On Deep Learning

川口賢司

今回の留学報告書では、ディープラーニング（深層学習）の論文を簡単に紹介してみようと思います。ディープラーニングは実用が進んでいる分野ですが、一方で、理論的な説明ができない現象が多い分野でもあります。今回紹介する論文は、ディープラーニングの理論に関するものです。

This memo ***informally*** introduces some context and implications of the following technical report at MIT:

Deep Learning without Poor Local Minima [1]

(1) A very rough, informal summary of the context

In deep learning literature, poor local minima have been considered to be **not** a big problem by many practitioners since a long time ago. But, there is not yet a complete picture of why this is the case. This paper [1] provides a theory to explain the reason. It also partially provides a justification of why we should care about saddle points in a sense.

(2) Abstract [1]

In this paper, we prove a conjecture published in 1989 and also partially address an open problem announced at the Conference on Learning Theory (COLT) 2015. Under the independence assumption adopted from recent work, we prove the following statements for an expected loss function of a deep nonlinear neural network: 1) the function is non-convex and non-concave, 2) every local minimum is a global minimum, 3) every critical point that is not a global minimum is a saddle point, and 4) there exist "bad"" saddle points for the deep networks (with more than three layers) whereas there is no bad saddle point for the shallow networks (with three layers). Moreover, we prove that the same four statements hold for deep linear neural networks with any depth, any widths and no unrealistic assumptions. As a result, we present an instance, for which we can answer to the following question: how difficult to directly train a deep model in theory? It is more difficult than the classical machine learning models (because of the non-convexity), but not too difficult (because of the nonexistence of poor local minima). Furthermore, the existence of bad saddle points would suggest an open problem. We note that even though we have advanced the theoretical foundations of deep learning, there is still a gap between theory and practice.

(3) "No poor local minima" does not mean "training is easy"

Even if there is no poor local minimum, we can have bad saddle points, plateaus, and the ill-conditioning problem. Indeed, these problems can create training behaviors in practice that can be miss-diagnosed as being trapped around poor local minimum even if there is no poor local minimum.

(4) Beyond over-parameterization assumption and linearity assumption

Many great new theoretical papers are around now on the topic of local minima. These theoretical papers often make some simplification assumptions. Major simplification assumptions are *over-parameterization assumption* and *linearity assumption*.

Some people would claim that the linearity assumption is more reasonable than over-parameterization (because over-parameterization results in a simple model that would fail on the unseen new data especially in high-dimension) and others would claim the opposite (because linearity assumption results in a simple model). But, both of these assumptions are meant to be unreasonable in different ways and such claims seem to be not constructive.

This new paper, Deep Learning without Poor Local Minima [1], uses a weaker assumption than the previous papers and provides several theoretical results about local minima and saddle points.

References

[1] Kawaguchi, Kenji. Deep Learning without Poor Local Minima. Massachusetts Institute of Technology, Technical Report, MIT-CSAIL-TR-2016-005.