EFFECT OF LOCAL SEARCH WITH REAL DATA ON GENERALIZATION IN DEEP LEARNING

YeonJu Lee [1] Sukho Lee [2]

1) *Division of Applied Mathematical Sciences, Korea University, Sejong 30019, KOREA*

2) *Division Computer Engineering, Dongseo University, Busan 47011, KOREA*

Corresponding Author : Sukho Lee, petra@gdsu.dongseo.ac.kr

ABSTRACT

One of the mystery in deep learning is why deep learning generalizes well in spite of the large number of parameters. The famous experiment of Zhang et al. has shown that the number of parameters in deep neural networks is so large that the deep neural network can memorize all the training dataset. This has aroused a lot of arguments on what could be the true reason for generalization. It has also been shown that the stochastic gradient descent method has an inherent property of regularization. However, since the stochastic gradient descent method is a local search method, there still exists the question why it does not end up into a meaningless local minimum but into a good minimum with enough generalization power. It is believed that this is due to the fact that the local search is performed on a energy landscape constructed from real data, so that the energy landscape has in fact a wide local minimum with generalization property. In this talk, I summarize some of the important recent researches, and suggest that for a better understanding we should work on the energy landscape which changes dynamically according to the batch change.

LOCAL SEARCH WITH REAL DATA IN DEEP LEARNING

The generalization gap in machine learning is

$$Generalization\ Gap = R[f_{A(S_m)}] - \hat{R}_{S_m}[f_{A(S_m)}] \leq \sqrt{\frac{N}{m}} \tag{1}$$

Here, $\hat{R}_{S_m}[f_{A(S_m)}]$ is the empirical risk of f, defined by $\hat{R}_{S_m}[f_{A(S_m)}] = \frac{1}{m}\sum_{i=1}^{m} \ell(f(x_i), y_i)$, where the function $f_{A(S_m)}$ is learned from a dataset S_m of m data with an algorithm A, and f is a true function which can ideally map an input x to a desired output y. The generalization gap is bounded by $\sqrt{\frac{N}{m}}$, where N is the effective capacity, for which the number of parameters can be used as an estimate. Normally, it is known that the effective capacity is vacuous for deep neural network models, since the number of parameters N is too large. Several deep neural network compression methods have been suggested that can compress the model so that the bound for generalization gap can be reduced. It can be shown that the effective capacity reduces if the weights in the neural network are updated by the following equation:

$$w''_{i,j,k} = w'_{i,j,k} - \alpha(\Delta_{i,j,k} - \frac{\partial L}{\partial w_{i,j,k}}). \tag{2}$$

where,

$$\begin{aligned}\Delta_{i,j,k} \quad &= p_i q_j \frac{\partial L}{\partial t_k} + q_j t_k \frac{\partial L}{\partial p_i} + p_i t_k \frac{\partial L}{\partial q_j} \\ -\alpha(p_i \frac{\partial L}{\partial q_j}\frac{\partial L}{\partial t_k} &+ q_j \frac{\partial L}{\partial p_i}\frac{\partial L}{\partial t_k} + t_k \frac{\partial L}{\partial p_i}\frac{\partial L}{\partial t_k}) + \alpha^2 \frac{\partial L}{\partial p_i}\frac{\partial L}{\partial q_j}\frac{\partial L}{\partial t_k}.\end{aligned} \tag{3}$$

and p, q, and t are the decomposed 1-D vectors of the 3-D convolution filters. This will reduce the vacuous bound to a certain extend. The reason that this kind of simple first-order approximated gradient flow can find a good local minimum is that the noise in it is possible to escape every saddle point in the landscape, and that the landscape itself is believed to be close to convex if the landscape is composed by data from the real world. However, in fact, the landscape changes its shape for every batch in the training dataset. Therefore, even if the landscape is not convex and the gradient computed from a particular landscape has the possibility to fall into a bad local minimum, the gradient computed by the next landscape can help the gradient to escape from this bad local minimum. This iterated gradient based update forces the solution to result in a common minimum shared by all the landscapes. We will show some experiments which supports these facts.

ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Science, ICT and Future Planning (No. NRF-2016R1D1A3B03931875)

REFERENCES

1. Kenji Kawaguchi., "Deep learning without poor local minima", *In Advances in Neural Information Processing Systems 2016a*, 2016.
2. Lei Wu, Zhanxing Zhu, et al., "Towards understanding generalization of deep learning: Perspective of loss landscapes", *arXiv preprint arXiv:1706.10239*, 2017.
3. Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, "Understanding deep learning requires rethinking generalization", *International Conference on Learning Representations (ICLR 2017)*, 2017.
4. Chiyuan Zhang, Samy Bengio, Moritz Gintare Karolina Dziugaite and Daniel M Roy, "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data", *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.