Fashion Clothing Category Identification

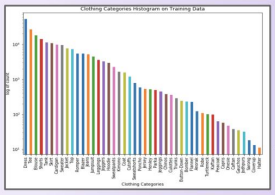
Abhishek Rawat, Dibyajyoti Ghosh, Shweta Karwa {arawat, dibghosh, skarwa}@stanford.edu

Problem

Given an image of a person wearing a clothing item automate determination of the item type and category. Online shopping for fashion items is a complex multi-step process. Part of the problem lies in incorrect annotations associated with a particular item like mismatches in type of clothing and its category.

Dataset

We are using Deep Fashion dataset [1] which has around 290,000 clothing images. Each image is annotated with one of 46 categories, like dress, T-shirt, coats, shorts, etc. Each category is of one of the 3 types: upper body clothing, lower body clothing and full body clothing.



Total samples: 289222
Training samples: 209222
Test samples: 40000
Validation samples: 40000

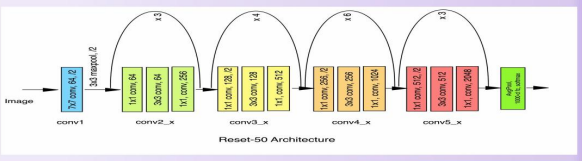
The y-axis in the above diagram is log of the count. Implying there is a huge discrepancy in the number of images we have for each category. To prevent this data imbalance we randomly chose ~6000 images of each category for training. We also considered creating a model only for upper body garments.

Network Architecture

We trained Fashion data on mainly two types of networks:

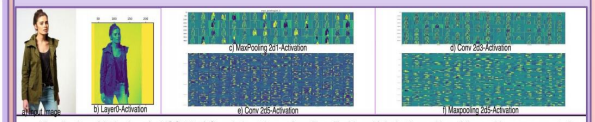
1. VGG-16 baseline network
2. Resnet-50 network with optimizations

We experimented with hyperparameter search for Resnet-50, to improve upon the loss and accuracy. Optimizations were done using a) gradient clipping, b) early stopping, c) RMS-Prop, d) Adam optimizer.



Visualization

Visualizing intermediate activations indicates how CNN layers transform the input. Input image (a) is transformed initially linearly (b), followed by #n convolution filters. Initial layer filters in (c,d) are doing edge detection, separating object from background, segment detections etc. Later layer filters in (e, f) are building more conceptual than basic visual feature maps. Hence the sparsity of activations [4] increases in later layers owing to absence of features detected by complex feature filters.

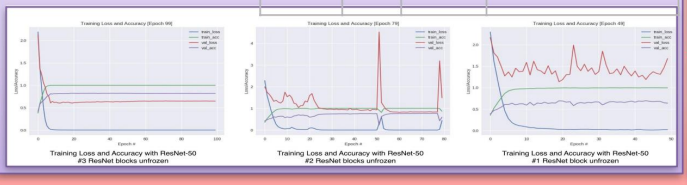


Visualization of Activations for VGG-16. a) Sample input image from DeepFashion, b) Activations with padding and image augmentation, c) #1 maxpool activations, d) #3 ConvActivations, e) #5 Conv Activations, f) #5 MaxPool Activations

Experiments and Results

Resnet-50 did better than VGG-16 as it's a deeper-network that can learn more complex features. Accuracy increased with unfreezing more Resnet blocks, as more activation layers got to train for specific task [fashion data set]. Even the #epochs for converging were lesser. Absence of landmark and attention mechanism[2] led to lower accuracy than state-of-art.

Network	Accuracy	Loss	Hyperparameters Used
Resnet50	57.97%	2.16	#1 block trained, 200 epochs, no clipping or regularization
Resnet50	64.01%	1.68	#2 blocks trained, 80 epochs, L2 regularization 0.3,
Resnet50	74.50%	0.95	#3 blocks trained, Early stopping at 30 epochs,
VGG16	49.7%	1.81	



Future

The next step in this project is to attempt category classification using Attention along with landmarks. Attribute identification is also an extension as the attribute vectors are available in the dataset. For visualization, we would attempt visualization of a) heatmaps of class activations, b) convnet filters

References

[1] Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, CVPR, 2016
[2] Attentive Fashion Grammar Network for Fashion Landmark Detection and Clothing Category Classification. Wenguan Wang*1,2, Yuanlu Xu*2 , Jianbing Shen†1 , and Song-Chun Zhu2, CVPR 2017
[3] Fashion Landmark detection in the wild, Z. Liu, S. Yan, P. Luo, et. al., ECCV 2016
[4] Visualization: github

Fashion Clothing Category Identification

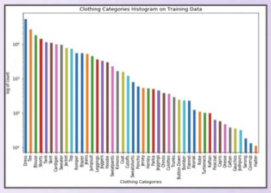
Abhishek Rawat, Dibyajyoti Ghosh, Shweta Karwa {arawat, dibghosh, skarwa}@stanford.edu

Problem

Given an image of a person wearing a clothing item automate determination of the item type and category. Online shopping for fashion items is a complex multi-step process. Part of the problem lies in incorrect annotations associated with a particular item like mismatches in type of clothing and its category.

