4 Contractive Autoencoders

The aim of a contractive autoencoder is to make the learned representation be robust towards small changes around its training examples. As with most other autoencoder variations, this is done by adding a penalty term to the cost function that we are trying to minimize, which penalizes the representation's sensitivity to the training input. The first part of the cost function is the mean squared error, for linear reconstructions, as usual but the Frobenius norm of the Jacobian matrix is also added.[2] We can calculate it from the formula:

$$||J_h(x)||_F^2 = \sum_{ij} \left(\frac{\delta h_j(x)}{\delta x_i} \right)$$

The formula contains a partial derivative of the activation value of a neuron with respect to the input value, and so it is possible to see how a large increase in the activation value will correspond to an increase in the Jacobian, penalizing the representation.

From this formula we can see that sparse autoencoders are likely to correspond to a contractive mapping while not explicitly learning it through their learning criterion. This is due to the low activation values of the neurons in sparse autoencoders being likely to occur in the left part of the sigmoid activation function, which is almost flat. The neurons will therefore have a small first derivative which corresponds to a small entry in the Jacobian.

Contractive autoencoders are also very similar to denoising autoencoders. Both encourage robustness but while denoising autoencoders encourage it with the reconstruction $(f \circ g)(x)$, contractive autoencoders do so with the encoder function $f(x)$. This is important when relying on the robustness of the encoder function rather than the reconstruction, for example for classification which only uses the encoder. Denoising autoencoders also obtain robustness stochastically, by randomly adding noise to the input, while contractive autoencoders obtain robustness analytically, by penalizing the magnitude of the first derivative. [1]

References

[1] Rifai et al. Contractive auto-encoders: Explicit invariance during feature extraction. *Proceedings of the 28th International Conference on Machine Learning*, 2011.

[2] Agustinus Kristiadi. Deriving contractive autoencoder and implementing it in keras, March 2017.