APPAREL CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORKS

Shivansh Srivastava

M.tech CDS
14521
shivansh2524@gmail.com

Udit Gupta

M.tech CDS
14454
uditgupta@iisc.ac.in

ABSTRACT

Apparel classification from images finds interesting applications in several domains such as e-commerce[1], online marketing[1] and video surveillance[2]. In this project, we proposed a deep learning architecture to classify an apparel from a given image using convolutional neural networks. We have also described several variants of our architecture in terms of trainable weights and pre-trained models used. The performance of all the variant models are reported over different modalities such as test set accuracy, confusion matrix etc. We have also reported some insights and inferences we gained by assessing combined performance of all the variants.

Index Terms— Apparel Classification, Convolutional Neural Networks, Deep Learning

1. INTRODUCTION AND MOTIVATION

With the advent of deep learning, there have been several attempts of using deep learning models to accomplish tasks related to fashion and apparels. Some of these tasks include attribute prediction, apparel classification[3], exact retrieval[4], e-commerce recommendation[1] and many more. The task of apparel classification can also be used as a part of the pipeline that accomplishes as different task. For example,identification of apparel class will lead to more relevant recommendations to a customer by an e-commerce company. Real-time clothing recognition is another application which has been shown to be useful in video surveillance systems[2]. In such systems apparel classification model can be one of the layers in the classification pipeline where information about individuals clothes can be used to identify crime suspects. The objective of this project is to classify clothes appearing in natural scenes which may contain images with cluttered backgrounds, occluded items, low lighting etc.

2. PROBLEM STATEMENT

Given an image consisting of a clothing item, our task is to predict the class of apparel (among 15 classes).

2.1. Dataset

For this project we have used 'Apparel classification with style dataset' which was made publicly available by [3]. The dataset consists of 71,093 training set images and 17,858 test set images. Each image is labeled as one of the 15 classes which represent the style of the apparel. The distribution of classes in this dataset is depicted in figure

Label	Images	Label	Images
Long Dress	12,622	Sweater	6,515
Coat	11,338	Short dress	5,360
Jacket	11,719	Shirt	1,784
Cloak	9,371	T-shirt	1,784
Robe	7,262	Blouses	1,121
Suit	7,573	Vest	938
Undergarment	6,927	Polo shirt	976
Uniform	4,194		

Fig. 1. Number of examples of each class in ACS dataset

2.2. Challenges

1. Clothing items in different classes may appear to be quite similar to each other. For example: a coat and a cloak or a T-shirt and a sweater.

2. The images can be occluded and the clothing item may only be partially visible.

3. Some images may contain too little information of the clothing item due to poor lighting or image being out of focus.

4. The background may contain clutter which makes distinguishing between clothing entity and background a difficult task.

5. Some categories have semantic meanings such as 'Uniform' class which may overlap with other classes such as suit or shirt.

3. RELATED WORK

Lukas Bossard et. al [3] proposed several classical machine learning models to classify the images into clothing categories on ACS Dataset.The models proposed by them were Random Forests, SVM and Transfer Forests. The main highlight of their method consisted of multiclass learner based on random forests that used various discriminative models as decision nodes. Their SVM baseline has an accuracy of 35.07% and the best transfer forest model obtained an accuracy of 41.3%. Using the same dataset,[5] have proposed a convolutional neural network which uses 5 convolutional layers and categorical cross entropy loss for the loss function. Their model gives top-1 accuracy of 41.1%. Another deep learning implementation for the similar task using same dataset was proposed by [6]. Their model used AlexNet[7] model pre-trained on ImageNet dataset with the addition of one fully connected layer. They proposed two different approaches for training their model, in which, in first approach only the parameters of last fully connected layers were trained and in the second approach, all the parameters of the model were trained. Accuracy of the model trained using first approach is 45.0% and 50.2%.