Dataset

We are using Deep Fashion dataset [1] which has around 290,000 clothing images. Each image is annotated with one of 46 categories, like dress, T-shirt, coats, shorts, etc. Each category is of one of the 3 types: upper body clothing, lower body clothing and full body clothing.



Total samples: 289222
Training samples: 209222
Test samples: 40000
Validation samples: 40000

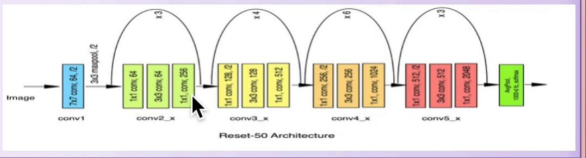
The y-axis in the above diagram is log of the count. Implying there is a huge discrepancy in the number of images we have for each category. To prevent this data imbalance we randomly chose ~6000 images of each category for training. We also considered creating a model only for upper body garments.

Network Architecture

We trained Fashion data on mainly two types of networks:

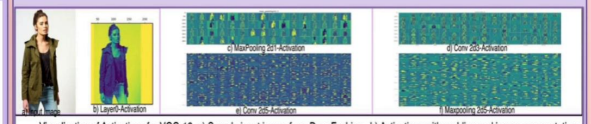
1. VGG-16 baseline network
2. Resnet-50 network with optimizations

We experimented with hyperparameter search for Resnet-50, to improve upon the loss and accuracy. Optimizations were done using a) gradient clipping, b) early stopping, c) RMS-Prop, d) Adam optimizer.



Visualization

Visualizing intermediate activations indicates how CNN layers transform the input. Input image (a) is transformed initially linearly (b), followed by #n convolution filters. Initial layer filters in (c,d) are doing edge detection, separating object from background, segment detections etc. Later layer filters in (e, f) are building more conceptual than basic visual feature maps. Hence the sparsity of activations [4] increases in later layers owing to absence of features detected by complex feature filters.

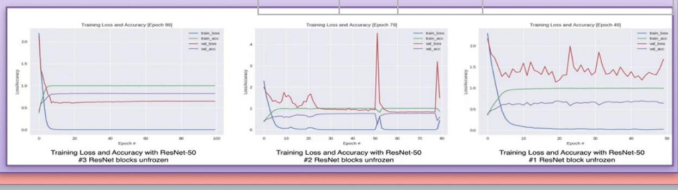


Visualization of Activations for VGG-16. a) Sample input image from DeepFashion, b) Activations with padding and image augmentation, c) #1 maxpool activations, d) #3 ConvActivations, e) #5 Conv Activations, f) #5 MaxPool Activations

Experiments and Results

Resnet-50 did better than VGG-16 as it's a deeper-network that can learn more complex features. Accuracy increased with unfreezing more Resnet blocks, as more activation layers got to train for specific task [fashion data set]. Even the #epochs for converging were lesser. Absence of landmark and attention mechanism[2] led to lower accuracy than state-of-art.

Network	Accuracy	Loss	Hyperparameters Used
Resnet50	57.97%	2.16	#1 block trained, 200 epochs, no clipping or regularization
Resnet50	64.01%	1.68	#2 blocks trained, 80 epochs, L2 regularization 0.3,
Resnet50	74.50%	0.95	#3 blocks trained, Early stopping at 30 epochs,
VGG16	49.7%	1.81	



Future

The next step in this project is to attempt category classification using Attention along with landmarks. Attribute identification is also an extension as the attribute vectors are available in the dataset. For visualization, we would attempt visualization of a) heatmaps of class activations, b) convnet filters

References

[1] Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, CVPR, 2016
[2] Attentive Fashion Grammar Network for Fashion Landmark Detection and Clothing Category Classification, Wenguan Wang∗1,2, Yuanlu Xu∗2 , Jianbing Shen†1 , and Song-Chun Zhu2, CVPR 2017
[3] Fashion Landmark detection in the wild, Z. Liu, S. Yan, P. Luo, et. al., ECCV 2016
[4] Visualization: github

Private video posted at: https://youtu.be/wfcadnAPUd0
Shared with Patrick Cho