MxNet Gluon

Basics, Computer Vision, NLP (and even more NLP)
Part VII (Machine Translation / Transformers)

Leonard Lausen
Haibin Lin
Alex Smola

aws

Outline

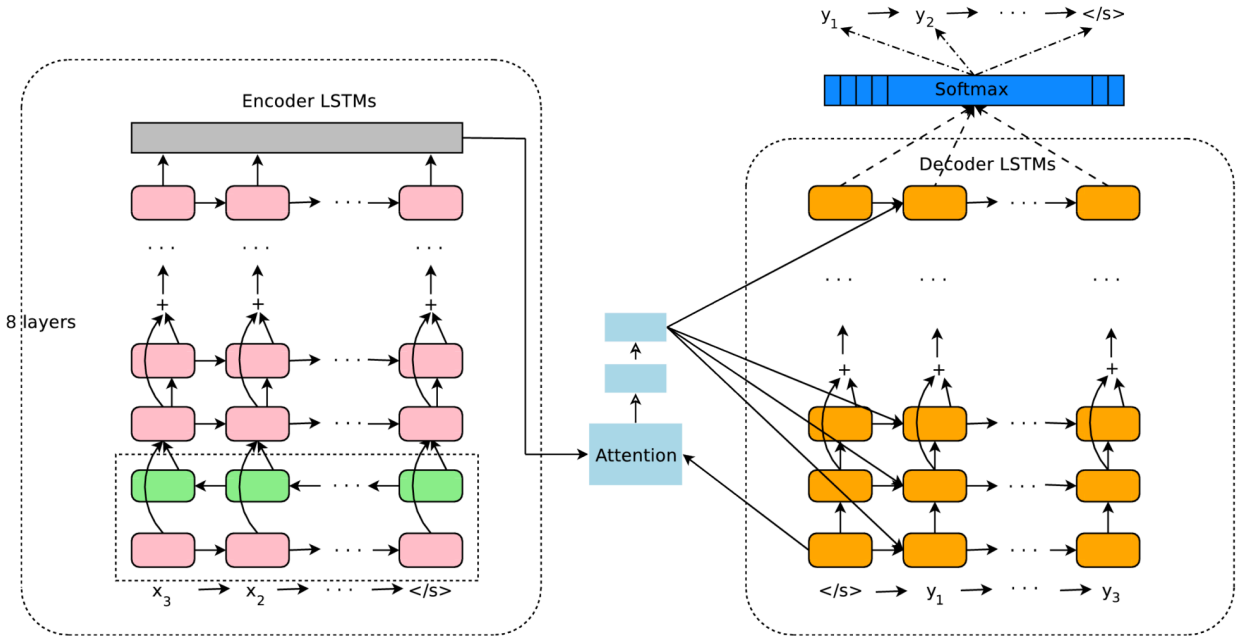
8:30-9:15	Installation and Basics (NDArray, AutoGrad, Libraries)
9:15-9:30	Neural Networks 101 (MLP, ConvNet, LSTM, Loss, SGD) - Part I
9:30-10:00	Break
10:00-10:30	Neural Networks 101 (MLP, ConvNet, LSTM, Loss, SGD) - Part II
10:30-11:00	Computer Vision 101 (Gluon CV)
11:00-11:30	Parallel and distributed training
11:30-12:00	Data I/O in NLP (and iterators)
12:00-13:30	Break
13:30-14:15	Embeddings
14:15-15:00	Language models (LM)
15:00-15:30	Sequence Generation from LM
15:30-16:00	Break
16:00-16:15	Sentiment analysis
16:15-17:00	Transformer Models & machine translation
17:00-17:30	Questions

Sequence translation models

Neural Machine Translation

- Need encoder of sequence
  - Words / characters to embedding
  - Embed entire sequence
- Attention for deciding where to position the decoder
- Again, LSTM stack for the decoded sequence
- Encoding / decoding via subwords rather than char / word

