Backpropagation Cheatsheet

Batches are stored on matrices with one sample per raw. 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 raw vector (dimension $1 \times n_x$). It is the same for intermediate activation. Thus, the weight marix $\mathbf{W}^h$is of size $n_h \times n_x$. The bias vector $\mathbf{b}^h$ is a raw vector (dimension $1 \times n_h$).

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

Forward

Elementwise

$$\begin{cases} \tilde{h}_i = \sum_{j=1}^{n_x} W^h_{i,j}\ x_j + b^h_i \\ h_i = \tanh(\tilde{h}_i) \\ \tilde{y}_i = \sum_{j=1}^{n_h} W^y_{i,j}\ h_j + b^y_i \\ \hat{y}_i = \mathrm{SoftMax}(\tilde{y}_i) = \dfrac{e^{\tilde{y}_i}}{\sum_{j=1}^{n_y} e^{\tilde{y}_j}} \end{cases}$$

Vector

$$\begin{cases} \tilde{\mathbf{h}} = \mathbf{x}{\mathbf{W}^h}^\top + \mathbf{b}^h \\ \mathbf{h} = \tanh(\tilde{\mathbf{h}}) \\ \tilde{\mathbf{y}} = \mathbf{h}{\mathbf{W}^y}^\top + \mathbf{b}^y \\ \hat{\mathbf{y}} = \mathrm{SoftMax}(\tilde{\mathbf{y}}) \end{cases}$$

Vector per batch

$$\begin{cases} \tilde{\mathbf{H}} = \mathbf{X}{\mathbf{W}^h}^\top + \mathrm{repmat}_{N\ \mathsf{raw}}(\mathbf{b}_h) \\ \mathbf{H} = \tanh(\tilde{\mathbf{H}}) \\ \tilde{\mathbf{Y}} = \mathbf{H}{\mathbf{W}^y}^\top + \mathrm{repmat}_{N\ \mathsf{raw}}(\mathbf{b}_y) \\ \hat{\mathbf{Y}} = \mathrm{SoftMax}_{\mathsf{line}}(\tilde{\mathbf{Y}}) \end{cases}$$

Loss

$$\begin{cases} \ell(\mathbf{y}, \tilde{\mathbf{y}}) = -\sum_{i=1}^{n_y} y_i \log \hat{y}_i = -\sum_{i=1}^{n_y} y_i \tilde{y}_i + \log \sum_{j=1}^{n_y} e^{\tilde{y}_j} \\ \mathcal{L}(\mathbf{Y}, \hat{\mathbf{Y}}) = -\dfrac{1}{N} \sum_{k=1}^{N} \sum_{i=1}^{n_y} Y_{k,i} \log \hat{Y}_{k,i} = -\mathrm{mean}_{\mathsf{col}}(\mathrm{sum}_{\mathsf{raw}}(\mathbf{Y} \odot \log \hat{\mathbf{Y}})) \end{cases}$$

Backward

Elementwise

$$\begin{cases} \delta^y_i = \dfrac{\partial \ell}{\partial \tilde{y}_i} = \hat{y}_i - y_i \\ \dfrac{\partial \ell}{\partial W^y_{i,j}} = \delta^y_i\ h_j \\ \dfrac{\partial \ell}{\partial b^y_i} = \delta^y_i \\ \delta^h_i = \dfrac{\partial \ell}{\partial \tilde{h}_i} = (1 - h_i^2) \sum_{j=1}^{n_y} \delta^y_j W^y_{j,i} \\ \dfrac{\partial \ell}{\partial W^h_{i,j}} = \delta^h_i\ x_j \\ \dfrac{\partial \ell}{\partial b^h_i} = \delta^h_i \end{cases}$$

Vector

$$\begin{cases} \nabla_{\tilde{\mathbf{y}}} = \hat{\mathbf{y}} - \mathbf{y} \\ \nabla_{\mathbf{W}^y} = {\nabla_{\tilde{\mathbf{y}}}}^\top \mathbf{h} \\ \nabla_{\mathbf{b}^y} = \nabla_{\tilde{\mathbf{y}}}^\top \\ \nabla_{\tilde{\mathbf{h}}} = (\nabla_{\tilde{\mathbf{y}}} \mathbf{W}^y) \odot (1 - \mathbf{h}^2) \\ \nabla_{\mathbf{W}^h} = {\nabla_{\tilde{\mathbf{h}}}}^\top \mathbf{x} \\ \nabla_{\mathbf{b}^h} = \nabla_{\tilde{\mathbf{h}}}^\top \end{cases}$$

Vector per batch

$$\begin{cases} \nabla_{\tilde{\mathbf{Y}}} = \hat{\mathbf{Y}} - \mathbf{Y} \\ \nabla_{\mathbf{W}^y} = {\nabla_{\tilde{\mathbf{Y}}}}^\top \mathbf{H} \\ \nabla_{\mathbf{b}^y} = \mathrm{sum}_{\mathsf{raw}}(\nabla_{\tilde{\mathbf{Y}}})^\top \\ \nabla_{\tilde{\mathbf{H}}} = (\nabla_{\tilde{\mathbf{Y}}} \mathbf{W}^y) \odot (1 - \mathbf{H}^2) \\ \nabla_{\mathbf{W}^h} = {\nabla_{\tilde{\mathbf{H}}}}^\top \mathbf{X} \\ \nabla_{\mathbf{b}^h} = \mathrm{sum}_{\mathsf{raw}}(\nabla_{\tilde{\mathbf{H}}})^\top \end{cases}$$