**ADITYA AMIN**

**ASSIGN : NLP-05**

1. What are Sequence-to-sequence models?

Sequence-to-sequence (Seq2Seq) models, also known as encoder-decoder models, are a type of neural network architecture designed to process sequences of varying lengths and generate variable-length output sequences. They are commonly used in tasks such as machine translation, text summarization, speech recognition, and conversation generation.

1. What are the Problem with Vanilla RNNs?

Vanishing Gradient Problem: Vanilla RNNs are prone to the vanishing gradient problem, where the gradients during backpropagation diminish or explode exponentially over long sequences.

Short-Term Memory: Vanilla RNNs have limited memory and struggle to retain information from earlier time steps. They are more focused on recent inputs and tend to forget relevant information from distant time steps.

Lack of Parallelization: Vanilla RNNs process sequences sequentially, one time step at a time. This sequential nature makes it challenging to parallelize the computations across time steps, which results in slower training and inference times.

Difficulty with Variable-Length Sequences: Vanilla RNNs require fixed-length input sequences, which can be problematic when dealing with variable-length sequences commonly found in natural language processing tasks.

1. What is Gradient clipping?

Gradient clipping is a technique used during the training of neural networks, including recurrent neural networks (RNNs), to mitigate the problem of exploding gradients. It involves limiting the magnitude of the gradients during backpropagation to prevent them from becoming too large.

1. Explain Attention mechanism

The attention mechanism is a key component in many sequence-to-sequence models, such as neural machine translation and text summarization, that allows the model to focus on relevant parts of the input sequence when generating the output sequence. It enables the model to dynamically weigh the importance of different input elements during the decoding process.

1. Explain Conditional random fields (CRFs)

Conditional Random Fields (CRFs) are a type of probabilistic graphical model used for sequential or structured prediction tasks. They are widely used in natural language processing (NLP) and computer vision for tasks such as named entity recognition, part-of-speech tagging, semantic segmentation, and handwriting recognition.

1. Explain self-attention

Self-attention, also known as intra-attention or scaled dot-product attention, is a mechanism used in neural networks, particularly in transformer models, to capture dependencies between different elements within a sequence or set of inputs.

1. What is Bahdanau Attention?

Bahdanau Attention, also known as additive attention, is an attention mechanism introduced by Dzmitry Bahdanau et al. in the context of neural machine translation. It is a popular variant of attention used in sequence-to-sequence models.

1. What is a Language Model?

A language model is a statistical model that is designed to capture the probability distribution of sequences of words or tokens in a language. It is trained on large amounts of text data and learns the patterns, relationships, and probabilities of word sequences.

1. What is Multi-Head Attention?

Multi-head attention is an extension of the self-attention mechanism used in transformer models. It allows the model to attend to different information at different positions within a sequence simultaneously.

1. What is Bilingual Evaluation Understudy (BLEU)

Bilingual Evaluation Understudy (BLEU) is a metric used for evaluating the quality of machine-generated translations compared to reference translations. It is widely used in the field of machine translation to assess the performance of translation systems.