Sequence to sequence learning with neural networks

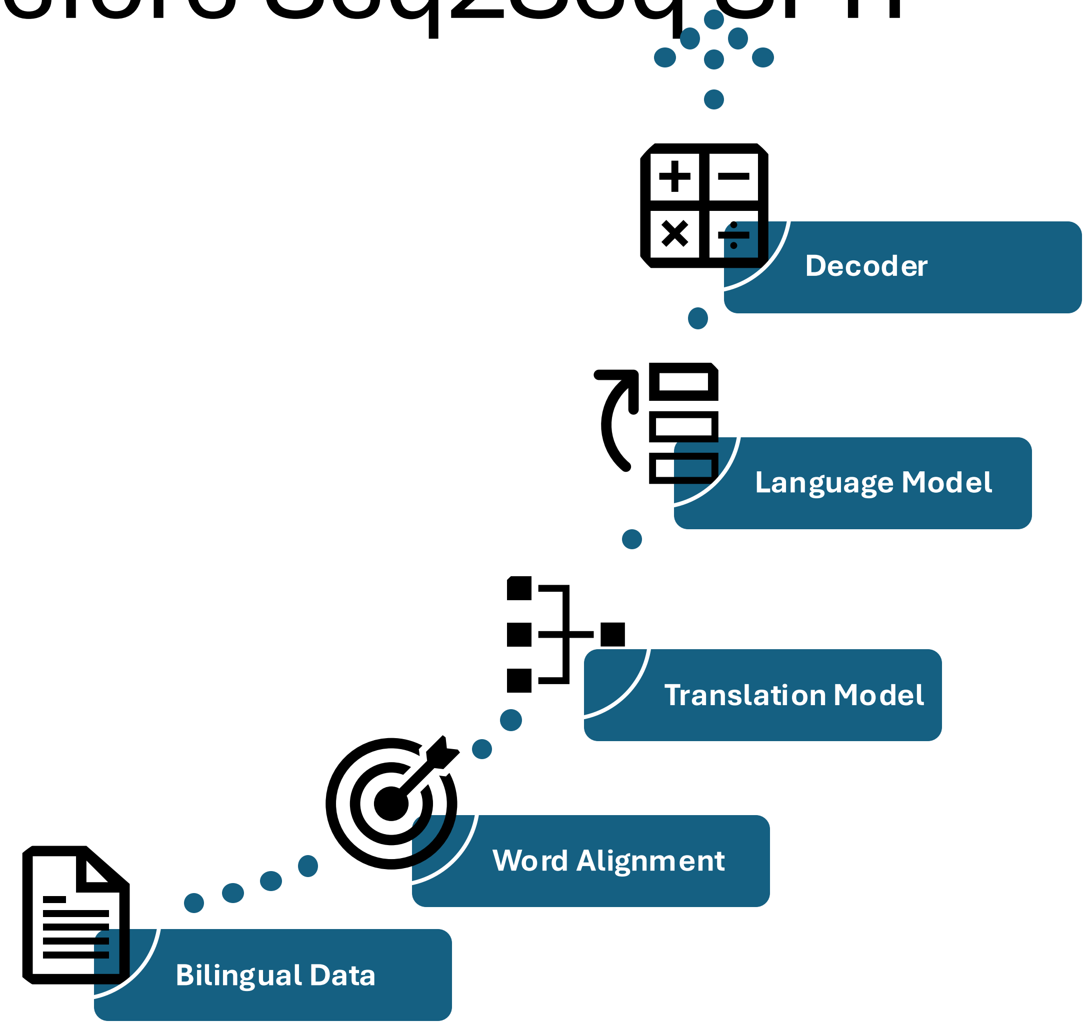
Ilya Sutskever, Oriol Vinyals, and Quoc V. Le., 2014

Presentation by
Lina Ben-Younes
Mattéo Debart
Nathan Janiec
Paul Massey
Samuel Sithakoul

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In Proceedings of Ilya Sutskever, Oriol Vinyals, and Quoc V. Le.

