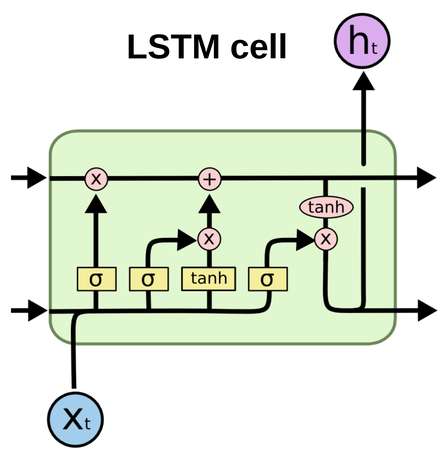
Comparison of LSTMs and Transformers in Sequence Modelling

Ashwin Raaghav Narayanan, Neeratyoy Mallik

Introduction

- The introduction of Transformers revolutionized NLP problems by dismantling the incumbent LSTMs to be new state-of-the-art.
- In this work, we compare their vanilla attention mechanisms to study how they differ across various sequential tasks.
- We attempt to show the trade-offs these modules can provide and how they *learn to remember*.

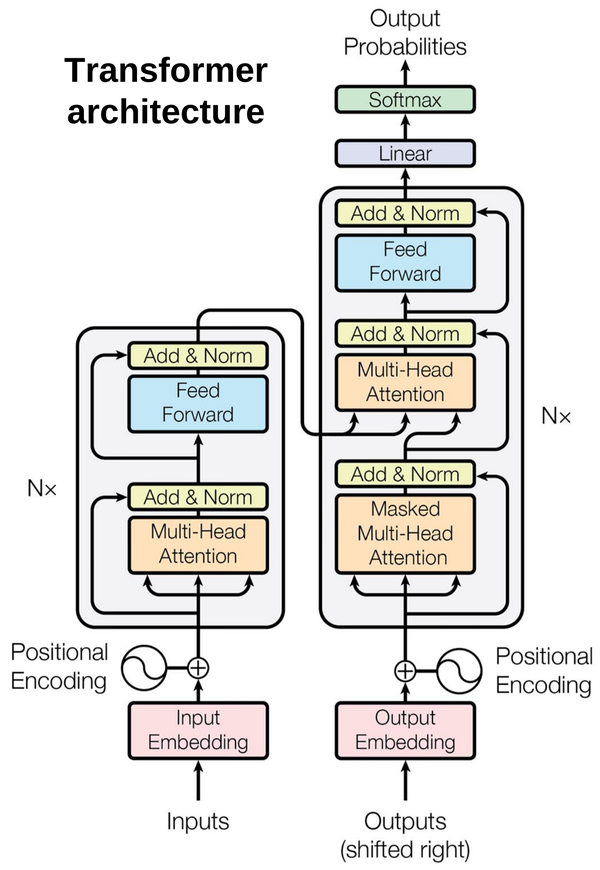


LSTMs

- Uses recurrence-based attention to remember patterns from sequential non i.i.d. data over extended time lag.
- Introduced novel learnable gating mechanism that regulates the flow of information through time.

Transformers

- Uses a global self-attention mechanism to create representations relating across time.
- First model to not rely on any sequence aligned recurrence or convolution mechanism.



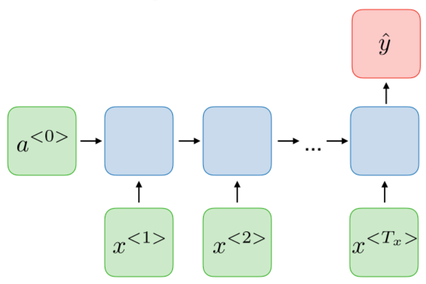
Experiments

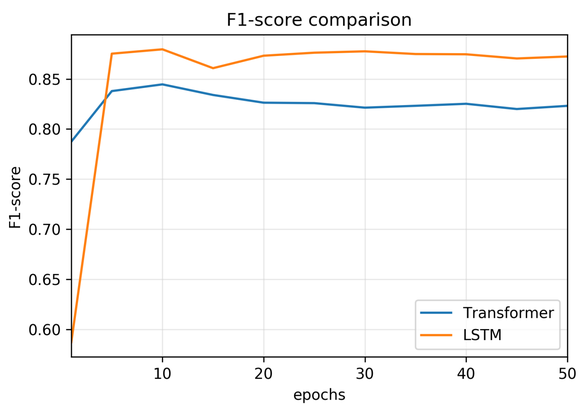
For fair comparison,

- Transformer's advantage of a global attention view is balanced by using a Bidirectional LSTM.
- Layer Normalization is applied before all the LSTM gates' non-linear activation.
- Only one layer was used for both LSTM and Transformer

