Applying Robust Estimations for Mini-batch Sample Selection in Deep Learning

Aurora Cobo Aguilera[1] Antonio Artés Rodríguez[1] Fernando Pérez Cruz[2]
[1]Universidad Carlos III de Madrid, Spain {acobo, antonio}@tsc.uc3m.es
[2]Swiss Data Science Center, Switzerland fernando.perezcruz@sdsc.ethz.ch

Introduction

- Neural Networks sometimes require long training times when the graph architecture is some how complex.
- We propose an intelligent mini-batch selection method motivated by a recently proposed robust regularization for risk minimization.
- The original regularized estimation is based on an alternative measure of the variance, achieving an optimal and computationally efficient solution.
- We try to reduce the computational cost or improve the performance through this step in a state-of-the-art model in Deep Learning.

Keywords

Deep learning, Convolutional Neural Networks, Mini-batch selection, robust regularized empirical risk.

Objectives

Study the robust approach for sample selection in deep learning to get:

- a better performance in those samples that are considered out-layers,
- an improvement in the performance of the overall model, or
- a faster convergence of the model to an optimal solution.

Variance-based regularization

[Namkoong and Duchi (2016)] proposed an alternative to risk minimization and stochastic optimization that provides an optimal and computationally efficient solution. Particularly, it is based on a robust regularization of the empirical risk by adding a variance term.
They are motivated by the bias-variance tradeoff in statistical learning in order to minimize a quantity trading between approximation and estimation error. Moreover, they provide a tractable convex formulation -whenever the loss function is convex- that approximates closely to the penalized risk.

We can consider the alternative in the study as a **min-max problem**, that is, an optimization with two steps.

- First, **the minimization of the risk**, $\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} z_i(\theta, x_i)$, where z is the loss function, θ its parameters, n the number of samples and x_i each one of them.
- Second, **the maximization of the robust objective**, $\max_{p} \sum_{i=1}^{n} p_i z_i$, where p_i is the weight associated to each sample, so the samples with higher contribution to the loss function are the more valuable in the model.

As a constraint, they propose the equation 1, where ρ is a parameter to select the confidence level.

$$p \in \mathcal{P}_n = \left\{ p \in \mathbb{R}_+^n : \frac{1}{2} \|np - \mathbf{1}\|_2^2 \leq \rho, \langle \mathbf{1}, p \rangle = 1 \right\} \quad (1)$$

They give a number of theoretical guarantees and empirical evidences in order to show the optimal performance of the estimator with faster rates of converge and the improvement of out-of-sample test performance.

Application on CNNs

In a classification problem, the robust approach can be applied as an alternative to improve the performance on unusual classes, where the traditional empirical risk minimization would sacrifice the accuracy to focus on the common classes. More specifically, in the image classification task, Convolutional Networks (ConvNets) are one of the most popularly chosen models, with competitive results in international problems proposed to the scientific community as the **ImageNet Large-Scale Visual Recognition Challenge** (ILSVRC) [O. Russakovsky, J. Deng et al.].

Because of this reason, we propose its application to a state-of-the-art **Convolutional Neural Network** (CNN) called VGG [K. Simonyan and A. Zisserman (2015)]. This CNN is based on a study of increasing the depth of the network using an architecture with very small convolutional filters. They show a significant improvement in performance when the number of layers increase from 11 to 19 and prove their results in different datasets as the one of the ILSVRC. In this study, we are considering the network with 11 layers and a simpler dataset of 27000 images of cats and dogs for a binary classification problem.

The main idea can be resume in the following line.

- Taking advantage of the robust regularization approach in order to select the samples of the mini-batch in a intelligent way.

Mini-batch selection method

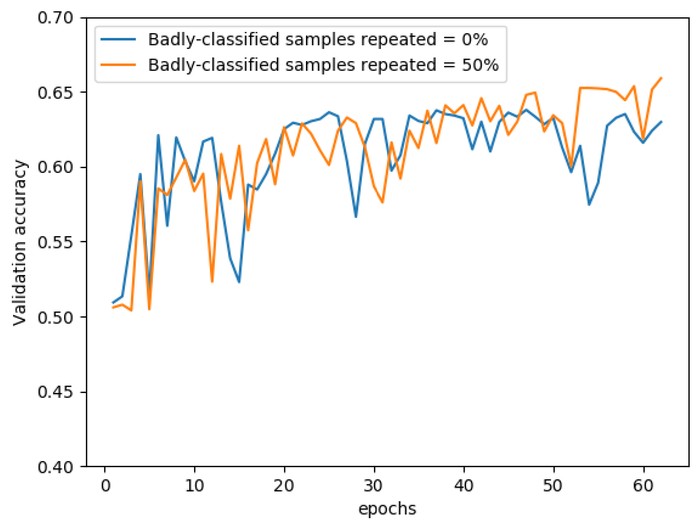
The way of selecting the samples of the mini-batch at each iteration of the training of the network can be seen with different approaches:

1. Repeat a percentage of the worst performed samples from one mini-batch to the next one.
2. Repeat a percentage of the worst performed samples from one epoch to the next one.
3. Re-sample a percentage of the worst performed samples each 10 epochs.

In every case, the repeated samples are those ones that get the worst score, that is, the higher value in the loss function at the previous iteration, epoch or stage, respectively.

Preliminary results

We present some preliminary results based on the method number 1 for the mini-batch selection.



The following table presents the test accuracy after a specified number of epochs for a regular model and another one with 50% of badly-classified repeated samples in the mini-batch.

Percentage	10 epochs	40 epochs
0 %	57.69%	**63.24%**
50 %	57.35%	**64.39%**

Conclusions

Although we are presenting some preliminary results, it can be seen how the mini-batch selection method makes the convergence of the algorithm faster, with a higher difference in the performance as the number of epochs increases.

References

H. Namkoong and J. C. Duchi. Variance-based Regularization with convex Objectives. *arXiv:1610.02581*, October, 2016.
O. Russakovsky, J. Deng and others, Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, *In Proc. ICLR*, 2015.