution Network (DIEN) only take advantage of product attributes and context review, which are basically all text information. There are also some studies focusing on the use of image features for fashion product recommendation, they try to extract features from images and recommend products based on their similarity. However, for DIEN. It works not well when there's little interactions between users and items, the model can not find user's interest effectively. These image based recommendation systems, on the other hand, they tend to ignore an important factor: user's interest. We propose a new system trying to find user's interest by introducing the visual information of the product and our image based deep interest attention model based on DIEN. Product attributes and user preference can both be represented by introducing their visual information, we can model the similar attribute items in the same place and find user interests more effectively. We conducted some experiments on the public dataset to compare our methods with the DIEN and some other existing methods. Our experimental results demonstrate that visual information effectively aids click rate prediction and achieved better recommendation results.

Index Terms—fashion recommendation, image feature, user interest, attention

I. Introduction

Nowadays, people's demand for buying fashion products online is increasing explosively. The recommendation system works by recommending the products and information we want from many of the candidates. More and more company focus on Identifying the products a specific user may like to increase the profit. There are generally three types of Recommendation models: collaborative filtering recommendation models, content-based recommendation models and hybrid recommendation models.

However, different from general products, users pay more attention to a product's appearance when making decisions in the fashion domain, it's hard to train a good recommendation model in the fashion domain just based on collaborative filtering. More and more similar fashion products are recommended at the same time nowadays, which reduces the user's purchase intention, so it's also important to get a good handle on user interest.

DIN proposes a local activation unit for adaptive learning of user interest representation from the historical behavior of a particular ad [1]. Based on DIN, researcher proposes a new model: DIEN [2] to learn the interest evolving process, it focus on interest evolving phenomenon. Both DIN and DIEN follow a similar Embedding and Multilayer Perceptron (MLP) paradigm: mapping all the input features into low dimensional embedding vectors, it will not only reduce the size of learning parameters but also lighten the burden of computation and storage. However, this method is hard to discover and generate meaningful embedding vectors to track user local preferences when the dataset is sparse.

Apparently, analysis of Integrating visual information into recommendations and finding user' interest in fashion products from implicitly data is necessary to achieve more accurate recommendations with better quality. However, there is little research about extracting users' interest with the help of visual information. To address the problems mentioned above, in this paper, we propose a novel fashion recommendation system, firstly we utilize two additional image feature extraction methods to generate image feature vector, then combine it with the Interest Extraction Module (IEM) to help our model extract user interest from the feature vector based user behavior data we created.

We collect data from the Amazon dataset. Extensive experiments evaluating our proposed framework on the dataset we built show that our method improves the performance significantly by incorporating image features and auxiliary interest attention module. We use Interest Extraction Module to track user interest hidden in user behavior data, We visualize user behavior relevance relationship in the experiment section and compare it to our previous work to demonstrate that our approach improves the model's ability to express itself in a high sparsity dataset and yields better user interests representation.

To summarize, our main contributions are as follows:

- We introduce a novel deep interest based recommendation approach that incorporates visual information

Supervisor: Prof. Kaoru Uchida.

into predictors of user's click rate while extends to high sparsity dataset.
- We build the new image feature based dataset from Amazon dataset images and experiments on both high sparsity and low sparsity dataset revealing our method's effectiveness.

II. Related Works

A. Image Based Recommendation

As a result of the wide use of the recommendations, many effective methods have been proposed [3]. To expand recommendation model expressiveness, many researchers work on introducing visual signals from the underlying data to find user preferences more accurate and generate visual characteristics of products. Now, most image based recommendation systems use Convolutional Neural Networks (CNN) to extract image features and introduce other algorithms for similarity calculation. Shankar et al [4] proposed a unified CNN structure model to learn embeddings of products to generate recommendations based on visual similarity.

