Noname manuscript No.
(will be inserted by the editor)

Entropy loss function for unsupervised deep learning

Dylan M. Paiton

Abstract Here we describe a loss function for training a typical deep network in an unsupervised way. The objective of the loss is to encourage the deep network to minimize the entropy of its output while still conveying information. This should result in an intelligent clustering of object categories without the need of human generated object labels.

Keywords entropy · deep network · loss functions · unsupervised learning

1. A Few Notes

The input to the Entropy function is X, a vector of N neurons.
All sums are to be calcualted for every neuron in the network, N.

2. Useful Derivations

The following are a handful of derivatives that have to be taken multiple times for the entropy gradient calculation.

$$\frac{\partial \sum_i X_i}{\partial X_k} = \frac{\partial X_k}{\partial X_k} + \frac{\partial \sum_{i \neq k} X_i}{\partial X_k} = 1 \tag{1}$$

$$\frac{\partial \sum_i e^{-\beta X_i}}{\partial X_k} = \frac{\partial e^{-\beta X_k}}{\partial X_K} + \frac{\partial \sum_{i \neq k} e^{-\beta X_i}}{\partial X_k} = -\beta e^{-\beta X_k} \tag{2}$$

DM Paiton
Vision Science Graduate Group
Redwood Center for Theoretical Neuroscience
University of California, Berkeley
E-mail: dpaiton@berkeley.edu

$$\begin{aligned}\frac{\partial \sum_i \beta X_i e^{-\beta X_i}}{\partial X_k} &= \frac{\partial \beta X_k e^{-\beta X_k}}{\partial X_k} + \frac{\partial \sum_{i \neq k} \beta X_i e^{-\beta X_i}}{\partial X_k} \\ &= \frac{\partial \beta X_K}{\partial X_k} e^{-\beta X_k} + \beta X_k \frac{\partial e^{-\beta X_k}}{\partial X_k} \\ &= \beta e^{-\beta X_k} + \beta X_k (-\beta e^{-\beta X_k}) \\ &= \beta e^{-\beta X_k}(1 - \beta X_k)\end{aligned} \tag{3}$$

$$\frac{\partial \ln \sum_j e^{-\beta X_j}}{\partial X_k} = \frac{1}{\sum_j e^{-\beta X_j}} \frac{\partial \sum_j e^{-\beta X_j}}{\partial X_k} = \frac{-\beta e^{-\beta X_k}}{\sum_j e^{-\beta X_j}} \tag{4}$$

3. Forward Function - Entropy Computation

Our loss should include a term for minimizing entropy as well as a term for preserving information.

First the equation for entropy. We define our output probability from the network as

$$q_i = \frac{e^{-\beta X_i}}{\sum_j e^{-\beta X_j}}. \tag{5}$$

This can then plug into the entropy equation:

$$H = -\sum_i q_i log(q_i), \tag{6}$$

which expands to:

$$H = \sum_i \left(\frac{e^{-\beta X_i}}{\sum_j e^{-\beta X_j}} \left(\beta X_i + \ln \sum_j e^{-\beta X_j} \right) \right)$$

$$H = \sum_i \frac{\beta X_i e^{-\beta X_i}}{\sum_j e^{-\beta X_j}} + \sum_i \frac{e^{-\beta X_i}}{\sum_j e^{-\beta X_j}} \ln \sum_j e^{-\beta X_j}$$

$$H = \overbrace{\sum_i \frac{\beta X_i e^{-\beta X_i}}{\sum_j e^{-\beta X_j}}}^{\text{Left Term}} + \overbrace{\frac{\ln \sum_j e^{-\beta X_j}}{\sum_j e^{-\beta X_j}} \sum_i e^{-\beta X_i}}^{\text{Right Term}}. \tag{7}$$

This is the only part of the forward function. I'll add the term for perserving information later.

4. Backward Function - Entropy Gradient

The gradient of the entropy forward function is the partial derivative of equation 7 with respect to an individual element, X_k. First we will take the derivative of the left term, as defined in 7, then we will take the derivative of the right term, then we will combine them.

4.1 Left Term Derivative

$$\frac{\partial H}{\partial X_k} = \frac{\sum_j e^{-\beta X_j} \frac{\partial \sum_i \beta X_i e^{-\beta X_i}}{\partial X_k} - \sum_i \beta X_i e^{-\beta X_i} \frac{\partial \sum_j e^{-\beta X_j}}{\partial X_k}}{\left(\sum_j e^{-\beta X_j}\right)^2} + \dots$$

$$\frac{\partial H}{\partial X_k} = \frac{\sum_j e^{-\beta X_j} \beta e^{-\beta X_k} (1 - \beta X_k) + \beta^2 e^{-\beta X_k} \sum_i X_i e^{-\beta X_i}}{\left(\sum_j e^{-\beta X_j}\right)^2} + \dots$$

$$\frac{\partial H}{\partial X_k} = \frac{\beta e^{-\beta X_k} (1 - \beta X_k)}{\sum_j e^{-\beta X_j}} + \frac{\beta^2 e^{-\beta X_k} \sum_i X_i e^{-\beta X_i}}{\left(\sum_j e^{-\beta X_j}\right)^2} + \dots$$

$$\frac{\partial H}{\partial X_k} = \frac{\beta e^{-\beta X_k}}{\sum_j e^{-\beta X_j}} \left(1 - \beta X_k + \frac{\beta \sum_i X_i e^{-\beta X_i}}{\sum_j e^{-\beta X_j}}\right) + \dots$$

$$\frac{\partial H}{\partial X_k} = \beta q_k \left(1 - \beta X_k + \beta \sum_i X_i q_i\right) + \dots$$

$$\frac{\partial H}{\partial X_k} = \beta q_k - \beta^2 X_k q_k + \beta^2 q_k \sum_i X_i q_i + \dots$$

4.2 Right Term Derivative

$$\frac{\partial H}{\partial X_k} = \dots + \sum_i e^{-\beta x_i} \left(\frac{\sum_j e^{-\beta X_j} \left(\frac{-\beta e^{-\beta X_k}}{\sum_j e^{-\beta X_j}}\right) + \beta e^{-\beta X_k} \ln \sum_j e^{-\beta X_j}}{\left(\sum_j e^{-\beta X_j}\right)^2} \right) + \frac{\ln \sum_j e^{-\beta X_j}}{\sum_j e^{-\beta X_j}} \left(-\beta e^{-\beta X_k}\right)$$

$$\frac{\partial H}{\partial X_k} = \dots + \frac{-\beta e^{-\beta X_k}}{\sum_j e^{-\beta X_j}} + \frac{\beta e^{-\beta X_k} \ln \sum_j e^{-\beta X_j}}{\sum_j e^{-\beta X_j}} - \frac{\beta e^{-\beta X_k} \ln \sum_j e^{-\beta X_j}}{\sum_j e^{-\beta X_j}}$$

$$\frac{\partial H}{\partial X_k} = \dots - \beta q_k \tag{8}$$

4.3 Combine Left and Right Derivatives

$$\frac{\partial H}{\partial X_k} = \beta q_k - \beta^2 X_k q_k + \beta^2 q_k \sum_i X_i q_i - \beta q_k$$

$$\frac{\partial H}{\partial X_k} = \beta^2 X_k q_k + \beta^2 q_k \sum_i X_i q_i$$

$$\frac{\partial H}{\partial X_k} = \beta^2 q_k \left(X_k + \sum_i X_i q_i \right) \tag{9}$$

where q_i is given in equation 5.