UE21CS343BB2

Topics in Deep Learning

Dr. Shylaja S S
Director of Cloud Computing & Big Data (CCBD), Centre for Data Sciences & Applied Machine Learning (CDSAML)
Department of Computer Science and Engineering
shylaja.sharath@pes.edu

PES UNIVERSITY

Ack:Divya K,
Teaching Assistant

Overview of lecture

- Introduction
- Vanishing Gradients
- Exploding Gradients
- Mathematical Insight
- Solutions

Introduction

- In the previous lectures, we learnt about the sequence learning problem, how RNNs solve the sequence learning problem and how BPTT can help in updating the weights.

- While RNNs provide a solution to deal with sequence learning problems, RNNs face challenges like:
    - ➢ Vanishing Gradients
    - ➢ Exploding Gradients

- Let us look at the causes, consequences and solutions to these challenges.

Vanishing Gradients

- The vanishing gradient problem occurs when the gradients of the loss function with respect to the parameters of the network become very small during training.

- Vanishing gradient makes it difficult for the network to learn long-term dependencies in the data, due to gradients diminishing over time

- The vanishing gradients are a result of the repeated multiplication of small gradient values during backpropagation through time(BPTT).

Exploding Gradients

- The exploding gradient problem occurs when the gradients of the loss function with respect to the parameters of the network become very large during training.

- Exploding gradients can cause instability in the training process, leading to divergent behavior and loss of meaningful learning.

- The exploding gradients are a result of the repeated multiplication of large gradient values during backpropagation through time(BPTT).

Mathematical Insight

From the previous lecture on BPTT, we derived the equation:

$$\frac{\partial \mathscr{L}_t(\theta)}{\partial W} = \frac{\partial \mathscr{L}_t(\theta)}{\partial s_t} \sum_{k=1}^{t} \frac{\partial s_t}{\partial s_k} \frac{\partial^+ s_k}{\partial W}$$

Let us focus on the term $\frac{\partial s_t}{\partial s_k}$:

$$\frac{\partial s_t}{\partial s_k} = \frac{\partial s_t}{\partial s_{t-1}} \frac{\partial s_{t-1}}{\partial s_{t-2}} \cdots \frac{\partial s_{k+1}}{\partial s_k}$$
$$= \prod_{j=k}^{t-1} \frac{\partial s_{j+1}}{\partial s_j}$$

Let us focus on one such term in the product $\frac{\partial s_j}{\partial s_{j-1}}$:

$$a_j = W s_{j-1} + b + Ux_j$$
$$s_j = \sigma(a_j)$$

$$a_j = [a_{j1}, a_{j2}, a_{j3}, \ldots a_{jd},]$$
$$s_j = [\sigma(a_{j1}), \sigma(a_{j2}), \ldots \sigma(a_{jd})]$$

$$\frac{\partial s_j}{\partial s_{j-1}} = \frac{\partial s_j}{\partial a_j} \frac{\partial a_j}{\partial s_{j-1}}$$

Mathematical Insight

$$\frac{\partial s_j}{\partial a_j} = \begin{bmatrix} \frac{\partial s_{j1}}{\partial a_{j1}} & \frac{\partial s_{j2}}{\partial a_{j1}} & \frac{\partial s_{j3}}{\partial a_{j1}} & \dots \\ \frac{\partial s_{j1}}{\partial a_{j2}} & \frac{\partial s_{j2}}{\partial a_{j2}} & \ddots & \\ \vdots & \vdots & \vdots & \frac{\partial s_{jd}}{\partial a_{jd}} \end{bmatrix}$$

$$= \begin{bmatrix} \sigma'(a_{j1}) & 0 & 0 & 0 \\ 0 & \sigma'(a_{j2}) & 0 & 0 \\ 0 & 0 & \ddots & \\ 0 & 0 & \dots & \sigma'(a_{jd}) \end{bmatrix}$$

$$= diag(\sigma'(a_j))$$

$$\frac{\partial a_j}{\partial s_{j-1}} = W$$

$$\frac{\partial s_j}{\partial s_{j-1}} = \frac{\partial s_j}{\partial a_j}\frac{\partial a_j}{\partial s_{j-1}}$$

$$= diag(\sigma'(a_j))W$$

We are interested in the magnitude of $\frac{\partial s_j}{\partial s_{j-1}}$, if it is small (or large) $\frac{\partial s_t}{\partial s_k}$ and hence $\frac{\partial \mathcal{L}_t}{\partial W}$ will vanish (or explode).

Mathematical Insight

$$\left\| \frac{\partial s_j}{\partial s_{j-1}} \right\| = \left\| diag(\sigma'(a_j))W \right\|$$
$$\leq \left\| diag(\sigma'(a_j)) \right\| \|W\|$$

$\because \sigma(a_j)$ is a bounded function (sigmoid, tanh) $\sigma'(a_j)$ is bounded

$$\sigma'(a_j) \leq \frac{1}{4} = \gamma \text{ [if } \sigma \text{ is logistic]}$$
$$\leq 1 = \gamma \text{ [if } \sigma \text{ is tanh]}$$

$$\left\| \frac{\partial s_j}{\partial s_{j-1}} \right\| \leq \gamma \|W\|$$
$$\leq \gamma\lambda$$

W is also a bounded, let's call it λ

$$\left\| \frac{\partial s_t}{\partial s_k} \right\| = \left\| \prod_{j=k+1}^{t} \frac{\partial s_j}{\partial s_{j-1}} \right\|$$
$$\leq \prod_{j=k+1}^{t} \gamma\lambda$$
$$\leq (\gamma\lambda)^{t-k}$$

- If $\gamma\lambda < 1$ the gradient will vanish
- If $\gamma\lambda > 1$ the gradient could explode

Solutions

- One simple way of avoiding the vanishing and exploding gradients is to use truncated back-propagation where we restrict the product to τ terms (which is lesser than t-k)

- Gradient Clipping: Gradient clipping involves normalizing the gradients to ensure they do not surpass a predefined threshold which helps to mitigate exploding gradient problem. For e.g.: if we set the threshold as 0.7, then we keep the gradients in the -0.7 to +0.7 range. If the gradient value drops below -0.7, then we change it to -0.7, and similarly if it exceeds 0.7, then we change it to +0.7.

- Use of LSTMs and GRUs: Architectures such as LSTM and GRU, which use gating mechanisms to control the flow of information through the network help to mitigate the vanishing gradient problem.

Acknowledgements

https://nptel.ac.in/courses/106106184

https://www.deeplearning.ai/courses/deep-learning-specialization/

https://youtube.com/playlist?list=PLKnIA16_RmvYuZauWaPlRTC54KxSNLtNn&si=a2L-j8rAkG15EWKY

UE21CS343BB2
Topics in Deep Learning

Dr. Shylaja S S
Director of Cloud Computing & Big Data (CCBD), Centre for Data Sciences & Applied Machine Learning (CDSAML)
Department of Computer Science and Engineering
shylaja.sharath@pes.edu

Ack:Divya K,
Teaching Assistant