19

Advanced Sequence Learners

First, second, third…

19.1
Training **LSTMs**.

Input and remembering long-term dependencies

- Typically, input to an LSTM is a tensor of shape: [batch_size, time_steps, features] where each timestep is a part of the sequence with its features
- Recall that LSTMs mitigate this by their gating mechanism, which maintains a more constant error that can be backpropagated through time and layers

Training Process

- **Backpropagation Through Time (BPTT):** LSTMs are trained using a variant of backpropagation called BPTT, where gradients are calculated and propagated back through time to update the weights
- **Gradient Clipping:** Often used to prevent exploding gradients in LSTMs by setting a threshold value

Optimization and Regularization

- Algorithms like Adam, RMSprop are typically used for better convergence in training LSTM networks
- Dropouts are applied to inputs and recurrent connections to prevent overfitting
- Choosing the right number of layers, units, and learning rate is crucial and often requires extensive experimentation

19.2
Bidirectional LSTMs.

What is it?

- A Bidirectional Long Short-Term Memory (Bi-LSTM) is an extension of the traditional Long Short-Term Memory (LSTM) network
- It improves an LSTM's understanding of context in sequence learning applications

Structure and Functioning

- A Bi-LSTM consists of two LSTM layers that process the data in opposite directions: one forward and one backward
- The forward LSTM layer processes the sequence from start to end, while the backward LSTM layer processes it from end to start

Structure and Functioning

- This dual structure allows the network to capture information from both past (backward) and future (forward) states of the sequence
- At any given point in the sequence, the Bi-LSTM has complete, contextual information about all points before and after it

Output and Insight

- The outputs of the forward and backward LSTMs are combined at each time step. This combination can be done in various ways, such as concatenation, summing, or averaging
- By processing the sequence in both directions, Bi-LSTMs provide a richer understanding of the context, which is particularly beneficial for tasks where the meaning of a sequence element can be significantly influenced by elements both before and after it

Training Similarities & Differences

- Like standard LSTMs, Bi-LSTMs are trained using Backpropagation Through Time
- The gradients from both directions are calculated separately and then combined to update the model parameters
- The training is computationally more intensive than standard LSTMs due to the doubled number of LSTM layers

19.3
Sequence to Sequence.

What is Sequence to Sequence (seq2seq)

- Sequence-to-Sequence (seq2seq) models are a category of neural network architectures designed to transform a given sequence of elements, such as words in a sentence, into another sequence
- These models are particularly effective in tasks where the input and output sequences can be of different lengths

Encoder-Decoder Architecture

- **Encoder**: Processes the input sequence and compresses the information into a context vector (also known as the state vector). This part of the model is typically an RNN or one of its variants like LSTM or GRU
- **Decoder**: Takes the context vector and generates the output sequence. It also usually employs RNNs, LSTMs, or GRUs

Context Vector

- This is a fixed-length representation of the entire input sequence and acts as the bridge between the encoder and decoder
- It's supposed to capture the essence of the input sequence

Working Mechanism

- The input sequence is fed into the encoder, which processes it one element at a time
- After processing the entire input sequence, the encoder's final state is used as the context vector
- The decoder generates the output sequence element by element, often using a special start-of-sequence token as the initial input

Training

- Seq2seq models are typically trained end-to-end on paired sequences (e.g., an English sentence and its French translation)
- The training objective is to maximize the likelihood of the correct output sequence given the input sequence

Applications

- Machine Translation: Translating text from one language to another
- Speech Recognition: Converting spoken language into text
- Text Summarization: Creating a short summary from a longer text document
- Chatbots: Generating conversational responses in dialogue systems

END.