**Assignment\_8**

1. What are the pros and cons of using a stateful RNN versus a stateless RNN?

**Ans: Setting an RNN to be stateful means that it can build a state across its training sequence and even maintain that state when doing predictions. The benefits of using stateful RNNs are smaller network sizes and/or lower training times**.

1. Why do people use Encoder–Decoder RNNs rather than plain sequence-to-sequence RNNs for automatic translation?

**Ans: seq-2-seq RNNs translate one word at a time. encoder-decoder RNNs read & translate a sentence at a time.**

1. How can you deal with variable-length input sequences? What about variable-length output sequences?

**Ans: 1. set sequence\_length parameter when calling static\_rnn() or dynamic\_rnn().**

**2. pad smaller input/output to make them same size as largest input/output**

1. What is beam search and why would you use it? What tool can you use to implement it?

**Ans: In computer science, beam search is a heuristic search algorithm that explores a graph by expanding the most promising node in a limited set. Beam search is an optimization of best-first search that reduces its memory requirements.**

1. What is an attention mechanism? How does it help?

**Ans: In neural networks, attention is a technique that mimics cognitive attention. The effect enhances some parts of the input data while diminishing other parts — the thought being that the network should devote more focus to that small but important part of the data**.

**This step captures the attention mechanism.**

1. **Compute the sum (or product) of the encoder's outputs and decoder states.**
2. **Pass the generated output through a fully-connected network.**
3. **Apply softmax activation to the output. ...**
4. **Create the context vector by computing the weighted sum of attention weights and encoder's outputs.**
5. What is the most important layer in the Transformer architecture? What is its purpose?

**Ans: The most important part here is the “Residual Connections” around the layers. This is very important in retaining the position related information which we are adding to the input representation/embedding across the network.**

**Multi-head attention plays a crucial role in the recent success of Transformer models, which leads to consistent performance improvements over conventional attention in various applications. The popular belief is that this effectiveness stems from the ability of jointly attending multiple positions.**

1. When would you need to use sampled softmax?

**Ans: Sampled softmax aims to approximate a full softmax during model training. Rather than computing the loss over all classes, only the positive class and a sample of m negative classes are considered. Each negative class is sampled with probability qi with replacement.**