Neural Machine Translation from Scratch

Bc. Josef Kotoun, Santosh Kesiraju Ph.D.
xkotou06@fit.vutbr.cz

Overview

- Implement Seq2Seq neural machine translation model based on transformers architecture
- Implement greedy and beam searching decoding algorithms
- Compare and evaluate the translation outputs using BLEU, chrF2 metrics
- Compare inference speed of decoding algorithms

Dataset

- European Parliament Proceedings Parallel Corpus [Koehn, 2005]
- A corpus of parallel text in 21 European languages from the proceedings of the European Parliament
- Used Czech – English pairs to train English to Czech translator
- Approximately 650 000 pairs of sentences

Transformer based translation

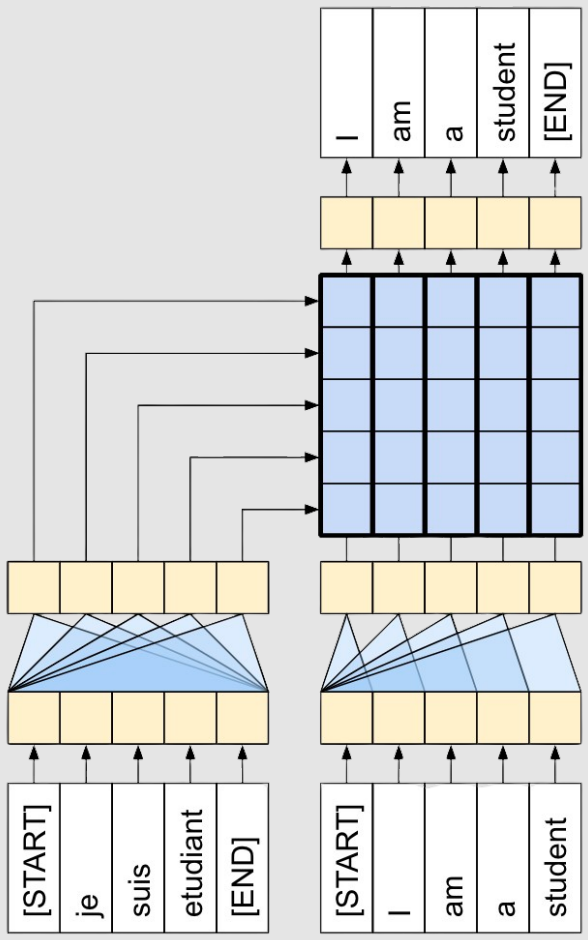

Figure: Transformer based translator training [Tensorflow authors, 2023]

Implementation

- WordPiece sub-word tokenization
- Transformer blocks used from Keras
- Trained on NVIDIA A40 48GB GPU provided by Cesnet Metacentrum

Results

- The translated sentences in Czech are in most cases properly syntactically structured
- Translations in most cases capture semantic meaning of original English sentence
- Due to used dataset, translations of formal language sentences are better than translations of informal simple sentences
- Beam search decoding gives better results than Greedy search, but Greedy search is faster

Example of Test Split Sentences Translations

English Sentence	Czech Translation
This is unacceptable: I shall vote against.	To je nepřijatelné: hlasuji proti.
I voted in favor of this report with conviction.	Hlasoval jsem ve prospěch této zprávy.
We have already heard about China.	O Číně jsme již slyšeli.
That is why, if we enter a new era of bilateral trade agreements, these should include guarantees and principles, some of which are mentioned in Mr. Martin's report.	To je důvod, proč v době, kdy zahájíme novou éru dvoustranných obchodních dohod, by měly být tyto záruky a zásady, z nichž některé jsou zmíněny ve zprávě pana Martina.

References

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. Phuket, Thailand.

Tensorflow authors (2023). Tensorflow documentation. online.