4. PROPOSED ARCHITECTURES

Following the work of [6], our proposed approach uses a base network such as VGG16[8] or ResNet50[9] which is pretrained on ImageNet dataset. In the context of fashion images, such ImageNet pre-trained models have been frequently used[4][1]. The output layer of the base network is replaced by two fully connected layers of size 1024x1 and 128x1.As a convention for this report, we will denote these layers as FCa and FCb respectively. After the 128x1 fully connected layer, output layer of 15 nodes is placed which uses softmax function. The loss function used in the proposed network is categorical cross entropy loss. The proposed model is our generic model, and we make subtle changes in the generic model in terms of trainable weights and base network model. Different variants of generic proposed model are based on certain insights gained from related works and performance of previously tested variants of the same generic model. The architecture of proposed generic model is depicted in Fig. 2.

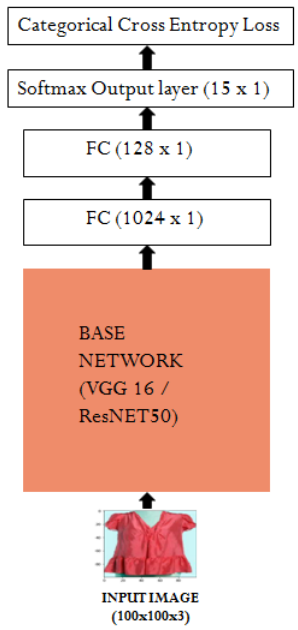

Fig. 2. Generic Proposed Model with fully connected and output layer

4.1. MODEL VARIANT 1: Fine tuned Fully connected layers + VGG16 pretrained on imageNet

In this variant, we are using VGG16 pre-trained on ImageNet as our base model and its output layer is replaced by FCa and FCb which are same as described in generic model description. The motivation for this variant arises from the question whether features learnt by VGG16 are good enough to give reasonable accuracy for the task of apparel classification. Good performance of this model would mean that features already learnt by pre-trained can be used for apparel classification task. In this model, only weights of FCa and FCb are trainable and no updates are performed on the weights of base network.

4.2. MODEL VARIANT 2: Fine tuned fully connected layers+ VGG16 with last two layers trainable

This model variant differs from model variant 1 just by the fact that no layers of VGG16 network were trainable in case of model variant 1 whereas last two layers (where 1st layer is considered to be input layer) of VGG16 base network are kept trainable in model variant 2. Hence, as the training progresses, the weights of the last two layers of base network are also fine tuned to our dataset. The motivation behind fine tuning of last few layers arises from the fact that the deep architectures designed for the tasks related to fashion and apparel images, have shown good performance when these architectures were designed in a way to give equal importance to fine grained features which are learnt in shallower layers and coarse grained features which are learnt in deeper layers [1].

4.3. MODEL VARIANT 3: Fine Tuned fully connected layers + ResNet50

This model variant has structure similar to model variant 1 expect the base network used in this case is ResNet50 pre-trained on ImageNet dataset. Only FCa and FCb are trainable. All the layers of base network are non trainable and are

not updated during the training process. This change in base network is to check whether ResNet50 network has learnt better embeddings on ImageNet dataset compared to VGG16 for the problem of apparel classification. This is anticipated because of superior performance of ResNet50 architecture on classification problem of ImageNet. Also, Due to structure of ResNet50, it is anticipated that the embeddings of ResNet50 may contain more information about fine grained details

4.4. MODEL VARIANT 4: Fine tuned fully connected layers+ VGG16 with last two layers trainable.

The motivation is similar to model variant 2. The only difference between model variant 2 and this model is that last five layers of base network are trainable in this model whereas only last two layers of base network were trainable in case of model variant 2. All other layers were kept frozen.

5. RESULTS

The performance of each model variant is measured by using test set which has 17858 images. The test set accuracies of each model variant is reported. Apart from test set accuracy, the learning process of each model variant is shown by plotting validation set and training set accuracies against number of epochs.Since the number of training images for each of the classes are unbalanced, test set accuracy would not be an appropriate measure for getting a complete picture of model performance. To get a better idea about performance of each model, we have shown confusion matrix over test set for each model variant, and Precision, Recall and F1 Score values for each class for a given model variant.

Each of the models is trained for 5 epochs and all the images in the dataset are resized to 100x100 pixels unless otherwise stated.