Table of contents

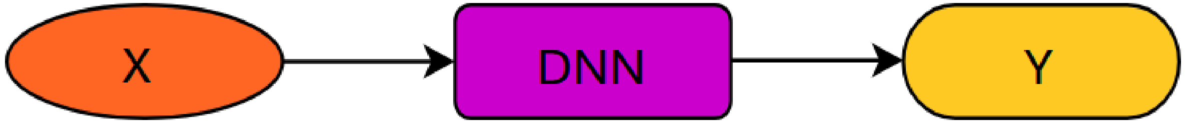
Context – before Seq2Seq SMT



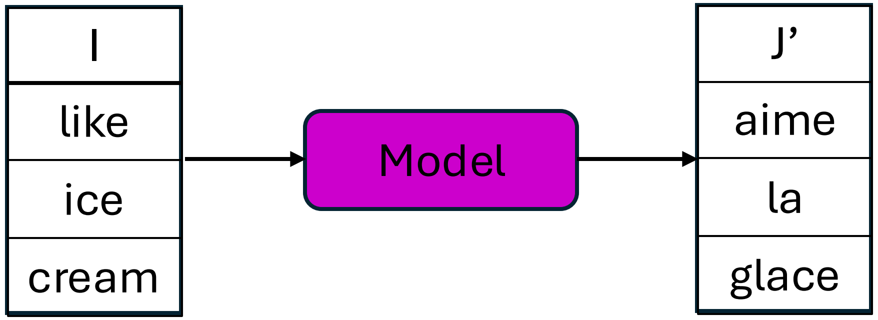
Context: Challenges

- Word Order
- Contextual Understanding
- Long-Distance Dependencies
- Manual feature engineering
- Handling variable-length inputs and outputs

Sequence-to-Sequence



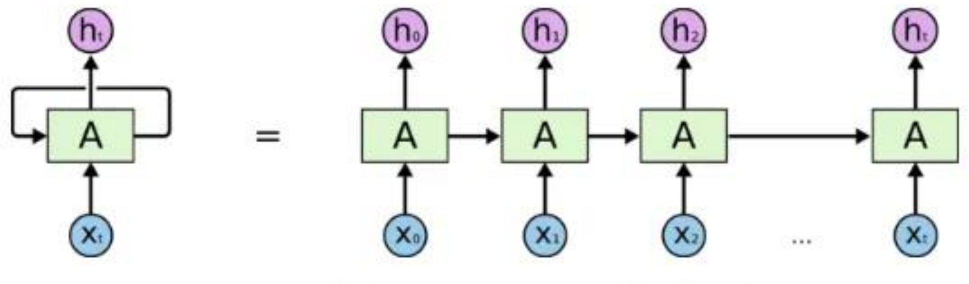
Traditional approach:
- X: Input = Image,Vector…
- Y: Output = Scalar (Regression), Probability (Classification)
- DNN: Deep Neural Network



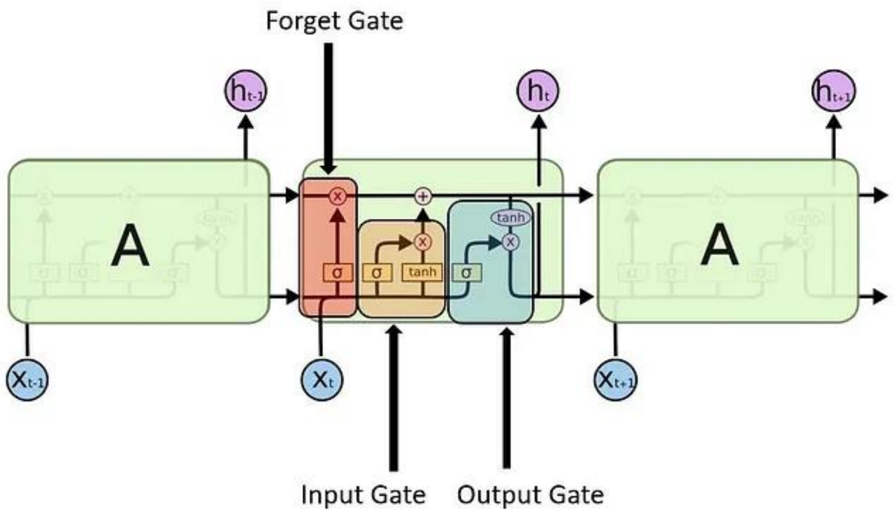
Machine Translation:
- Word order important but….
- Order not used with DNNs

How to use **sequential** information to train a model ?

