Less complex Multi-head Attentive Convolutional Neural Network for Landmark-free Clothing Category Classification Using MobileNet

Nimasha Fernando and Amirthalingam Ramanan
Department of Computer Science, Faculty of Science, University of Jaffna

Introduction

In recent years, the style industry has emerged collectively for the betterment of the world's economy. With this advancement, research in this area has become more popular these days.

Recent studies have shown that the use of landmark information has gained great success. However, the landmark annotation is time-consuming and also suffers from inter-and intra-individual variability. Moreover, many recent studies have focused on VGG-16 based architectures which have too many parameters to learn, and the network architecture weights themselves are quite large. To conquer these complications, this study utilizes an attentive landmark-free deep learning-based network for fashion clothes category classification networks that can be trained end-to-end.

Contribution

- Some image datasets do not contain information about landmark annotations, but this proposed model is a landmark-free model, so it can be used with any image set that does not contain landmark annotations.
- The proposed model is implemented based on MobileNet architecture which is 32 times smaller and 10 times faster than the VGG-16 architecture. This saves the memory, storage as well as reduces the number of learning parameters in the training process.
- Multi-head attention mechanism has been used in order to increase the final accuracies of the proposed model.

Methodology

- In the proposed model, the initial convolutional operations are identical to the MobileNet architecture. Since MobileNet is a light weighted architecture it helps to reduce the complexity of the proposed model.
- In order to extend the ultimate accuracies, we have added the multi-head attention mechanism. This mechanism is more efficient as it performs multiple self-attention mechanism in parallel. Also, with the help of self-attention this multi-head attention enhances the model performance by calculating all the pairing covariances between all the pixels, considering each pixel in the feature map as a random variable. Thus it helps to increase the final accuracies.
- Finally we use two fully-connected layers to perform the classification.

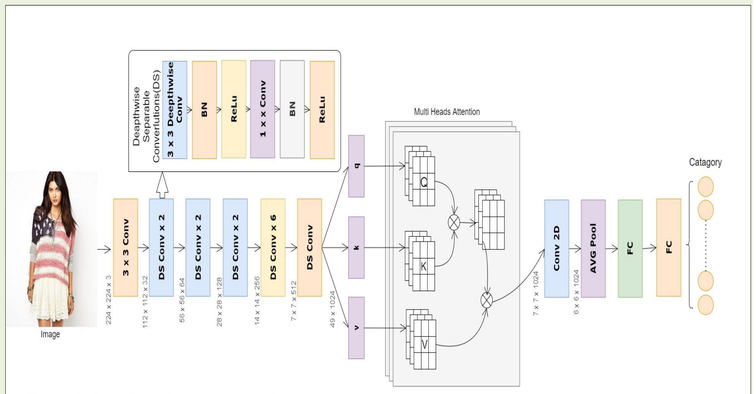
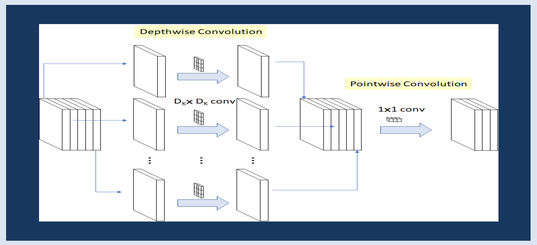


Figure 1: Overview of the proposed framework

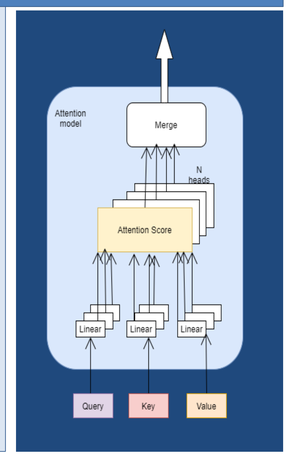
MobileNet Architecture

The MobileNet model is based on depth-wise separable convolution and it splits the standard convolution into two parts, the first part for filtering and the second separate layer for combining. This factorization reduces the computation and model size.



Multi-head attention

- Multi-head attention module repeats the self-attention computation multiple times in a parallel way.
- Self-attention helps to leverage the global information while calculating the destination pixels.



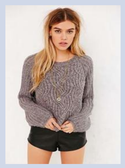
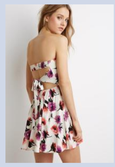
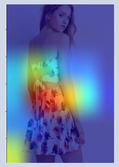
Experimental Setup

- The structure is based on the MobileNet network.
- Each image is resized to 224×224.
- Training, Validation, Testing = 209222; 40000; 40000
- Optimizer: SGD
- Batch size: 16
- Learning rate: 0.0001 and drop by 0.9 while validation plateaus.
- Loss: Categorical cross-entropy loss

Data Set

DeepFashion

- Images: 800, 000
- Annotated:209222
- Categories: 50
- Landmarks: 8
- Attributes: 1000

Image	Class activation maps of the inputted images	Top-5 Category Predictions
		Blouse Kaftan Henley Top Robe
		Kaftan Robe Romper Blouse Tee
		Poncho Caftan Jeggings Coverup Dress

Performance of category classification

Authors	Base Architecture	Top-5 accuracy
Liu *et al.*, (2016) [3]	VGG-16	90.17
Liu *et al.*, (2018) [4]	VGG-16	95.78
Lee *et al.*, (2019) [5]	VGG-16	95.26
Shajini *et al.*, (2020) [6]	VGG-16	96.20
MobileNet without attention mechanism	MobileNet	89.13
MobileNet with a multi-head attention mechanism (Ours)	MobileNet	92.08

Number of parameters required

Parameters	MobileNet without attention	MobileNet with attention	VGG-16
Total parameters	4,253,864	14,703,738	138,357,544
Trainable parameters	4,231,976	13,061,050	138,357,544

Discussion and Conclusion

Compared to the existing models although this study has poor performance, it utilizes a less complex model that requires less memory, and as it is a landmark-free model we can use it with any clothing image set even it does not contain the landmarks information.

By leveraging the multi-head attention, it gives a greater power of discrimination to the self-attention mechanism and by attaching self-attention, we can realize global reference during the model training process. Also, the model will be more reasonable with a good bias-variance trade-off. We demonstrated these experiments on the DeepFashion benchmark dataset.

References

[1] Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M. and Adam, H.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. (2017).

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I.: Attention is all you need. In Advances in neural information processing systems, pp. 5998-6008, (2017).

[3] Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X.: DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations , In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1096-1104, (2016).

[4] Liu, Jingyuan, and Hong Lu.: Deep fashion analysis with feature map up sampling and landmark-driven attention. Proceedings of the European Conference on Computer Vision (ECCV) Workshops, (2018).

[5] Lee, S., Eun, H., Oh, S., Kim, W., Jung, C. and Kim, C.: Landmark-free clothes recognition with a two-branch feature selective network. pp.745-747, (2019).

[6] Shajini, M. and Ramanan, A.: An improved landmark-driven and spatial-channel attentive convolutional neural network for fashion clothes classification, The Visual Computer, Springer, pp. 1–10. (2020)