5.1. MODEL VARIANT 1: Fine tuned Fully connected layers + VGG16 pretrained on imageNet

Test set accuracy of 43.92% is achieved.

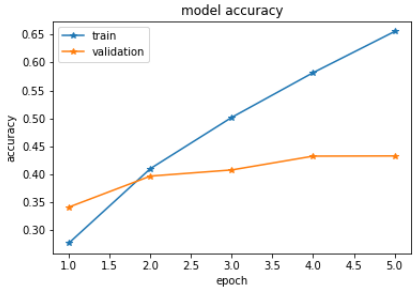

Fig. 3. Variation of Accuracy during training for Model Variant 1

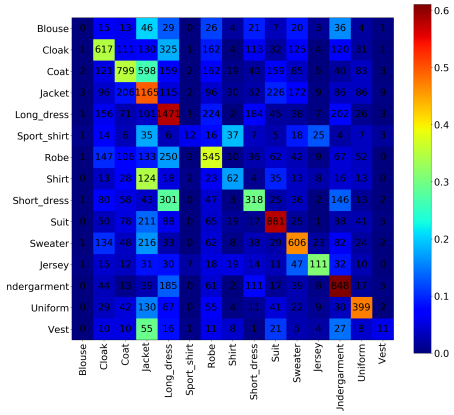

Fig. 4. Confusion matrix for Model Variant 1

The confusion matrix over test set shown in Fig. 4 shows that Model 1 performs well on some of classes such as long_dress, suits and undergarments. Although, the model performs quite poorly on most of the semantically related classes such as blouse, vest and Shirt. This gives an indication that network is not able to distinguish between fine grained details that would be important in classifying these closely related classes. This gives us motivation for fine tuning weights that are related to finer level activations. To do so, in the next model variant, we unfreeze the last 2 layers (Input layer is considered as 1st layer) of base network and make their weights trainable in training process. The classification report containing precision, recall and F1 score of each class is given in Fig. 5.

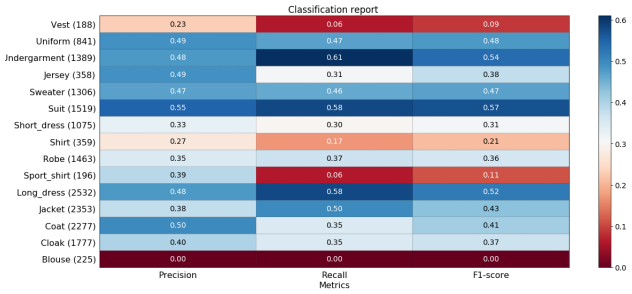

Fig. 5. Precision, Recall and F1 scores for each class for Model Variant 1

5.2. MODEL VARIANT 2: Fine tuned fully connected layers+ VGG16 with last two layers trainable

Test set accuracy of 47.14% is achieved.

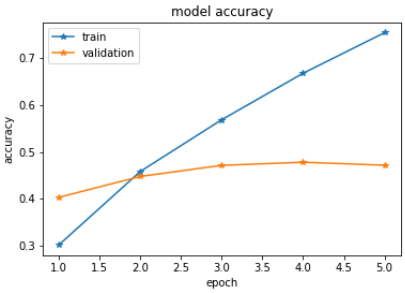

Fig. 6. Variation of accuracy during training for Model Variant 2

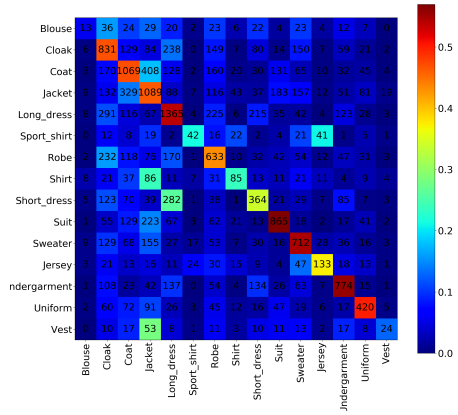

Fig. 7. Confusion matrix for Model Variant 2

Several improvement in performance is observed such as blouse, cloak and coat. This shows that fine tuned of features at a finer level helps the network to classify some categories with more accuracy which was not achievable in model1. The classification report containing precision, recall and F1 score of each class is given in Fig. 8.

