Predicting Fashion Styles using a Multimodal Approach with Instant Onboarding

Manav Kedia **Vinitha Ravichandran** **Li jen Tu**

Guoyong Li **Dandi Wang**

ABSTRACT

Fashion recommendation systems should not just be collaborative but should also take content into account. We propose a novel fashion style predictor based on a multimodal set of features including visual, textual and user information to train a classifier. While most style predictors have a long onboarding process where they ask a series of questions and risk incorrect self assessment by the user, ours is an almost instant onboarding process. We latently infer a user's styles using her rich social network data that is already available.

ACM Reference format:
Manav Kedia, Vinitha Ravichandran, Li jen Tu, Guoyong Li, and Dandi Wang. 2016. Predicting Fashion Styles using a Multimodal Approach with Instant Onboarding. In *Proceedings of ACM Conference, Washington, DC, USA, July 2017 (Conference'17),* 7 pages.
DOI: 10.475/123_4

1 INTRODUCTION

With the advent of Internet, people are moving towards their smart phones and computers to shop for their needs. The fashion industry is no different. There are several websites such as Macy's [1] that suggest clothing recommendation to users based on their previous purchases using collaborative techniques [6], however these do not capture the content of the clothing items or fashion or trend. Recommending visually similar clothing outfit from a given image query is also not enough [7] since it cannot capture the high-level fashion semantics that clothes have. In particular, we believe that having style in a fashion recommendation model is necessary because it captures more high level semantics than simple low level features such as color combination, shape, contrast, etc. Two images which are definitely not of the same style could be visually similar. There are systems that help put together heterogenous recommendations using non metric systems [8] [9].

People have tried to predict fashion styles before [12], but it only has 5 styles and uses SIFT features which have been replaced by the much recent CNN features in the vision community.

Reported by Vogue, romantic, elegant and classic are the top fashion trends during Fall 2016 Couture collection. These fashion styles rely heavily on specific visual details such as nipped waist, lapel collar, matched with high-waistlines dress. People have tried to analyse fashion by adding occasion and scenario elements both visually [14] and using text [19]. [10] proposes a data driven model which has large online clothing images to build a recommendation system. A recent work by [11] proposes to appreciate the aesthetic effects of upper-body menswear. However, fashion collocation and variety are essential elements of fashion that cannot be ignored. [15] proposes a multimodal solution to fashion styles where their two modes are the upper and lower body fashion clothes. They map visual features to a space of fashion styles, however being trained on images of models doing ramp walks, the generalization is poor.

Some work has been done in the area that suggest recommendations based on personal style such as topshop [2]. For many applications such as Stylebook[3] the onboarding process involves taking pictures of every piece of clothing before making recommendations. All of these systems put the users through an arduous task of answering long questionnaire sometimes as many as 30 questions before predicting the user's style. Even an expert human stylist would ask the user several questions to be able to learn the user's fashion taste. This leads to a poor user experience. In addition, a user's self evaluation in anticipation of these questions might be incorrect which would lead to poor results. Our work latently captures a user's fashion style using his/her images on social media to make the onboarding process instant and automatic.

We take a data-driven approach to the problem of predicting fashion styles. Social media is a rich source of fashion data. With social applications like Facebook and Instagram, several users post hundreds of photos. Also, few fashionistas

[1] macys.com

[2] topshop.com

[3] stylebook.com

Conference'17, Washington, DC, USA
2016. 123-4567-24-567/08/06…$15.00
DOI: 10.475/123_4