Recently, some researchers start to apply deep learning directly to image recommendation systems, Yufeng Duan et al [5] propose a model that integrates the CNN extracted image shape feature into the probabilistic matrix factorization (PMF). Min Hou et al [6] propose a Semantic Attribute based Recommender system. They use the attributes collected from the image to represent users and products, then project all the entities into the same space to capture the user's preferences. Xu Chen et al [7] proposed a fashion recommendation system utilize both user review data and region-level visual features. But there's little research about joint modeling of the image features and user.

B. Interest Based Recommendation

Besides good image content representations, the user interest also plays a vital role for personalized recommendation, user interest aware recommendation uses many kinds of user behavior related data to model characteristics of items and user preferences. Xu Chen et al [7] use the review data as one weak supervision signal to track user interest, Zhou et al [1] propose Deep Interest Network (DIN) to adaptively learn the representation of user interests, they concentrate on paying more attention to higher relevance product in user behavior data and extract users interest based on this. Zhou et al [1] propose another model: Deep Interest Evolution Network (DIEN) to learn the interest evolving process, they found that in the e-commerce platform, user interest always evolves over time dynamically, their interests are diverse. So, based on DIN, they add a new interest evolving layer in Interest Extraction Module modeling interest evolving process that is relative to the target product. By combining the Gate Recurrent Unit (GRU) and attention mechanism, the interest extraction module of DIEN is formed. See Figure 1 in this section.

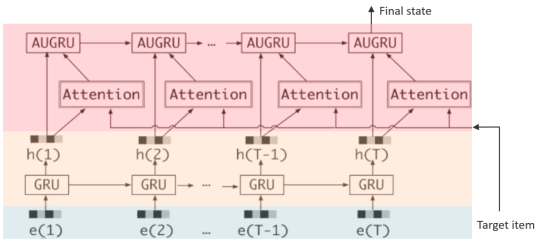

Figure 1. DIEN Interest Extraction Module

C. Class Activation Mapping

Zhou et al [8] proposed the Class Activation Map (CAM). CAM is widely used to interpret the results predicted by the model. It can visualize the heat trying to capture what the model considers to be the most significant results. The last convolution layer of the convolution neural network contains the most spatial and semantic information, so CAM replaced the full connection layer behind with the Global average pooling (GAP) layer, replaced the value of the entire feature map with the mean value of all pixels in the feature map, then utilize it as the input of the softmax classification layer, after training, it is used to weight and sum all the feature maps, and finally generate a visual heat map. Selvaraju et al [9] proposed Gradient-weighted Class Activation Map (Grad-CAM), it is more general than CAM. The problems with CAM were the need to modify the network structure and retrain, and Grad-CAM avoided these problems perfectly.

III. Proposed Scheme

The strategy of our proposal is to incorporate visual information into a deep interest based recommendation model to address the problems.

We first describe the two methods to extract features from an image. Then, we develop a new user interest attention module based on Interest Extraction Module (IEM) and introduce these image features as a weak supervision signal to enhance the model learning process. Finally, we present the overall optimization objective.

A. Image Feature Representation

There are many methods to extract the feature from the image. The point is extracting the feature that is applicable to Recommendation System (RS) and good enough to represent the image. Many image based recommendation system use the last fully connected layer in the classification model as a feature of the image, and most of the classification model based on Convolutional Neural Network (CNN), Recently, we found liu et al [10] proposed the model: Swin transformer with Hierarchical Transformer structure whose representation is computed with Shifted windows. Swin transformer got a better

performance than the classical CNN based model in many image-process tasks and was quickly used in various studies, so we use the pretrained swin transformer model to extract the feature from the product images in our work. Besides utilizing swin transformer, we train a classification model on the e-commerce product image dataset and utilize Grad-CAM results to extract the category aware image features. Former research about studying user interest are always pays more attention to category similarity rather than product similarity, a product may get a higher attention weight as long as it belongs to the same category as the target product. However, there are still huge differences between these same category products. Therefore, it is necessary to introduce visual information for each category. The formulations of Grad-CAM are listed as follows:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha^c_k A^k\right) \tag{1}$$

