Attention Is All You Need

Lee Haesung
Soynet Internship

NIPS'17

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

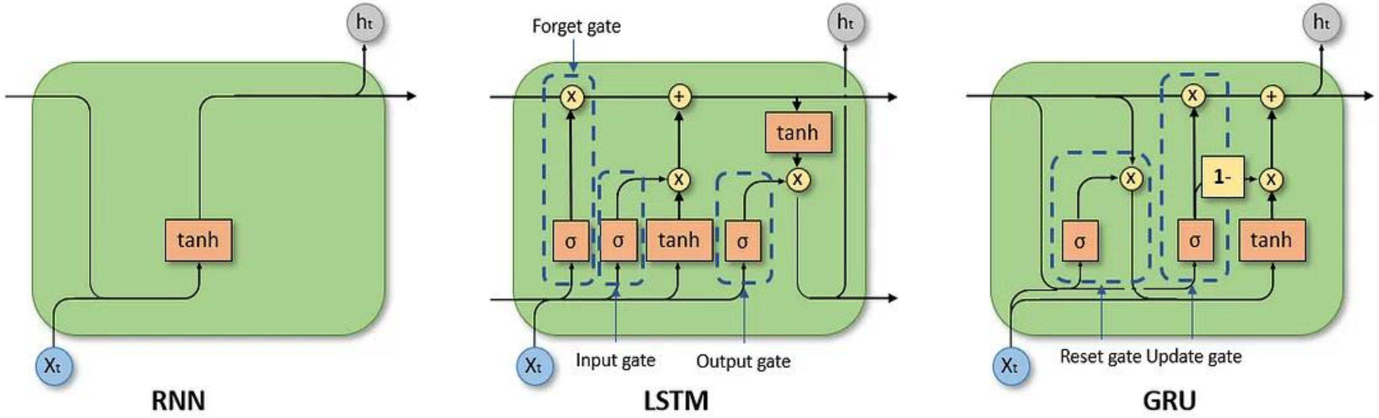
†Work performed while at Google Brain.
‡Work performed while at Google Research.

Overview

- Introduction (Motivation)
- Background (RNN, Attention)
- Model Architecture (Transformer)
- Self-Attention
- Training & Results
- Discussion

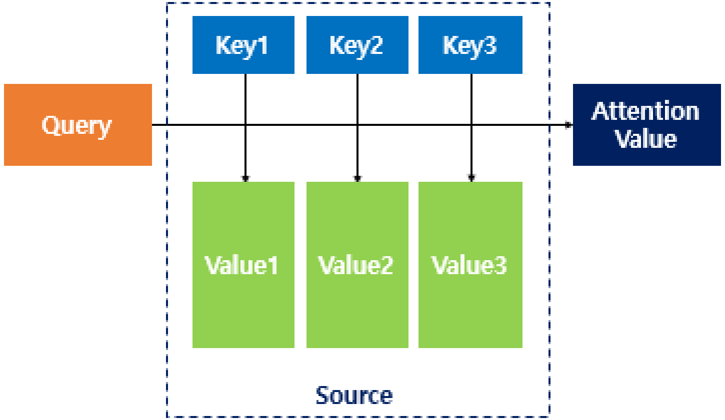
Introduction – Sequential Nature

- **Recurrent model (Recurrent Neural Network, RNN)**
  - Generate a sequence of hidden states h_t, as a function of the previous hidden state h_{t-1} and the input for position t
  - Should wait h_{t-1} for h_t for sequence information
  - precluding parallelization



Introduction – Attention

- **Modelling of dependencies without regard to their distance in the input or output sequences**
  - Mostly used with RNN (not now)

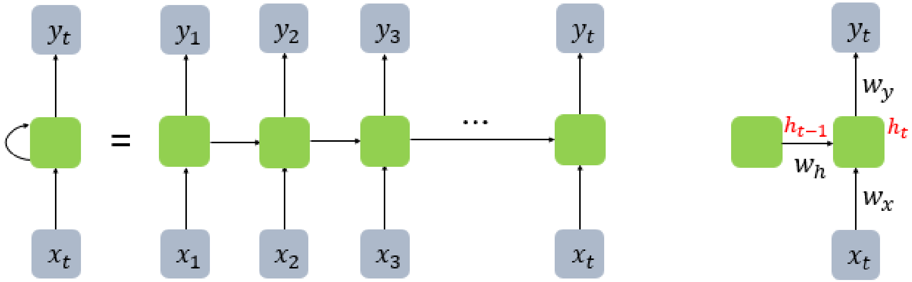


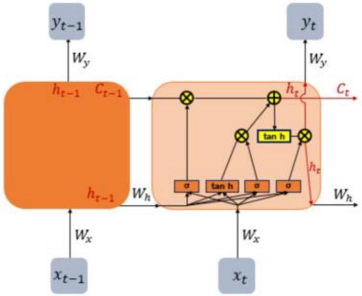
Introduction - Transformer

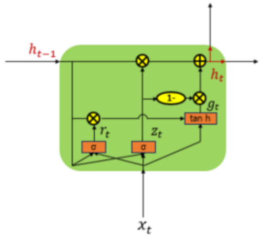
- **Model architecture**
  - Eschewing recurrence
  - Relying entirely on an attention mechanism to draw global dependencies between input and output.

