Nirbhay Sharma (B19CSE114)

Regularizing Neural Networks via Adversarial Model Perturbations

---

Summary

The authors proposes a regularization scheme which utilizes perturbations in model parameters to yield a more robust and generalized model which is robust to overfitting. The recent research shows the effectiveness of certain regularizer lies in its ability to reach flat minima. Thus flat minima becomes a crucial aspect in designing any regularizer techniques. To the same end, the authors proposes a novel loss function ($\mathcal{L}_{AMP}(\theta)$) which acts as a regularizer loss function. The core idea behind this loss function is to reach the flat minima. The significance of flat minima is such that, even if the test data has some slight difference in distribution than train data, due to flat minima it eventually reaches to minima thus flat minima helps in improving the robustness of the model. The empirical risk loss function which is naive loss function is prone to overfitting thus authors tries to make a robust empirical loss function by introducing model perturbations. The empirical loss function is as follows,

$$L_{EMP}(\theta) = \frac{1}{|D|}\sum_{i=1}^{|D|} l(x_i, y_i; \theta)$$

where D is the dataset and (x_i, y_i) are i^{th} datapoint in the dataset. and l is any general loss function for any task such as image classification. Then, authors tried to formulate the $L_{AMP}(\theta)$ loss function as follows.

Consider ϕ to be model's weight space, ϵ as the positive small number, $\mu \in \phi$. Authors define $B(\mu, \epsilon)$ as norm ball in ϕ with center μ and radius ϵ and can be represented as

$$B(\mu, \epsilon) = \{\theta \in \phi : ||\theta - \mu|| \leq \epsilon\}$$

The target of L_{AMP} is to keep the model parameters under this norm ball with radius ϵ and center as $\mu = 0$. Thus the L_{AMP} can be formulated as

$$L_{AMP}(\theta) = max_{\delta \in B(0,\epsilon)}(\frac{1}{|D|}\sum_{i=1}^{|D|} l(x_i, y_i; \theta + \delta))$$

The target is to minimize the $L_{AMP}(\theta)$ loss function.

The authors performed experiments with different datasets such as SVHN, CIFAR10, CIFAR100 as well and on different models such as preresnet, wideresnet etc. We choose the new dataset as GTSRB [3] dataset which has 43 classes in total. The results are shown below on one of the paper's dataset and on another dataset which is not included in the paper.

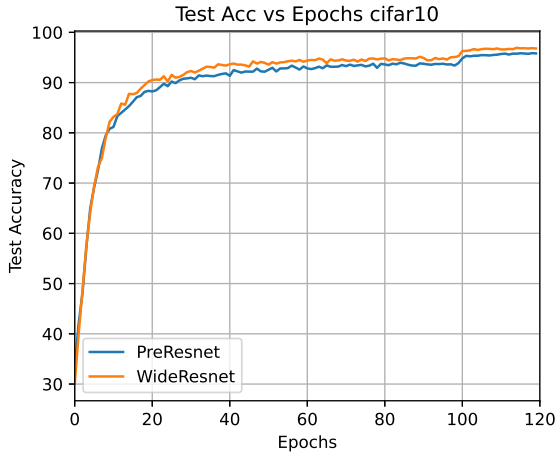
The explaination to why $L_{AMP}(\theta)$ loss function tries to find flatter minima lies in the fact that, it poses a penalty on the gradient norm and tries to keep it under L_2 norm ball of radius ϵ. Thus it not only tries to minimize the loss function but also tries to find the minima which has small gradient norm near the minima as well. small gradient norm near minima essentially imples falt minima.

Results

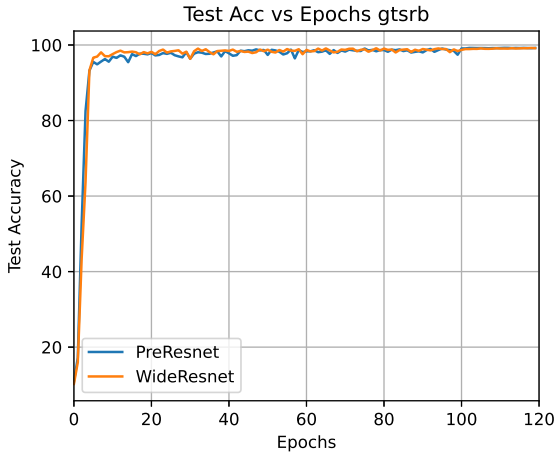
The results are presented and visualized in Table and Fig. a and Fig. b respectively as shown below. The same set of hyperparameters are being used in order to reproduce the results. The results clearly shows the generalization capability of the approach.

Dataset	Model	Test Accuracy	Train Accuracy
CIFAR10	Preresnet	95.86	98.87
CIFAR10	WideResnet	96.79	99.87
GTSRB	Preresnet	99.25	100
GTSRB	WideResnet	99.22	100

The above table clearly shows the effectiveness of using AMP regularizer and there is not considerable difference between train and test accuracy. The $L_{AMP}(\theta)$ loss effectively regularize the network and eventually reached the flat minima thus reducing the testing error by a significant amount. The regularizer not only works with different dataset (CIFAR10, GTSRB) but it also shows consistent performance with other networks as well such as (PreResnet, WideResnet) etc. Thus this approach is highly effective in terms of reularizing network and achieves flatter minima to achieve low test error. Fig. a and Fig. b also clearly shows the performance improvement in terms of test accuracy as the epochs are advanced. The test accuracy is continuously increasing indicating the consitent improvement in performace every epoch.



(a) Test results on CIFAR10



(b) Test results on GTSRB

References

1. code github
2. regularizing neural network via adversarial model perturbation
3. gtsrb dataset