A represents for the feature map. In our thesis, we use the feature map output from the last convolution layer of classification model. k represents the k channel in the feature map A, c Represents category, A^k represents the data of channel k in feature map A, α^c_k is the calculated weight for A^k

$$\alpha^c_k = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{2}$$

y^c represents the network's predicted score for category c (score), A^k_{ij} represents the data of the feature map A in channel k with coordinates at the position of $\{i, j\}$, Z is the product of the width and height of the feature map. Then we take the absolute value of the obtained α^c_k and rank all the channel's feature map by this value. Finally we flatten and concatenate the top10 weight feature map data as the image feature we extracted by utilizing Grad-CAM. The feature extraction process is shown in Figure 2 .

To summarize, our main feature extraction methods are as follows:

- We introduce swin transformer to effectively extract features from image.
- We use Grad-CAM to locate the area that the classification model focuses on and extract the category aware image features in the process of generating heatmaps.

For each product in our dataset, we create these two kind corresponding feature vectors. These feature vectors then be used as input to our recommendation model.

B. Image Based Deep Interest Attention

Our core approach consists of two parts: The first part is introducing the image feature into our model to directly generate a product space that captures the visual similarity. These image features we extracted contain

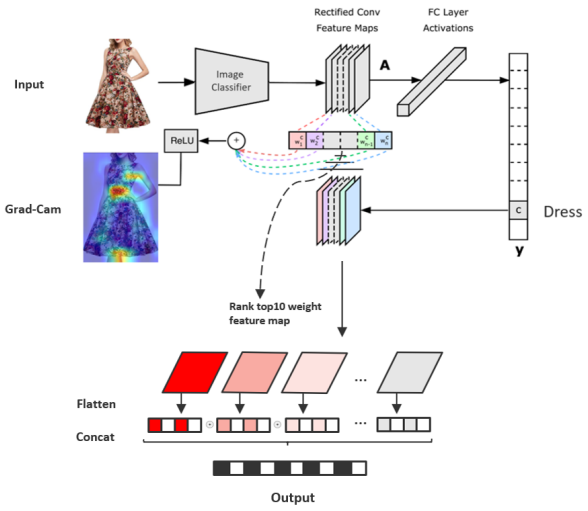

Figure 2. Grad CAM Extractor

a complex combination of colors and patterns. We use Euclidean distance between features of two images to measure the similarity between the images. The second part is adding an Auxiliary attention module to improve the model learning ability: track the various interests of users and interests change tendency. In addition to this, the addition of new modules enables the model to learn user interest features from different perspectives, further improving the expressiveness of the model. In this paper, we denote the name of the model as: IBDIM. The model structure is shown in Figure 3 in this section.

C. Loss Function

Based on the DIEN structure, we add one auxiliary attention loss, which benefits from the auxiliary attention module for stable training. There are a total of three loss functions in our model.

- Negative log-likelihood loss: It's widely used in Click-Through-Rate models, which uses model prediction results and the label of target item to calculate Cross-Entropy cost. It plays the role of supervising overall prediction in our model:

$$\begin{aligned} L_{\text{target}} = &-\frac{1}{N}\sum_{(\mathbf{x},y)\in\mathcal{D}}^{N}(y\log p(\mathbf{x}) \\ &+(1-y)\log(1-p(\mathbf{x}))) \end{aligned} \tag{3}$$

  D is the training set we created, N is training set size, (x, y) represents the sample from D, the concat of different fields' vectors from User, Context, Target product, User Behavior, Auxiliary attention module form x_u, x_c, x_t, x_b, x_a respectively, x = [x_u, x_c, x_t, x_b, x_a] ∈ D, y is the label of target product and y ∈ {0, 1}, it represents whether the user clicks target item. p(x) is model prediction result.

