Transfer Learning for Multi-label Street Fashion Images

Yiying Hu
yhu1102@stanford.edu

Motivation

We trained a deep neural network to predict the clothing attributes of fashion images using transferred learning on top of pre-trained VGG-16. The motivation of training this model is to create “search engine” for fashion images, the possible use case will be for fashion journal editors to find the similar outlook from their own database of images and illustrate for blog readers. The commercial value from making the fashion photos able to ‘talk’ about their attributes is tremendous. If we are able to label the street snapshots of their color pattern quickly, then the advertiser can quickly and automatically search for ones with the similar labels and display links for recommendation, thus improving the ads conversion rate.

Data

We used clothing attributes dataset of Research from Stanford University Data and More from Stanford's Cutting Edge Researchers

Originally there are 1856 pictures [6], and each image is processed with data augmentation, resized to the resolution of 224x224x3 and normalized by dividing 225 and between 0 and 1

The training and testing split is 67% and 33%. Therefore, the shape for training is a tensor of (13676, 224, 224, 3) and test is (6736, 224, 224, 3).

Transfer Learning Architecture

Transfer learning model is illustrated below:
the network is stacking two layers on top of VGG-16 in order to obtain a nx1 output (n is number of predicted attributes as needed), and the first 12 layers are fixed and the rest layers are allow to be fine-tuned.

Result and Analysis

The network is stacking two layers on top of VGG-16 in order to obtain a nx1 output (n is number of predicted attributes as needed), and the first 12 layers are fixed and the rest layers are allow to be fine-tuned.

Experiment Group No.	y dimension	k (loss func weight for pos)	learning rate	F1 (train+test)	Train_Loss	Test_Loss	Test_Accuracy
g1	11	2	1.00E-07	0.44	6.08	6.38	0.61
g2	11	5	1.00E-07	0.46	9.97	10.43	0.62
g3	11	10	1.00E-07	0.40	14.28	14.90	0.63
g4	20	1.5	1.00E-07	0.49	5.23	5.48	0.54
g5	20	2	1.00E-08	0.25	16.90	17.72	0.02
g6	20	2	1.00E-07	0.51	6.09	6.39	0.59
g7	20	2	1.00E-06	0.53	9.52	14.16	0.02
g8	20	3	1.00E-07	0.52	7.79	7.92	0.67
g9	20	4	1.00E-07	0.51	8.96	9.27	0.60
g10	20	5	1.00E-08	0.34	22.46	23.69	0.15
g11	20	5	1.00E-07	0.50	16.49	22.95	0.07
g12	20	5	1.00E-06	0.51	15.52	24.16	0.01
g13	20	10	1.00E-07	0.44	14.15	15.03	0.34
g14	20	10	1.00E-07	0.43	23.26	33.65	0.07

Discussion and Future Work

We found that higher dimension of classification output doesn’t necessarily means worse performance. However, the intuition is that with more dimension, the output labels tend to have correlation with each other, which provides the training model more noise. It might worth testing the performance of reducing the output label dimension by PCA to see whether the model prediction result will change.

The result complies with the intuition that fine-tuned model works better than original VGG. In the future, it might be worthwhile to explore a more sophisticated method to perform multi-label classification by loss function variation, input data processing especially the output labels fed into the model, and etc, to prioritize the most important output attributes.

Reference

[1] Karen Simonyan: “Very Deep Convolutional Networks for Large-Scale Image Recognition”, 2014; arXiv:1409.1556. [2] Han Xiao, Kashif Rasul: “Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms”, 2017; arXiv:1708.07747. [3] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang: “DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations”, 2016
[4] Tong He: “FashionNet: Personalized Outfit Recommendation with Deep Neural Network”, 2018; arXiv:1810.02443. [5] Jiang Wang, Yang song, Thomas Leung, Chuck Rosenberg, Jinbin Wang, James Philbin, Bo Chen: “Learning Fine-grained Image Similarity with Deep Ranking”, 2014; arXiv:1404.4661. [6] Huizhong Chen, Andrew Gallagher, and Bernd Girod, “Describing Clothing by Semantic Attributes”, European Conference on Computer Vision (ECCV), October 2012.