Learning about the Attention Mechanism and the Transformer model

Baptiste AMATO Alexis DUROCHER Gabriel HURTADO
Alexandre JOUANDIN Vincent MAROIS

Georgia Tech

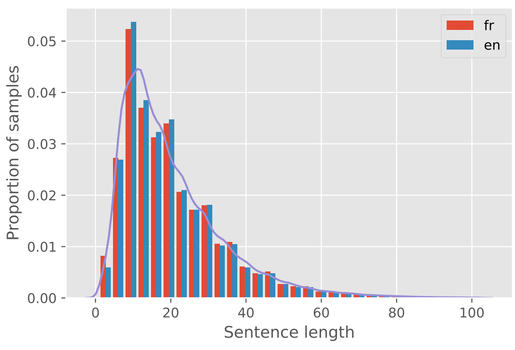
Motivation

- Transformer [VSP+17] model is a SoTA Machine Translation model,
- No recurrence, only uses the Attention Mechanism [BCB14],

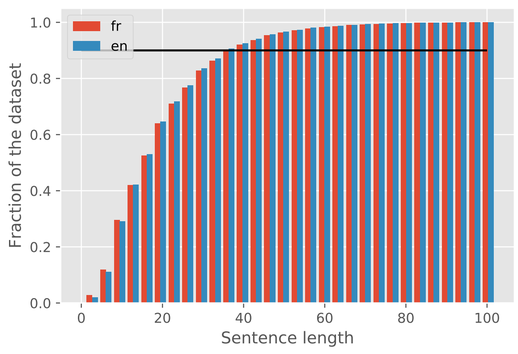
⇒ Can we reproduce the paper's results with our implementation and what can we learn about the model?

The Dataset

- IWSLT 2016 TED talk translation task (French → English),
- 220k train samples, 1025 validation, 1305 test,
- Avg. sentence length: 20 (train) – 21 (val) – 19 (test).



(a) Sentence Length Distribution.



(b) Cumulated Distribution.

The Transformer Model

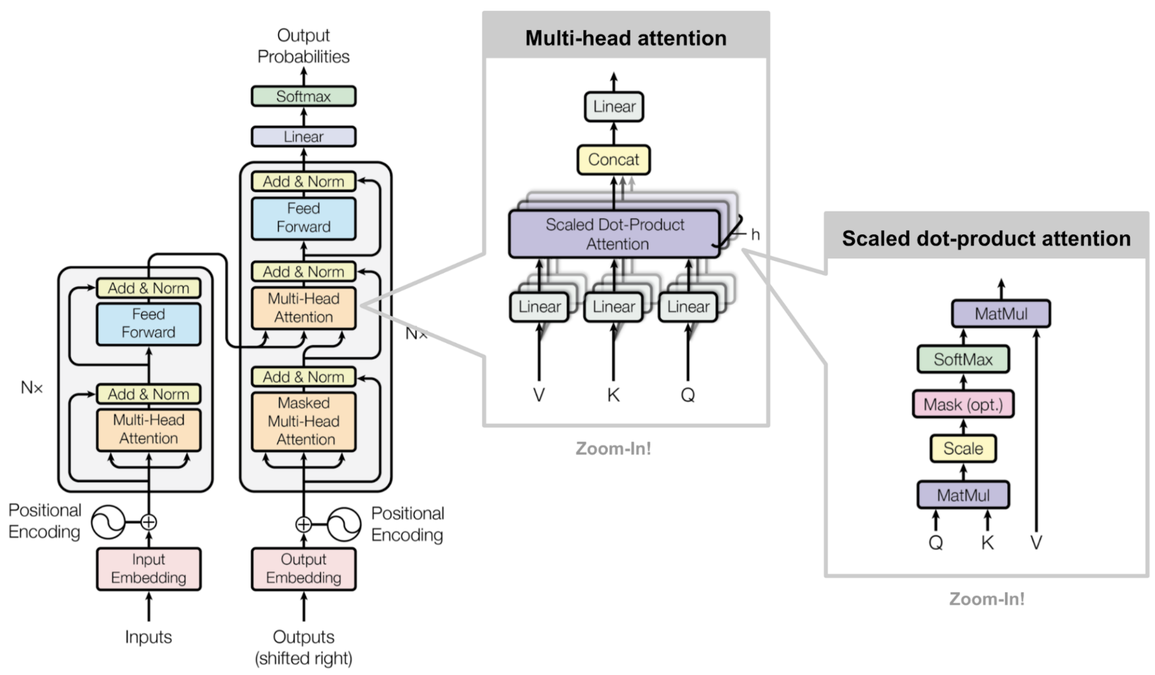


Figure 2: The Transformer model Architecture and the Attention Heads.

- Encoder-Decoder architecture,
- Less computation-heavy than RNNs for translation,
- Multi-Head Attention: Allows model to jointly attend to information from different representation subspaces.

References

[BCB14] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*, 2014.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Experiments

- Model converging on the IWSLT dataset,
- Early inference tests not satisfying: Further training & Beam Search should help.