(a) Beach (b) Blouse (c) Dark (d) Pink (e) Sporty

Figure 1: Examples of images belonging to different styles in our dataset

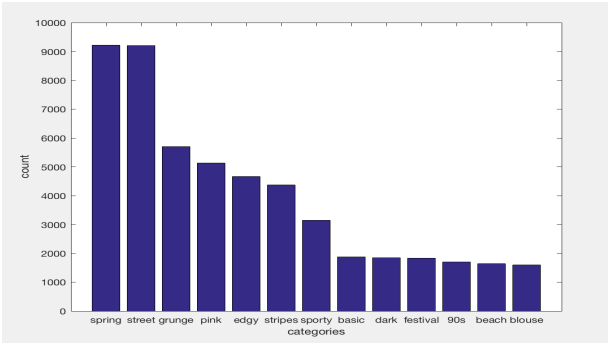

Figure 2: Number of entries of the top 13 styles

tag their pictures with hashtags that indicate the style of the clothes, occasion of use etc. All of these information provide useful insights regarding the fashion style of a person. Our system aims to exploit this information to predict the fashion style of a person without explicitly asking the users for any kind of inputs. We train a multimodal (image, hashtags and user features) classifier on data crawled from lookbook [4] which is like a fashion diary with the latest fashion trends and is popular amongst the fashion bloggers. The images on lookbook are taken from instagram which makes our model social-media ready, which is unlike any previous work. Thus, a system built around our model can predict the style of *any* user given his/her instagram profile.

We summarize our contribution as follows:

- We introduce a dataset of fashion images tagged with styles crawled from lookbook. In addition, we also crawl the hashtags associated with the images and details of the user.
- A novel fashion recommender system that takes a multimodal input to train classifiers to predict styles.
- Our work eliminates the problem of cumbersome onboarding and faulty self-evaluations by users. We latently infer the style of a user without any user intervention.

[4]lookbook.nu

Table 1: Top 13 styles on lookbook

Beach	Blouse	Spring	Grunge	Basic
Street	Pink	Edgy	Stripes	Sporty
90s	Dark	Festival		

The rest of the paper is divided as follows: Section 2 presents our dataset, Section 3 talks about our method and the features that we used, Section 4 presents our evaluations results and Section 5 summarizes our paper and future work.

2 DATASET COLLECTION

We crawled lookbook[4] to build our dataset. Lookbook.nu is a fashion community site that allows users to share their looks from various social media (Facebook, Twitter, Instagram, etc). The website is quite popular amongst the fashion bloggers and has been described as Fashion Diaries by Wikipedia. Thus, because of its popularity and social media appeal it makes a good fit for building our dataset.

The site separates its content into three main categories: trendings, styles, and occasions. Each of these categories have subcategories which we define as styles. The styles have been annotated by experts in lookbook and can be assumed to be noise-free. There are 121 styles in total ranging from vastly different styles such as formal and casual. These styles form the Fashion Semantic Space for our purposes. Depending on the popularity of the style, each style may contain 100 to 10,000+ entries. We define an entry as a set of image and its associated hashtags. Additional features such as the user age, gender, location, brands, items were also crawled when available. After removing incomplete entries we ended up with 95,459 entries. The dataset has a total of 6863 unique users. Of the 6863 unique users, 1520 reported they were female and 169 reported they were male.

For our experiments, we decided to use only 13 of the 121 styles because these 13 categories had over 1500 entries each