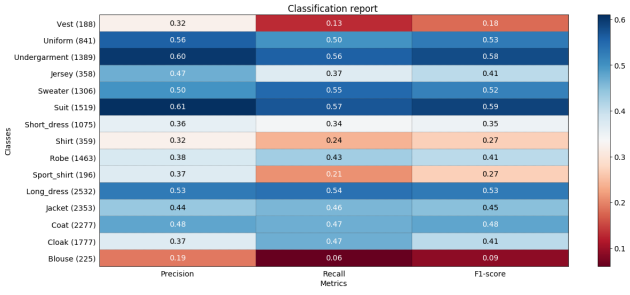

Fig. 8. Precision, Recall and F1 scores for each class for Model Variant 2

5.3. MODEL VARIANT 3: Fine Tuned fully connected layers + ResNet50

Test set accuracy of 47.48% is achieved.

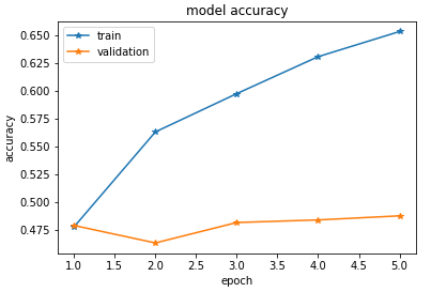

Fig. 9. Variation of accuracy during training for Model Variant 3

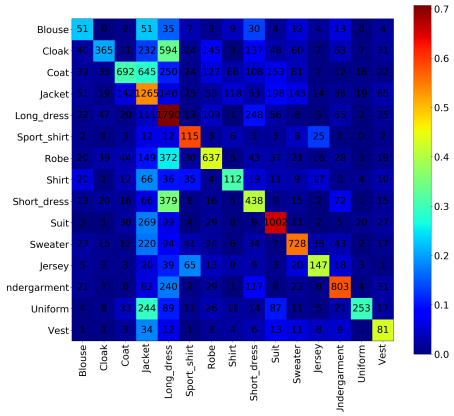

Fig. 10. Confusion matrix for Model 3

The performance of this model is comparable with model 2. The performance on some of the visually similar classes, such as blouse and vest, which are expected to use fine grained details for their distinction from each other, is improved. Although, this network has shown depreciation in performance over some classes such as coat and cloak, on which model 2 was giving better performance. Another issue with this network was high training time due to large number of layers. It is for this reason, we havent considered any other variants with ResNet50 as our base network. The classification report containing precision, recall and F1 score of each class is given in Fig. 11.

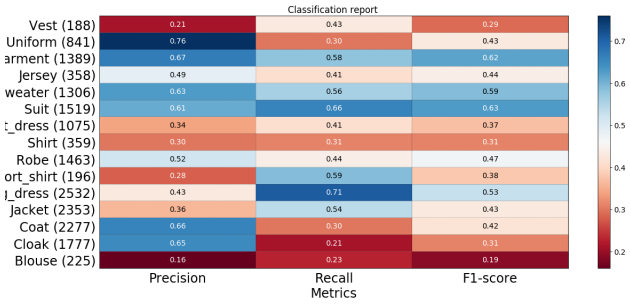

Fig. 11. Precision, Recall and F1 scores for each class for Model Variant 3

5.4. MODEL VARIANT 4: Fine tuned fully connected layers+ VGG16 with last two layers trainable.

Test set accuracy of 50.24% is achieved.

