NEURAL MACHINE TRANSLATION

Presented by Celina Agostinho

MACHINE TRANSLATION MARKET

Compound annual growth rate of 7.1% from 2020 to 2024

Higher accuracy levels --> increased employment in diverse areas of business

Growing volumes of Big Data, demand for content localization, and need for cost-effective and high-speed translation drive growth of MT market

Source: https://www.marketwatch.com/press-release/machine-translation-market-2020-industry-size-share-future-challenges-revenue-demand-industry-growth-and-top-players-analysis-and-forecast-by-360-market-updates-2020-07-08?tesla=y

SEQ2SEQ LEARNING

Introduced by Google in 2014
(https://arxiv.org/pdf/1409.3215.pdf)

Use cases: MT, chatbots, text summarization, question answering, speech recognition...

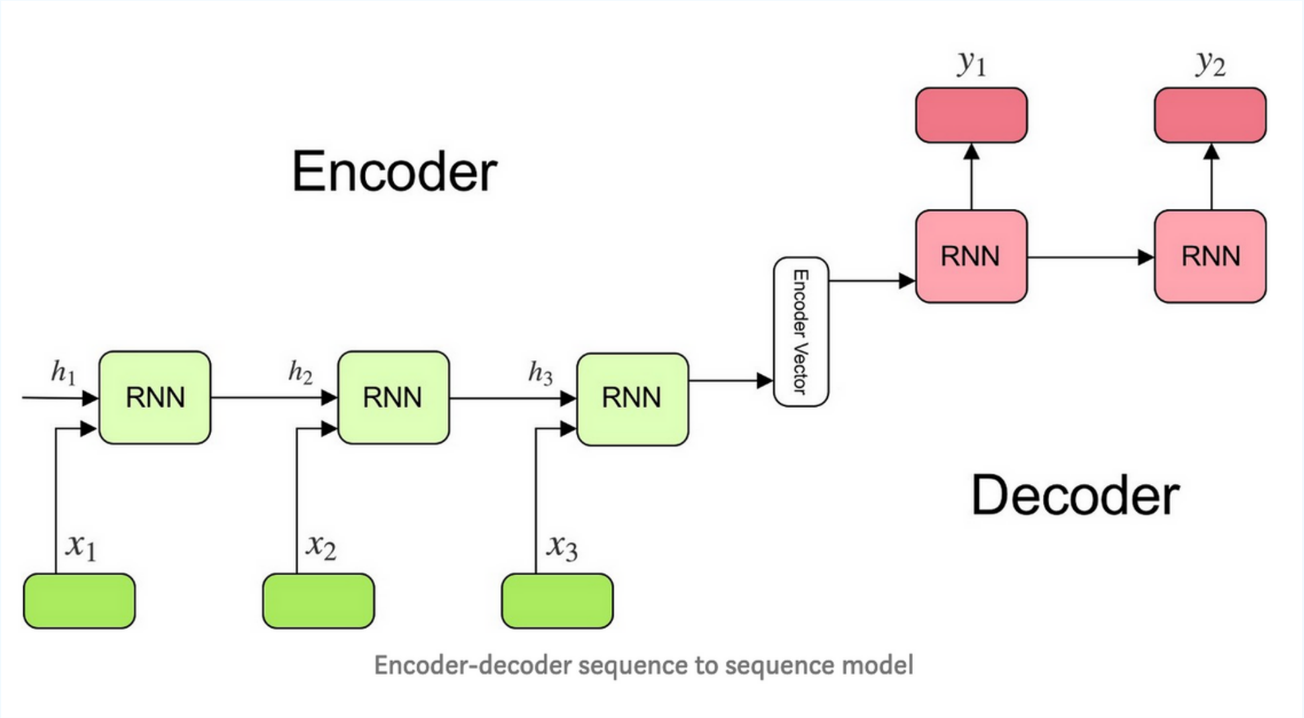
Map an input in one domain (e.g. sentences in Spanish) to an output in another domain (e.g. sentences in English) where the length of the input and output may differ.

