Contents

1 weight normalization

1.1 motivations

- the practical success of first-order gradient based optimization is highly dependent on the curvature of the objective that is optimized.
- If the condition number of the Hessian matrix of the objective at the optimum is low, the problem is said to exhibit pathological curvature.
  - when the condition number is low, the first-order gradient descent is hard to make progress.
- There may be multiple equivalent ways of parameterizing the same model.
  - some are much easier to optimize than others.
- Finding good ways of parameterizing neural networks: improve the conditioning of the cost gradient for general neural network.
  1. preconditioning.
  2. change the parameterization of the model to give gradients that are more like the whitened natural gradients.

  The later one is the way weight normalization chooses.

1.2 weight normalization

The computation of each neural is the weighted sum of input features followed by an element-wise nonlinearity, formulated as follows:

$$y = \phi(\mathbf{w} \cdot x + b) \tag{1}$$

weight normalization

1.2.1 forward pass

1. explicitly reparameterize each weight vector $\mathbf{w}$ in terms of a parameter vector $\mathbf{v}$ and a scalar parameter g.
2. perform stochastic gradient in the new parameters $\|\mathbf{v}\|$ and g.

$$\mathbf{w} = \frac{g}{\|\mathbf{v}\|}\mathbf{v} \tag{2}$$

1. the parameterization has the effect of fixing the Euclidean norm of the weight vector $\mathbf{w}$.
2. $\|\mathbf{w}\| = g$ independent of $\|\mathbf{v}\|$.

1.2.2 gradients

$$\nabla_g L = \frac{\nabla_w L \cdot \mathbf{v}}{\|\mathbf{v}\|}$$

$$\nabla_{\mathbf{v}} L = \frac{g}{\|\mathbf{v}\|} \nabla_{\mathbf{w}L} - \frac{g \nabla_g L}{\|\mathbf{v}^2\|} \mathbf{v}$$