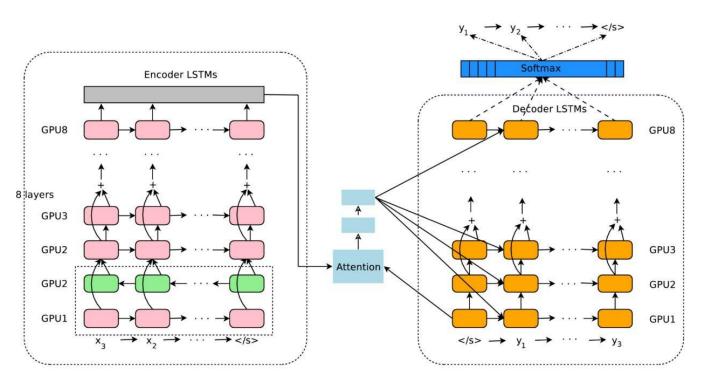
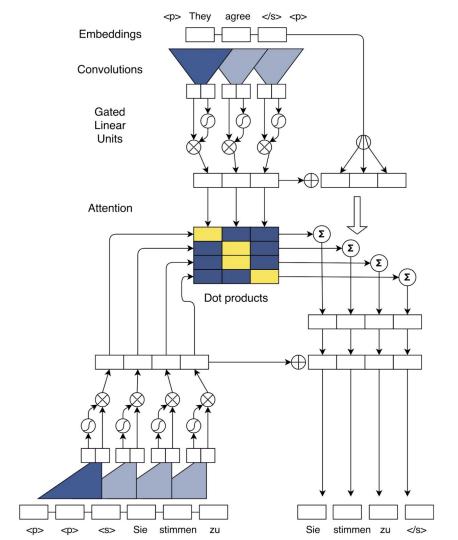
Encoder / Decoder Architectures

Google's Neural Machine Translation System



- Model is deep
- Bi-directional encoder for first layer
- Parallelism



Convolutional Sequence to Sequence Learning

- 9x times faster than RNN
- CNNs can outperform RNNs on big benchmarks
- Multi-step attention

Output Probabilities Softmax Linear Add & Norm Feed Forward Add & Norm Add & Norm Multi-Head Feed Attention Forward N× Add & Norm N× Add & Norm Masked Multi-Head Multi-Head Attention Attention Positional Positional Encoding Encoding Input Output Embedding Embedding Outputs Inputs (shifted right)

Attention Is All You Need

- Outperforms all previous SotA in quality and speed of training
- Encoder is highly parallel