From DNN to RNN/LSTM



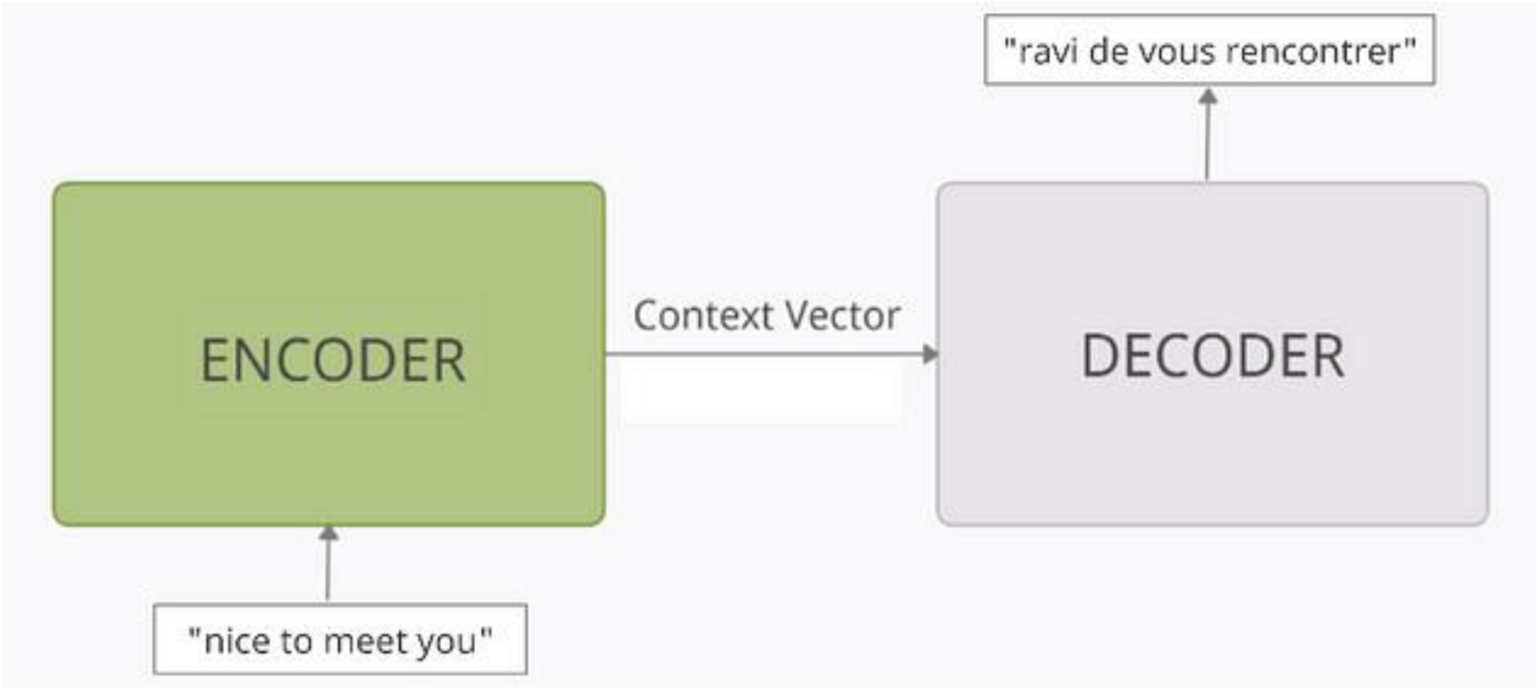
<u>RNN layer (Recurrent Neural Network):</u>
Use previous information and current input for prediction