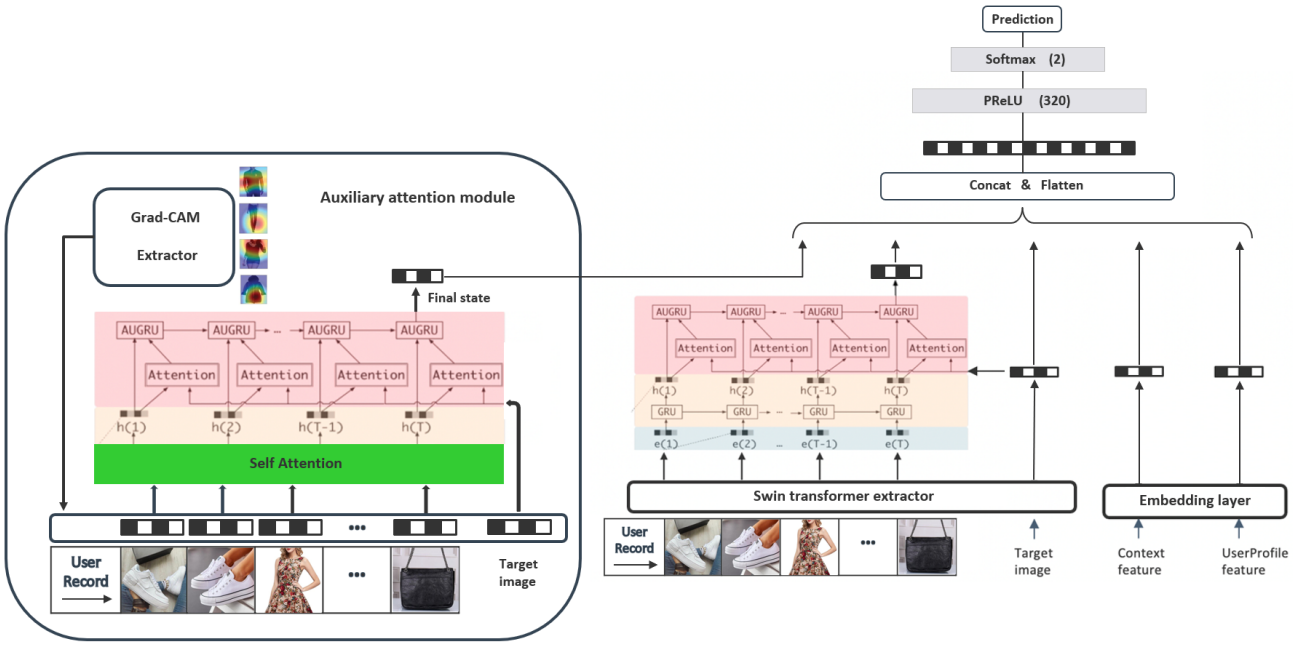

Figure 3. Structure of our model IBDIM. We divide the input of the model into three parts: the first part uses Embedding representation vector for users, the second part and the third part both use GRU and attention-based interest extraction unit to process interest sequence data to produce final vector containing interest state information of users. The difference between the two parts is the different types of input features and the different focus of interest extraction: the second part uses a sequence of image features extracted by swin transformer, which focuses on modeling time-series information of user behavior data, and the third part uses a sequence of image features extracted by Grad-CAM, which focuses on modeling the relationships within user behavior data.

- Auxiliary loss: The Negative log-likelihood function we use in the model has a weak supervisory effect in the process of final interest's state generation. The behaviors in the Interest extraction layer are consecutive. We use two parts to calculate the auxiliary loss.

$$L_{aux} = -\frac{1}{N}(\sum_{i=1}^{N}\sum_{t}\log\sigma(\mathbf{h}_t^i, \mathbf{e}_b^i[t+1]) + \log(1-\sigma(\mathbf{h}_t^i, \hat{\mathbf{e}}_b^i[t+1]))) \quad (4)$$

  1) Utilize behavior b_{t+1} to supervise the learning process of interest state h_t.
  2) Utilize the negative product feature (negative instance that samples from product set except the clicked item) as negative sample.

