

Parts of Deep Learning

Konstantin Burlachenko (burlachenkok@gmail.com)

Last update: 9 MAY 2022

Contents

1 Introduction	1
2 AlexNet	1
3 Some exotic types of convolution in convolution Neural Nets	3
3.1 Dilated convolution	3

1 Introduction

This document is a draft which I have created for purpose to systematize used tricks in various application around deep learning models. The all original papers are driven firstly by practice. This is not a note about that principle and it's pros. and cons. For some areas it's probably fine, for some does not.

2 AlexNet

The AlexNet paper "ImageNet Classification with Deep Convolutional Neural Networks" [1] is arguably the paper that brought deep learning to the forefront. After reading the paper [1] we can obtain the following main ten ideas from it.

First idea Authors faced with situation that before them people work with smaller tasks and people provides in common the following suggestion in the area of computer vision: "*... collect larger datasets, learn more powerful models, and use better techniques for preventing overfitting...*". For example before AlexNet was a work of LeCun, 1998 [lecun1998gradient] and model with LeNet-5 which uses *60K parameters*. Authors tried go beyond that.

Second idea Architecture presented at [1], Fig.1 carefully transform original input image. With several steps spatial dimension is reduced with 1 strided convolution (stride 4), and 3 max-pools. From another side channel dimension is increasing. The work empirically try to loose information very carefully. Finally input image is converted into long 4096 vector. During inference images are projected into that space. Finally there is a softmax into 1000 classes. What has been observed practically the importance of the depth of the network (in terms number of layers) which seems matters for vision tasks.

Third Idea Authors have made the hugest up to date (at moment of 2012) model with **60 million parameters** to train a models with using 1.2 millions samples. In fact 1.2 million samples is still nothing for huge dimensional models. And due to that image classification task is not so easy due to a that inherited difficulty.

Fourth idea Authors use Relu activation function, also known as $\max(0, x)$ or x_+ for people from Optimization community. Original motivation of the authors was that it was practically impossible to train with sigmoid in that time, due to practically observing slower training time in case of using sigmoid activation function.

Fifth idea Authors split layers across several GPUs by hands. And there is an additional trick as authors said the inter-GPU communication happened in a specific layers.

Sixth idea Local Response Layer (LRL) try to “normalize” along channel (i.e. in each specific location for each feature map). Authors stated that for the task it appends generalization ability. Inspiration for this author obtained from research how real neurons behave.

Seventh idea This paper make huge impact and momentum into Deep Learning technics modeling approach for tackle image classification tasks.

Eight idea Network is trained on RAW RGB images with central crops which after uniform rescaling into $256 \times H$ or $W \times 256$ are cropped by central crop via obtaining 256×256 . There is no use of SIFT or any specific Computer Vision algorithms for feature detection / important points detection.

Ninth idea Data Augmentation generates artificially trained images on the fly. One extra thing authors found via PCA/SVD the need space for images and append unbiased noise for new images via adding random variables in that space.

Tenth idea Use CNN layers for architecture by itself and build computation graph with that primitives. Use *DropOut* regularization mechanism which in the same time can reduce compute time and prevent over fitting. Where *Dropout* is in fact a sampling mechanism across Neural Network architectures by itself.

3 Some exotic types of convolution in convolution Neural Nets

3.1 Dilated convolution

...