Sequence Labelling

IMDb Sentiment Analysis

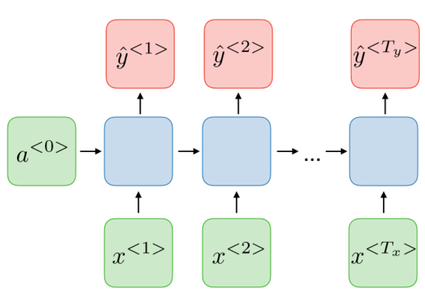


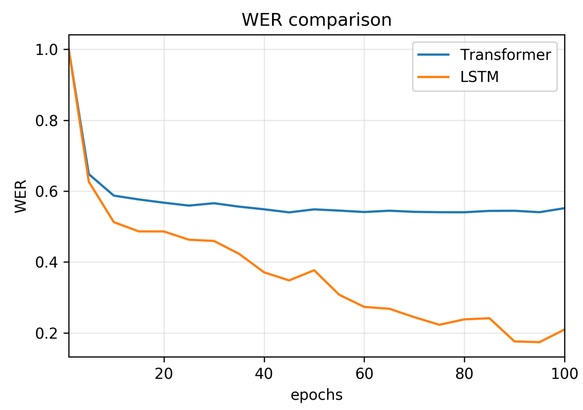


After 10 epochs	**LSTMs**	**Transformers**
# parameters	440,577	473,089
Test F1-score	0.88	0.86
Wall-clock(s)	41,759	681

Sequence-to-Sequence same

bAbi Task 2

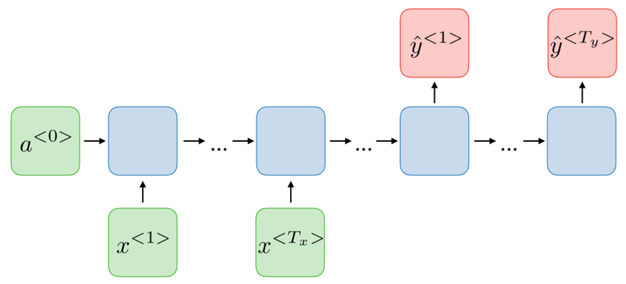


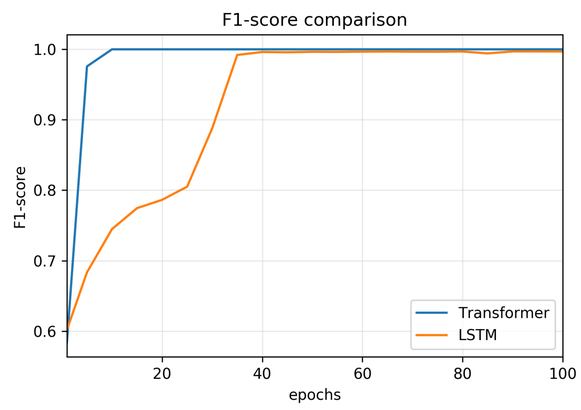


After 100 epochs	**LSTMs**	**Transformers**
# parameters	180,772	174,692
Test WER	0.54	0.69
Wall-clock(s)	31,606	1,780

Sequence-to-Sequence different

n-copy Task



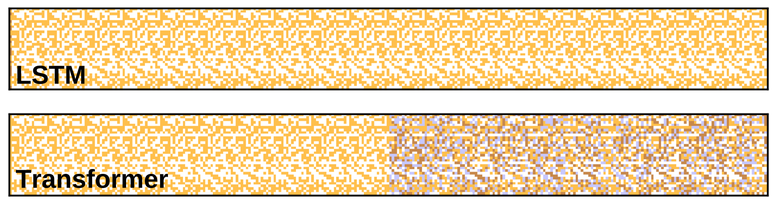


After 100 epochs	**LSTMs**	**Transformers**
# parameters	27,522	13,642
Test F1-score	0.99	1.0
Wall-clock(s)	14,297	503

Insights

- Transformers run faster, except when inferencing with a decoder.
- LSTMs and Transformers show similar performance when output doesn't depend on retaining information from specific timesteps.
- LSTM's recurrence and forget gates allow learning of a repetitive structure.
- Transformers lose context beyond sequence length it has encountered.
- "Transformer is the first transduction model relying entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution" -- Attention Is All You Need by Vaswani et al

Evaluation on a longer output



Source code: http://tiny.cc/lstm_transformer

As part of final evaluation for Deep Learning Lab, SS 19

ALBERT-LUDWIGS-UNIVERSITÄT FREIBURG
UNI FREIBURG