Google Neural Machine Translation

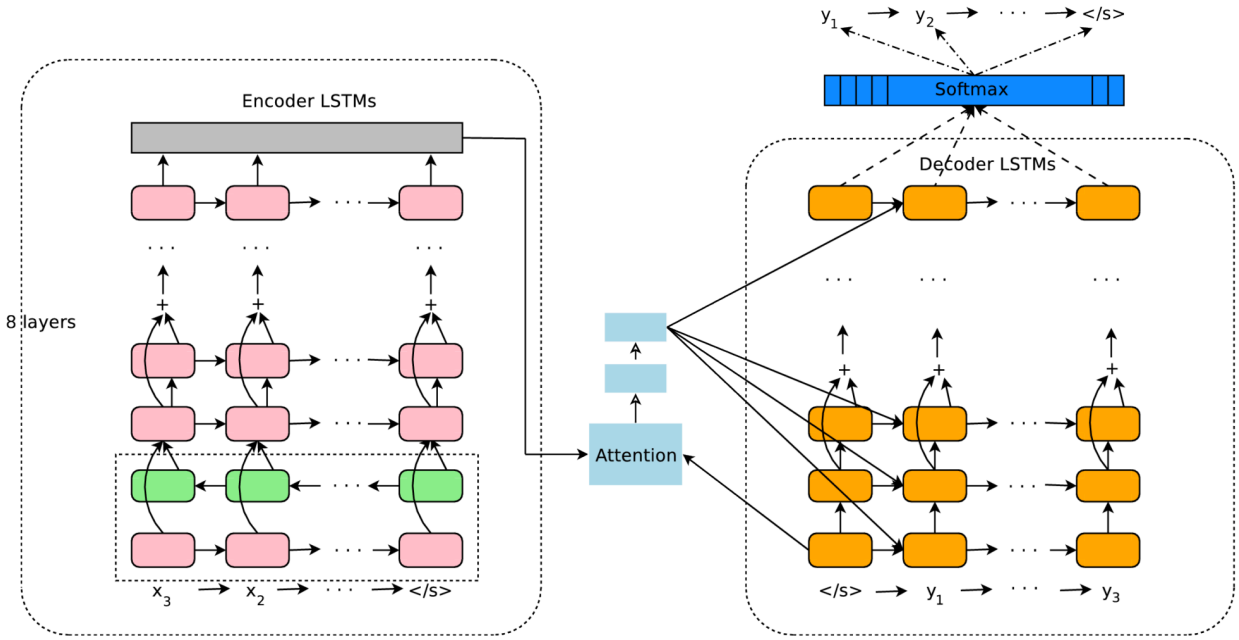


Encoder: Bidireciontal LSTM + LSTM + Residual

Decoder: LSTM + Residual + MLP Attention

Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv preprint arXiv:1609.08144* (2016).

Google Neural Machine Translation



Encoder: Bidireciontal LSTM + LSTM + Residual

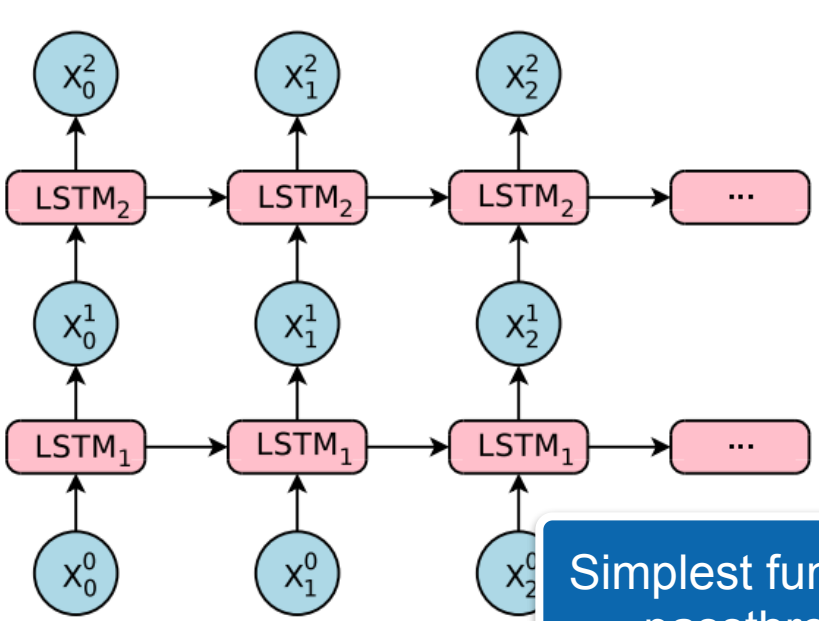
Decoder: LSTM + Residual + MLP Attention

