Spectral Norm Regularization for Improving the Generalizability of Deep Learning

Authors

Kefang Liu
Tommy He
Shangyang Shang

Emails

kefang.liu@mail.mcgill.ca
tommy.he@mail.mcgill.ca
shangyang.shang@mail.mcgill.ca

1. Introduction

A central question in deep learning, despite its massive success, is still what precisely allows it to generalize so well, though there has been progress [2-4]. Yoshida and Miyato [1] approach the problem of generalizability of deep learning by considering its sensitivity to input perturbation. They note that adversarial training is designed to achieve insensitivity to perturbation of training data. However, insentivity to perturbation of test data is most important for generalizability [5][6]. So, they propose using spectral norm regularization as a simple yet effective means to achieve both. Using practical datasets, they experimentally confirm that spectral norm regularization helps models generalize better than other baseline methods. They also show that insensivity to perturbation of test data is indeed an important factor for generalizability, and that spectral norm regularization does so by reducing the spectral norms of the weight matrices.

2. Spectral Norm Regularization

2.1 Basic Idea

Consider a feed-forward neural network with L layers where $x^l \in \mathbb{R}^{nl}$ is the vector of features of the l-th layer with. Then, $x^l = f^l(W^l x^{l-1} + b^l)$ for $l = 1,2, \dots, L$. We have nonlinear activation functions $f^l: \mathbb{R}^{n0} \to \mathbb{R}^{nl}$, weights $W^l \in \mathbb{R}^{n_l \times n_{l-1}}$ and biases $b^l \in \mathbb{R}^{n_l}$. Let $\theta = \{W^l, b^l\}_{l=1}^{L} \in \mathbb{R}^{n_l \times n_{l-1}}$ denote the model parameters and the model function as $f^l: \mathbb{R}^{n0} \to \mathbb{R}^{nl}$.

To make our model insensitive to input perturbation, we aim to minimize $||f\ (x + \xi) - f(x)||_2$. Notice that neural networks are nonlinear only due to the activation functions, which tend to be piecewise linear such as ReLU and maxpooling. Therefore, if we consider a small enough neighborhood of x, then we can effectively take f_θ to be linear. Let's represent the locally linear map as the linear map $x \mapsto W_\theta x$ + $b_\theta x$. Now, we have

$$\frac{||f\ (x + \xi) - f\ (x)||_2}{||\xi||_2} = \frac{||W_\theta\ (x + \xi) + b - W_\theta x + b_\theta||_2}{||\xi||_2}$$

$$= \frac{||W_\theta \xi||_2}{||\xi||_2}$$

$$\leq \sigma(W_\theta)$$

where $\sigma(W_\theta)$ is the spectral norm of W_θ. By definition, the spectral norm of a matrix $A \in \mathbb{R}^{m \times n}$ is

$$\sigma(A) = max_{\xi \in \mathbb{R}^{m \times n},\ \ \xi \neq 0} \frac{||W_\theta \xi||_2}{||\xi||_2}$$

which corresponds to the largest singular value. Therefore, x_θ is insensitive to perturbations in x if the spectral norm of W_θ is small. By this argument, it seems that we should train θ to reduce the spectral norm of W_θ.

Let us assume for now that each f^l is a ReLU [7]. Then, each f^l is a diagonal matrix $D_\theta^l \in \mathbb{R}^{n_l \times n_l}$ where an element in the diagonal is 1 if and only if the corresponding element x^{l-1} is positive, or else it is 0. Then, we have

$$W_\theta = D_\theta^L W^l \cdots D_\theta^1 W^1$$

Now, note that $\sigma\left(D_\theta^l\right) \leq 1 \forall l: 1 \leq l \leq L$. Therefore, by submultiplicativity of the spectral norm,

$$\sigma(W_\theta) = \sigma(D_\theta^L)\sigma(W^l) \cdots \sigma(D_\theta^1)\sigma(W^l) \leq \prod_{l=1}^{L} \sigma(W^l)$$

So, it makes sense for us to bound the spectral norm of W_l for all l: $1 \leq l \leq L$, which gives some theoretical motivation for spectral norm regularization.