Background – RNN

- **Sequence Modeling and Transduction Problems**
  - Language Modeling, Machine Translation
  - Time Series Prediction
- **Recurrent Neural Network**
  - RNN (vanila)
  - Long Short-Term Memory
  - Gated Recurrent Unit





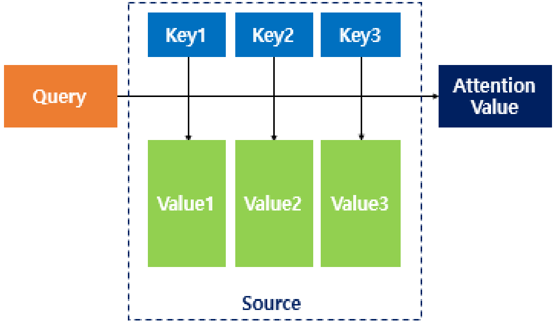


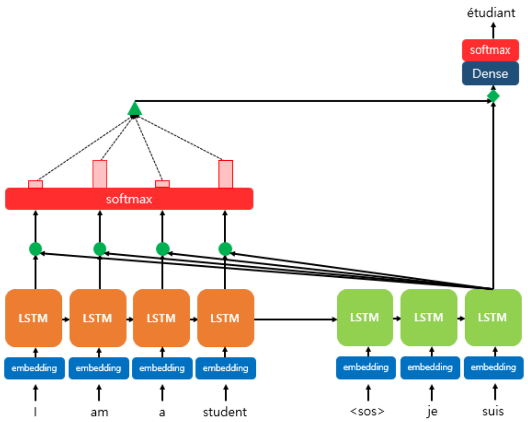
Background - Attention

- **RNN – Drawbacks**
  - Gradients Vanishing
  - Lose of Information
- **Attention Mechanism**
  - Calculate current state with every previous states
  - Find out important states for now

Background – Attention (Cont'd)

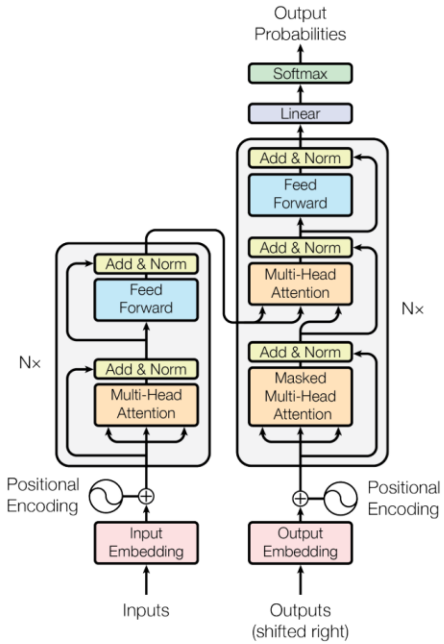
- **Query, Key, and Value**
  - Query($Result_{t-1}$)
  - Key($Probability_{(t-1),i} = SoftMax(Hidden_i * Result_{t-1}\,)$)
  - Value($Probability_{(t-1),i} * Hidden_i$)





Model Architecture

- **Encoder and Decoder Stacks**
- **Attention**
- **Position-wise Feed-Forward Networks**
- **Embeddings and Softmax**

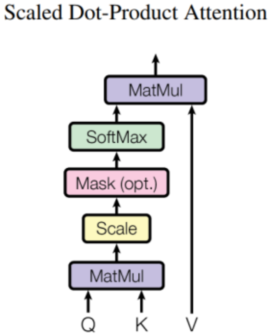


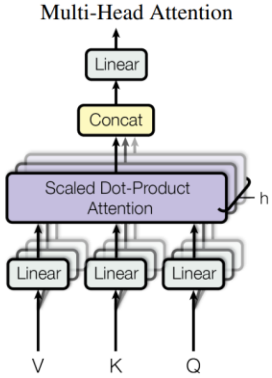
Model Architecture – Encoder/Decoder

- **Encoder**
  - Six identical layers
  - Each layer consists of self attention and position-wise fully connected feed-forward network sublayers
    - Each sublayer got residual connection by itself
    - Layer batch normalization
- **Decoder**
  - Six identical layers
  - Each layer consists of two sublayers as same as encoder and one additional sublayers
    - Perform attention mechanism with output from encoder
    - Modify existed attention layer to masked one

Model Architecture - Attention

- **Multi-Head Attention**
  - Multiple queries at once
- **Application**
  - Encoder
    - Query – Certain position
    - Value – Every positions except query
    - Key – Same position with value
  - Decoder
    - Query – Certain position
    - Value – Every positions except query
    - Key – Same position with value
  - Encoder-Decoder
    - Query – Output of previous decoder layer
    - Value – Output of encoder
    - Key - Same position with value





Self Attention

- **Convolution**
- **Self Attention**
  - Computational Complexity
  - Parallelism
  - Path Length between Long-Range Dependencies
    - Length of the paths forward and backward signals have to traverse in the network matters

Training & Result

- **Dataset**
  - WMT 2014 English-German dataset
- **Hardware**
  - 8 NVIDIA P100 GPUs
- **Optimizer**
  - Adam

Discussion

Q&A