Insights, conclusion and Critics on Paper (Understanding Deep Learning requires rethinking generalization) best paper award at ICLR 2017

Insights

ABSTRACT

1)The paper lights on how the traditional approaches fail to explain why large neural networks generalize well in practice.

2)What is it, that distinguishes neural networks that generalize well from those that don't?

3)In this work, we problematize the traditional view of generalization by showing that it is incapable of distinguishing between different neural networks that have radically different generalization performance.

Randomized test: -

In a first set of experiments, they trained several standard architectures on a copy of the data where the true labels were replaced by random labels.

The following observations were made

1.the neural network easily contoured random labels

2.As after training the model, it can be visualized that there was 0 training error no correlation between test and train labels (due to random labeling of data).

3. Randomization can considerably hike the generalized error without any change in the model, hyperparameter or the optimizer.

Further increasing the noise into the CNN deteriorated, some generalized error and it can be understood that neural network can fit the random noise with 0 training error and are able to capture the remaining signal in the data.

The explicit regularization (weight decay, dropout Regularization and data augmentation)

It can be summed up that explicit regularization may improve the generalization performance but it is not a method itself to be muffled on a model to drop generalization error.

However, it was found that explicit regularization works differently in deep learning models as it works like a parameter tuning which improves the model by improving final test error

The role of implicit regularization

As SGD always converges to a solution for linear model as well some models work great by using gaussian kernel, it is needed to investigate the paradigm that these algorithm use.

Related works: -

As uniform stability of algorithm Is independent of labeling of data. Hence, we can't use the same concept for distinguish between model with low generalized error and model with high generalized error.

Effective capacity of neural network: -

For understanding the effective model capacity of ANN, a model is trained with both the true labels and random labels respectively.

For the 2nd case learning loss is found to be higher and learning becomes impossible.

For extracting the conclusion, different level of randomization was done and with one of the best convolution neural networks like Alex net as a result test error growth and relative convergence slowdown was seen.

Fitting random labels and pixels: -

The same data was modified with these types

True label, Partially corrected label, Random label, Shuffled pixels, Random pixels, Gaussian Distributed pixels.

The learning was done on Inception model. And it was observed that random labels were fitted easily without any prior change in learning rates schedule. No hyperparameter tuning was done when switching from true to random labels. The graph visualizes the learning curve for each of the modified data.

Partially corrupted labels: -

For inspecting the behavior again, the model was trained using different label of corruptions rising from 0 to 1(no corruption to complete random labels). It was observed the network fit corrupted training set perfectly for all the cases.

As we moved to higher level of corruption a hike in generalization error was observed as it converged to ~90%.

Implications

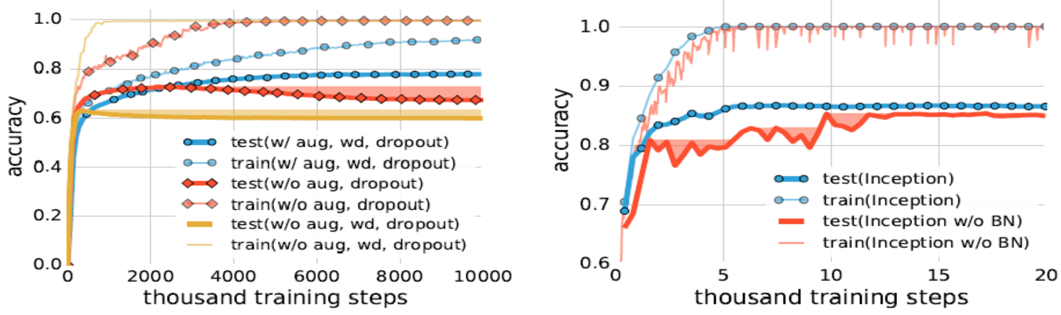
Rademacher complexity and VC-dimension: -

Rademacher complexity is used to measure richness of a class and we can bound the generalization error of a hypothesis in terms of its empirical error and the Rademacher complexity of the class of loss function (source: http://www.cs.cmu.edu/~ninamf/ML11/lect1117.pdf). As our neural networks fit the training set with random labels perfectly, H ≈ 1 was expected for corresponding model class H. But this upper bound does not lead to useful generalization bounds in realistic settings.

So, both VC-Dimensions and Rademacher complexity were not able to explain the generalization behavior that was to be expected.

Uniform stability: -

uniform stability of an algorithm defines how well the algorithms performs after changing one or two examples from training set. And it is done with some notion of stability .As It is possible to define weaker notions of stability. The weakest stability measure was found to be correlated with bounding generalization error. But it is difficult to utilize the notion efficiently. Hence it couldn't justify generalization behavior clearly.