- Auxiliary attention loss: The Auxiliary attention module we add in the model contains the same auxiliary supervisor with DIEN Interest Extract module in the first layer consecutive behavior process part.

$$L_{auxatt} = -\frac{1}{N}(\sum_{i=1}^{N}\sum_{t}\log\sigma(\mathbf{h}_t^i, \mathbf{a}_b^i[t+1]) + \log(1-\sigma(\mathbf{h}_t^i, \hat{\mathbf{a}}_b^i[t+1]))) \quad (5)$$

We use these three loss functions to compose the total loss function in our model, the value of parameters α and β in this thesis we both set to 1:

$$L = L_{\text{target}} + \alpha * L_{aux} + \beta * L_{auxatt} \quad (6)$$

IV. Experiments

A. Dataset and Preprocessing

We build both high sparsity dataset and low sparsity dataset to verify the effect of our model. The statistics of these two datasets are shown in Table 1.

Table I
Statistics of two datasets in this paper

Dataset	User	products	Samples	Sparsity
Fashion	406	31	812	93.5484%
Clothing	50000	147248	100000	99.9986%

We collect the data from the public dataset: Amazon Dataset, it includes reviews, product metadata and links. We use two subsets of Amazon dataset: Amazon Fashion and Clothing, Shoes and Jewelry dataset. The high sparsity dataset we build in this paper called Clothing dataset, the method we preprocess the data is similar with DIEN, only one part is different during the construction of our high sparsity dataset: we randomly select a certain number of users from the whole dataset and collect all the records related with these users to build our own dataset, so there

are far more products in our dataset than users, it makes the dataset highly sparse. The low sparsity dataset we build in this paper called Fashion dataset, the data is directly from 5-core Amazon Fashion, we retained all the users in this small dataset, 5-core means that each of the remaining users and items have at least 5 reviews each.

We collect product images and extract image features from the Amazon Clothing Shoes and Jewelry dataset. To generate the proper format for our recommendation model, we also collect all the records together for each user and sort the reviews by time to generate the sequence behavior data. Our prediction purpose is the same with DIEN: Assuming one user has N-behaviors, we set the last behavior as the target product. Our goal is to use former N-1 behaviors to predict whether this user will interact with the target product. In order to introduce the visual information, we download the corresponding image by the image URL for each product, then we use pretrained Swin transformer and Grad-cam based method to extract features from the image and save them with PKL format, all the features we collected from image for each product, their size is [1, 1000].

B. Training Details

We trained our recommendation model on the Google Colab platform, using a 16 GB RAM graphics card with TensorFlow 1.4.

- Classification models: We use the timm [11] pretrained swin transformer model to build our swin transformer feature extractor. In order to build our Grad-CAM image feature extractor Firstly, we use the seresnet model structure for image classification and train it on our online-shops dataset for 60 epochs, the online-shops dataset includes 10 categories product image in the Street-to-online-shops dataset [12] and one jewelry category product image we collect from google image, the classification model we trained get a best accuracy at 81.59%.
- Optimization: We use the Adam optimizer with β_1=0.9, β_2=0.999 to optimize the Cross-Entropy loss function of our recommendation model.
- Hyper-parameters: In order to better compare the effect of various methods, we set basically the same training parameters for each methods, we train all model for 8 epochs with a initial learning rate of 0.001, the learning rate declines by half with each batch of training. Because of the limitation of GPU memory, the batch size we set in baseline work DIEN is 128, and in our model batch size is set to 32.
- MLP parameters: There are 2 fully connected layers used in last MLP of our model for final prediction, the dimensions area 320, 2 respectively, in order to prevent quick over-fitting, we use dropout function with rate=0.2 here: randomly discard 20% of neural.

C. Qualitative Evaluation

We follow the recommendation evaluation metrics in the DIEN model: AUC and Accuracy. In this paper, we use the best metrics obtained by the models. AUC is a widely used measure in the field of Click-Through-Rate (CTR) prediction, and we usually use a variation of user weighted AUC, it concentrate more on the goodness of orders within users. The AUC calculation method is as follows:

$$\text{AUC} = \frac{\sum_{i=1}^{n} \#\ \text{impression}_i \times \text{AUC}_i}{\sum_{i=1}^{n} \#\ \text{impression}_i},$$

Table II
Results (AUC) on two datasets

Model	Fashion	Clothing
DIEN	0.9762	0.7163
IBDIM	0.9824	0.8242

Accuracy is the proportion of click markers correctly predicted by the model for the test data.

