Object Recognition and Computer Vision - Assignment 3 report

Hugo Cisneros
ENS Paris-Saclay
hugo.cisneros@ens-paris-saclay.fr

Abstract

This report presents my approach for object recognition and computer vision's 3rd assignment. The task consisted in building a classifier and applying it on the provided subset of the Caltech-UCSD Birds-200-2011 bird dataset. The final solution I have selected consists in fine-tuning a 152 layers ResNet [1] model with heavy data augmentation.

1. Summary of the approach

I started to work from the provided default script on building a conventional convolutional neural network architecture [2] with varying number of filters, layers and other parameters for classification. After increasing the depth of the base network significantly, improvement in accuracy started to appear on the validation set. However, as those improvements began to be satisfying, the training time of the network became very large.

Since the challenge was constrained in terms of time and resources (two weeks of training and testing and no particular computational resources provided), I eventually chose to move to a transfer learning based approach for building the classifier. Those methods benefit from a large amount of information contained in a network already trained on a very complex dataset such as ImageNet. The weights of such a network are made available through common deep learning libraries and can for example be loaded in one line of code in Pytorch.

Transfer learning is particularly adapted to tasks where the studied dataset is similar but much smaller than the original dataset the network was trained on. State of the art architectures couldn't be trained on datasets with thousands of sample but the complex patterns they are already able to recognize can be slightly adapted to a particular dataset. Features from a CNN are often very powerful for classifying, even without further training as empirically shown in [3].

2. Methods

The retained solution uses the 152 layer ResNet [1] pretrained on ImageNet, available through Pytorch. During fine-tuning of the network on the dataset, some layer are "frozen", that is they aren't updated in the optimization algorithm. This is done to speed up the training process and is based on the assumption that the features extracted from the "frozen" layers can generalize to datasets similar to the original training set.

The best solution was obtained by freezing all layers except the last block of convolutions and replacing the final fully connected layer with a layer with as many outputs as classes in the dataset and training the network on the dataset for 40 epochs. Other methods were tried such as using other architecture or replacing the fully connected layer by an other classification method such as SVM or Logistic regression but this yielded poorer results.

Some data augmentation transformations were applied to the input data to increase generalization capabilities. Random crops, rotations, flips, shear transformations, brightness and contrast changes were applied, their effects can be visualized on Figure 1.

Figure 1. Sample input batch from the dataset. Random rotation, shear, brightness and color changes have been applied to augment the data. Note that as a result, birds are partly occluded in some images.

3. Results

After training the network for 40 epochs, the classifier achieved approximately 90% accuracy on the validation set and 78% on the public subset of the test set.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[2] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time-series. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*. MIT Press, 1995.

[3] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. CNN features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382, 2014.