Sources: https://blog.keras.io/a-ten-minute-introduction-to-sequence-to-sequence-learning-in-keras.html
& https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346

SEQ2SEQ LEARNING

ENCODER AND DECODER

Sequence-to-sequence learning has three parts: an encoder, an intermediate (encoder) vector and a decoder.



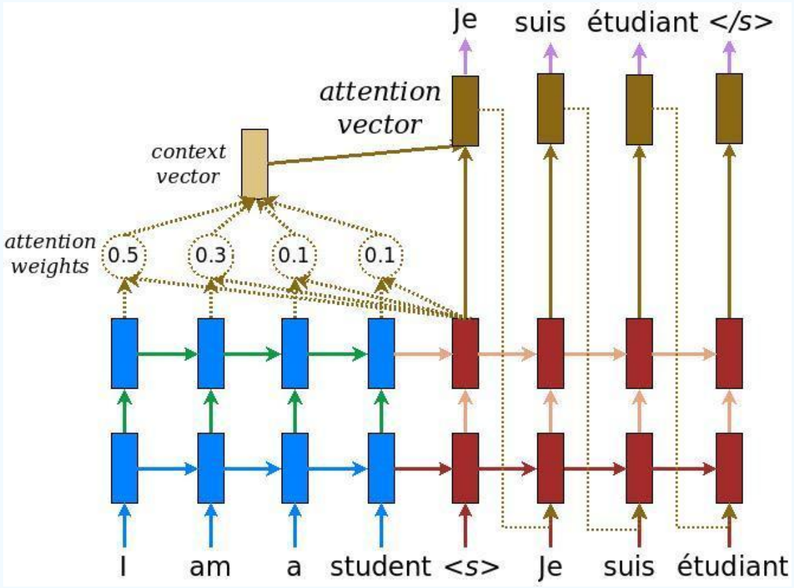
Encoder-decoder sequence to sequence model

Source: https://towardsdatascience.com/understanding-encoder-decoder-sequence-to-sequence-model-679e04af4346

SEQ2SEQ LEARNING

ATTENTION MECHANISM

Each input word is assigned a weight which is then used by the decoder to predict the next word in the sentence. The model learns what to attend to based on the input sentence and what it has produced so far.



Sources: http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp/ & https://www.tensorflow.org/tutorials/text/nmt_with_attention & https://machinetalk.org/2019/03/29/neural-machine-translation-with-attention-mechanism/

SEQ2SEQ LEARNING

ISSUES

- Names of people, places, dates, etc.
- Repetition of words
- Translation of long sentences
- Attention: computationally expensive

Sources: http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp/
& https://www.aclweb.org/anthology/I17-1003.pdf

DATASET

SPANISH - ENGLISH

Dataset: parallel text translation corpus provided by http://www.manythings.org/anki/

```
May I borrow this book?    ¿Puedo tomar prestado este libro?
```

A parallel text translation corpora is a structured set of translated texts between two languages.

Source: https://lionbridge.ai/datasets/25-best-parallel-text-datasets-for-machine-translation-training/ & https://www.tensorflow.org/tutorials/text/nmt_with_attention

RESULTS

```
Input: <start> vete a la puerta a recibir el paquete . <end>
Predicted translation: go to the door to get a parcel . <end>
```

```
Input: <start> no fuimos al cine porque ya era tarde <end>
Predicted translation: we didn t go to the movies because it was late . <end>
```

```
Input: <start> madrid es una ciudad que no me agrada . <end>
Predicted translation: excuse me a city where i don t like . <end>
```

```
Input: <start> las primeras dosis de la vacuna llegaran a espana a finales de ano <end>
Predicted translation: the first impressions came to end to the end of the year out of year . <end>
```

DIFFICULTIES

- Computing power
- Time of training
- Saving and loading trained model
- Getting the model to work on Flask app

NEXT STEPS

- Deploy in Heroku
- Set up Google Cloud VM to increase computing power
- Train on a larger dataset
- Add languages
- Train for specific purposes (legal documents, political speeches, scientific reports, etc.)
- Integrate sentiment analysis, NER and topic identification

QUESTIONS? COMMENTS?

GitHub: cfmago

cfm.agostinho@gmail.com