Contents

1 A Bayesian Perspective on Generalization and Stochastic Gradient Descent

The question proposed in [1]: why large neural networks generalize well in practice, and the neural network can easily memorize the random labeled training data. canbe understood by the Bayesian model comparison theroy.

First consider a simple classification model M with a single parameter ω.

The authors prove that the Bayesian evidence can be approximated by (detail proof canbe found in section 2):

$$p(y|x;M) \approx \exp\left\{-\left(C(\omega_0) + \frac{1}{2}\ln\left(\frac{c''(\omega_0)}{\lambda}\right)\right)\right\} \tag{1}$$

From equation (1), the evidence is controlled by:

1. the value of the cost function at the minimum
2. the logarithm of the ration of the curvature about this minimum compared to the reguarization constant

For a model contains p parameters (given by [3]):

$$p(y|x;M) \approx \exp\left\{-\left(C(\omega_0) + \frac{1}{2}\sum_{i=1}^{p}\ln\frac{\lambda_i}{\lambda}\right)\right\} \tag{2}$$

where:

- $C(\omega;M)$: the L_2 regularized cross entropy, or cost funcions
- λ is the regularization conefficient
- ω_0 is the minimum of cost function
- λ_i is the eigenvalue of cost function's Hessian matrix

Insights from (1) and (2):

- the second term is often called "Occam factor"
  - it enforces Occam's razor: when two models describe the data equally well, the simpler model is ususlly better.
  - it describes the fraction of the prior parameter space consistent with the data
- minima with low curvature are simple, because the parameters do not have to be fine-tuned to fit the data.

1.1 Some findins suggested in the paper:

- generalization is strongly correlated with the Bayesian evidence: the weighted combination of the depth of a minimum (the cost function) and its breadth (the Occam factor).

- the gradient drives the SGD towards deep minima, while noise drives the SGD towards the broad minima.
- the test performance shows a peak at an optimal batch size which balances these competing contributions to the evidence.
- the SGD noise scale: $g = \epsilon(\frac{N}{B} - 1) \approx \epsilon\frac{N}{B}$, where N is the number of training samples, B is size of mini-batch, ϵ is the learning rate.
  - when we vary the batch size or the training set size, we shuld keep the noise scale fixed, which implies that $B_{opt} \propto \epsilon N$
  - progressively growing the batch size as new training data is collected.
- when using SGD with momentum, the noise scale : $g \approx \frac{\epsilon N}{B(1-m)}$, where m is momentum.

1.2 References

1. Zhang C, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization[J]. arXiv preprint arXiv:1611.03530, 2016.
2. Everything that Works Works Because it's Bayesian: Why Deep Nets Generalize?
3. Kass R E, Raftery A E. Bayes factors[J]. Journal of the american statistical association, 1995, 90(430): 773-795.