Deep Learning Assignment 4

Ali Abbasi - 98105879

February 2, 2023

1

1.1

This is an instance of Gated Recurrent Unit (GRU) cell.

If z_t is zero, then last hidden state of the cell doesn't affect the next hidden value and $h_t = \tilde{h}_t$. And if z_t is one, then the hidden state doesn't change from the previous time and $h_t = h_{t-1}$. So z_t is the update gate, averaging between last and the new values of the hidden state.

r_t on the other hand, is the reset gate, controlling the amount of information that is used from the previous hidden state to calculate the new one (or the new candidate hidden state). If $r_t = 0$ then the new candidate hidden state is calculated solely from the current input and the previous hidden state is ignored.

1.2

Using the formula we had for calculating gradients in RNNs:

$$\frac{\partial L_j}{\partial W} = \sum_{k=1}^{j} \frac{\partial L_j}{\partial \hat{y}_j} \frac{\partial \hat{y}_j}{\partial h_j} \frac{\partial h_j}{\partial h_k} \frac{\partial h_k}{\partial W}$$

$$\text{where } \frac{\partial h_j}{\partial h_k} = \prod_{m=k+1}^{j} \frac{\partial h_m}{\partial h_{m-1}}$$

In GRU, the gradients back propagate through the hidden states very well:

$$\frac{\partial h_m}{\partial h_{m-1}} = z_m + (1 - z_m) \frac{\partial \tilde{h}_m}{\partial h_{m-1}}$$

So if $z_m \approx 1$ then $\frac{\partial h_m}{\partial h_{m-1}} \approx 1$ and if $z_m \approx 0$ then $\frac{\partial h_m}{\partial h_{m-1}} \approx \frac{\partial \tilde{h}_m}{\partial h_{m-1}} = W^{(hh)^T} \left[(1 - \tilde{h}_m^2) \odot r_t \right]$. And in both cases, a significant amount of the gradient is passed back to the previous hidden state. Thus, the gradient vanishing problem is less severe in GRU.