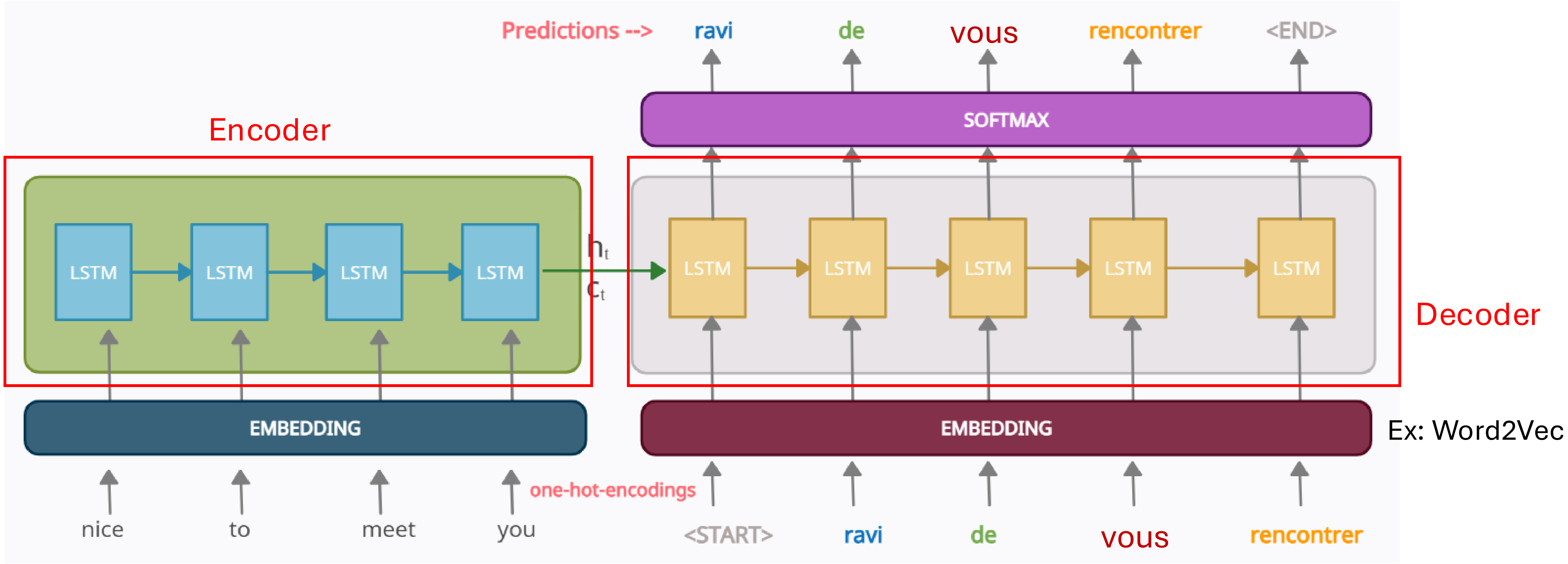
<u>LSTM layer (Long-Short Term Memory):</u>
Improvement to consider long term dependencies and avoid forgetting

Seq2Seq General idea

“Our method uses a multilayered Long Short-Term Memory(LSTM) to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector.”

