**Assignment - 5**

1. What are Sequence-to-sequence models?

Ans: Sequence-to-sequence (Seq2Seq) models are a type of neural network architecture designed to map sequences from one domain to sequences in another domain. They are commonly used in tasks such as machine translation, text summarization, and speech recognition, where the input and output are both variable-length sequences.

1. What are the Problem with Vanilla RNNs?

Ans: The problem with Vanilla RNNs is vanishing and exploding gradients. In Vanilla RNNs, during backpropagation through time (BPTT), gradients tend to either vanish (become very small) or explode (become very large) as they are propagated through many time steps. This makes it difficult for Vanilla RNNs to learn long-range dependencies in sequences.

1. What is Gradient clipping?

Ans: Gradient clipping is a technique used to prevent exploding gradients during training. It involves scaling down gradients if their norm exceeds a predefined threshold. By limiting the magnitude of gradients, gradient clipping helps stabilize the training process and prevents numerical overflow.

1. Explain Attention mechanism

Ans: Attention mechanism is a mechanism used in neural networks to selectively focus on specific parts of the input sequence when making predictions. Instead of processing the entire input sequence at once, the model dynamically attends to relevant parts of the input at each step of decoding. This allows the model to weigh the importance of different input elements and improve performance, especially in tasks involving long sequences.

1. Explain Conditional random fields (CRFs)

Ans: Conditional random fields (CRFs) are a type of probabilistic graphical model used for structured prediction tasks, such as sequence labeling and segmentation. CRFs model the conditional probability distribution of output sequences given input sequences, taking into account dependencies between neighboring output labels.

1. Explain self-attention

Ans: Self-attention is an attention mechanism where the attention weights are computed from the input sequence itself. Each input element attends to all other elements in the sequence, and the attention weights are determined based on the similarity between elements. Self-attention allows the model to capture long-range dependencies and learn contextual representations for each element in the sequence.

1. What is Bahdanau Attention?

Ans: Bahdanau Attention, also known as additive attention, is an attention mechanism introduced in the context of neural machine translation. It extends the basic attention mechanism by learning a set of parameters to compute attention weights based on the alignment between the decoder hidden state and the encoder hidden states. Bahdanau Attention allows the model to dynamically focus on different parts of the input sequence during decoding.

1. What is a Language Model?

Ans: A language model is a probabilistic model that assigns probabilities to sequences of words or characters in a language. It captures the statistical properties of natural language, such as word frequencies and dependencies between words, and can be used to generate text, perform language modeling tasks, and evaluate the fluency of generated text.

1. What is Multi-Head Attention?

Ans: Multi-Head Attention is an extension of the self-attention mechanism that allows the model to focus on different parts of the input sequence simultaneously. Instead of computing a single set of attention weights, Multi-Head Attention computes multiple sets of attention weights in parallel, each representing a different attention "head". This allows the model to capture different aspects of the input sequence and learn more complex representations.

1. What is Bilingual Evaluation Understudy (BLEU)

Ans: Bilingual Evaluation Understudy (BLEU) is a metric used to evaluate the quality of machine-translated text by comparing it to reference translations. BLEU measures the overlap between n-grams (contiguous sequences of n words) in the candidate translation and the reference translations, with higher scores indicating better translation quality. It is widely used in machine translation research and is one of the most commonly used metrics for evaluating translation systems.