Figures and Description

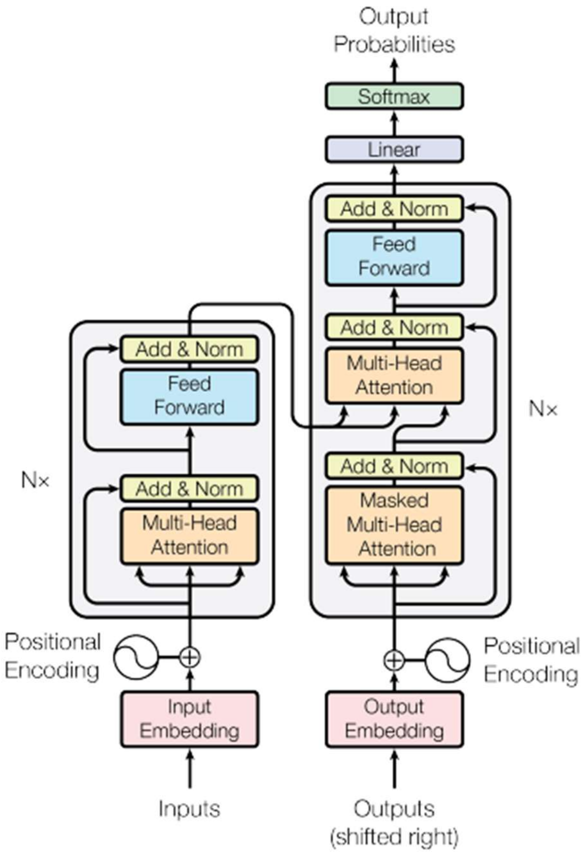

Figure 1, overall structure of Transformer

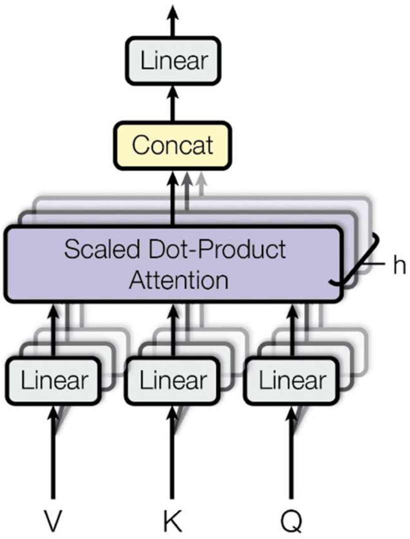

Figure 2, mechanism of multi-head attention. It means that cut the embedding vectors to several heads for efficient processing.

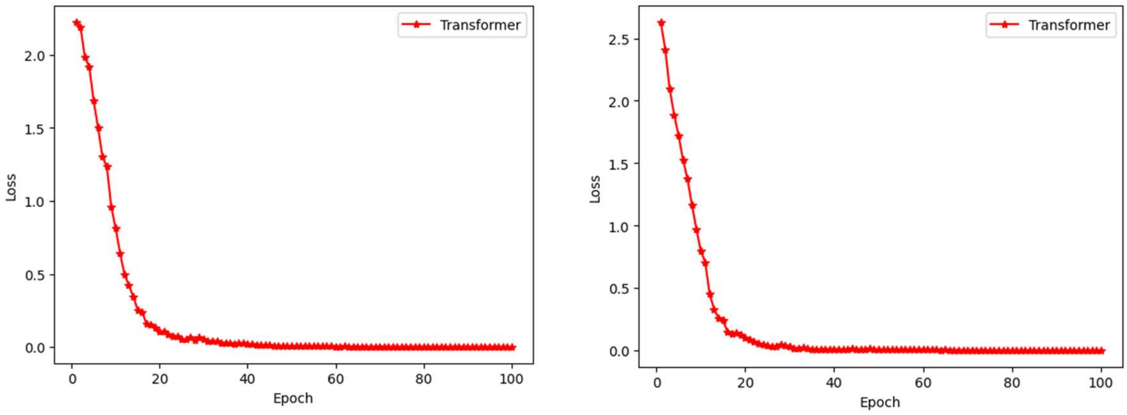

Figure 3, training loss of Transformer with different embedding size 512(left), 256(right). We can see that embedding size 256 is better than 512.