Intelligent Speech Distinguish Homework 02

Qiu Yihang

June.13-17, 2022

Backward Propagation in RNN.

Input Sequence: $\mathbf{x}$. Output Sequence: $\hat{\mathbf{r}}$. Label Sequence: $\mathbf{r}$ (with length T_r).

Number of categories: C. Activation function: $\sigma(z) = \frac{1}{1+e^{-z}}$.

Network Structure:

- Input Layer: $\mathbf{a}_t^{(\mathtt{in})} = \sigma\left(\mathbf{W}^{(\mathtt{in})}\mathbf{x}_t + \mathbf{b}^{(\mathtt{in})}\right)$.
- Hidden Layer (RNN): $\mathbf{h}_t = \sigma\left(\mathbf{U}\mathbf{a}_t^{(in)} + \mathbf{V}\mathbf{h}_{t-1} + \mathbf{b}_h\right)$, $\mathbf{o}_t = \sigma\left(\mathbf{W}\mathbf{h}_t + \mathbf{b}_o\right)$.
- Output Layer: $\mathbf{h}_t^{(\mathtt{out})} = \mathbf{W}^{(\mathtt{out})}\mathbf{o}_t + \mathbf{b}^{(\mathtt{out})}$, $\hat{\mathbf{r}}_t = \mathtt{softmax}\left(\mathbf{h}_t^{(\mathtt{out})}\right)$.
- Loss Function: (Cross Entropy Loss)
  $\mathcal{L} = \sum_{t=1}^{T_r} \mathcal{L}\mathbf{oss}(r_t, \hat{\mathbf{r}}_t) = -\sum_{t=1}^{T_r}\sum_{i=1}^{C} r_{t,i}\log\hat{\mathbf{r}}_{t,i} = -\sum_{t=1}^{T_r} r_t^T \log\hat{\mathbf{r}}_t$

Give the Backward Propagation of RNN.

Proof. Use $\delta\alpha$ to denote $\frac{\partial\mathcal{L}}{\partial\alpha}$. We know $\sigma'(z) = \sigma(z)\left(1-\sigma(z)\right)$.

Then we have

$$\delta\hat{\mathbf{r}}_{t,i} = \frac{\partial\mathcal{L}}{\partial\hat{\mathbf{r}}_{t,i}} = \frac{r_{t,i}}{\hat{\mathbf{r}}_{t,i}} \quad (1 \le i \le C) \quad \Longleftrightarrow \quad \delta\hat{\mathbf{r}}_t = r_t \oslash \hat{\mathbf{r}}_t \quad \text{(where } \oslash \text{ is element-wise division)}$$

$$\begin{aligned}
\delta\mathbf{h}_{t,i}^{(\mathtt{out})} &= \sum_{j=1}^{C}\delta\hat{\mathbf{r}}_{t,j}\cdot\mathtt{softmax}'_{\mathtt{i,j}}\left(\mathbf{h}_{t,i}^{(\mathtt{out})}\right) = -\sum_{j\ne i}\frac{r_{t,j}}{\hat{\mathbf{r}}_{t,j}}(-\hat{\mathbf{r}}_{t,i}\hat{\mathbf{r}}_{t,j}) - \frac{r_{t,i}}{\hat{\mathbf{r}}_{t,i}}\hat{\mathbf{r}}_{t,i}\left(1-\hat{\mathbf{r}}_{t,i}\right) \\
&= -r_{t,i} + \hat{\mathbf{r}}_{t,i}\sum_{j=1}^{C} r_{t,j} = \hat{\mathbf{r}}_{t,i} - r_{t,i} \quad \Longleftrightarrow \quad \delta\mathbf{h}_t^{(\mathtt{out})} = \hat{\mathbf{r}}_t - r_t \\
\delta\mathbf{o}_t &= \frac{\partial\mathbf{h}_t^{(\mathtt{out})}}{\partial\mathbf{o}_t}\delta\mathbf{h}_t^{(\mathtt{out})} = \mathbf{W}^{(\mathtt{out})}\delta\mathbf{h}_t^{(\mathtt{out})} \\
\delta\mathbf{b}^{(\mathtt{out})} &= \delta\mathbf{h}_t^{(\mathtt{out})} \\
\delta\mathbf{W}_{i,j}^{(\mathtt{out})} &= \mathbf{o}_{t,j}\cdot\delta\mathbf{h}_{t,i}^{(\mathtt{out})} \\
\delta\mathbf{y}_o &\triangleq \frac{\partial\mathcal{L}}{\partial\left(\mathbf{W}\mathbf{h}_t + \mathbf{b}_o\right)} \\
&= \sigma(\mathbf{W}\mathbf{h}_t + \mathbf{b}_o)\odot\left(\mathbf{1} - \sigma\left(\mathbf{W}\mathbf{h}_t + \mathbf{b}_o\right)\right)\odot\delta\mathbf{o}_t \quad \text{(where } \odot \text{ is element-wise product)}
\end{aligned}$$