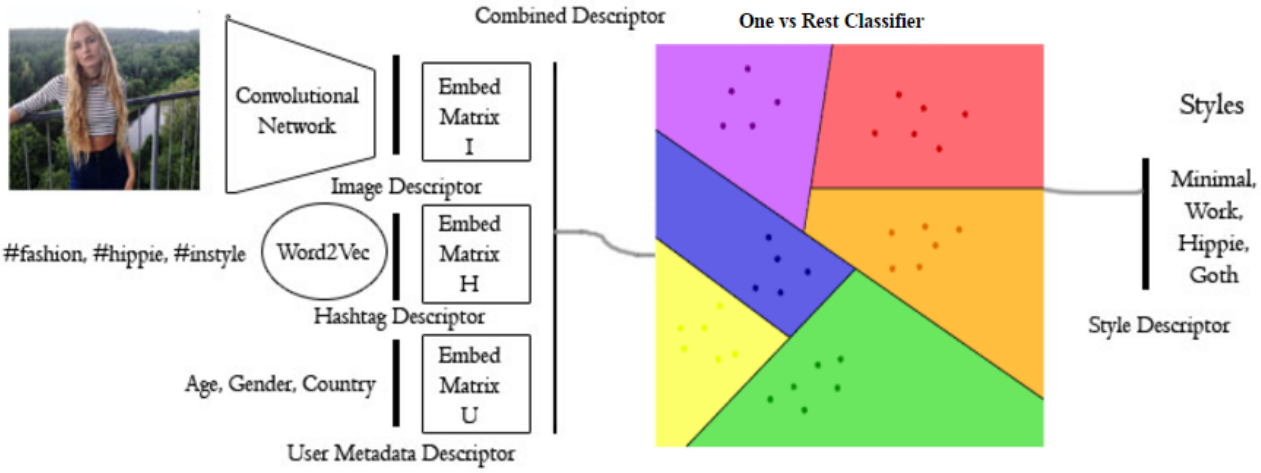

Figure 3: A diagram depicting our multimodal approach

and we did not want to overfit our model. Figure 1 shows examples of some of these styles. Table 1 shows a list of the 13 styles we selected and Figure 2 shows the count of entries in these 13 styles.

3 METHODOLOGY

Our model employs a multimodal solution to the problem of fashion style prediction. For each user image on lookbook, we use a combination of the image features, the associated hashtags features and the corresponding features of the user of the image to predict the style (which is also obtained from the annotations on lookbook). We train a OneVsRest classifier for each of the 13 style categories. Figure 3 shows a summary of our multimodal approach. We now describe the details of each feature and our model in this section.

Image Features

Immense progress has been made on methods to categorize image and video content. From recognizing objects [13] [4], to scenes [20], to activities [5], automatic techniques are beginning to successfully answer the question of what is shown in an image or video quite well. It has been applied for getting visual features from fashion images as well [15] [7].

We use a Convolutional Neural Network (CNN) model pre-trained on 1.2 millions labeled images from ImageNet. We used the VGG network [17]. Some basic preprocessing was done on each image to make it ready for the VGG network for obtaining image features such as resizing and subtracting the mean of the image. We used Keras which is a framework for TensorFlow [1] in Python for extracting the image features. We extract the learned representation from the second fully connected layer and use this 4096 dimensional vector as our visual representation.

Text Features

Each image in our dataset is associated with several hashtags which are manually tagged by the user of that image. We used a bag of words model for the hashtags. We decompose each hashtag into its constituent word using the longest matching algorithm. Each word is then embedded into a word space using word2vec [16]. The word vectors for each word for each hashtag are summed to get the resultant hashtag vector. The resultant dimension is 300-d.

$$\text{Hashtag Vector} = \frac{\sum_{h \in Hashtags} \sum_{word \in h} f(word)}{\text{total no. of words}}$$

User Features

For each image, we extracted the poster information to obtain our user features vector. We represent age as a discrete value, gender as Male (+1), Female (-1), Unknown (0), and a location vector that is the number of unique locations crawled from our users. Each position in the location vector represents a unique location. The position will have a 1 if they live in the corresponding location and 0 otherwise. We filtered out locations that have less then 10 examples. All location that were cities were changed to their corresponding state because styles within a state are similar. Thus, we obtained 189 unique location and when combined with age and gender results in a 191 dimension vector.

Table 2: 5-fold cross validation results

Features	Dimensions	Random Forest (n=10)	Random Forest (n=50)	Linear SVM
Image only	4096	15.17	18.98	14.90
User only	191	31.57	31.60	16.56
Text only	300	52.07	55.10	**71.47**
Text + User	491	51.89	55.66	69.66
Image + User	4287	15.05	19.41	19.14
Image + Text	4396	52.42	**61.11**	51.42
Image + Text + User	4587	**53.19**	60.94	51.79

Table 3: Individual accuracies of the styles

Style	Accuracy
Spring	77.68
Street	76.48
Grunge	73.73
Stripes	69.86
Pink	68.80
Beach	64.96
Sporty	53.01
Festival	51.26
Blouse	36.18
Basic	25.76
Dark	21.30
90s	03.10

Model

We used a OneVsRest classifier for each of the 13 style categories as shown in Table 1 which constitute a Fashion Semantic Space. We experimented with different choice of the classifier including a linear SVM [3] without stochastic gradient descent since the computations were tractable given the size of our dataset and RandomForest. For RandomForest we experimented with the number of trees which is a hyperparameter. We also experimented with the choice of features by including/excluding the three features described above. Whenever 2 or more features of dimensions d1 and d2 were included they were simply concatenated to obtain the resultant feature vector of dimension d1+d2. We believe that the set of Image + Hashtag + User feature would be the most discriminating for classifying styles. Results on the various feature and model combinations can be seen in the next section.

