Backpropagation Cheatsheet

Batches are stored as matrices with 1 example per line. $\odot$ is the element-wise product. $\text{repmat}_{N\ \text{lines}}(b^\top)$ broadcast a vector $b^\top$ of dimension $1 \times p$ N times to produce a matrix $N \times p$.

Forward

Elementwise

$$\begin{cases} \tilde{h}_i = \sum_j W_{h,ij} x_j + b_{h,i} \\ h_i = \tanh(\tilde{h}_i) \\ \tilde{y}_i = \sum_j W_{y,ij} h_j + b_{y,i} \\ \hat{y}_i = \text{SoftMax}(\tilde{y}_i) = \dfrac{e^{\tilde{y}_i}}{\sum_j e^{\tilde{y}_j}} \end{cases}$$

Vectoriel

$$\begin{cases} \tilde{h} = W_h x + b_h \\ h = \tanh(\tilde{h}) \\ \tilde{y} = W_y h + b_y \\ \hat{y} = \text{SoftMax}(\tilde{y}) \end{cases}$$

Vector per batch

$$\begin{cases} \tilde{H} = X W_h^\top + \text{repmat}_{N\ \text{lines}}(b_h^\top) \\ H = \tanh(\tilde{H}) \\ \tilde{Y} = H W_y^\top + \text{repmat}_{N\ \text{lines}}(b_y^\top) \\ \hat{Y} = \text{SoftMax}_{\text{line}}(\tilde{Y}) \end{cases}$$

Loss

$$\begin{cases} \ell(y, \tilde{y}) = -\sum_i y_i \log \hat{y}_i = -\sum_i y_i \tilde{y}_i + \log \sum_j e^{\tilde{y}_j} \\ \mathcal{L}(Y, \hat{Y}) = -\dfrac{1}{N} \sum_k \sum_i Y_{k,i} \log \hat{Y}_{k,i} = -\text{mean}_{\text{col}}(\text{sum}_{\text{line}}(Y \odot \log \hat{Y})) \end{cases}$$

Backward

Elementwise

$$\begin{cases} \delta_{y,i} = \dfrac{\partial \ell}{\partial \tilde{y}_i} = \hat{y}_i - y_i \\ \dfrac{\partial \ell}{\partial W_{y,ij}} = \delta_{y,i} h_j \\ \dfrac{\partial \ell}{\partial b_{y,i}} = \delta_{y,i} \\ \delta_{h,i} = \dfrac{\partial \ell}{\partial \tilde{h}_i} = (1 - h_i^2) \sum_j \delta_{y,j} W_{y,ji} \\ \dfrac{\partial \ell}{\partial W_{h,ij}} = \delta_{h,i} x_j \\ \dfrac{\partial \ell}{\partial b_{h,i}} = \delta_i^h \end{cases}$$

Vector

$$\begin{cases} \nabla_{\tilde{y}} = \hat{y} - y \\ \nabla_{W_y} = \nabla_{\tilde{y}} h^\top \\ \nabla_{b_y} = \nabla_{\tilde{y}} \\ \nabla_{\tilde{h}} = W_y^\top \nabla_{\tilde{y}} \odot (1 - h^2) \\ \nabla_{W_h} = \nabla_{\tilde{h}} x^\top \\ \nabla_{b_h} = \nabla_{\tilde{h}} \end{cases}$$

Vector per batch

$$\begin{cases} \nabla_{\tilde{Y}} = \hat{Y} - Y \\ \nabla_{W_y} = \nabla_{\tilde{Y}}^\top H \\ \nabla_{b_y} = \text{sum}_{\text{col}}(\nabla_{\tilde{Y}})^\top \\ \nabla_{\tilde{H}} = \nabla_{\tilde{Y}} W_y \odot (1 - H^2) \\ \nabla_{W_h} = \nabla_{\tilde{H}}^\top X \\ \nabla_{b_h} = \text{sum}_{\text{col}}(\nabla_{\tilde{H}})^\top \end{cases}$$