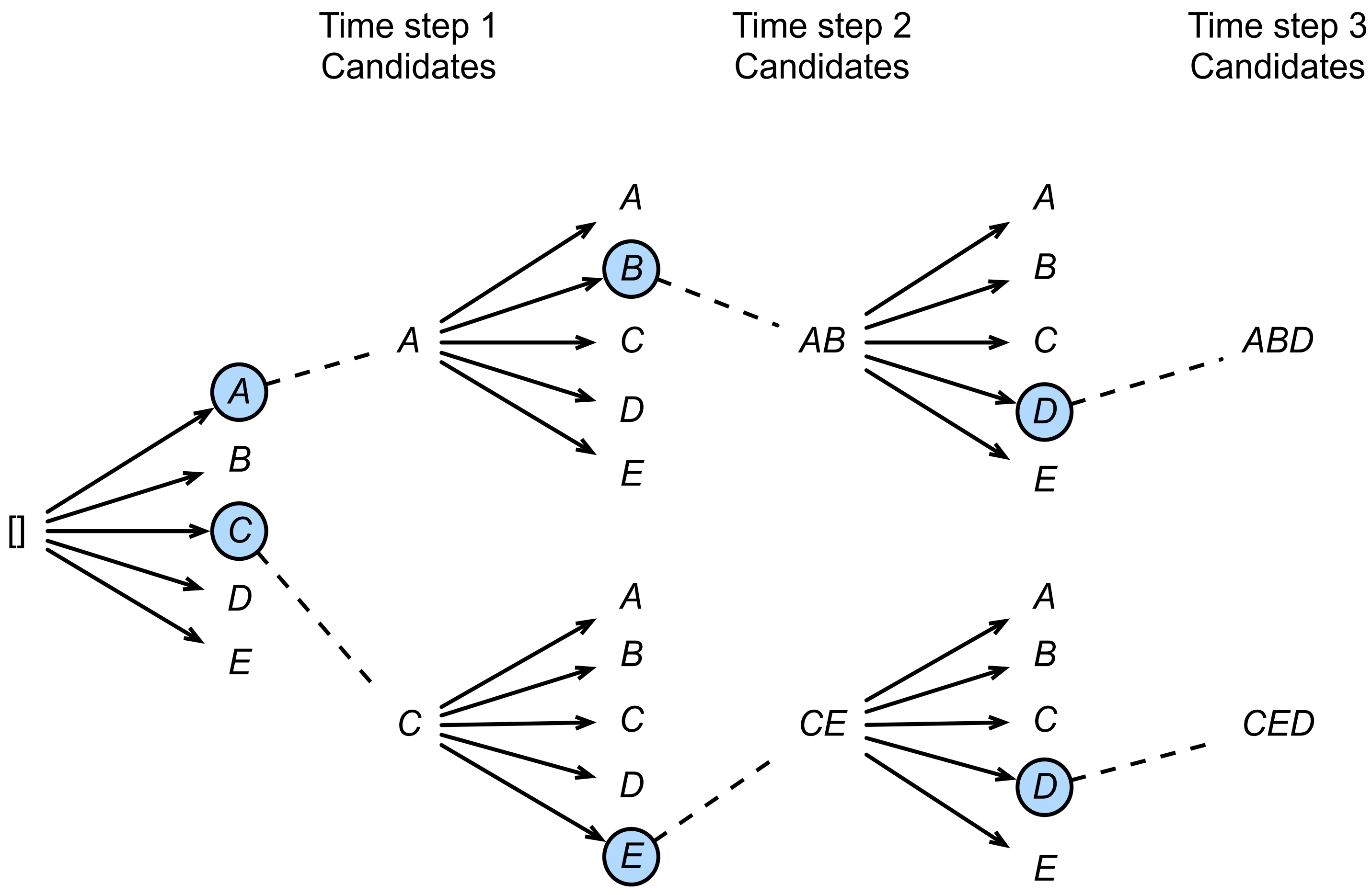


Seq2Seq architecture



Tries to predict next word with Context (Cell state c_t ,Hidden state h_t) and current word

