31/10/23

Deep Learning

References:
Chapters 9 & 10 in
Dive Into Deep Learning

- Recap: RNN
- Back Propagation Through Time (BPTT)
- Gated Recurrent Units (GRUs)

- RNN:
  - $z_m^{(t)} = \sigma\left(\underline{\alpha}_m^T \cdot \underline{x}^{(t)} + \underline{\gamma}_m^T \cdot \underline{z}^{(t-1)}\right)$ — (1)
  - $\hat{y}_k^{(t)} = g_k\left(\underline{\beta}_k^T \cdot \underline{z}^{(t)}\right)$ — (2)
  - $\theta = \{\underline{\alpha}_1, \underline{\alpha}_2, \ldots \underline{\alpha}_M, \underline{\beta}_1, \underline{\beta}_2 \ldots \underline{\beta}_K, \underline{\gamma}_1, \underline{\gamma}_2 \ldots \underline{\gamma}_M\}$

Aside:
$p(y^{(t)} | x^{(t)} \ldots x^{(1)})$
$\propto p(y^{(t)} | x^{(t)}, z^{(t-1)})$

  - Just as with ANNs, we use an iterative approach to update the model parameters

$$\theta^{(r)} = \theta^{(r-1)} - \eta \nabla_{\theta^{(r-1)}} R(\theta), \text{ where}$$

  - $\mathcal{D} = \{(\underline{x}^{(1)}, \underline{y}^{(1)}), (\underline{x}^{(2)}, \underline{y}^{(2)}), \ldots (\underline{x}^{(t)}, \underline{y}^{(t)}), \ldots (\underline{x}^{(T)}, \underline{y}^{(T)})\}$
  - $R(\theta) = \frac{1}{T} \sum_{t=1}^{T} d(y^{(t)}, \hat{y}^{(t)})$
  - $\frac{\partial R(\theta)}{\partial \gamma_{mp}} = \frac{1}{T} \sum_{t=1}^{T} \frac{\partial d(y^{(t)}, \hat{y}^{(t)})}{\partial \hat{y}^{(t)}} \cdot \frac{\partial \hat{y}^{(t)}}{\partial z_m^{(t)}} \cdot \frac{\partial z_m^{(t)}}{\partial \gamma_{mp}}$ — (3)
  - $\frac{\partial z_m^{(t)}}{\partial \gamma_{mp}} = \frac{\partial \sigma(\underline{\alpha}_m^T \cdot \underline{x}^{(t)} + \underline{\gamma}_m^T \cdot \underline{z}^{(t-1)})}{\partial \gamma_{mp}} + \frac{\partial \left[\sigma(\underline{\alpha}_m^T \underline{x}^{(t)} + \underline{\gamma}_m^T \cdot \underline{z}^{(t-1)})\right]}{\partial z_m^{(t-1)}} \cdot \frac{\partial z_m^{(t-1)}}{\partial \gamma_{mp}}$
  - $\frac{\partial z_m^{(t)}}{\partial \gamma_{mp}} = \frac{\partial \sigma(\underline{\alpha}_m^T \cdot \underline{x}^{(t)} + \underline{\gamma}_m^T \cdot \underline{z}^{(t-1)})}{\partial \gamma_{mp}} + \sum_{i=1}^{t-1} \frac{\partial z_m^{(i)}}{\partial \gamma_{mp}} \cdot \left[\prod_{j=i+1}^{t} \frac{\partial \sigma(\underline{\alpha}_m^T \underline{x}^{(j)} + \underline{\gamma}_m^T \cdot \underline{z}^{(j-1)})}{\partial z_m^{(j-1)}}\right]$ — (4)

$\frac{\partial f(x^{(t)}, y^{(t)})}{\partial t}$
0

- Plugging ④ into ③ gives us $\frac{\partial}{\partial \gamma_{mp}} R(\theta)$

- Takeaway: * Computing $\frac{\partial}{\partial \gamma_{mp}} R(\theta)$ is expensive especially when t (or T) is large.
  * We run into the issue of vanishing/exploding gradients.

- A workaround: Truncated BPTT. Only a fixed number of time-steps used in gradient computation.

- Gated Recurrent Units (GRUs)

$$z_m^{(t)} = s_m^{(t)} \cdot z_m^{(t-1)} + (1 - s_m^{(t)}) \tilde{z}_m^{(t)}$$

($s_m^{(t)}$: update gate; $\tilde{z}_m^{(t)}$: candidate z_m)

$$\tilde{z}_m^{(t)} = \sigma(\alpha_m^T \cdot \underline{x}^{(t)} + r_m^{(t)} \cdot \underline{\gamma}_m^T \cdot \underline{z}^{(t-1)})$$

($r_m^{(t)}$: reset gate)