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 the determination of the item type and category. Online shopping for fashion items is a complex multi-step process. Part of the problem lies in the 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 45 categories, like dress, T-shirt, coats, shorts, etc. Each category is one of the 3 types: upper body clothing, lower body clothing and full body clothing.

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 non-convolution filters. Initial layer filters in (c,d) are doing edge detection, separating object from background, segmenting 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.

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 Resnet50 model led to lower accuracy than state-of-art.

<table>
<thead>
<tr>
<th>Network</th>
<th>Accuracy</th>
<th>Loss</th>
<th>Hyperparameters Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet50</td>
<td>87.97%</td>
<td>2.16</td>
<td>#1 block trained, 200 epochs, no clipping or regularization</td>
</tr>
<tr>
<td>ResNet50</td>
<td>64.01%</td>
<td>1.68</td>
<td>#2 blocks trained, 30 epochs, L2 regularization 0.3.</td>
</tr>
<tr>
<td>ResNet50</td>
<td>74.50%</td>
<td>0.95</td>
<td>#3 blocks trained, Early stopping at 30 epochs,</td>
</tr>
<tr>
<td>VGG16</td>
<td>49.7%</td>
<td>1.81</td>
<td></td>
</tr>
</tbody>
</table>

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

[4] Visualization: github

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.
Fashion Clothing Category Identification
Abhishek Rawat, Dibyajyoti Ghosh, Shweta Karwa (arawat, dbghosh, skarwa)@stanford.edu

Problem
Given an image of a person wearing a clothing item, determine the 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.

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.

Experiments and Results

<table>
<thead>
<tr>
<th>Network</th>
<th>Accuracy</th>
<th>Loss</th>
<th>Hyperparameters Used</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resnet50</td>
<td>57.97%</td>
<td>2.16</td>
<td>#1 block trained, 200 epochs, no clipping or regularization</td>
</tr>
<tr>
<td>Resnet50</td>
<td>64.01%</td>
<td>1.68</td>
<td>#2 blocks trained, 50 epochs, L2 regularization 0.3</td>
</tr>
<tr>
<td>Resnet50</td>
<td>74.50%</td>
<td>0.95</td>
<td>#3 blocks trained, Early stopping at 30 epochs</td>
</tr>
<tr>
<td>VGG16</td>
<td>49.7%</td>
<td>1.81</td>
<td></td>
</tr>
</tbody>
</table>

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) heatmap b) image of class activations, b) ground filters.

Visualization
Visualizing intermediate activations indicates how CNN layers transform the input. Input image (a) is transformed initially linearly (b), followed by 4\(n\) 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.

References
[4] Visualization: github
Private video posted at: https://youtu.be/wfcadnAPUd0
Shared with Patrick Cho