Question 1 — What were the challenges with the previous architectures for sequence-to-sequence problems like encoder-decoder using the RNN?

Question 2— Why do we attention or how does it solve the problems of long-distance interaction and?

Question 3 — What is the interpretation of key, query and value in the attention.

Question 4— Why do we need multi-head attention?

Question 5— Why do we need to scale the dot product between the query and key vectors before the SoftMax operation?

Question 6 — Does attention layer adds any non-linearity in the neural network? If yes how else how to add the non-linearity when incorporating the attention layer?

Question 7 — What is the use or advantage of a residual layer or skip connections?