2.2 Details of Spectral Norm Regularization Gradient Descent

Suppose we have the training data $(x_i, y_i)_{i=1}^{K}$ for $x_i \in \mathbb{R}^{n_0}$, $x_i \in \mathbb{R}^{n_L}$. We define the loss over the training data to be $\frac{1}{K}\sum_{i=1}^{K} L(f_\theta(x_i), y_i)$ for some loss function L. Then, we consider minimizing the empirical risk

$$min_\theta(\frac{1}{K}\sum_{i=1}^{K} L(f_\Theta(x_i), y_i)) + (\frac{\lambda}{2}\sum_{l=2}^{L} \sigma(W^l)^2)$$

for regularization factor $\lambda \in \mathbb{R}_+$. We call the second term the spectral norm regularization, as its purpose is to decrease the spectral norm of the weight matrices. When $\sigma 1 > \sigma 2$, we have the gradient

$$\nabla \frac{\sigma(W^l)^2}{2} = \sigma_1 u_1 v_1^T$$

Note that we can assume $\sigma_1 > \sigma_2$ due to numerical errors. It is, however, computationally difficult to compute σ_1, u_1, v_1, so we will approximate this value by updating $u \leftarrow W^l v$, $v \leftarrow (W^l)^T u$, $\sigma \leftarrow \frac{||u||_2}{||v||_2}$ until it is sufficient. With this, σ, u, v will converge to σ_1, u_1, v_1 as long as $\sigma_1 > \sigma_2$, which we are assuming. The pseudocode for such an algorithm is provided below from [1].

Table 1. An Algorithm based on SGD with spectral norm regularization

```
for $l = 1\ to\ L$ do
$v^l \leftarrow$ a random Gaussian vector.
for each iteration of SGD do
     Consider a minibatch, $\{(x_{i1}, y_{i1}), \dots, (x_{ik}, y_{ik})\}$, from training data.
     Compute the gradient of $\frac{1}{K}\sum_{i=1}^{K} L(f_\Theta(x_i), y_i))$ with respect to Θ.
    for $l = 1\ to\ L$ do
          for a sufficient number of times do
               $u \leftarrow W^l v, v \leftarrow (W^l)^T u, \sigma \leftarrow \frac{||u||_2}{||v||_2}$
               Add $\lambda \sigma^l u^l (v^l)^T$ to the gradient of $W^l$
    Update Θ using the gradient.
```

3. Comparison with other Regularization Methods

In Chapter 1 and 2, it has been noted that spectral norm regularization is an effective method to achieve high generalizability for deep learning models. In this part, several popular generalization techniques will be introduced. The comparison between these regularization techniques help to further evaluate the effectiveness of SNR.

3.1 Weight Decay

Weight decay [8], also known as Frobenius norm regularization, is a famous regularization technique in deep learning area. It considers the following empirical risk minimization problem:

$$minimize \frac{1}{K}\sum_{i=1}^{K} L(f_\Theta(x_i), y_i) + \frac{\lambda}{2}\sum_{l=1}^{L} ||W^l||_F^2$$

$$||W^l||_F^2 = \sum_{i=1}^{\min\{n_{l-1}, n_l\}} \sigma^i(W^l)^2$$

where $\lambda \in \mathbb{R}_+$ is regularization factor, $\sigma^i(W^l)^2$ is the i-th squared singular value of matrix W^l.

It is clear that $||W^l||_F^2$, the Frobenius norm of W^l, reduced $\sigma^i(W^l)^2$. In other words, the trained weight matrix W^l shrinks in all the directions. Hence, it is of high possibility that the trained model loses important information about the input. On the other hand, W^l shrinks in limited directions when SNR is applied. It is because that SNR only focuses on the first singular value instead of all of them [9].

3.2 Adversarial Training

Adversarial training generalization considers the following problem[10]

$$minimize\ \alpha \cdot \frac{1}{K}\sum_{i=1}^{K} L(f_\Theta(x_i), y_i) + (1-\alpha)\frac{1}{K}\sum_{i=1}^{K} L(f_\Theta(x_i + \eta_i), y_i)$$

where

$$\eta_i = \epsilon \cdot \frac{g_i}{||g_i||_2}$$

$$g_i = \nabla_x L(f_\Theta(x_i), y_i)|_{x=x_i}$$

and α and $\epsilon \in \mathbb{R}_+$ are both regularization factors. It can be observed that adversarial training considers the perturbation towards the directions that affect the loss function most. Consider two directions x_1 and x_2. The perturbation changes drastically in the direction x_1 while it doesn't change toward the direction of x_2. x_1 gives a significant contribution to the empirical risk minimization problem while x_2 shows almost no impact. Thus, the model trained by adversarial training is insensitive to the perturbation to the training data. In contrast, the model trained by spectral norm regularization could perform insensitivity to the perturbation of training data and test data[11].

