**DL theory : Assingments-8**

1. Stateful RNNs maintain the hidden state from one input to the next, allowing them to "remember" previous inputs. This can be useful for tasks such as language modeling or speech recognition. However, stateful RNNs can also make it more difficult to parallelize training and may require more memory. Stateless RNNs do not maintain a hidden state, making them simpler and more easily parallelizable, but they do not have the ability to "remember" previous inputs.
2. Encoder-Decoder RNNs are used for automatic translation because they can handle variable-length input sequences and variable-length output sequences. The encoder processes the input sequence and compresses it into a fixed-length context vector, which is then used by the decoder to generate the output sequence. This architecture allows for more accurate translations than plain sequence-to-sequence RNNs.
3. Variable-length input sequences can be handled by using padding or by using an attention mechanism. Variable-length output sequences can be handled by using a special end-of-sequence token or by using a beam search algorithm.
4. Beam search is a technique used to generate a sequence of outputs by maintaining a "beam" of the k most likely next tokens at each step. It can be useful for tasks such as machine translation, where multiple possible translations may be correct. The TensorFlow library provides a tool to implement beam search.
5. An attention mechanism is a technique used to weight different parts of an input sequence when processing it. Attention mechanisms can help to improve the accuracy of models such as encoder-decoder RNNs and transformers by allowing the model to focus on the most relevant parts of the input.
6. The most important layer in the Transformer architecture is the multi-head self-attention layer. This layer allows the model to weigh different parts of the input and output sequences and to generate a context vector, which is used to generate the output sequence.
7. Sampled softmax is a technique used to speed up training of models with a large output vocabulary. It is done by sampling a subset of the output vocabulary and using that subset to estimate the full softmax. This technique can be useful when training models such as machine translation models, which have a very large output vocabulary.