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

ANS:

Statefulness of RNNs

The benefits of using stateful RNNs are smaller network sizes and/or lower training times. The disadvantage is that we are now responsible for training the network with a batch size that reflects the periodicity of the data, and resetting the state after each epoch.

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

ANS:

This two-step model, called an Encoderâ€“Decoder, works much better than trying to translate on the fly with a single sequence-to-sequence RNN (like the one represented on the top left), since the last words of a sentence can affect the first words of the translation, so you need to wait until you have heard the whole

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

ANS:

The most common way people deal with inputs of varying length is padding. You first define the desired sequence length, i.e. the input length you want your model to have. Then any sequences with a shorter length than this are padded either with zeros or with special characters so that they reach the desired length.

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

ANS:

Beam search is an algorithm used in many NLP and speech recognition models as a final decision making layer to choose the best output given target variables like maximum probability or next output character.

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

ANS:

Attention mechanisms are a type of neural network layer that can be added to deep learning models. They allow the model to focus on specific parts of input by assigning different weights to different parts of the input. This weighting is typically based on the relevance of each part of the input to the task at hand.

1. 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.

The Transformer architecture uses an encoder-decoder structure that does not rely on recurrence and convolutions to generate an output. The encoder maps an input sequence to a series of continuous representations.

The main architecture of the transformer consists of an encoder stack followed by a decoder stack built from multihead attention layers.

1. When would you need to use sampled softmax?

ANS:

Sampled softmax only makes sense if we sample(our V) less than vocabulary size. If your vocabulary(amount of labels) is small, there is no point using sampled\_softmax\_loss . Save this answer.