$$
\begin{aligned}
\delta\mathbf{h}_t &= \delta\mathbf{y}_o \frac{\partial\left(\mathbf{W}\mathbf{h}_t + \mathbf{b}_o\right)}{\partial\mathbf{h}_t} = \mathbf{W}\delta\mathbf{y}_o \\
\delta\mathbf{b}_o &= \delta\mathbf{y}_o \frac{\partial\left(\mathbf{W}\mathbf{h}_t + \mathbf{b}_o\right)}{\partial\mathbf{b}_o} = \delta\mathbf{y}_o \\
\delta\mathbf{W}_{i,j} &= \delta\mathbf{y}_{o,i} \frac{\partial\left(\mathbf{W}\mathbf{h}_t + \mathbf{b}_o\right)_i}{\partial\mathbf{W}_{i,j}} = \mathbf{h}_{t,j} \cdot \delta\mathbf{y}_{o,i} \\
\delta\mathbf{y}_h &\triangleq \frac{\partial\mathcal{L}}{\partial\left(\mathbf{U}\mathbf{a}_t^{(\mathtt{in})} + \mathbf{V}\mathbf{h}_{t-1} + \mathbf{b}_h\right)} \\
&= \sigma\left(\mathbf{U}\mathbf{a}_t^{(\mathtt{in})} + \mathbf{V}\mathbf{h}_{t-1} + \mathbf{b}_h\right) \odot \left(\mathbf{1} - \sigma\left(\mathbf{U}\mathbf{a}_t^{(\mathtt{in})} + \mathbf{V}\mathbf{h}_{t-1} + \mathbf{b}_h\right)\right) \odot \delta\mathbf{h}_t \\
\delta\mathbf{a}_t^{(\mathtt{in})} &= \mathbf{U}\delta\mathbf{y}_h \\
\delta\mathbf{b}_h &= \delta\mathbf{y}_h \\
\delta\mathbf{U}_{i,j} &= \mathbf{a}_{t,j}^{(\mathtt{in})} \cdot \delta\mathbf{y}_{h,i} \\
\delta\mathbf{V}_{i,j} &= \mathbf{h}_{t-1,j} \cdot \delta\mathbf{y}_{h,i} \\
\delta\mathbf{y}_a &\triangleq \frac{\partial\mathcal{L}}{\partial\left(\mathbf{W}^{(\mathtt{in})}\mathbf{x}_t + \mathbf{b}^{(\mathtt{in})}\right)} = \sigma\left(\mathbf{W}^{(\mathtt{in})}\mathbf{x}_t + \mathbf{b}^{(\mathtt{in})}\right) \odot \left(\mathbf{1} - \sigma\left(\mathbf{W}^{(\mathtt{in})}\mathbf{x}_t + \mathbf{b}^{(\mathtt{in})}\right)\right) \odot \delta\mathbf{a}^{(\mathtt{in})} \\
\delta\mathbf{b}^{(\mathtt{in})} &= \delta\mathbf{y}_a \\
\delta\mathbf{W}_{i,j}^{(\mathtt{in})} &= \mathbf{x}_{t,j} \cdot \delta\mathbf{y}_{a,i}
\end{aligned}
$$

In conclusion,

Define $\delta\mathbf{h}_t^{(\mathtt{out})} = \hat{\mathbf{r}}_t - r_t$.

Use $\mathbf{v}_i$ to denote the i-th element of vector $\mathbf{v}$.

Use $\mathbf{A}_{i,j}$ to denote the element in the i-th row and j-th column of matrix $\mathbf{A}$.

