Theano Computations

Jonathan Simon

January 25, 2016

MLP Activation Layer

The hidden Layer in Theano's MLP is computed as: $X \to f(XW + b) = A$, implying that the the activation vectors are rows, and the batch-wide unit activation are columns. For an activation layer with d hidden units and a minibatch of size t, the activation layer is given by the matrix:

$$A = \begin{bmatrix} a_{11} & \dots & a_{1d} \\ \vdots & & \vdots \\ a_{t1} & \dots & a_{td} \end{bmatrix} = \begin{bmatrix} | & & | \\ u_1 & \dots & u_d \\ | & & | \end{bmatrix} = \begin{bmatrix} \rule{1cm}{0.4pt} & o_1 & \rule{1cm}{0.4pt} \\ & \vdots & \\ \rule{1cm}{0.4pt} & o_t & \rule{1cm}{0.4pt} \end{bmatrix} \tag{1}$$

where:

$u_i = t$-dimensional timecourse of the i^{th} unit

$o_i = d$-dimensional state of the activation layer at the i^{th} observation

Covariance Computation

We can then compute the between-unit covariance using the formulas provided on the wikipedia pages:

1. `https://en.wikipedia.org/wiki/Sample_mean_and_covariance`
2. `https://en.wikipedia.org/wiki/Estimation_of_covariance_matrices`
3. `https://en.wikipedia.org/wiki/Covariance`

Consider the definition of covariance for a pair of random variables U_j and U_k representing the idealized scalar activations of the j^{th} and k^{th} hidden units respectively:

$$\operatorname{cov}(U_j, U_k) = \mathrm{E}\left[(U_j - \mathrm{E}[U_j])(U_k - \mathrm{E}[U_k])\right] \tag{2}$$

Substituting the **random variables** U_j and U_k for the **column vectors** u_j and u_k of scalar activation observations:

$$\operatorname{cov}(u_j, u_k) = \frac{1}{t-1}\Big(\big(u_j - \mathbb{1}\overline{u}_j\big)^\top \big(u_k - \mathbb{1}\overline{u}_k\big)\Big) \tag{3}$$

where the **scalar** $\overline{u}_*$ is defined as:

$$\overline{u}_* = \frac{1}{t}\sum_{i=1}^{t} \big[u_*\big]_i \quad \left(= \frac{1}{t}\sum_{i=1}^{t} \big[o_i\big]_* \right)$$

We can then extend this calculation to all pairs of units in the hidden layer to compute the covariance matrix:

$$\Sigma = \frac{1}{t-1}\Big(\big(A - \mathbb{1}\bar{o}\big)^\top \big(A - \mathbb{1}\bar{o}\big)\Big) \tag{4}$$

where the **row vector** $\bar{o}$ is defined as:

$$\bar{o} = \frac{1}{t}\sum_{i=1}^{t} o_i$$

Correlation Computation

We can also compute the between-unit pearson correlation using the formulas provided at:

1. `http://www.johndcook.com/blog/2008/11/05/how-to-calculate-pearson-correlation-accurate`
2. `https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient`

Consider the definition of correlation for a pair of random variables U_j and U_k representing the idealized scalar activations of the j^{th} and k^{th} hidden units respectively:

$$\rho_{jk} = \frac{\operatorname{cov}(U_j, U_k)}{\sigma_j \sigma_k} = \frac{\mathrm{E}\left[(U_j - \mathrm{E}[U_j])(U_k - \mathrm{E}[U_k])\right]}{\sqrt{\mathrm{E}\left[(U_j - \mathrm{E}[U_j])^2\right]}\sqrt{\mathrm{E}\left[(U_k - \mathrm{E}[U_k])^2\right]}} \tag{5}$$

Substituting the **random variables** U_j and U_k for the **column vectors** u_j and u_k of scalar activation observations:

$$r_{jk} = \frac{\big(u_j - \mathbb{1}\overline{u}_j\big)^\top \big(u_k - \mathbb{1}\overline{u}_k\big)}{\sqrt{\big(u_j - \mathbb{1}\overline{u}_j\big)^\top \big(u_j - \mathbb{1}\overline{u}_j\big)}\sqrt{\big(u_k - \mathbb{1}\overline{u}_k\big)^\top \big(u_k - \mathbb{1}\overline{u}_k\big)}} \tag{6}$$

where the **scalar** $\overline{u}_*$ is defined as:

$$\overline{u}_* = \frac{1}{t}\sum_{i=1}^{t}\left[u_*\right]_i \quad \left(= \frac{1}{t}\sum_{i=1}^{t}\left[o_i\right]_*\right)$$

Note that the normalizing factor of $\frac{1}{t-1}$ is absent in the above expression because it is canceled by the two occurances of $\sqrt{\frac{1}{t-1}}$ in the denominator.

Define the sample variance s_* as:

$$s_* = \sqrt{\frac{1}{t-1}\left(u_* - \mathbb{1}\overline{u}_*\right)^\top \left(u_* - \mathbb{1}\overline{u}_*\right)} \tag{7}$$

Then we can compute the **row vector** s as:

$$s = \begin{bmatrix} s_1 & \dots & s_d \end{bmatrix} = \sqrt{\frac{1}{t-1}\left(\mathbb{1}^\top \left(\left(A - \mathbb{1}\bar{o}\right) \odot \left(A - \mathbb{1}\bar{o}\right)\right)\right)} \tag{8}$$

Using the definition of the covariance matrix from the previous section, we can then write the correlation matrix as:

$$\rho = \frac{\Sigma}{\left(\mathbb{1}s\right) \odot \left(\mathbb{1}s\right)^\top} = \frac{\Sigma}{\left(\mathbb{1}s\right)^\top \odot \left(\mathbb{1}s\right)} \tag{9}$$

where the division is performed element-wise.

Additional optimizations which can be performed in the Theano code include:

1. Removing the $\frac{1}{t-1}$ scaling factors, since they are present in both numerator and denominator
2. Removing the square roots, since the final step involves element-wise squaring
3. Replacing the multiplications by $\mathbb{1}$ with broadcasting and axis-wise summation
4. Replacing the Hadamard products with element-wise squaring

Update: Implemented optimization 3 and 4, but optimizations 1 and 2 don't speed things up enough to make it worth the confusion.