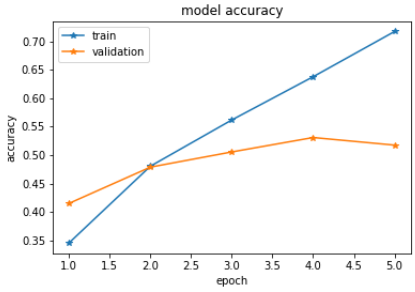

Fig. 12. Variation of accuracy during training for Model 4

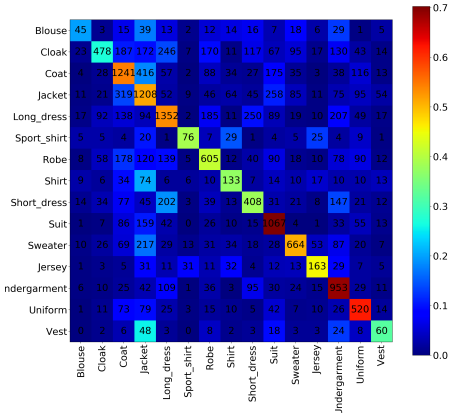

Fig. 13. Confusion matrix for Model variant 4

The performance of this model is better than all other variants discussed previously for almost all the classes. This again gives us evidence regarding importance of fine tuning our fine grained features. The classification report containing precision, recall and F1 score of each class is given in Fig. 14.

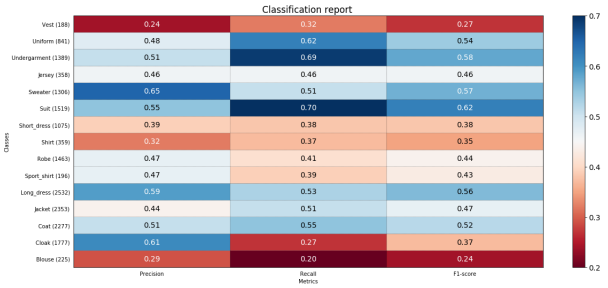

Fig. 14. Precision, Recall and F1 scores for each class for Model Variant 4

6. CONCLUSION

Some of the examples of correctly and incorrectly classified images by Model Variant 4 (Best performing model are given in Fig. 15 and Fig. 16 respectively.

Fig. 15. Correctly Classified Examples by Best Performing Model (Model Variant 4)

Fig. 16. Incorrectly Classified Examples by Best Performing Model (Model Variant 4)

6.1. Baseline Comparisons

The following table shows baseline comparisons with classification models used for the task of apparel classification on Apparel Classification with Style Dataset.

Performance Comparisons

Model	Test Set Accuracy
SVM (Bossard et. al)[3]	35.0%
Random Forests (Bossard et. al)[3]	38.3%
Shallow CNN (5 Convolutional Layers)[5]	41.1%
Transfer Forests (Bossard et. al)[3]	41.4%
Fine Tuned FC layers + VGG16 (Model Variant 1)	**43.92%**
Fine Tuned FC layers + VGG16 with 2 trainable layers (Model Variant 2)	**47.14%**
Fine Tuned FC layers + ResNet50 (Model Variant 3)	**47.48%**
Fine Tuned all layers Alexnet[6]	50.2%
Fine Tuned FC layers + VGG16 with 5 trainable layers (Model Variant 4)	**50.24%**

From the analysis of model architectures and their respective performance on test set, we are able to draw following conclusions and inferences:

1. Our best learning model surpasses all the baseline models discussed in section 4. Also, this accuracy was achieved only within 5 epochs.

2. Our best performing model is giving reasonable accuracy on all the classes except some misclassifications in case of short dress class in which most of the images in this class are classified as long dress. A similar case is observed in case of jacket and vest also. By visual analysis of dataset we observe that these two categories are semantically quite similar are expected to be difficult to classify even for a human based classifier agent.

3. Fine tuning parameters of pre-trained base networks is observed to give better performance. This is in coherence with literature pertaining to fashion datasets for tasks other than classification such as exact retrieval[4], recommendation[1] etc.

7. REFERENCES

[1] 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.

[2] M. Yang and K. Yu, "Real-time clothing recognition in surveillance videos," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pp. 2937–2940, IEEE, 2011.

[3] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool, "Apparel classification with style," in *Asian conference on computer vision*, pp. 321–335, Springer, 2012.

[4] 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 *ICCV*, pp. 3343–3351, 2015.

[5] R. Patki and S. Suresha, "Apparel classification using cnns,"

[6] B. Lao and K. Jagadeesh, "Convolutional neural networks for fashion classification and object detection," 2016.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.