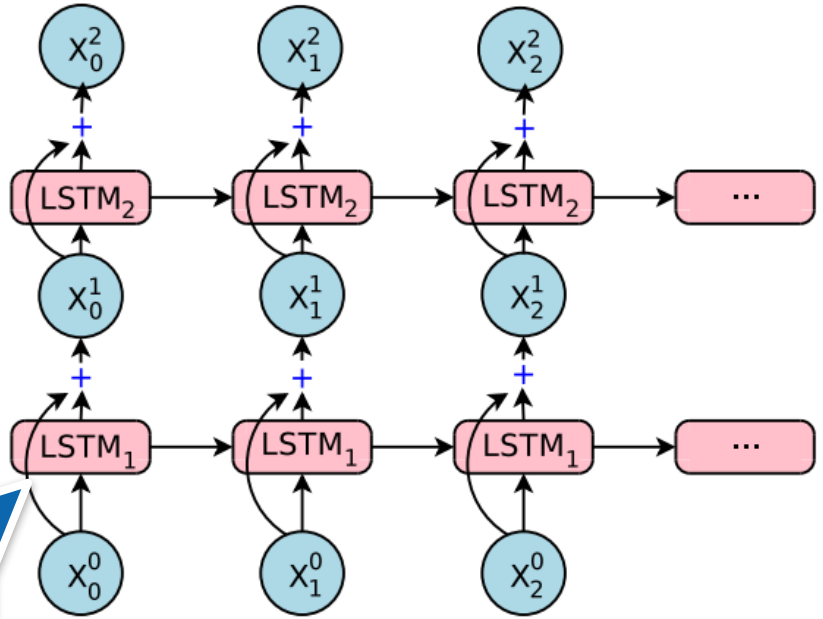
- Gluon-NLP:
  - BLEU **26.22** on IWSLT2015, 10 epochs, Beam Size=10
- Tensorflow/NMT:
  - BLEU **26.10** on IWSLT2015, Beam Size=10

Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." *arXiv preprint arXiv:1609.08144* (2016).