Beam search for Inference



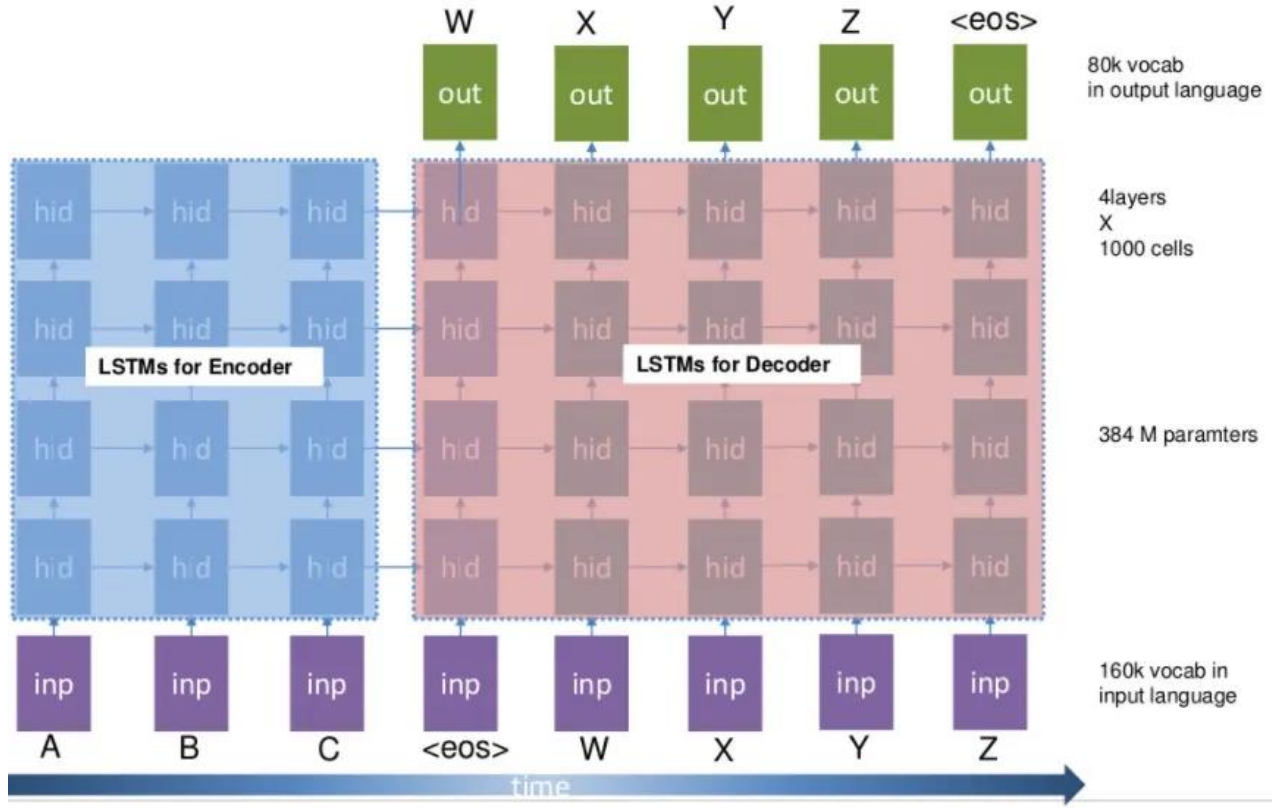
Other details

- Input sentences are reverted

  -> Better performances for short and long sentences

- Usage of Deep LSTMs

  -> Empirically: 10% gain in Perplexity for each layer

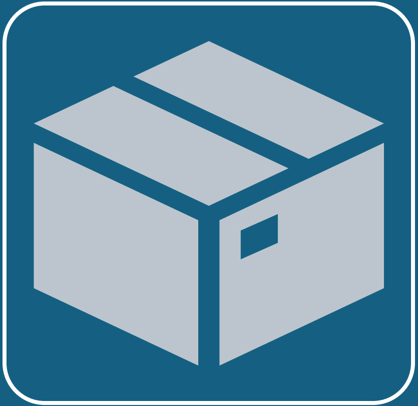


Training results

- Datasets of English to French translations with sentences of variable lengths
- Metric used: BLEU score (a kind of average of precision of n-grams)
- Closeness in semantic representation

Model	BLEU score
STM (Statistical Machine Translation)	33.3
LSTM from Seq2Seq	34.8
Seq2Seq with same Condition as STM	36.5
Best WMT (in 2014)	37.0
Best MT today (GPT4, DeepL...)	~50
Human translations	~30-70

Limitations



Compressing all the necessary details into a single vector.

- ⇒ **Attention Mechanism**: Attention allows the decoder to focus on specific parts of the input sequence during each decoding step, avoiding this bottleneck.

Stacked LSTMs, require sequential processing, meaning tokens are processed one by one. This results in slow training.

- ⇒ **Transformer Architecture**: positional encodings enabled parallelizable computations

Why it matters

- Founding approach for NLP and MTL
- Used in many LLM today with replacement of LSTM by Multi-Head attention for Transformer architecture: Bart, T5

Sources

- Chiusano, F. (2022, 20 septembre). A Brief Timeline of NLP - Generative AI - Medium. *Medium*. https://medium.com/nlplanet/a-brief-timeline-of-nlp-bc45b640f07d
- Schäferhoff, N. (2024, 9 juillet). The History of Google Translate (2004-Today) : A Detailed Analysis. TranslatePress. https://translatepress.com/history-of-google-translate/
- Hu, C. (2021, 14 décembre). What is the basic flow of statistical machine translation ? *Medium*. https://medium.com/mr-translator/what-is-the-basic-flow-of-statistical-machine-translation-44eaf069c50d