

Problems with RNN

Dr. Muhammad Safyan

Problem with RNN

- Suitable for sequential data
 - Text , Time series
- Not used too much
 - Suffer with 2 major problems
 - Problem of long term dependency
 - Unstable gradient.
- Start to forget with the time step.
 - Next word prediction.
 - Punjabi is spoken in Punjab. Lahore is beautiful city. But I could not enjoy because I don't understand Punjabi.
 - Vanishing Gradient Descent

Unstable Training

- **Stagnant Training:**
- Exploding Gradient problem
- Longer term having so much large number , dominate the short term and become finite.
- e.g. relu +ve term derivative.
- Learning rate is not proper
 - Gradient Clipping
 - Control learning rate
 - LSTM

- RNN unfold input times
- Its length depends upon values in time steps(100 time steps)
- In Back propagation your tried to minimize the loss.
- Its done with Gradient Descent formula
- W_i, w_h, w_o

$$\frac{\partial L}{\partial w_{in}} = \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial w_{in}} + \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial o_2} \frac{\partial o_2}{\partial w_{in}} + \frac{\partial L}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial o_3} \frac{\partial o_3}{\partial o_2} \frac{\partial o_2}{\partial o_1} \frac{\partial o_1}{\partial w_{in}}$$

Long Term Dependency Problem

- In long sequence, Gradient Descent of short term Dependency contribute more then long term dependency.

The diagram shows a sequence of partial derivatives representing the flow of gradients through a recurrent neural network over time. It is written on a dark background with white chalk-like text. The sequence starts with $\frac{\partial \mathcal{L}}{\partial \hat{y}}$ and $\frac{\partial \hat{y}}{\partial o_{100}}$ on the left. This is followed by $\frac{\partial o_{100}}{\partial a_99}$. A series of dots indicates intermediate steps. On the right, the sequence continues with $\frac{\partial o_2}{\partial o_1}$ and $\frac{\partial o_1}{\partial w_{in}}$. The diagram visually demonstrates that the gradient from a short-term dependency (like $\frac{\partial o_1}{\partial w_{in}}$) can be much stronger than the gradient from a long-term dependency (like $\frac{\partial \mathcal{L}}{\partial \hat{y}}$).

How to reduce this problem

- Proper Activation function
 - Relu, Leaky Relu
- Better weight Initiation
- Skip gram
- LSTM