aws

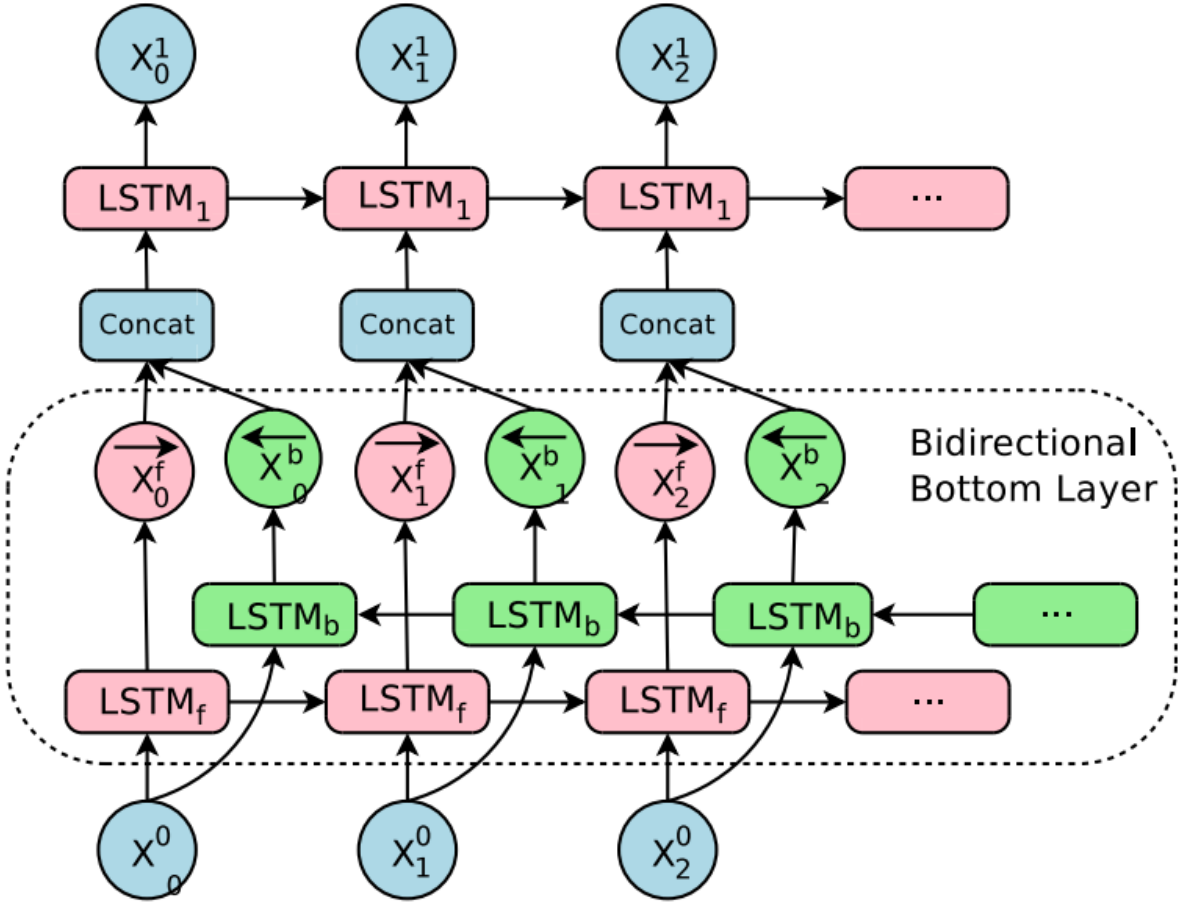
Detail - LSTM with Residual Connections





Simplest function is passthrough

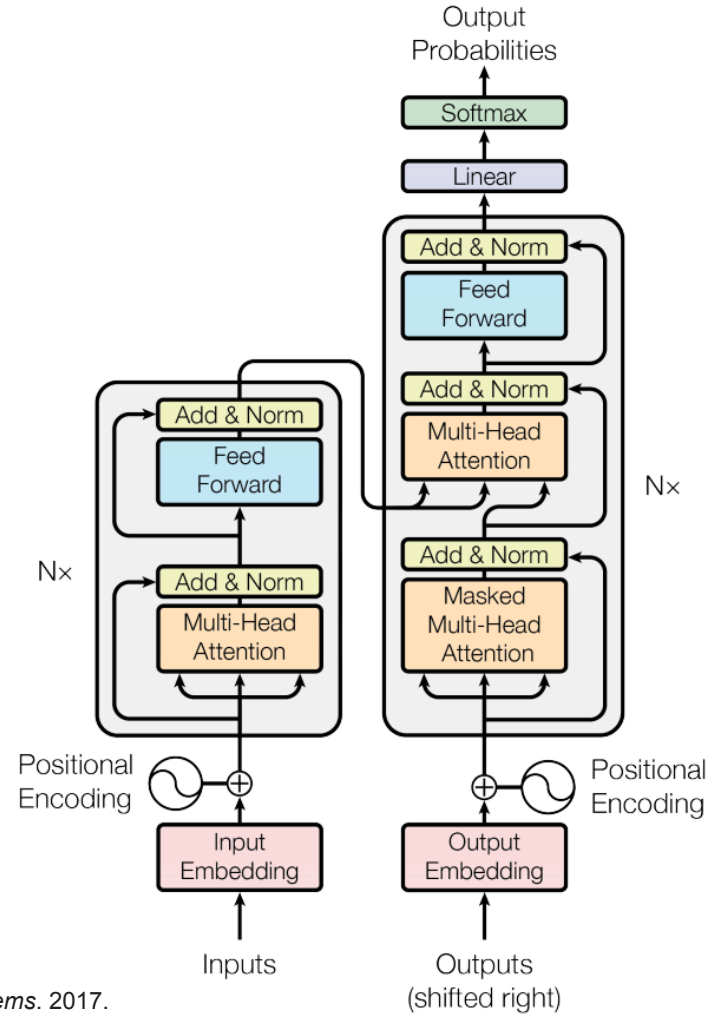
Detail - Hindsight is 20/20 - Bidirectional LSTM