4. Experiments

In chapter 3, two well-known regularization techniques, weight decay or adversarial training, are introduced. Also, it is clarified in theory level that spectral the models trained by norm regularization could perform better insensitivity to the perturbation than those trained by weight decay or adversarial training. This chapter focuses on the experiments where the effectiveness of spectral norm regularization over regularization techniques will be demonstrated.

Stochastic gradient descent (SGD) boosted by different regularization techniques is the fundamental method to train the models. Classification tasks are chosen to test the models. The mini-batch sizes are set as B=64 and B=4096 respectively in the small and large batch regimes

Four regularization techniques are compared in the experiments, including vanilla (no regularization), weight decay, adversarial training, and spectral norm regularization. The regularization factor, λ, is set as 10^{-4} in weight decay. The hyperparameters $\alpha = 0.5$ and $\epsilon = 1$ are assigned respectively. Finally, $\lambda = 0.01$ is selected in spectral norm regularization.

Four experiments are implemented in total. Different model and dataset settings are selected in each experiment, shown by Table 2. In experiment 1 and 3, the leaning rate of SGD is 0.01 for the small-batch regime and 0.1 for the large-batch regime. In experiment 2 and 4, the leaning rate of SGD is 0.01 for the small-batch regime and 1.0 for the large-batch regime.

Table 2. Model and Dataset Settings in the Experiments

Experiment Number	Model	Dataset
1	VGG Network	CIFAR-10
2	Network in Network (NIN)	CIFAR-100
3	Densely Connected Convolutional network (DenseNet)	CIFAR-100
4	DenseNet	STL-10

5. Results

In terms of the results, initially, by using the four different regularization techniques discussed above, in the batch size of 64 for small regime and 4096 for large regime, both the test accuracy and generalization gap were reported and compared among the four different models as well as the correlated datasets. For the performance of test accuracies, as indicated in Table 3 below, Spectral Norm Regularization (SNR) method achieved the largest test accuracies among the large batch regime tests and as well as when using NIN and CIFAR-100 DenseNet.

Table 3. Performance of Test Accuracies

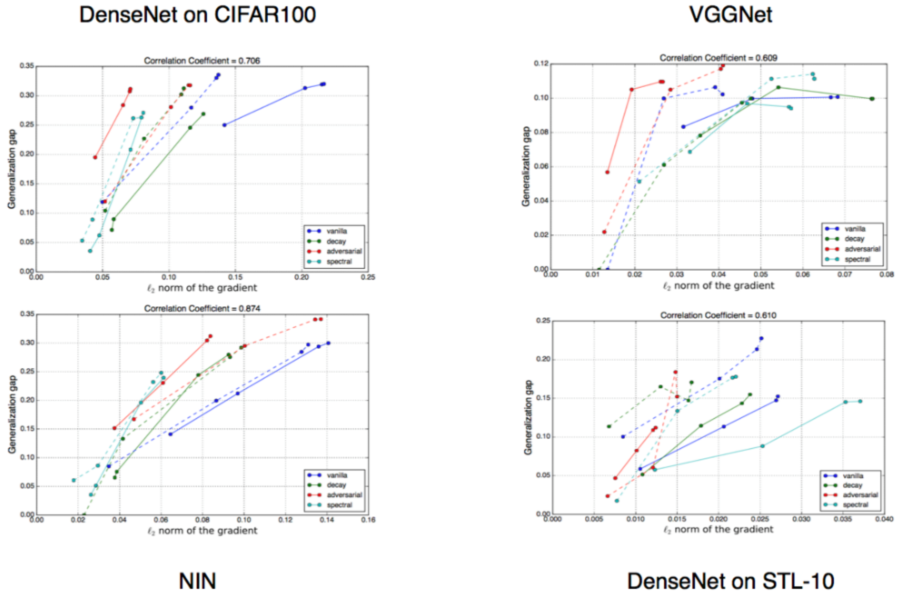
Model & Dataset	Batch Size	Vanilla	Weight Decay	Adversarial Learning	SNR
VGGNET (CIFAR-10)	64	0.898	0.897	0.884	0.904
	4096	0.858	0.863	0.870	0.885
NIN (CIFAR-100)	64	0.626	0.672	0.627	0.669
	4096	0.597	0.618	0.607	0.640
DenseNet (CIFAR-100)	64	0.675	0.718	0.675	0.709
	4096	0.639	0.671	0.649	0.697
DenseNet (STL-10)	64	0.724	0.723	0.707	0.735
	4096	0.686	0.689	0.676	0.697

For the generalization gap results, please note that generalization gap was computed through calculating the minimum difference between training and test accuracies when the magnitude of accuracies has reached a defined threshold value. We could observe from Table 4 below that, for the performance of Spectral Norm Regularization (SNR) method, it reached the smallest generalization gap though the comparison of 3 out of 4 models and datasets (VGGNET, NIN, and CIFAR-100 DenseNet).

