Assignment-2 Coarse and Fine grained Classification

Abhishek Kumar 18111002
Hemant Parihar 18111021

February 2019

Abstract

Recent advances in image classification includes many different types of methods like Pyramid Matching over set of features, VLAD, CNNs(Convolutional Neural Networks) and its variants like Bi linear models, Attention Based models, Region specific models etc. In this report we have implemented some of these methods and compared as to for what kind of task what kind of works and tried a bit of arguments as to why it happens. Specifically we have implemented VLAD based Classification, Residual Network CNNs, Bi-linear Dense CNNs, Broad Fine Classification and combination of VLAD with other feature descriptors obtained from different models.

1 Introduction

Image Classification is a task of classifying an image into some set of classes. VLAD gives a good hand made feature set but is not automatic learned based on problem, whereas CNNs learn representation best suited for problem of discrimination but does not generalizes very well.

Challenges that needed to be tackled.

1. **Small Datasize**: Due to small number of tuples we can do two things. First we can use some pretrained network to extract feature set that has been trained on the larger dataset and use it (Transfer learning), other we can use some other time of descriptor set that is robust to orientation, clutter, lighting etc (VLAD).

2. **Fine Grain Learning**: We need to learn good set of incremental discriminative features while doing the fine grain classification.

3. **Choice of hyperparameters and models**: Given such vast number of methods, choosing optimizers, model, losses, structure of network were all needed to be taken care of.

2 Models

We have used following models or feature representations.

2.1 Multiple VLAD Encoding

We tried to do the classification with **sift** features, but it was not giving good performance so we included **surf and orb** features as well to get a concatenated set of feature VLAD representation which worked pretty well in comparison to single feature representation.

2.2 Auto-encoder CNN

We tried to train auto-encoder CNNs from scratch hoping to get a good representation of images in a lower dimensional space from where we can also do classification with a fairly simple model and do the Reconstruction of input as well. It worked well for Broad Classification but didn't perform so well for fine grain classification.

2.3 Dense CNNs VGG16 : Multi Task Learning

We tried some dense networks like VGG16 and divide the branch of network in two parts one for doing broad classification and another for doing fine-grain classification. This was performing very well but since we trained more than some layers of pretrained network this did not generalize very well on the test dataset.

2.4 Residual Network Architecture (ResNet)

we directly took the ResNet18 architecture and applied it fine-grain classification task. This particular structure seemed to work better as compared to other network in terms of accuracy of fine grained classification but got saturated at around 70 percent or so.

2.5 Bi-linear Convolutional Network (BCNNs)

BCNN's gave the highest accuracy on the finegrained classification. The reason we think is because it tries to learn local descriptors of image which are invariant across different images and learnns this kind of representation automatically as a result of do product between output of two CNN Networks.

2.6 Combinations

We also tried to combine the feature representation that we got from these models mostly we concatenated VLAD encoding with the CNN Network intermediate structure to get a rich representation of image which helped a bit.

3 Preprocessing

The image due to high variance in pixel values due to intensities, scaling etc, needs following preprocessing steps before we can give it to the networks for training to make the learning easier and generalize well on both train and test dataset.

1. Resize of image to a fixed size
2. Flipping along horizontal axis for pretarined networks.
3. Center Cropping of images
4. Random Cropping of images (Data Augmentation)
5. Normalization of images for training to avoid gradient related problems and good generalizations.

Flow of Prediction

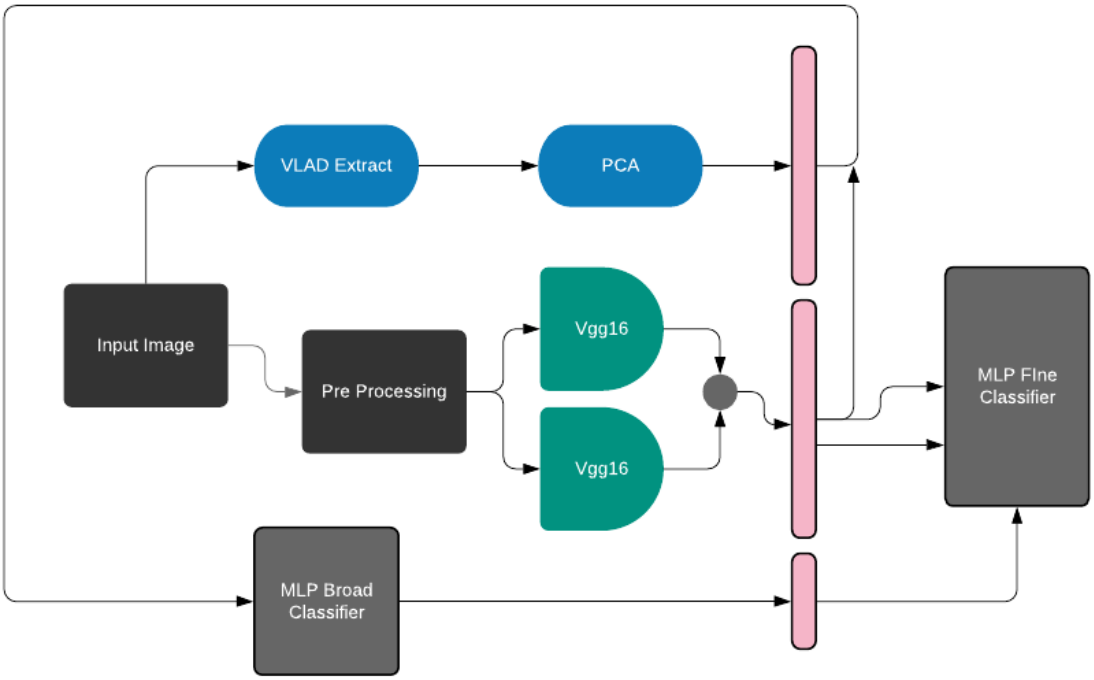

Figure 1: Model

Jeffry's Divergence

Instead of minimizing the crossentropy loss we tried the jeffry's divergence by adding to the loss a new loss function (pairwise confusion Loss [2]) which is euclidean distance between predicted labels and shuffled labels with a mask lambda.

4 Results and Conclusion

We ran all these models to observe the following results over the given dataset.

ACC(%)	Broad Classification	Fine Grain Classification
Multi VLAD	83.5	60.1
Autoenc-Classify	98.7	29
MultiTask VGG16	97.01	10.9
ResNet18	NA	67.3
BCNNs	99	89
ResNet + VLAD	97.8	69.5
BCNN + VLAD	99.8	92.37

We conclude by saying that the Bi-linear CNN Network works better and the with the inclusion of VLAD features it becomes more richer. Our accuracies achieved are not that high but we believe training on more augmented data can increase the performance a bit more.

5 References

1. Bilinear CNNs for Fine-grained Visual Recognition, Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji Transactions on Pattern Analysis and Machine Intelligence, 2017

2. Pairwise Confusion for Fine-Grained Visual Classification, Abhimanyu Dubey1 , Otkrist Gupta , Pei Guo , Ramesh Raskar , Ryan Farrell , and Nikhil Naik,

3. Deep Residual Learning for Image Recognition, Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Computer Vision and Pattern Recognition (cs.CV)

4. NetVLAD: CNN architecture for weakly supervised place recognition, Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, Josef Sivic, IEEE Computer Vision and Pattern Recognition (CVPR) 2016

5. See Better Before Looking Closer: Weakly Supervised Data Augmentation Network for Fine-Grained Visual Classification Tao Hu1 , Honggang Qi1 1 University of Chinese Academy of Sciences, Beijing, China, 2019