4 EVALUATION

To evaluate our model quantitatively, we did a five-fold cross-validation on our dataset. The dataset was split into five equal sized parts. One part is used for the test set while the remaining four samples will be used to train the model. This process will be repeated five times so each subsample will be used once as test data. This gives us the accuracy of our model. This process was repeated for all combinations of features and models as shown in Table 2. We experimented with different combinations of features to understand their relative significance for the task of fashion style prediction.

If we randomly assign every image to one of the 13 categories we would end up with an accuracy of 7.69 %. All our features perform better than this random baseline.

From table 2, it can be seen that image or user features alone do not produce good results. Whereas, text features only perform better in general. Both the classifiers, i.e. SVM and RandomForest perform almost equally on all features with some noticeable differences in Text only and Image + Text + User. On increasing the hyperparameter n which is the number of trees for RandomForest, the accuracy increases across all features, however further increase in n led to a decrease in accuracy. For RandomForest the best performer was Image + Text + User whereas for Linear SVM the best performer was Text only. The high performance of text features can be attributed to the fact that the hashtags might contain the style words which leads to better performance.

However, for the remaining section we use the model obtained from random forest (n=50) with all 3 features since it is more general and captures more information than a text would in a normal situation. Also its performance is comparable to the best performance. Table 3 shows the accuracies of the individual styles.

For a qualitative evaluation of our approach we show the confusion matrix in Figure 4 obtained from the above model. From the figure we see that the diagonals have high values which implies high individual accuracies for the styles. 90's style has a high confusion with the grunge class. Edgy(actual) and street(predicted), basic(actual) and street(predicted), and blouse(actual) and spring(predicted) are amongst the highest confused pair of classes. These classes have a very fine line of distinction and it is very easy to categorize one as the other.

Figures 5 and 6 shows pairs of images. Each pair consists of the misclassified image (left) and a representative image

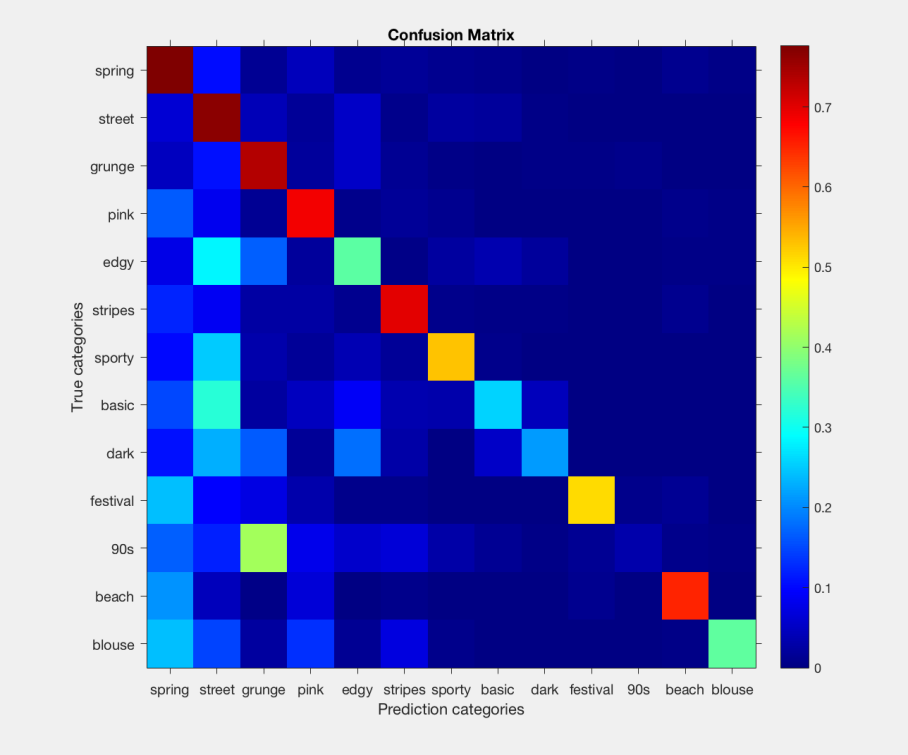

Figure 4: Confusion Matrix of our classification results

(a) 90s (b) Grunge

Figure 5: The left image belongs to 90s class but is misclassified as grunge. The right image shows a representative image of the grunge class

