Towards an hybrid neural machine translation ?

Steve Edgar Feze feugang
University of yaoundé 1 – University of Bergen
www.fezu23@yahoo.fr, www.sfe012@uib.no

Machine translation as we know it today is based on artificial neural networks, namely recurrent neural networks and the Transformer. Inspired by the functioning of biological neurons, Franck Rasenblatt invented the perceptron or artificial neuron 1957. It's in fact an abstract mathematical model of the human brain neuron. Today, neural networks are used to build and train models to perform specific tasks. The neural network is an algorithm used in deep learning to predict output values. In machine translation, it's used to predict translations. To do this, it needs to be trained through a dataset consisting of the source elements (source sentences) and their prediction (target sentences or labels). The training of a neural network is an iterative game of propagation and back-propagation. Propagation allows the network to predict outputs which are then compared with the labels to find the error. This error is then propagated in the opposite direction of the network to modify the network parameters in oder to minimise the error. This cycle is repeated several times up to maximum error reduction.

1) From recurrent neural networks to LSTM and GRU

Neural machine translation is based on the encoder-decoder architecture. This machine translation technique was initially based on recurrent neural networks (RNN), which construct a contextual vector representation of the source sentence from the encoder and predict the target words from the decoder based on the previous ones for RNN and from the previous and the next one for Bi-RNNs (bidirectional RNN).

However, during back-propagation this architecture véry often leads to implosion of the network parameters, therefore blocking learning. Due to the risk of implosion during model training, Hochreiter and Schmildhuber (1997) proposed to add to the RNN a memory called LSTM (Long short term memory). Its task is to select the information that it considers relevant for the processing of a data item.

To alleviate the learning that had become too computationally expensive and slow in processing information, Cho et al. (2014) developed the GRU (Gated recurrent unit). The GRU is a simpler architecture that produces the same results as the LSTM. This is also the view of

Bardet (2021) who thinks that the GRU produces the same results in machine translation as the LSTM.

Although these selection and information retention mechanisms (LSTM and GRU) have solved the problem of implosion of learning during back propagation. It is clear that, from the information forgetting (LSTM), reset and update (GRU), the recurrent neural network loses information throughout the processing of a sequence. Therefore, the output of a RNN may not fully represent the sentence or sequence as originally intended. Furthermore, the RNN provides a sequential representation of the sentence, so the relationships between the words in a sentence are not always linear. To overcome this limitation, Badhanau et al (2015) propose to add another neural network to the RNN, which they will call the attention mechanism.

1) From attention mechanism to transformer

The attention mechanism is a small neural network located between the decoder and the encoder. Its mission, as Bardet (2021) points out, is to assign weights or degrees of relevance to the vectors coming from the encoder and representing the words of the source sentence. The objective of this vector weighting is to provide the decoder with the most relevant vector for the prediction of a target word. Henceforth, to generate a word in the target language, the decoder points to the previously generated target word or target vector and to the vector with more weight at the time (t) of the prediction of the current word. Machine translation researchers, notably Badet (2021), believe that the attention mechanism implicitly captures syntactic dependencies between source words.

While it is true that this mechanism has considerably improved the quality of translations, there is no explicit guarantee that it captures syntactic dependency relations and provides the decoder with the word that syntactically governs the target to be predicted and that transfers all its morphosyntactic properties to the target. Furthermore Wasvani et al (2017) found that the attention mechanism did not capture long distance dependencies, especially when sentences are long. To solve this problem, they will develop a neural network that relies primarily on attention. This is the Transformer.

The general principle of this new architecture is that all words in the source sentence must have attention on themselves (self-attention) and attention on all other words in the sentence (multi-attention). The implementation of this architecture mainly involves the generation of the keys (k) of the words of the source sentence, and the generation of the values (v) of these words. Each value corresponds to one key. This association makes it possible to

build a dictionary of keys-values. Then queries (Q) are generated. With the help of these K, V and Q, Transformer generate auto and multi-attention.

Although it must be acknowledged that the transform represents the state of the art of neural architectures for machine translation today, it must be noted that this architecture still encounters difficulties in handling dependencies within the sentence. This limitation of neural translation models is very often observed from the lack of cohesion and coherence between words in the translated text. Allowing these models to explicitly process morphosyntactic data in sentences could make them more efficient.

Bibliography

Badhanau et al (2015). Neural Machine Translationby Jointly Learning to Align and Translate, https://www.researchgate.net/publication/265252627_Neural_Machine_Translation_by_Jointly_Learning_to_Align_and_Translate, consulté le 25/ 07/ 2021

Bardet, A. (2021). *Architectures neuronales multilingues pour le traitement automatique des langues naturelles*, Informatiques et langage, Université de Maine, Français.NMT.

Cho, K. et al. (2014). « Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation *»*, *Computation and Language*, https://aclanthology.org/D14-1179.pdf , consulted on 24/05/2021

Hochreiter, S. and Schmildhuber, J. (1997). « Long short-term memory » in *Neural Computation*, https://www.bioinf.jku.at/publications/older/2604.pdf, consulted on 18/07/2021

Wasvani, A. et al. (2017). *Attention Is All You Need*, https://papers.nips.cc/paper/2017/file/-Paper.pdf, consulted le 13/19/2021.