Transformers
aka
Do we really need LSTMs?

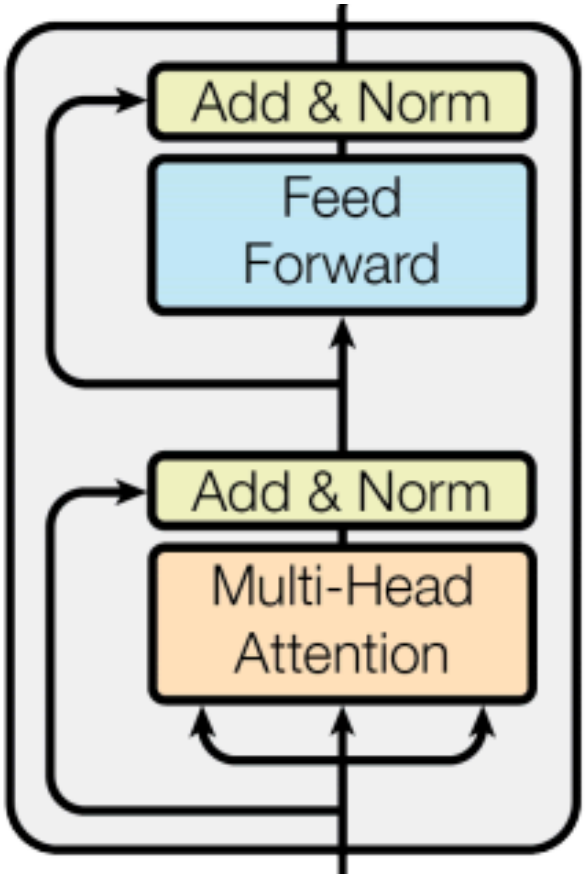
Transformer

- **Encoder**
  - 6 layers of self-attention+ffn
- **Decoder**
  - 6 layers of masked self-attention
  - output of encoder + ffn
- Our implementation:
  - BLEU **27.51** on WMT2014en_de,
- Tensorflow/t2t:
  - BLEU **26.55** on WMT2014en_de



Vaswani, Ashish, et al. "Attention is all you need." *Advances in Neural Information Processing Systems*. 2017.

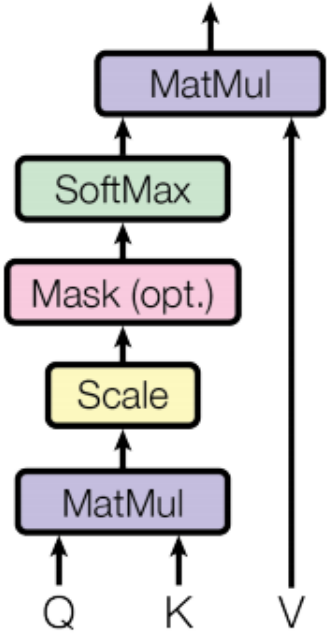
Transformer



Vaswani, Ashish, et al. "Attention is all you need." *Advances in Neural Information Processing Systems*. 2017.