(a) Basic (b) Street

Figure 6: The left image belongs to basic class but is misclassified as street. The right image shows a representative image of the street class

of the incorrect predicted class (right). As can be seen from the images that it is visually very difficult even for humans to predict the correct class of the image and the confusion is likely very high amongst some of the style classes. Tables 4 and 5 show the corresponding top predictions for Figures 5 and 6 made by our model. From the tables we can see that the second best prediction is in fact the true class which shows the robustness of our model.

5 CONCLUSION AND FUTURE WORK

To summarize our work, we used a multimodal approach to tackle the problem of fashion style classifcation using visual, textual and user-based features. This combination of features had comparable performance to the best case of text-only. This combination adds extra information and is more general than textual features only. Our approach uses social media

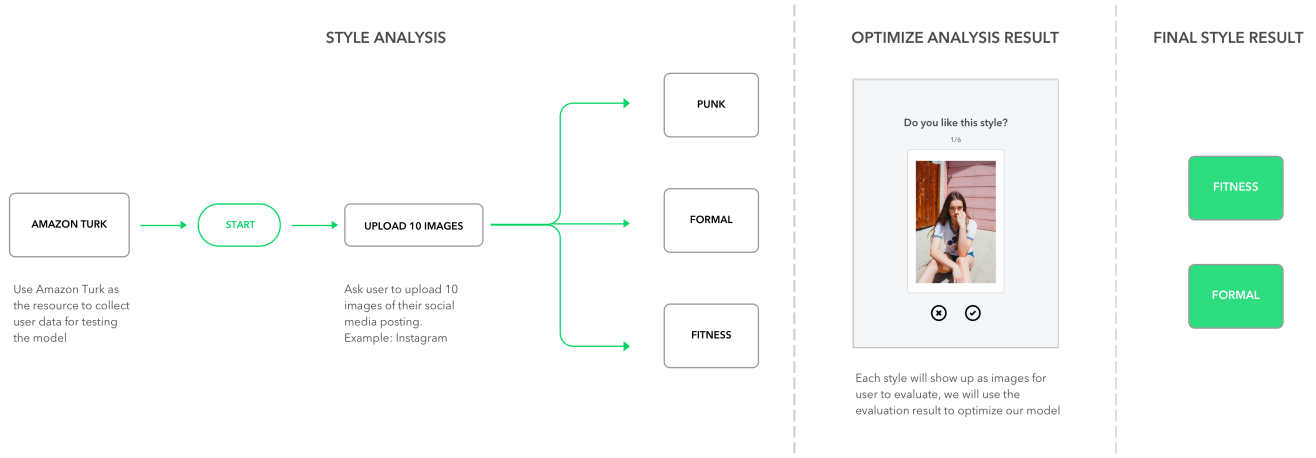

Figure 7: Proposed style quiz

Table 4: Top predicted styles for Fig 5 (left)

Style	Prediction Confidence
grunge	76.36
90s	7.27
spring	5.45
festival	3.63
others	7.2

Table 5: Top predicted styles for Fig 6 (left)

Style	Prediction Confidence
street	69.56
basic	8.69
grunge	8.69
dark	4.34
sporty	4.34
stripes	4.34

data which is readily available to provide instant onboarding to users by latently inferring the user's fashion style.

Our work has several potential applications. It can be used to detect the most trending fashion styles on Instagram or any other fashion websites. Since users search for clothes using high-level style semantics such as classic, beach, streetwear, etc our model is a good fit for retrieving relevant clothing from e-commerce websites. Finally, our model could be used to build a complete recommendation system by replacing collaborative recommendations by style based content recommendations on Amazon and other e-commerce websites for clothing.