Then the gradient of all parameters trainable during the training process are as follows.

$$
\begin{cases}
\delta\mathbf{b}^{(\mathtt{out})} = \delta\mathbf{h}_t^{(\mathtt{out})} & \text{(a vector)} \\
\delta\mathbf{W}_{i,j}^{(\mathtt{out})} = \mathbf{o}_{t,j} \cdot \left(\delta\mathbf{h}_t^{(\mathtt{out})}\right)_i & \text{(a number) for any element } \mathbf{W}_{i,j}^{(\mathtt{out})} \text{ in } \mathbf{W}^{(\mathtt{out})} \\
\delta\mathbf{b}_o = \sigma'(\mathbf{W}\mathbf{h}_t + \mathbf{b}_o) \odot \delta\mathbf{o}_t & \text{(a vector)} \\
\delta\mathbf{W}_{i,j} = \left(\mathbf{h}_t\right)_j \cdot \left(\delta\mathbf{b}_o\right)_i & \text{(a number) for any element } \mathbf{W}_{i,j} \text{ in } \mathbf{W} \\
\delta\mathbf{b}_h = \sigma'\left(\mathbf{U}\mathbf{a}_t^{(\mathtt{in})} + \mathbf{V}\mathbf{h}_{t-1} + \mathbf{b}_h\right) \odot \mathbf{W}\delta\mathbf{b}_o & \text{(a vector)} \\
\delta\mathbf{U}_{i,j} = \left(\mathbf{a}_t^{(\mathtt{in})}\right)_j \cdot \left(\delta\mathbf{b}_h\right)_i & \text{(a number) for any element } \mathbf{U}_{i,j} \text{ in } \mathbf{U} \\
\delta\mathbf{V}_{i,j} = \left(\mathbf{h}_{t-1}\right)_j \cdot \left(\delta\mathbf{b}_h\right)_i & \text{(a number) for any element } \mathbf{V}_{i,j} \text{ in } \mathbf{V} \\
\delta\mathbf{b}^{(\mathtt{in})} = \sigma'\left(\mathbf{W}^{(\mathtt{in})}\mathbf{x}_t + \mathbf{b}^{(\mathtt{in})}\right) \odot \mathbf{U}\delta\mathbf{b}_h & \text{(a vector)} \\
\delta\mathbf{W}_{i,j}^{(\mathtt{in})} = \left(\mathbf{x}_t\right)_j \cdot \left(\delta\mathbf{b}^{(\mathtt{in})}\right)_i & \text{(a number) for any element } \mathbf{W}_{i,j}^{(\mathtt{in})} \text{ in } \mathbf{W}^{(\mathtt{in})}
\end{cases}
$$

(where $\sigma'(\mathbf{z}) \triangleq \sigma(\mathbf{z}) \odot (\mathbf{1} - \sigma(\mathbf{z}))$)

The backward propagation is as follows.

(where η is the learning rate)

$$\begin{cases} \mathbf{b}^{(\texttt{out})} & \leftarrow \mathbf{b}^{(\texttt{out})} - \eta \cdot \delta \mathbf{b}^{(\texttt{out})} \\ \mathbf{W}_{i,j}^{(\texttt{out})} & \leftarrow \mathbf{W}_{i,j}^{(\texttt{out})} - \eta \cdot \delta \mathbf{W}_{i,j}^{(\texttt{out})} \\ \mathbf{b}_o & \leftarrow \mathbf{b}_o - \eta \cdot \delta \mathbf{b}_o \\ \mathbf{W}_{i,j} & \leftarrow \mathbf{W}_{i,j} - \eta \cdot \delta \mathbf{W}_{i,j} \\ \mathbf{b}_h & \leftarrow \mathbf{b}_h - \eta \cdot \delta \mathbf{b}_h \\ \mathbf{U}_{i,j} & \leftarrow \mathbf{U}_{i,j} - \eta \cdot \delta \mathbf{U}_{i,j} \\ \mathbf{V}_{i,j} & \leftarrow \mathbf{V}_{i,j} - \eta \cdot \delta \mathbf{V}_{i,j} \\ \mathbf{b}^{(\texttt{in})} & \leftarrow \mathbf{b}^{(\texttt{in})} - \eta \cdot \delta \mathbf{b}^{(\texttt{in})} \\ \mathbf{W}_{i,j}^{(\texttt{in})} & \leftarrow \mathbf{W}_{i,j}^{(\texttt{in})} - \eta \cdot \delta \mathbf{W}_{i,j}^{(\texttt{in})} \end{cases}$$

$\blacksquare$