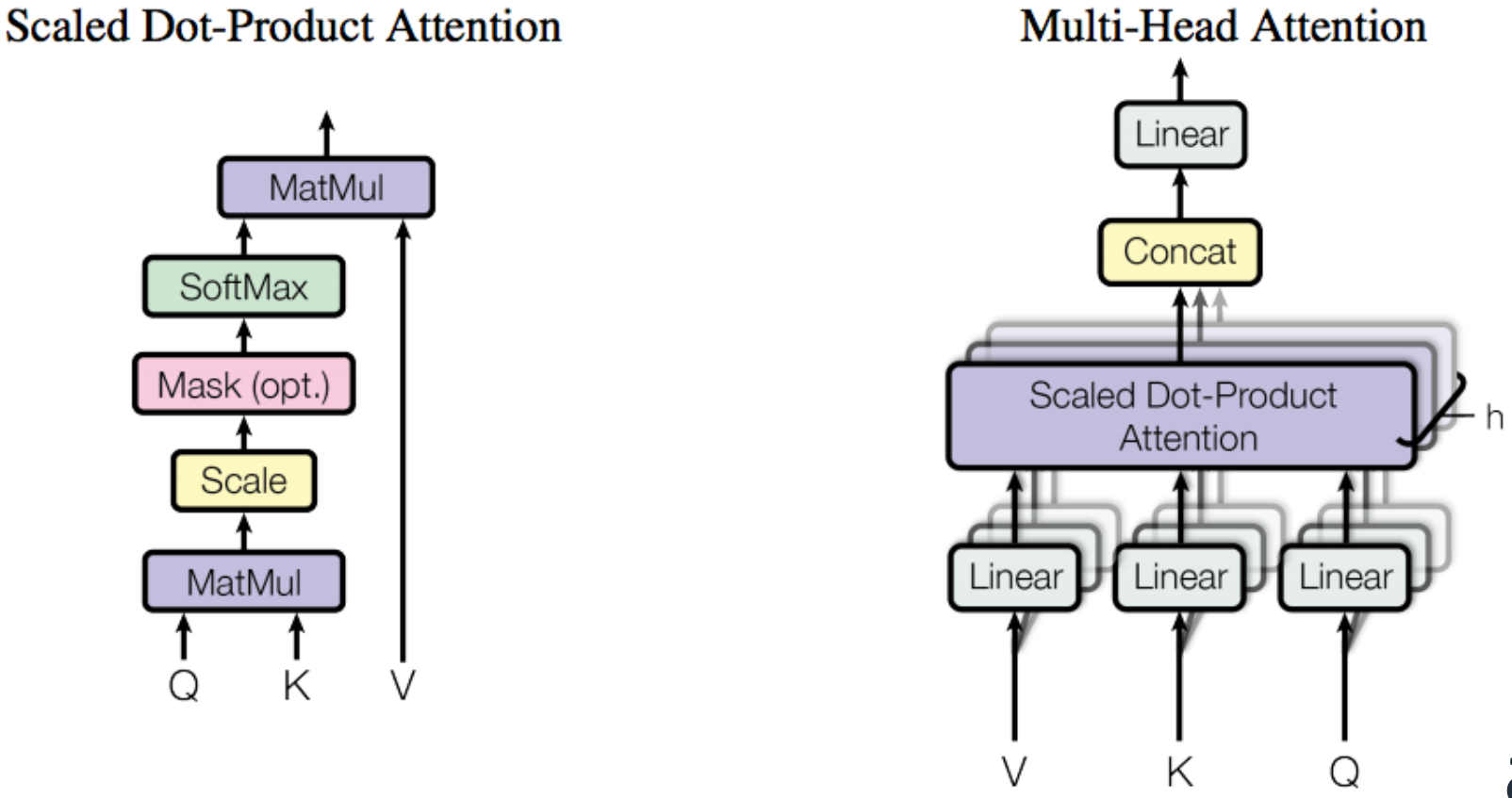
Self Attention Module

$$\text{Attention}(Q, K, V) = \text{softmax}\left(d_k^{-\frac{1}{2}} QK^\top\right) V$$

