HW1 Part One: Backpropogation

[redacted]

1. Given a standard BatchNorm layer, please calculate the gradients of the output $y_i = BN_{\gamma,\beta}(x_i)$ with respect to the parameters of γ, β shown in Figure 3. (5 points)

$$\frac{\partial y_i}{\partial \gamma} = \hat{x}_i$$

$$\frac{\partial y_i}{\partial \beta} = 1$$

2. Given a softmax function, please calculate the gradients of the output of a softmax function with respect to its input. (5 points)

For a vector with length n:

$$S(x_i) = \mathrm{Softmax(x_i)} = \frac{\mathrm{exp(x_i)}}{\sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{exp(x_j)}}$$

$\therefore$

$$\frac{\partial S(x_i)}{\partial x_i} = \frac{\exp(x_i)(\Sigma_{j=1}^n \exp(x_j)) - \exp(x_i)^2}{(\Sigma_{j=1}^n \exp(x_j))^2} = S(x_i) - S(x_i)^2.$$

$$\frac{\partial S(x_i)}{\partial x_k} = \frac{-\exp(x_i)\exp(x_k)}{(\Sigma_{j=1}^n \exp(x_j))^2} = -S(x_i) \cdot S(x_k). \quad (i \neq k)$$

$\therefore$

$$\frac{\partial S(x_i)}{\partial x_k} = \begin{cases} S(x_i) - S(x_i)^2 & i = k \\ -S(x_i) \cdot S(x_k) & i \neq k \end{cases}$$

3. Finish the detailed feed-forward computations of a batch samples $(\mathbf{x}, \mathbf{y}_A, \mathbf{y}_B)$ during a training iteration, coming with final predictions $(\hat{\mathbf{y}}_A \; and \; \hat{\mathbf{y}}_B)$ of Task A and Task B.

Using the denotations given above, the computations are as follows:

$$\mathbf{z}_{1A} = \theta_{1A}\mathbf{x} + \mathbf{b}_{1A}$$

$$\mathbf{h}_{1A} = \sin \mathbf{z}_{1A}$$

$$\mathbf{z}_{DP} = M \circ \mathbf{h}_{1A}$$

$$\hat{\mathbf{y}}_A = \theta_{2A}\mathbf{z}_{DP} + \mathbf{b}_{2A}$$

$$\therefore \hat{\mathbf{y}}_A = \theta_{2A}(M \circ \sin(\theta_{1A}\mathbf{x} + \mathbf{b}_{1A})) + \mathbf{b}_{1B}$$

While the computations of $\hat{\mathbf{y}}_B$ are as follows:

$$\mathbf{z}_{1B} = \theta_{1B}\mathbf{x}$$

$$\mathbf{z}_{BN} = \mathbf{z}_{1B} - \mu + \mathbf{b}_{1B} = \mathbf{z}_{1B} - \frac{1}{m}\sum_{i=1}^{m} \mathbf{z}_{1B}^{i} + \mathbf{b}_{1B}$$

$$\mathbf{h}_{BN} = \mathrm{ReLU}(\mathbf{z}_{\mathrm{BN}})$$

$$\mathbf{z}_{2B} = \theta_{2B}(\mathbf{h}_{BN} \oplus \hat{\mathbf{y}}_A) + \mathbf{b}_{2B}$$

$$\hat{\mathbf{y}}_B = \mathrm{Softmax}(\mathbf{z}_{\mathrm{2B}})$$

4. Use the backpropagation algorithm we have learned in class and give the gradients of the overall loss in a mini-batch with respect to the parameters at each layer.

For mini-batch i, residual

$$\delta_{2B}^{i} = \frac{\partial \mathcal{L}}{\mathbf{z}_{2B}^{i}} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}_B}\frac{\partial \hat{\mathbf{y}}_B}{\partial \mathbf{z}_{2B}^{i}} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}_B^{i}}\frac{\partial \hat{\mathbf{y}}_B^{i}}{\partial \mathbf{z}_{2B}^{i}}$$

