1. What are Sequence-to-sequence models?

Sequence-to-sequence models, often referred to as Seq2Seq models, are neural network architectures designed for processing sequences of data and generating sequences as output. These models consist of two main components: an encoder, which processes the input sequence and encodes it into a fixed-size context vector, and a decoder, which generates the output sequence from the context vector. Seq2Seq models are used in various natural language processing tasks, such as machine translation, text summarization, and chatbot responses.

1. What are the Problem with Vanilla RNNs?

* **Vanishing Gradients**: Vanilla RNNs suffer from the vanishing gradient problem, which makes it challenging for the network to capture long-range dependencies in sequences.
* **Exploding Gradients**: Conversely, they can also experience exploding gradients, leading to instability during training.
* **Short-term Memory**: Vanilla RNNs have difficulty retaining information over long sequences.
* **Lack of Parallelism**: They process sequences sequentially, limiting parallelism and efficiency.

1. What is Gradient clipping?

Gradient clipping is a technique used during training to prevent exploding gradients. It involves setting a threshold value, and if the gradient of a parameter exceeds this threshold, it is scaled down to limit its magnitude. Gradient clipping helps stabilize training and ensures that gradients do not become too large, which can lead to instability.

1. Explain Attention mechanism

Attention mechanisms are components in neural networks that allow models to focus on specific parts of input data when making predictions. In natural language processing, attention mechanisms enable models to weigh the importance of different words in the input when generating output sequences, improving the model's ability to capture relevant information and generate contextually accurate responses.

1. Explain Conditional random fields (CRFs)

Conditional Random Fields are a type of graphical model used in various sequence labeling tasks, such as part-of-speech tagging and named entity recognition. CRFs model the conditional probability distribution over label sequences given input observations. They take into account dependencies between neighboring labels, making them particularly useful for structured prediction tasks.

1. Explain self-attention

Self-attention is an attention mechanism where an element in a sequence focuses on other elements within the same sequence. It is used in models like the Transformer to capture relationships and dependencies within a sequence. Self-attention mechanisms compute weighted representations of all elements in the input sequence based on their relevance to a particular element.

1. What is Bahdanau Attention?

Bahdanau Attention, also known as additive attention, is an attention mechanism introduced in the context of machine translation. It allows the model to learn contextually relevant alignments between words in the source and target languages. Unlike simple dot-product attention, Bahdanau Attention uses learned alignment scores to capture complex relationships.

1. What is a Language Model?

A language model is a statistical model that assigns probabilities to sequences of words or tokens in a language. It estimates the likelihood of observing a particular sequence of words. Language models are used in natural language processing tasks such as text generation, speech recognition, and machine translation.

1. What is Multi-Head Attention?

Multi-Head Attention is an extension of the attention mechanism in the Transformer model. It involves using multiple sets of attention weights to compute multiple context vectors in parallel. This allows the model to capture different types of dependencies and relationships in the input data, making it more expressive and capable of handling diverse patterns.

1. What is Bilingual Evaluation Understudy (BLEU)

BLEU is a metric used to evaluate the quality of machine-generated text, such as translations, by comparing them to reference (human-generated) text. BLEU measures the precision of n-grams (word sequences) in the generated text compared to the reference. It is commonly used in machine translation evaluations to assess the quality of translations.