Generalization Ideas in Deep Learning

Marcel Bruckner

Seminar: Optimization and Generalization in Deep Learning

Abstract

Trying to understand the generalization abilities of deep neural networks several capacity measures are experimentally explored. Following the thoughts of [3] different norm-based capacity measures over the network weights and the sharpness as the robustness to perturbations on the parameter space are investigated. The measures are used in different experiments and the results are plotted against each other.

1 Introduction

Huge empirical success of deep neural networks
Complex, non-convex optimization problem
Simple stochastic gradient descent (SGD) methods achieve good generalization
Even in this highly over parameterized setting
Many local minima with different generalization behavior

2 What is a matrix norm and which do we use

2.1 What is a matrix norm

What does it do to the function

2.2 Why do we use matrix norms as measures for capacity bounds

how controls the norm the model capacity

2.3 Which norms do we use

- $l2$ norm with capacity proportional to $\frac{1}{\gamma^2_{margin}} \prod_{i=1}^{d} 4 \left\|W_i\right\|_F^2$ (1)
- $l1$-path norm with capacity proportional to $\frac{1}{\gamma^2_{margin}} \left|\sum_{j \in \prod_{k=0}^{d}[h_k]} \left|\prod_{i=1}^{d} 2W_i[j_i, j_{i-1}]\right|\right|^2$ (2)
- $l2$-path norm with capacity proportional to $\frac{1}{\gamma^2_{margin}} \sum_{j \in \prod_{k=0}^{d}[h_k]} \prod_{i=1}^{d} 4h_i W_i^2[j_i, j_{i-1}]$ (3)
- spectral norm with capacity proportional to $\frac{1}{\gamma^2_{margin}} \prod_{i=1}^{d} h_i \left\|W_i\right\|_2^2$ (4)

3 What is sharpness

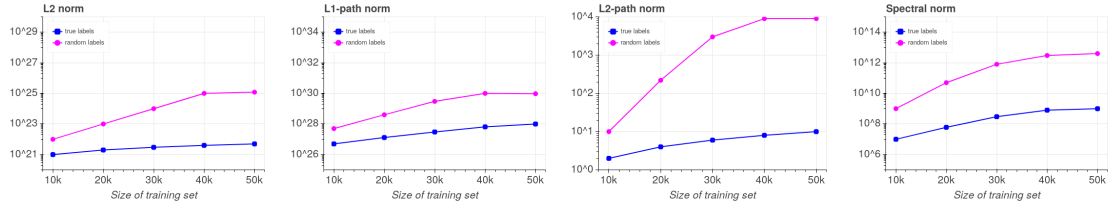

Figure 1: The proposed norms calculated on the network weights of the VGG model [6] after training on random (magenta) and true (blue) labeled subsets of CIFAR-10 are displayed. We see only a small increase of the measures on true labels but a huge increase using random labels and the measures on random labels to be bigger for every subset.

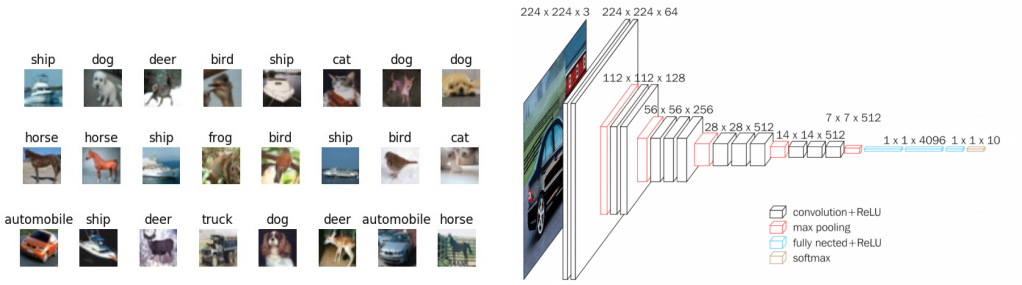

Figure 2: Left: A subset of the CIFAR-10 small image data set with ten classes. Right: The used implementation of the VGG-16 network with batch normalization.

4 Empirical study

In this section we discuss the experimental results using the proposed measures and their implications on different phenomena seen in practice regarding the generalization behavior.