$\therefore$

$$\delta_{2B,j}^{i} = \sum_{k=1}^{b} \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}_{B,k}^{i}}\frac{\partial \hat{\mathbf{y}}_{B,k}^{i}}{\partial \mathbf{z}_{2B,j}^{i}}$$

$$= \frac{1}{m}\sum_{k\neq j} \frac{\mathbf{y}_{B,k}^{i}}{\hat{\mathbf{y}}_{B,k}^{i}} \cdot \hat{\mathbf{y}}_{B,k}^{i} \cdot \hat{\mathbf{y}}_{B,j}^{i} - \frac{1}{m} \cdot \frac{\mathbf{y}_{B,j}^{i}}{\hat{\mathbf{y}}_{B,j}^{i}}[\hat{\mathbf{y}}_{B,j}^{i} - (\hat{\mathbf{y}}_{B,j}^{i})^2]$$

$$= \frac{1}{m}\sum_{k=1}^{b} \mathbf{y}_{B,k}^{i}\hat{\mathbf{y}}_{B,j}^{i} - \mathbf{y}_{B,j}^{i} = \frac{1}{m}(\hat{\mathbf{y}}_{B,j}^{i} - \mathbf{y}_{B,j}^{i})$$

$\therefore$

$$\delta_{2B}^{i} = \frac{1}{m}(\hat{\mathbf{y}}_{B}^{i} - \mathbf{y}_{B}^{i})$$

For the task A route:

The residual for the $\mathrm{FC_{2A}}$ layer is

$$\begin{aligned}\delta_{2A}^{i} &= \delta_{2B}^{i}\frac{\partial \mathbf{z}_{2B}}{\partial \hat{\mathbf{y}}_{A}^{i}} + \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}_{A}^{i}} \\ &= \theta_{2B}^{T}\delta_{2B}^{i} + \frac{1}{m}(\hat{\mathbf{y}}_{A}^{i} - \mathbf{y}_{A}^{i})\end{aligned}$$

Residual for the DP layer

$$\delta_{DP}^{i} = \delta_{2A}^{i}\frac{\partial \hat{\mathbf{y}}_{A}^{i}}{\partial \mathbf{z}_{DP}^{i}} = \theta_{2A}^{T}\delta_{2A}^{i}$$

Residual for the $\mathrm{FC_{1A}}$ layer

$$\delta_{1A}^{i} = \delta_{DP}^{i}\frac{\partial \mathbf{z}_{DP}^{i}}{\partial \mathbf{h}_{1A}^{i}}\frac{\partial \mathbf{h}_{1A}^{i}}{\partial \mathbf{z}_{1A}^{i}} = (\delta_{DP}^{i} \circ M) \circ \cos(z_{1A}^{i})$$

Thus,

$$\frac{\partial \mathcal{L}}{\partial \theta_{2A}} = \sum_{i=1}^{m}\delta_{2A}^{i}\frac{\partial \hat{\mathbf{y}}_{A}^{i}}{\partial \theta_{2A}} = \sum_{i=1}^{m}\delta_{2A}^{i}(\mathbf{z}_{DP}^{i})^{T}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}_{2A}} = \sum_{i=1}^{m}\delta_{2A}^{i}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{1A}} = \sum_{i=1}^{m}\delta_{1A}^{i}\frac{\partial \mathbf{z}_{1A}^{i}}{\partial \theta_{1A}} = \sum_{i=1}^{m}\delta_{1A}^{i}(\mathbf{x}^{i})^{T}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}_{1A}} = \sum_{i=1}^{m}\delta_{1A}^{i}$$

For the task B route:

Residual for the BN layer

$$\delta_{BN}^{i} = \delta_{2B}^{i}\frac{\partial \mathbf{z}_{2B}^{i}}{\partial \mathbf{h}_{BN}^{i}}\frac{\partial \mathbf{h}_{BN}^{i}}{\partial \mathbf{z}_{BN}^{i}} = \theta_{2B}^{T}\delta_{2B}^{i} \circ \mathrm{sgn}(\mathbf{z}_{\mathrm{BN}}^{\mathrm{i}})$$

Residual for the $\mathrm{FC_{1B}}$ layer

$$\delta_{1B}^{i} = \sum_{l=1}^{m} \delta_{BN}^{l} \frac{\partial \mathbf{z}_{BN}^{l}}{\partial \mathbf{z}_{1B}^{i}} = -\frac{1}{m} \sum_{l \neq i} \delta_{BN}^{l} + (1 - \frac{1}{m}) \delta_{BN}^{i}$$

Thus,

$$\frac{\partial \mathcal{L}}{\partial \theta_{2B}} = \sum_{i=1}^{m} \delta_{2B}^{i} (\mathbf{h}_{2B}^{i} \oplus \hat{\mathbf{y}}_{B}^{i})^{T}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}_{2B}} = \sum_{i=1}^{m} \delta_{2B}^{i}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{b}_{1B}} = \sum_{i=1}^{m} \delta_{BN}^{i}$$

$$\frac{\partial \mathcal{L}}{\partial \theta_{1B}} = \sum_{i=1}^{m} \delta_{1B}^{i} (\mathbf{x}^{i})^{T}$$