Improving Neural Networks using Gradient Space

Rita Aleksziev
Hungarian Academy of Sciences
Institute for Computer Science and Control

Objectives

We investigate the gradient space of Neural Networks in order to:

- Find "important" edges
- Improve performance
- Understand the underlying structure

Introduction

When training an Artificial Neural Network with a gradient-based method, we calculate (or at least approximate) the partial derivatives of the loss function with respect to each parameter of the network in every backward step. Over the parameter space and the loss function we can often determine a smooth manifold [1]. In our ongoing project we investigate the space of the tangent bundle of the manifold in order to understand the behavior of certain networks better and take advantage of specific Riemannian metrics with unique invariance properties [2, 3, 4].

Using the gradients of a pre-trained network, one of our goals is to figure out if the gradient space holds information about which parts of the network are more "important" than others. Finding these parameters could make further training faster and more efficient.

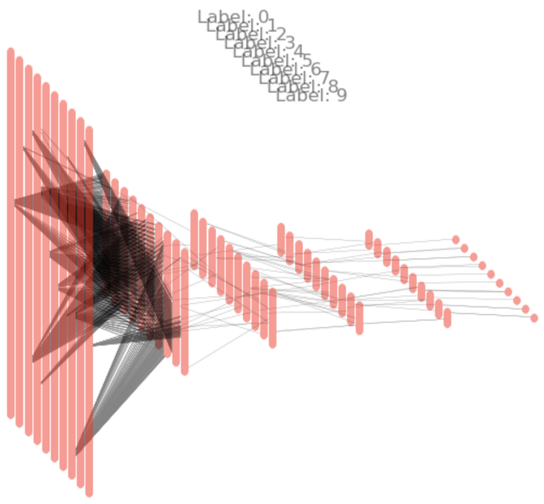

Figure 1: Important edges in an MLP based on the gradients

Our other approach is to extract information about the data directly from the gradient space that the model was not able to find. We do this by applying certain machine learning models on the set of the gradients.

Research [5]

In our ongoing project we use certain machine learning models to assess the expressive power of the gradient spaces of certain CNN's and MLP's, moving the optimization from the parameter space to the gradient space.
In our experiments we test the hypothesis that for a given neural network pre-trained for a classification problem, a quadratic separator on the gradient space can outperform the original ("base") network. We wish to find the proper separator by training a two-layer NN having a block-like structure demonstrated in Figure 2.

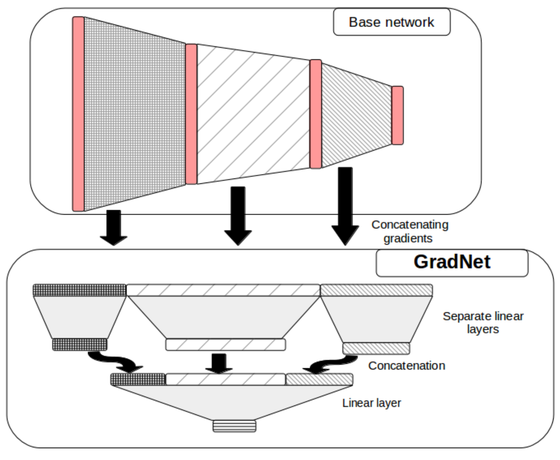

Figure 2: GradNet

This model (that we call GradNet) is capable of capturing connections between gradients from adjacent layers of the base network.

Methodology

In order to find the optimal architecture, the right normalization process and regularization technique, and the best optimization method, we experimented with a large number of setups.

