Understanding deep learning requires rethinking generalization

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals

In Proceedings of the 5th International Conference on Learning Representations, 2017

ML Journal Club

February 22, 2018

Traditional statistical learning theory

Vapnik-Chervonenkis-dimension

Rademacher complexity

Uniform stability

Effective capacity of NN

Question

- Role of SGD as implicit regularizer
- DNN = brute force?

Experiments

- Feed Forward NN – Input=images
- Random labels: Continuum between no label noise and completely random ones
- Randomization of input images

Results

Learning random labels

- Tested networks were always able to fit.
- No need to change the learning rate schedule
- Once the fitting starts, it converges quickly.
- It converges to (over-)fit the training set perfectly.

Continuous corruption

- Slower convergence
- Converge to 90% test error.

Implications: Challenge for traditional approaches

- Rademacher complexity: $\mathcal{R}_n(H) = 1$
- Similar result for VC-dimension
- Uniform stability: weakest form still equivalent to bounding generalization/does not take data into account

Role of regularization

Basic ideas

- Regularizers: mitigate overfitting when more parameters than datapoints (confine learning to a subset of the hypothesis space).
- Explicit regularizers (e.g. norm penalization): effective Rademacher complexity highly reduced.

Role of regularization

In deep learning?

- In DNN: still able to fit the random training set even with dropout, weight decay. (Exception: AlexNet with weight decay) (table 1).
- Still able to fit the true labels and generalize well, without regularization (table 1).
- While regularization is important, bigger gains can be achieved by simply changing the model architecture (Inception vs winner ILSVRC 2012).
- Difficult to say that regularizers = fundamental phase change in the generalization capability of deep nets.

Role of regularization

Implicit regularizers?

- Explicit = Dropout, weight decay, data augmentation, etc.
- Implicit = early stopping, batch normalisation, SGD = regularization as an unattended consequence
- SGD = regularizer, confirmed by *Empirical Analysis of Deep Network Loss Surfaces, ICLR 2017*
- Again no crucial role

Finite-sample expressivity

Population-level results

- Expressivity of NN: most results at population level.
- Possible to transfer population-level results to finite-size samples but requirement of unrealistic size.

Finite-size samples

- Sample size = n; dimension =d
- A two-layer NN (ReLu) with 2n + d weights can represent any function

Implicit regularization in linear models

- Not easy to understand generalization source for linear models as well.
- Quality of a global minimum? Curvature? In linear case, the curvature of all optimal solutions is the same.
- SGD? $\rightarrow$ unique solution, equivalent to minimum $\updownarrow_2$-norm solution

Conclusion

- Simple experiments to show effective capacity of several NNs, large enough to shatter the training data.
- Consequence: rich enough to memorize training data.
- Consequence: Challenge to statistical learning theory: how to explain generalization ability of NN?

- Optimization empirically easy even without generalization
- Consequence: reasons for easy optimization and generalization must be different.