Information theory perspective of Deep Learning

Filip Boltuzic

2018
March

Understanding Neural networks via Information theory is done in [Schwartz-Ziv and Tishby, 2017]. They calculate the mutual information from a known distribution they attempt to sample (so they **do** know the mutual information). They show that in their toy model, which is a simple feedforward up-to 50 layered neural network, there exist two phases in training a neural network: *empirical error minimization* (ERM) and *representation compression*. They call the second phase the *stochastic relaxation* phase and argue this phase is the inefficient part of the DNN training, since it seems to behave like a random/Wiener process and could be done much more efficiently using other, more simple, approaches. Strong conclusions from this paper are:

- More than half of the training time in the neural network could be more efficiently spent (stochastic relaxation period)
- Adding hidden units speeds up the training time for good generalization
- Compression phase of a layer is shorter when it starts from a previously compressed layer (layers closer to the output are done "faster".

More to come when the review process for a counter paper `https://openreview.net/pdf?id=ry_WPG-A-` gets published

References

[Schwartz-Ziv and Tishby, 2017] Schwartz-Ziv, R. and Tishby, N. (2017). Opening The Black Box Of Deep Neural Networks Via Information.