Neural Machine Transalation:

RNN structure:

A picture containing text, whiteboard

Description automatically generated

The encoder encodes hidden state at each step. the previous hidden state and the current word is used to predict the next hidden state. The decoder decodes the hidden state at each step. The current hidden state and the current word are used to predict the next translated word?

Problems with RNN:

Cannot learn long range dependencies.

Attention:

Focuses on specific part of the context.

Attention mechanism is performed with transformer archetechture:

Diagram

Description automatically generated

It inputs the whole sentense and the translated whole sentence, and outputs the probability over all input embeddings. Position encoding is also added as arguments, but are not as important as the input sentences.

Multi-head attention block accepts three arguments: values, keys and queries. Values and keys are outputted by encoding part of sentence; queries are outputted by the encoding part of the output sentence.

Where softmax over dot product of Q and K gives a distribution of similarity(with Q) over different values, and is used to select values.

Encoder of the original sentence