1. **What are Sequence-to-sequence models?**

Sequence to Sequence models is a special class of Recurrent Neural Network architectures that we typically use (but not restricted) to solve complex Language problems like Machine Translation, Question Answering, creating Chatbots, Text Summarization, etc.

1. **What are the Problems with Vanilla RNNs?**

* Training RNNs
* The vanishing or exploding gradient problem
* RNNs cannot be stacked up
* Slow and Complex training procedures
* Difficult to process longer sequences

1. **What is Gradient clipping?**

Gradient Clipping clips the size of the gradients to ensure optimization performs more reasonably near sharp areas of the loss surface. It can be performed in a number of ways. One option is to simply clip the parameter gradient element-wise before a parameter update. Another option is to clip the norm || g || of the gradient g before a parameter update:

if ||g||>v then g←gv||g||

where v is a norm threshold.

1. **Explain Attention mechanism.**

* Attention is proposed as a solution to the limitation of the Encoder-Decoder model which encodes the input sequence to one fixed length vector from which to decode the output at each time step.
* In attention when the model is trying to predict the next word it searches for a set of positions in a source sentence where the most relevant information is concentrated.

1. **Explain Conditional random fields (CRFs).**

Conditional random fields (CRFs) are a class of statistical modeling methods often applied in pattern recognition and machine learning and used for structured prediction. Whereas a classifier predicts a label for a single sample without considering "neighboring" samples, a CRF can take context into account.

1. **Explain self-attention.**

The attention mechanism allows output to focus attention on input while producing output while the self-attention model allows inputs to interact with each other (i.e calculate attention of all other inputs wrt one input.

1. **What is Bahdanau Attention?**

* Conventional encoder-decoder architectures for machine translation encoded every source sentence into a fixed-length vector, irrespective of its length, from which the decoder would then generate a translation. This made it difficult for the neural network to cope with long sentences, essentially resulting in a performance bottleneck.
* The Bahdanau attention was proposed to address the performance bottleneck of conventional encoder-decoder architectures, achieving significant improvements over the conventional approach.

1. **What is a Language Model?**

A language model in NLP is a probabilistic statistical model that determines the probability of a given sequence of words occurring in a sentence based on the previous words. It helps to predict which word is more likely to appear next in the sentence.

1. **What is Multi-Head Attention?**

* In the Transformer, the Attention module repeats its computations multiple times in parallel. Each of these is called an Attention Head.
* The Attention module splits its Query, Key, and Value parameters N-ways and passes each split independently through a separate Head.
* All of these similar Attention calculations are then combined together to produce a final Attention score.
* This is called Multi-head attention and gives the Transformer greater power to encode multiple relationships and nuances for each word.

1. **What is Bilingual Evaluation Understudy (BLEU).**

* The Bilingual Evaluation Understudy Score, or BLEU for short, is a metric for evaluating a generated sentence to a reference sentence.
* A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0.
* The score was developed for evaluating the predictions made by automatic machine translation systems. It is not perfect, but does offer 5 compelling benefits:
  + It is quick and inexpensive to calculate.
  + It is easy to understand.
  + It is language independent.
  + It correlates highly with human evaluation.
  + It has been widely adopted.