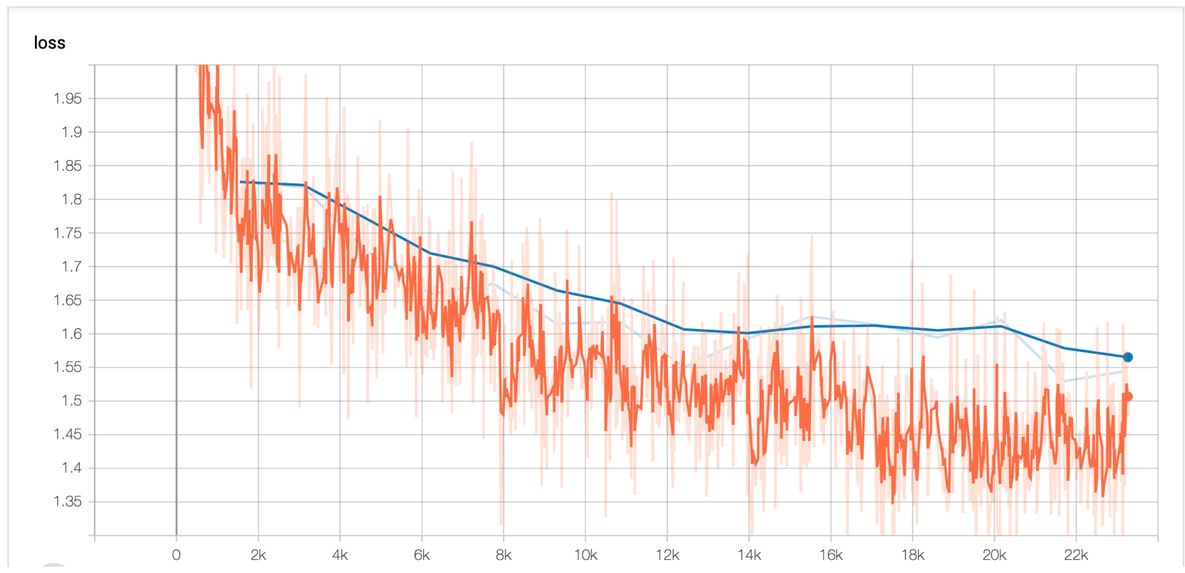


Figure 3: Training and Validation loss on 90% of the IWSLT dataset (15 epochs).

Memory Use Analysis

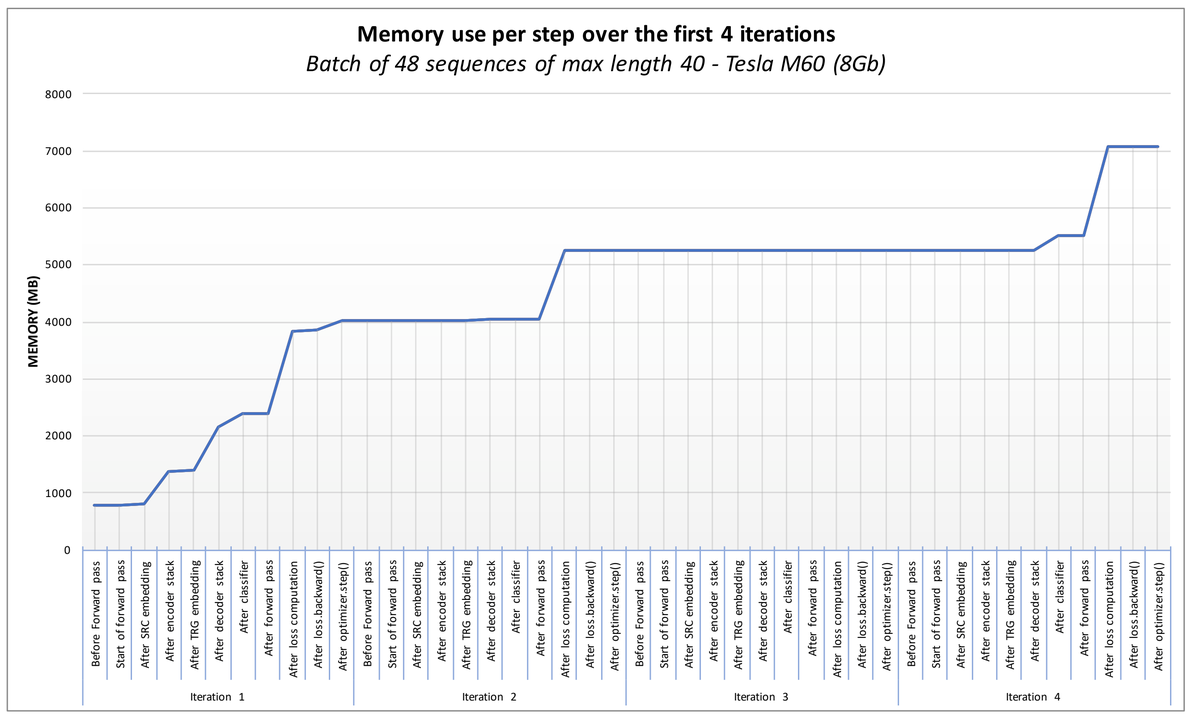


Figure 4: GPU Memory Use over the 1^{st} 4 iterations.

- Initial increase of memory use, particularly when computing loss,
- Stabilization over epoch at ∼6 Gb,

⇒ `PyTorch` most likely optimizing in the background.

Challenges

- Heavy model (65M parameters) & Aggregation of multiple, fine-tuned specifications ⇒ Non-trivial training,
- Non-intuitive training behavior: "No recurrence", but stack of layers and use of subsequent masking on an additional dimension,
- Inference is nonetheless step-by-step,
- Question of reproducibility and transparency remains open.