深度学习

2020 年春季学期

作业 1

徐荣琛 2019214518 软件学院 2020 年 3 月 9 日

1.1. **Block One: Gradients of some basic layers (30 points)**

(i) Given a BatchNorm layer, please calculate the gradients of the output $y_i = \mathbf{BN}_{\gamma,\beta}(x_i)$ with respect to the parameters of γ, β shown in Figure 4. **(10 points)**

(ii) Given a dropout layer, please calculate the gradients of **the output of a dropout layer** with respect to **the input of a dropout layer**. **(10 points)**

(iii) Given a Softmax function, please calculate the gradients of **the output of a Softmax function** with respect to **the input of a Softmax function**. **(10 points)**

解. (i) $\frac{\partial y_i}{\partial \gamma} = \hat{x}_i$, $\frac{\partial y_i}{\partial \beta} = 1$

(ii) 设 dropout 层的输入为 $\mathcal{DI}$（n_{1a} 维），其中第 i 个值为 $\mathcal{DI}_i$，设 dropout 层的输出为 $\mathcal{DO}$（n_{1a} 维），其中第 j 个值为 $\mathcal{DO}_j$，则：

$$\frac{\partial \mathcal{DO}_j}{\partial \mathcal{DI}_i} = \begin{cases} 0 & i \neq j \\ 0 & i = j \wedge r_i < p \\ 1/(1-p) & i = j \wedge r_i \geq p \end{cases}$$

(iii) 设 Softmax 函数的输入为 $\mathcal{SI}$（n_{yb} 维），其中第 i 个值为 $\mathcal{SI}_i$，设 Softmax 函数的输出为 $\mathcal{SO}$（n_{yb} 维），其中第 j 个值为 $\mathcal{SO}_j$，则：

- 若 $i = j$：

$$\begin{aligned} \frac{\partial \mathcal{SO}_j}{\partial \mathcal{SI}_i} &= \frac{\partial \frac{e^{\mathcal{SI}_i}}{\sum_k e^{\mathcal{SI}_k}}}{\partial \mathcal{SI}_i} = \frac{e^{\mathcal{SI}_i} \cdot (\sum_k e^{\mathcal{SI}_k}) - e^{\mathcal{SI}_i} \cdot e^{\mathcal{SI}_i}}{(\sum_k e^{\mathcal{SI}_k})^2} \\ &= \mathcal{SO}_j \cdot (1 - \mathcal{SO}_j) \end{aligned}$$

- 若 $i \neq j$：

$$\frac{\partial \mathcal{SO}_j}{\partial \mathcal{SI}_i} = \frac{\partial \frac{e^{\mathcal{SI}_j}}{\sum_k e^{\mathcal{SI}_k}}}{\partial \mathcal{SI}_i} = \frac{-e^{\mathcal{SI}_j} \cdot e^{\mathcal{SI}_i}}{(\sum_k e^{\mathcal{SI}_k})^2} = -\mathcal{SO}_i \cdot \mathcal{SO}_j$$

1.2. **Block Two: Feed-forward and back-propagation of the multi-task network (30 points)**

(i) Finish the detailed **feed-forward computations** of a batch samples $(\boldsymbol{x}, y_a, y_b)$ during a training iteration, coming with final predictions $(\hat{y}_a, \hat{y}_b)$ of Task A, Task B. **(10 points)**

(ii) Use the back-propagation algorithm we have learned in class and give **the gradients of the overall loss function with respect to the parameters at each layer** corresponding to a batch of samples. **(20 points)**

解.