

Backpropagation Cheatsheet

Batches are stored on matrices with one sample per row. For example, the input matrix \mathbf{X} is of size $N \times n_x$ for N examples each of dimension n_x . Each example \mathbf{x} is thus a row vector (dimension $1 \times n_x$). It is the same for intermediate activation. Thus, the weight matrix \mathbf{W}^h is of size $n_h \times n_x$. The bias vector \mathbf{b}^h is a row vector (dimension $1 \times n_h$).

\odot is the elementwise product ('Hadamard'). $\text{sum}_{\text{row}}(\mathbf{X})$ with \mathbf{X} of size $N \times n_x$ performs the sum for each row and outputs a column vector of size N . $\text{repmat}_{N \text{ row}}(\mathbf{b})$ repeat the row vector \mathbf{b} (dimension $1 \times p$) N times in row to output a matrix of size $N \times p$.

Forward

Elementwise

Vector

Vector per batch

$$\left\{ \begin{array}{l} \tilde{h}_i = \sum_{j=1}^{n_x} W_{i,j}^h x_j + b_i^h \\ h_i = \tanh(\tilde{h}_i) \\ \tilde{y}_i = \sum_{j=1}^{n_h} W_{i,j}^y h_j + b_i^y \\ \hat{y}_i = \text{SoftMax}(\tilde{y}_i) = \frac{e^{\tilde{y}_i}}{\sum_{j=1}^{n_y} e^{\tilde{y}_j}} \end{array} \right. \quad \left\{ \begin{array}{l} \tilde{\mathbf{h}} = \mathbf{x} \mathbf{W}^{h\top} + \mathbf{b}^h \\ \mathbf{h} = \tanh(\tilde{\mathbf{h}}) \\ \tilde{\mathbf{y}} = \mathbf{h} \mathbf{W}^{y\top} + \mathbf{b}^y \\ \hat{\mathbf{y}} = \text{SoftMax}(\tilde{\mathbf{y}}) \end{array} \right. \quad \left\{ \begin{array}{l} \tilde{\mathbf{H}} = \mathbf{X} \mathbf{W}^{h\top} + \text{repmat}_{N \text{ row}}(\mathbf{b}_h) \\ \mathbf{H} = \tanh(\tilde{\mathbf{H}}) \\ \tilde{\mathbf{Y}} = \mathbf{H} \mathbf{W}^{y\top} + \text{repmat}_{N \text{ row}}(\mathbf{b}_y) \\ \hat{\mathbf{Y}} = \text{SoftMax}_{\text{line}}(\tilde{\mathbf{Y}}) \end{array} \right.$$

Loss

$$\left\{ \begin{array}{l} \ell(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_{i=1}^{n_y} y_i \log \hat{y}_i = - \sum_{i=1}^{n_y} y_i \tilde{y}_i + \log \sum_{j=1}^{n_y} e^{\tilde{y}_j} \\ \mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}}) = - \frac{1}{N} \sum_{k=1}^N \sum_{i=1}^{n_y} Y_{k,i} \log \hat{Y}_{k,i} = - \text{mean}_{\text{col}}(\text{sum}_{\text{row}}(\mathbf{Y} \odot \log \hat{\mathbf{Y}})) \end{array} \right.$$

Backward

Elementwise

Vector

Vector per batch

$$\left\{ \begin{array}{l} \delta_i^y = \frac{\partial \ell}{\partial \tilde{y}_i} = \hat{y}_i - y_i \\ \frac{\partial \ell}{\partial W_{i,j}^y} = \delta_i^y h_j \\ \frac{\partial \ell}{\partial b_i^y} = \delta_i^y \\ \delta_i^h = \frac{\partial \ell}{\partial \tilde{h}_i} = (1 - h_i^2) \sum_{j=1}^{n_y} \delta_j^y W_{j,i}^y \\ \frac{\partial \ell}{\partial W_{i,j}^h} = \delta_i^h x_j \\ \frac{\partial \ell}{\partial b_i^h} = \delta_i^h \end{array} \right. \quad \left\{ \begin{array}{l} \nabla_{\tilde{\mathbf{y}}} = \hat{\mathbf{y}} - \mathbf{y} \\ \nabla_{\mathbf{W}^y} = \nabla_{\tilde{\mathbf{y}}}^\top \mathbf{h} \\ \nabla_{\mathbf{b}^y} = \nabla_{\tilde{\mathbf{y}}}^\top \\ \nabla_{\tilde{\mathbf{h}}} = (\nabla_{\tilde{\mathbf{y}}} \mathbf{W}^y) \odot (1 - \mathbf{h}^2) \\ \nabla_{\mathbf{W}^h} = \nabla_{\tilde{\mathbf{h}}}^\top \mathbf{x} \\ \nabla_{\mathbf{b}^h} = \nabla_{\tilde{\mathbf{h}}}^\top \end{array} \right. \quad \left\{ \begin{array}{l} \nabla_{\tilde{\mathbf{Y}}} = \hat{\mathbf{Y}} - \mathbf{Y} \\ \nabla_{\mathbf{W}^y} = \nabla_{\tilde{\mathbf{Y}}}^\top \mathbf{H} \\ \nabla_{\mathbf{b}^y} = \text{sum}_{\text{row}}(\nabla_{\tilde{\mathbf{Y}}})^\top \\ \nabla_{\tilde{\mathbf{H}}} = (\nabla_{\tilde{\mathbf{Y}}} \mathbf{W}^y) \odot (1 - \mathbf{H}^2) \\ \nabla_{\mathbf{W}^h} = \nabla_{\tilde{\mathbf{H}}}^\top \mathbf{X} \\ \nabla_{\mathbf{b}^h} = \text{sum}_{\text{row}}(\nabla_{\tilde{\mathbf{H}}})^\top \end{array} \right.$$