Table 4. Performance of Generalization Gap

Model & Dataset	Batch Size	Threshold	Vanilla	Weight Decay	Adversarial Learning	SNR
VGGNET	64	0.88	0.079	0.074	0.109	0.068
(CIFAR-10)	4096	0.85	0.092	0.064	0.064	0.045
NIN	64	0.62	0.231	0.120	0.253	0.090
(CIFAR-100)	4096	0.59	0.205	0.119	0.196	0.090
DenseNet	64	0.67	0.317	0.080	0.299	0.095
(CIFAR-100)	4096	0.63	0.235	0.111	0.110	0.051
DenseNet	64	0.70	0.063	0.073	0.069	0.068
(STL-10)	4096	0.67	0.096	0.057	0.015	0.042

Moreover, for the input perturbation sensitivity of the test data, the performance of Spectral Norm Regularization (SNR) method towards four different models and datasets is shown in Figure 1 below through the plot of generalization gap versus the gradient of loss function. For the test data, to have a relatively lower sensitivity to the fluctuation in the input, both the generalization gap as well as the loss function gradient should be in the relatively small magnitude [2]. Thus, when applying SNR method, on the CIFAR-100 DenseNet and STL-10 DenseNet, the input perturbation was kept at a lower level.


DenseNet on CIFAR100

VGGNet

NIN

DenseNet on STL-10

Figure 1. Relationship of Generalization Gap and Gradient of Loss Function

6. Conclusion and Future Work

To conclude, through the theoretical study and experiments, compared with baseline pathways including Vanilla Problem, Weight Decay, as well as Adversarial Learning, the Spectral Norm Regularization (SNR) method showed it capability of reducing the input perturbation sensitivity and better generalizability.

In terms of the future work, in the original SNR paper, the authors pointed out that clearer theoretical understanding of SNR method on generalization and the discussion of potential effect of prevent neural networks from fitting noises by using SNR method should be taken into the scope. Moreover, we also would like to propose that more different models and datasets are to be tested. And different batch sizes need to be tried as well including medium sized regime.

7. Self-Grading/Assessment

For this group project, we evaluate our work as E (excellent) as although there are a lot of potential challenges, including the schedule conflict, unfamiliarity of the topics, as well as lack of source codes, we managed to overcome those and come up with a presentation and report which we suppose worth re-using for future students.

During this group work, each of the group members contribute equally and generously by finishing all the assigned tasks in time. We did not encounter any problem through this cooperating process. And comparing with the initial project proposal, more theoretical study was done instead of coding according to our accessible sources.

References

[1] Yoshida, Y., & Miyato, T. (2017). Spectral Norm Regularization for Improving the Generalizability of Deep Learning. arXiv. https://doi.org/10.48550/ARXIV.1705.10941

[2] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning - generalization gap and sharp minima. In ICLR, 2017.

[3] Bartlett, Peter L., Dylan J. Foster, and Matus J. Telgarsky. "Spectrally-normalized margin bounds for neural networks." Advances in neural information processing systems 30 (2017).

[4] Sankaranarayanan, Swami, et al. "Regularizing deep networks using efficient layerwise adversarial training." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.

[5] Zhang, Guodong, et al. "Three mechanisms of weight decay regularization." arXiv preprint arXiv:1810.12281 (2018).

[6] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, Yuichi Yoshida. "Spectral Normalization for Generative Adversarial Networks". ICLR2018.

[7] Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).

[8] Bartlett, Peter L., Dylan J. Foster, and Matus J. Telgarsky. "Spectrally-normalized margin bounds for neural networks." Advances in neural information processing systems 30 (2017).

[9] Loshchilov, Ilya, and Frank Hutter. "Fixing weight decay regularization in adam." (2018).

[10] Nasr, Milad, Reza Shokri, and Amir Houmansadr. "Machine learning with membership privacy using adversarial regularization." Proceedings of the 2018 ACM SIGSAC conference on computer and communications security. 2018.

[11] Sato, Motoki, Jun Suzuki, and Shun Kiyono. "Effective adversarial regularization for neural machine translation." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019.