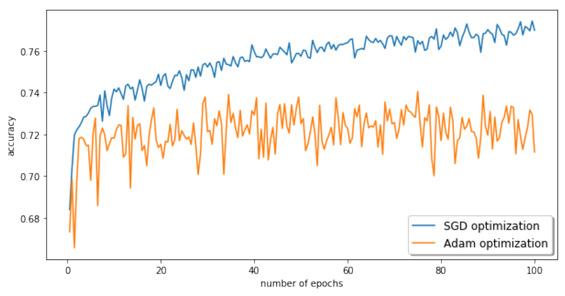

Figure 3: Optimization

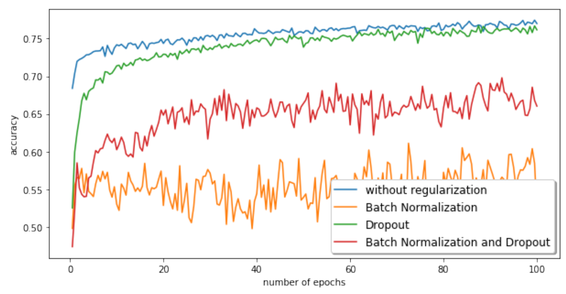

Figure 4: Regularization

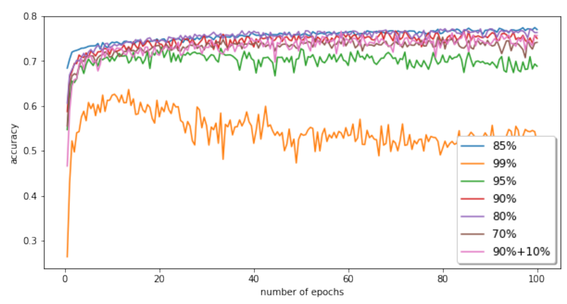

Figure 5: Percentile of gradients to be chosen

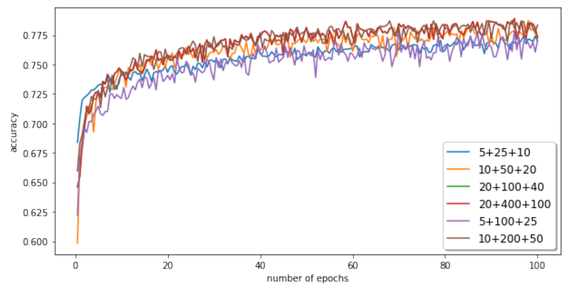

Figure 6: Structure of GradNet

Experiments

Table 1 shows how much our method improves the performance of a network having undergone different levels of pre-training.

CIFAR		
Original	Improved	Gain
0.79	0.8289	+4.9%
0.76	0.8201	+7.9%
0.74	0.8066	+9%
0.72	0.7936	+10.2%
0.68	0.7649	+12.5%
0.65	0.7511	+15.5%
0.62	0.7274	+17.3%
0.55	0.7016	+27.5%
0.51	0.6856	+34.4%
0.49	0.678	+38.3%

MNIST		
Original	Improved	Gain
0.92	0.98	+6.5%
0.96	0.9857	+2.7%
0.9894	0.9914	+0.2%

Table 1: Performance measure of the improved networks.

Conclusion

Our joint research with Bálint Daróczy shows that better performance can be achieved with this technique than with regular training. Our results are preliminary and further investigation is needed into the approximation of the invariant Hessian/Fisher metrics on the gradient space.

Future work

- Investigating the underlying manifold and squeezing
- Expectation Maximization and Cencov characterization
- Finding invariant metric and generalization to Finsler
- Tensor networks and quantum Fisher metric
- Recurrent nets → time-dependent manifolds

References

[1] Yann Ollivier.
Riemannian metrics for neural networks i: feedforward networks.
Information and Inference: A Journal of the IMA, 4(2):108–153, 2015.

[2] NN Cencov.
Statistical decision rules and optimal inference.
Transactions of Mathematical, 53, 1982.

[3] LL Campbell.
An extended Čencov characterization of the information metric.
Proceedings of the American Mathematical Society, 98(1):135–141, 1986.

[4] Balint Daroczy, David Siklosi, Robert Palovics, and Andras A Benczur.
Text classification kernels for quality prediction over the c3 data set.
In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW'15*, pages 1441–1446, 2015.

[5] Balint Daroczy, Rita Aleksziev, and Andras Benczur.
Expressive power of outer product manifolds on feed-forward neural networks.
Under submission.

MTA SZTAKI
Hungarian Academy of Sciences
Institute for Computer Science and Control