Resumo – ImageNet Classification with Deep Convolutional Neural Networks

O artigo descreve uma rede neural profunda conhecida como AlexNet, empregada na classificação das imagens da competição LSVRC ImageNet. Este é um banco de mais de 15 milhões de imagens de alta definição rotuladas em 22.000 categorias. A ILSVRC começou em 2010 e usa um subconjunto de 1,2 milhões de imagens de treinamento, 50.000 de validação e 150.000 de teste, e mil categorias. Há versões mais novas, porém sem imagens de teste. O desempenho é medido por taxas de acurácia top-1 e top-5.

O pré-processamento das imagens consiste em recortar quadrados e reescalar para 256x256, além de centralizar os valores RGB pelas médias do conjunto de treinamento.

A rede consiste de camadas convolucionais: a primeira camada transforma imagens 224 x 224 x 3 em 96 kernels de 11 x 11 x 3, a segunda em 256 kernels de 5 x 5 x 48, a terceira em 384 kernels de 3 x 3 x 256, a quarta em 384 kernels de 3 x 3 x 192 e a quinta em 256 kernels de 3 x 3 x 192. Ao final, camadas completamente conectadas: duas com 4096 neurônios cada, a última com 1000, e softmax. O processamento das camadas escondidas é dividido entre duas GPUs de forma paralela, se comunicando apenas entre a segunda e a terceira camada convolucional, e na entrada das camadas completamente conectadas. Este uso de duas GPUs melhorou o top-1 em 1,7% e top-5 em 1,2%, e treinamento ficou um pouco mais rápido.

Emprega a função de ativação ReLU em todas as camadas escondidas. As vantagens citadas são o número menor de épocas necessárias para reduzir o erro abaixo de 25%, bem como a menor saturação. Emprega normalização de resposta local, calculada sobre alguns kernel maps adjacentes entre a primeira e segunda camada convolucional, obtendo melhora de top-1 em 1,4% e top-5 de 1,2%. Camadas de pooling na CNN resumem a saída da vizinhança de um mesmo kernel map, reduzindo o top-1 em 0,4% e top-5 em 0,3%.

Para reduzir o overfitting, ocorre aumento da base de entrada, processado na CPU, por extração de pedaços de 224 x 224 das imagens de 256 x 256, por reflexão horizontal desses pedaços, bem como alterando as intensidades dos valores RGBs, usando transformações lineares baseadas no PCA de cada imagem. Além disso, utiliza dropout nas duas últimas camadas escondidas, desligando os neurônios com uma probabilidade de 50%.

A aprendizagem emprega gradiente estocástico descendente com batch de 128 amostras, taxa de aprendizado inicial de 0,01 e divide por 10 quando o erro de validação para de cair, o que ocorre três vezes no treinamento. Emprega também momento de 0,9, com decaimento de pesos de 0,0005. Os pesos multiplicativos são inicializados conforme uma distribuição gaussiana de média 0 e desvio 0,01, e os pesos de bias são inicializados com 1 na segunda, quarta e quinta camadas convolucionais e nas duas camadas cheias escondidas, os demais com 0. O treinamento abrange 90 épocas, levando entre cinco e seis dias em duas GPUs GTX 580, com 3GB de memória cada.

Faz observações adicionais nos resultados do treinamento: a especialização de uma GPU no processamento de cores, enquanto a outra fica bem menos sensível a cores; a razoabilidade das top-5 previsões; bem como a similaridade euclidiana de imagens de mesma categoria na última camada escondida.

A AlexNet ganhou a ILSVRC-2012, com acurácia top-5 de 15,3%, bem abaixo da taxa de 26,2% do segundo lugar, estabelecendo-se como um marco na visão computacional, influenciando o uso de redes convolucionais, de GPUs, e de redes cada vez mais profundas.