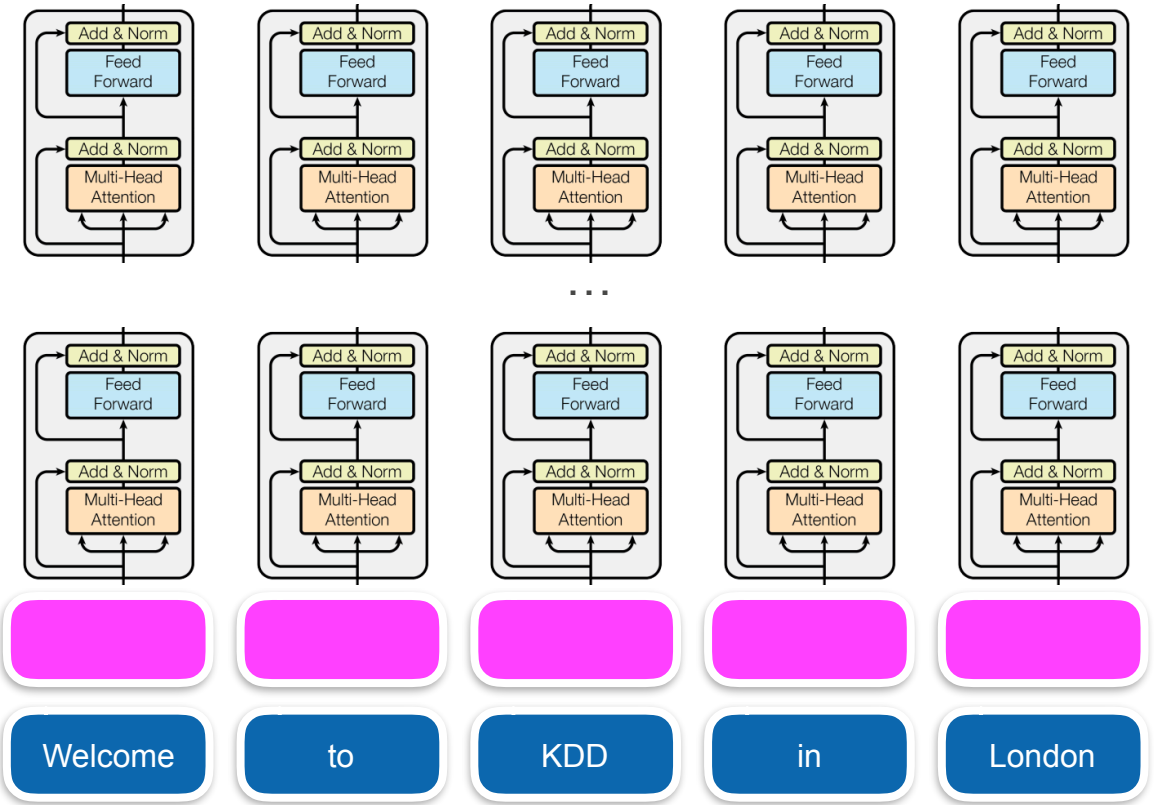


- Q - ‘queries’
  (usually learned parameters)
- K - ‘keys’
  (can be embeddings themselves)
- V - ‘values’
  (can be embeddings, too, i.e. K=V)

