Summary of the Paper

“NEURAL MACHINE TRANSLATION”

Link:- 1409.0473

Problem:-

This paper addresses the problem of fixed-length vectors being a bottleneck in improving the performance of encoder-decoder architectures for neural machine translation. The paper proposes extending the encoder-decoder model to allow it to automatically search for relevant parts of the source sentence when predicting a target word, rather than compressing the entire source sentence into a fixed-length vector. This is particularly problematic for long sentences, where compressing all necessary information into a fixed-length vector can lead to performance deterioration.

Solution :-

RNNsearch encodes the source sentence into a sequence of vectors and allows the decoder to focus on different parts of the source sentence at each decoding step. This is achieved through an attention mechanism that computes a context vector as a weighted sum of the encoder's hidden states, where the weights are determined by the relevance of each source word to the current target word being generated.

- **Encoder**: A bidirectional RNN processes the source sentence, generating a sequence of hidden states that capture contextual information from both directions.
- **Attention Mechanism**: At each decoding step, the decoder computes attention scores for each encoder hidden state, which are then normalized using a SoftMax function to obtain attention weights. These weights are used to compute a context vector, representing the relevant information from the source sentence for generating the next target word.
- **Decoder**: The decoder uses the context vector, along with its previous hidden state and the previous target word, to generate the next word in the target sequence.

Results:-

The results show that the RNNsearch model outperforms the conventional RNN Encoder-Decoder model, especially for long sentences. The RNNsearch model is as good as the conventional phrase-based translation system (Moses) when only considering sentences with known words.