Table III
Results (Accuracy) on two datasets

Model	Fashion	Clothing
DIEN	0.9432	0.6492
IBDIM	0.9659	0.7414

D. Ablation Experiments

There are two important components in our model, for better analysis their impact on the final performance. We conduct an ablation study to confirm the benefits of our proposed auxiliary attention module and the image features we used in the model.

The first test component is the auxiliary attention module. We remove the swin transformer image feature extractor in our model input layer, we denote it as IBDIM(-stf). The other variation is removing the auxiliary attention module. The only difference with DIEN is that we introduce the visual information to replace the product embedding part, we denote it as IBDIM(-aam).

The comparison results on AUC and prediction accuracy are presented in Table IV.

Table IV
Ablation results (AUC) of two variants in this paper

Variants	AUC	Accuracy
DIEN	0.7163	0.6492
IBDIM(-stf)	0.7537	0.6789
IBDIM(-aam)	0.8170	0.7296
IBDIM	0.8242	0.7414

.

Figure 4. Illustration of adaptive attention weight. The orange flag represents the attention allocation value obtained using the DIEN and the red flag represents the attention allocation value obtained using our model.

The results in Table IV prove that both these two variants are helpful for improving the performance of the model, while by incorporating auxiliary attention module and image features together, the final model relatively performed better than both of these two variants.

E. Visualize Analysis

Finally, we visualize the attention weight in the Interest extraction module to reveal the effectiveness of the new model. We compared the attention allocation accuracy in user behaviors with respect to a target product. See Figure 4 in this section Compared with DIEN, behaviors with high relevance to target product allocated with a higher score.

V. Conclusion

In this paper, we propose a new system trying to find users' interest from clicked product history by introducing the visual information of the product and our image based on a deep interest attention mechanism based on DIEN. The detailed explanation of two variations experimental results also supports our hypothesis that utilizing image features can expand recommendation model expressiveness.

In conclusion, our method is superior to conventional methods under high sparsity conditions. But there are still some problems for future research. In the future, we plan to study the relationships between recommendation results and visual explanations and try to find more effective methods to introduce visual signals into recommendation, reduce the model parameters size and calculation. Based on which, we can not only get a better recommendation accuracy but also train our model with a larger dataset and less time.

References

[1] G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai, "Deep interest network for click-through rate prediction," in Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018, pp. 1059–1068.

[2] G. Zhou, N. Mou, Y. Fan, Q. Pi, W. Bian, C. Zhou, X. Zhu, and K. Gai, "Deep interest evolution network for click-through rate prediction," in Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, 2019, pp. 5941–5948.

[3] W. Yu, H. Zhang, X. He, X. Chen, L. Xiong, and Z. Qin, "Aesthetic-based clothing recommendation," in Proceedings of the 2018 world wide web conference, 2018, pp. 649–658.

[4] D. Shankar, S. Narumanchi, H. Ananya, P. Kompalli, and K. Chaudhury, "Deep learning based large scale visual recommendation and search for e-commerce," arXiv preprint arXiv:1703.02344, 2017.

[5] R. Saga and Y. Duan, "Apparel goods recommender system based on image shape features extracted by a cnn," in 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC). IEEE, 2018, pp. 2365–2369.

[6] M. Hou, L. Wu, E. Chen, Z. Li, V. W. Zheng, and Q. Liu, "Explainable fashion recommendation: A semantic attribute region guided approach," arXiv preprint arXiv:1905.12862, 2019.

[7] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha, "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation," in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 765–774.

[8] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.

[9] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," arXiv preprint arXiv:2103.14030, 2021.

[11] R. Wightman, "Pytorch image models," https://github.com/rwightman/pytorch-image-models, 2019.

[12] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 3343–3351.