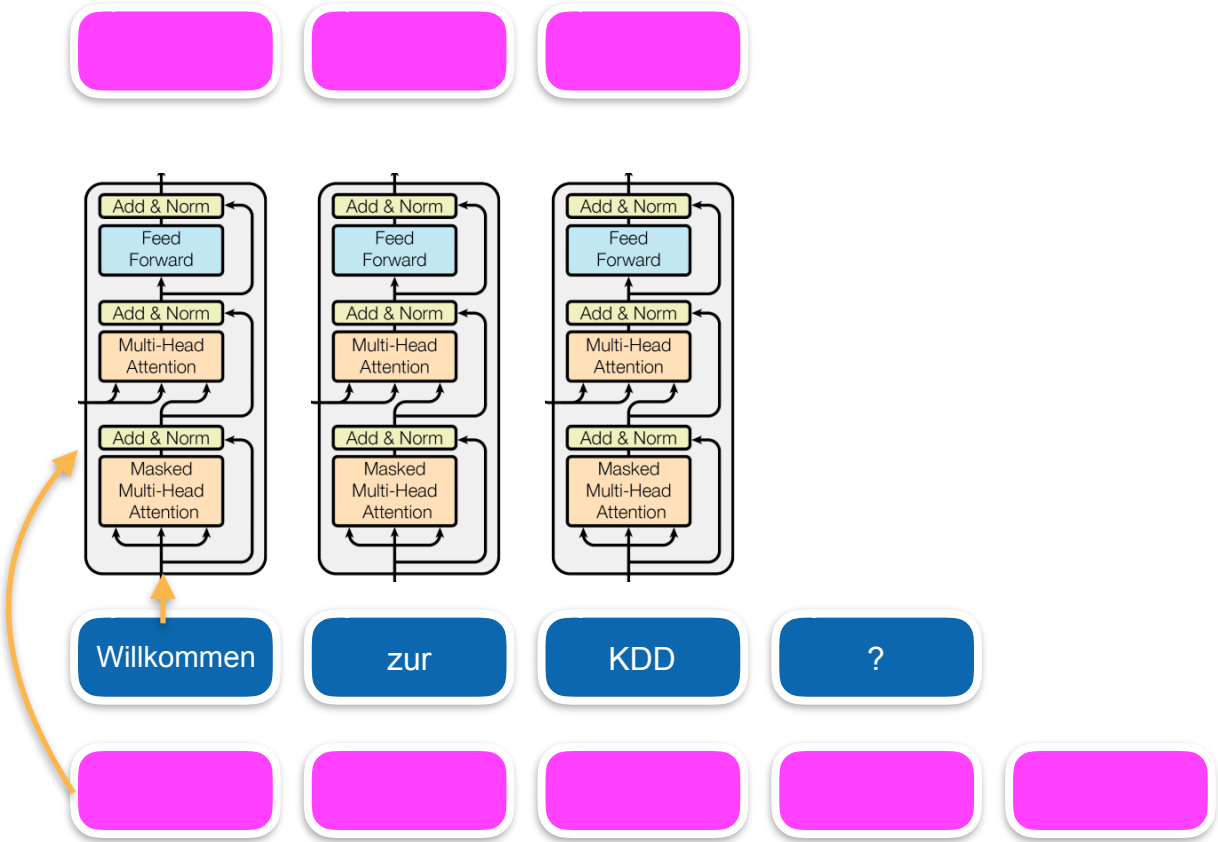
Multi-Head Attention



Encoding the input



Decoding the output



… then beam search to decode

Resources

- Deep Learning — the Straight Dope
  https://gluon.mxnet.io/
- A 60-minute Gluon Crash Course
  https://gluon-crash-course.mxnet.io/
- GluonCV
  http://gluon-cv.mxnet.io/
- GluonNLP
  https://gluon-nlp.mxnet.io/
- MXNet User Forum
  http://discuss.mxnet.io/
- MXNet Documentation
  https://mxnet.apache.org/

aws