(a) Inception on ImageNet (b) Inception on CIFAR10

The role of regularization: -

A comparison of network architecture is done with different types of regularizers.

It was observed that

a) Early stopping can potentially improve generalization when the regularizers are absent (as the test accuracy reached ≈ 0.7 in upper bound of 2000 steps and then it decreased slightly).
b) On CIFAR10 early stopping was not proved necessary, but batch normalization stabilized the training process and outcome projected a slight gain in generalization.

From the table 2 In appendix: -

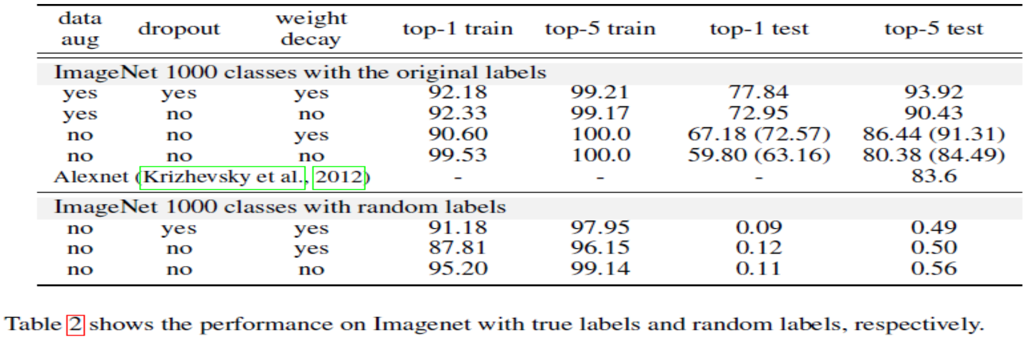
On ImageNet dataset top-1 accuracy was about 72 % without regularizations.

Using data Augmentation Inception achieved a accuracy of about 72.95 which shows that augmentation was more powerful that weight decay.

But in the case of Inception, it achieved a loss of 19.62% without regular while ILSCRC 2012 achieved a loss of only 16.4%.

It was concluded that bigger gains can be achieved by changing model architecture and regularizers are not always fundamentally proportional to increase in generalization capacity of deep nets.

They can help us in gaining a slightly better generalization capacity when tuned correctly.



data aug	dropout	weight decay	top-1 train	top-5 train	top-1 test	top-5 test
ImageNet 1000 classes with the original labels						
yes	yes	yes	92.18	99.21	77.84	93.92
yes	no	no	92.33	99.17	72.95	90.43
no	no	yes	90.60	100.0	67.18 (72.57)	86.44 (91.31)
no	no	no	99.53	100.0	59.80 (63.16)	80.38 (84.49)
Alexnet (Krizhevsky et al., 2012)			-	-	-	83.6
ImageNet 1000 classes with random labels						
no	yes	yes	91.18	97.95	0.09	0.49
no	no	yes	87.81	96.15	0.12	0.50
no	no	no	95.20	99.14	0.11	0.56

Table 2 shows the performance on Imagenet with true labels and random labels, respectively.

Finite sample expressivity: -

Analyzing final sample expressivity of Neural network for simplifying the things.

Expressive power of NN on a finite sample size of n

Prove shows that: -

For every k >= 2; there exists neural network with ReLU activations of depth k;

width O(n=k) and O(n + d) weights that can represent any function on a sample of size n in d

dimensions.

Implicit regularization: -

it is not necessarily easy to understand the source of generalization for linear models either.
In linear cases curvature of all optimal solution is same.

SGD probably gives hints to the solution. If we consider the solution of SGD we can come across a single equation having (XX_T_ = y) a unique solution and kernel trick is derived.

Fitting the training label yields in excellent results for the convex models.

Unfortunately, the minimum norm notion is also not predictive of generalization performance. As l2-norm for MNIST with no preprocessing is approx. 220 and with wavelet preprocessing the norm jumps up to 390. And a 2% drop in test error is found.

Conclusion: -

The effective capacity of neural network is quite sufficient to fit the entire dataset.

Training time and optimization on random labels was found to be less than the expected.

A precise formal measure to explain generalization is yet to be discovered.

The Paper lights on an important property of neural networks that deep nets architectures always fit the data set even it consist of random data points and even a lot of noise.

Bigger gains can be achieved by changing model architecture and regularizers are not always fundamentally proportional to increase in generalization capacity of deep nets but can help us in gaining a slightly better generalization capacity when tuned correctly.

This paper has a huge impact on my personal thinking of regularization and further open the doors for the world to generalization deserves more attention.

Critics: -

The paper was not able to justify perfectly how to control the generalization error with regularization and what about nonrandom data fitting. (As the paper has already won the best paper award at ICLR 2017 it’s so difficult to criticize it)