We identify several areas of improvement in the future. The current VGG network has been trained to detect objects and is unable to capture the semantics of fashion images as is evident from the low accuracies of image only feature from Table 2. However it could be fine-tuned on our dataset for style classification to generate more semantic image features for the domain of fashion. Currently, we combine the different features using a simple concatenation, however the underlying spaces are not the same and the features should be transformed to a common space. To this end we propose to use the embedding techniques proposed in [2]. Our style embedding model would learn a d-dimensional embedding space for images, hashtags and styles.

Currently our model implements a classification loss. However, we think for the task of style prediction a ranking loss would make more sense as could be seen from Tables 4 and 5. Some style categories are not very distinct from each other with only subtle differences, hence a weighted ranked list of styles for a particular image would be more insightful than a hard classification. As seen from the confusion matrix and Fig 5, the task of style classification is more involved than a simple one class classification. We propose to extend this work to the case of multi-class classification where each image can belong to more than one category of style.

It would be interesting to validate our fashion style predictor model with the help of a user study conducted on MTurk. In the study, pictures of users on Instagram along with the corresponding hashtags and basic user information such as gender, age and location could be collected. We could compare the results of our model on user input against a simple style quiz. In the style quiz the user selects a bunch of thumbnails of people wearing fashion outfits that closely

matches with their style and taste [18]. See Fig 7 for our proposed style quiz.

ACKNOWLEDGMENTS

The authors would like to thank Ranjitha Kumar and Kristen Vaccaro for their guidance and support. We really enjoyed CS598RK!

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and others. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Emily Denton, Jason Weston, Manohar Paluri, Lubomir Bourdev, and Rob Fergus. 2015. User conditional hashtag prediction for images. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1731–1740.

[3] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of machine learning research* 9, Aug (2008), 1871–1874.

[4] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[5] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. 2014. R-cnns for pose estimation and action detection. *arXiv preprint arXiv:1406.5212* (2014).

[6] KANG Hanhoon and Seong Joon Yoo. 2007. Svm and collaborative filtering-based prediction of user preference for digital fashion recommendation systems. *IEICE transactions on information and systems* 90, 12 (2007), 2100–2103.

[7] Ruining He, Chunbin Lin, and Julian McAuley. 2016. Fashionista: A fashion-aware graphical system for exploring visually similar items. In *Proceedings of the 25th International Conference Companion on World Wide Web*. International World Wide Web Conferences Steering Committee, 199–202.

[8] Ruining He, Charles Packer, and Julian McAuley. 2016. Learning Compatibility Across Categories for Heterogeneous Item Recommendation. *arXiv preprint arXiv:1603.09473* (2016).

[9] Yang Hu, Xi Yi, and Larry S Davis. 2015. Collaborative fashion recommendation: a functional tensor factorization approach. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 129–138.

[10] Vignesh Jagadeesh, Robinson Piramuthu, Anurag Bhardwaj, Wei Di, and Neel Sundaresan. 2014. Large scale visual recommendations from street fashion images. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1925–1934.

[11] Jia Jia, Jie Huang, Guangyao Shen, Tao He, Zhiyuan Liu, Huan-Bo Luan, and Chao Yan. 2016. Learning to Appreciate the Aesthetic Effects of Clothing.. In *AAAI*. 1216–1222.

[12] M Hadi Kiapour, Kota Yamaguchi, Alexander C Berg, and Tamara L Berg. 2014. Hipster wars: Discovering elements of fashion styles. In *European conference on computer vision*. Springer, 472–488.

[13] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[14] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. 2012. Hi, magic closet, tell me what to wear!. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 619–628.

[15] Yihui Ma, Jia Jia, Suping Zhou, Jingtian Fu, Yejun Liu, and Zijian Tong. 2017. Towards Better Understanding the Clothing Fashion Styles: A Multimodal Deep Learning Approach. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[17] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[18] Kristen Vaccaro and Ranjitha Kumar. 2017. An Experimentation Engine for Data-Driven Fashion Systems. *AAAI 2017* (2017).

[19] Kristen Vaccaro, Sunaya Shivakumar, Ziqiao Ding, Karrie Karahalios, and Ranjitha Kumar. 2016. The Elements of Fashion Style. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 777–785.

[20] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.