ImageNet Classification with Deep Convolutional Neural Networks

Link

AlexNet

Summary

- We need a large model capacity to learn about thousands of objects from millions of images. However the immense complexity of object recognition means that this problem cannot be specified even with a large dataset like ImageNet. Our model needs to have lots of prior knowledge to compensate for the data we don't have. CNN constitute one such class of models which have strong and mostly correct prior belief about the nature of image. Compared to standard feedforward neural networks with similar number of layers CNN have much fewer connections and parameters, so they are easier to train while providing almost as good performance.

- ImageNet dataset has over 15 million labeled high resolution images belonging to roughly 22000 categories. ILSVRC(Imagenet Large-Scale Visual Recognition Challenge) uses a subset of ImageNet with roughly 1000 images in each of 1000 categories. In all, there are roughly 1.2 million training images, 50,000 validation images and 150,000 test images.

- The images were each down-sampled to a fixed resolution of 256×256. No image prepossessing was performed other than subtracting mean pixel value over training set from each pixel.

- The network consists of five convolution layer followed by three fully connected layers.

- They use ReLU non linear activation. Compared to sigmoid or tanh units deep neural networks with ReLU trains several times faster.

- They use local response normalization layer followed by the ReLU non linearity in some layers. This improves generalization ability.

- Instead of using non-overlapping pooling they use overlapping pooling with pooling window 3 and stride 2. This reduces top-1 error by 0.4%. Models with overlapping pooling find it slightly more difficult to overfit.

- To prevent overfitting they adopt two schemes

  1. Artificially enlarging dataset using image augmentation is the simplest method to reduce overfitting. We generate image by translation and horizontal reflection. By extracting 224×224 patches from the 256×256 image crop, they generate 2048 image per sample$(32 * 32 * 2)$. At test time prediction is made by extracting five 224×224 images(four corners and center patch) as well as their horizontal reflection(10 patches in all) and averaging the network's prediction. We also augment the data by altering RGB intensities in training images.
  2. Using dropout reduces the amount of overfitting substantially while roughly doubling number of iterations needed to converge. Dropout reduces complex co-adaptation of neurons and forces them to learn more robust features in conjunction with many different subset of neurons.

- They trained the model with SGD with momentum 0.9 and weight decay 0.0005. The learning rate was initialized at 0.01 and divided by 10 when validation error stopped improving. They initialized neuron biases in second, fourth and fifth convolution layers as well as the fully connected layers with constant 1. This accelerates early stages of training by proving ReLU with positive inputs. Neuron biases in other layers are initialized with constant 0.

- In ILSVRC-2012 competition their best model obtained 15.3% error rate and won first place, significantly better than second best entry with error rate 26.2%.

- The activation of the final 4096 dimensional fully connected layer can be used to interpret the network's visual knowledge. If two images produce feature activation with small Euclidean separation we can say the higher layer network considers them to be similar. This network's feature produces small distance between same label images even when the input image don't necessarily have small Euclidean distance.