4.1 Difference between true and random labeled data

It is possible to obtain zero training error on random labels using the same architecture for which training with real labels leads to good generalization. We would expect the networks learned using real labels (and which generalizes well) to have much lower complexity, under the suggested measure, than those learned using random labels (and which obviously do not generalize well). [3]

Figure 1 shows the results of [3]. We see that training the network on true labels (blue) only results in small increases of the norm-based capacity measures in comparison to the increases when training on random labels (magenta). As the network learns the true functional dependence between the input and output using true labels it only requires small increases in complexity as the subset size increases, whereas it requires more capacity for every newly seen data point when using random labels as it needs to memorize the data point.
We also see that the norm-based capacity measures are all bigger for random labeled data as the network has to memorize the data whereas the dependence between the input and output can be learned with lower capacity on true labeled data.

4.1.1 Problems reproducing results

Following the vague instructions of [3] the VGG-16 network with batch normalization (Fig: 2) is trained on different subsets of the CIFAR-10 (Fig: 2) data set. For every subset a copy with the labels replaced by random labels has been created and the model is trained on both the true and random labeled subset.
The huge scope for interpretation of the approach used in [3] caused the non reproducibility of the results. The problems we face are:

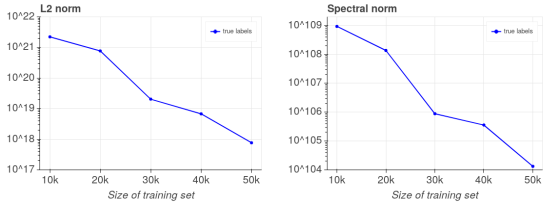

Figure 3: The $l2$ and spectral norm calculated on the network weights of the VGG-16 model after training on subsets of CIFAR-10 are displayed. The results show no reproducibility of the results presented by [3].

Reproduction of network training Neyshabur et al. [3] only stated that their results were produced using VGG. As VGG is a class of CNN architectures with varying depths and optional batch normalization a grid search over the implementations is performed. Considering the capacity vs. training time trade-of the best results were achieved using VGG-16 with batch normalization. Furthermore, [3] didn't provide information on the used solving strategy. Therefore, we tried simple stochastic gradient descent and Adam but stuck with the later as of the faster convergence.

Calculation of norms The calculation of path norms is trivial for the fully connected case but not for convolutional layers. As we need to enumerate every path through the network we need a way to associate the layer neurons of succeeding layers when applying a convolution. In the short time of this paper we couldn't come up with a feasible implementation which made it impossible to reproduce the path norms.

Training on random labeled data Following the missing details of the network implementation and solving strategy a training on random labeled data was not possible. The network didn't overcome a 1/10 test accuracy on random labels which represents a random choice of the result by the network out of the 10 classes in the CIFAR-10 data set.

Training time The short scope of the seminar enclosing this paper combined with training times ranging from roughly five hours on the smallest subset ranging up to over one day for the biggest one resulted in only small search intervals for the grid search used on the hyperparameters as well the different algorithmic and architectural choices.

4.2 Difference between local minima

When training the same architecture, with the same training set, using two different optimization methods (or different algorithmic or parameter choices), one method results in better generalization even though both lead to zero training error. We would expect to see a correlation between the complexity measure and generalization ability among zero-training error models. [3]

4.3 Implications of different hidden unit sizes

Increasing the number of hidden units, thereby increasing the number of parameters, can lead to a decrease in generalization error even when the training error does not decrease. We would expect to see the complexity measure decrease as we increase the number of hidden units. [3]

5 A few remarks

Each report should include an introduction describing the problem, the motivations and a brief outline. The main approach should then be described and discussed in separate sections, followed by experimental results (when applicable) and conclusions.

- Please use citations when appropriate. Again, you are not expected to read through all the references appearing in your assigned paper. Add your citations in bibtex format into the file egbib.bib. An example is [3].
- You can use the theorem environment to write theorems. An example:

  Theorem 1. *Let p be a prime number. Then, for any $a \in \mathbb{N}$, $a^p - a$ is evenly divisible by p. More formally,*

  $$a^p \equiv a \,(\mathrm{mod}\, p)\,. \tag{5}$$

- Please keep all your formulas numbered.
- The report should be 4 to 6 pages long (not including citations).
- Reports must be in English.
- Please do not change the layout (*e.g.*, do not change page margins, font size, etc.).

References

[1] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, pages 224–240, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[2] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning, 2017.

[3] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5947–5956, 2017.

[4] Behnam Neyshabur, Ruslan Salakhutdinov, and Nathan Srebro. Path-sgd: Path-normalized optimization in deep neural networks. *CoRR*, abs/1506.02617, 2015.

[5] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. *CoRR*, abs/1503.00036, 2015.

[6] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[7] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530, 2016.