PAY LESS ATTENTION WITH LIGHTWEIGHT AND DYNAMIC CONVOLUTIONS

Authors: Alexei Baevski et al,
Facebook AI research

Presented by
Krishna Bairavi Soundararajan
Graduate Intern, Mayo Clinic

Attention is all you need!

Transformer Architecture

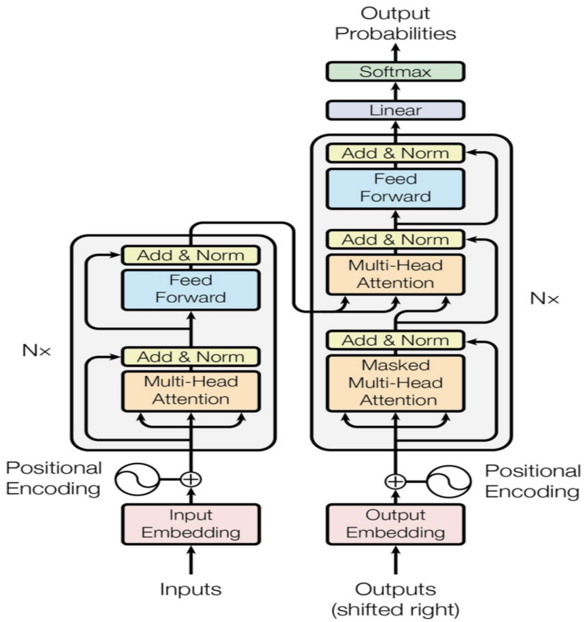
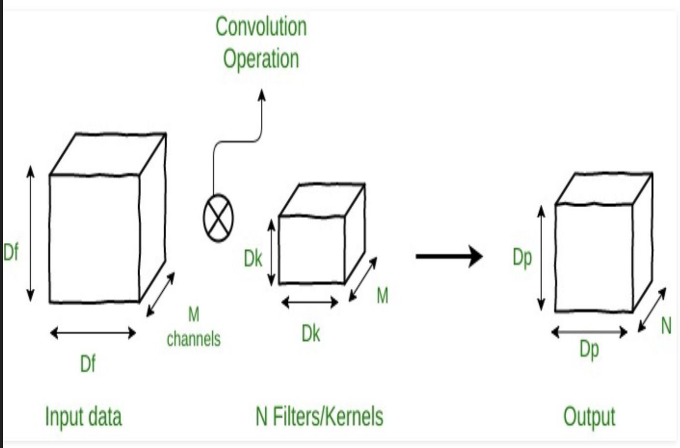


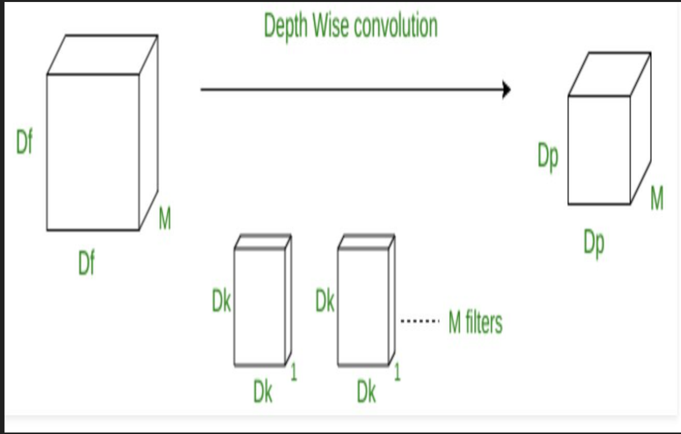
Figure 1: The Transformer - model architecture.

Depth-wise convolutions

- Less number of parameters to adjust
- Reduces Overfitting
- Computationally cheaper [2]

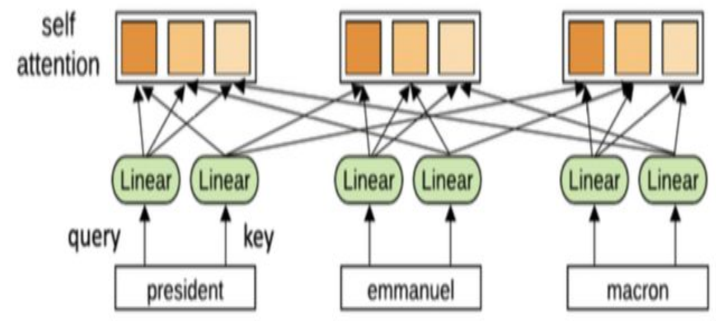


Normal Convolution

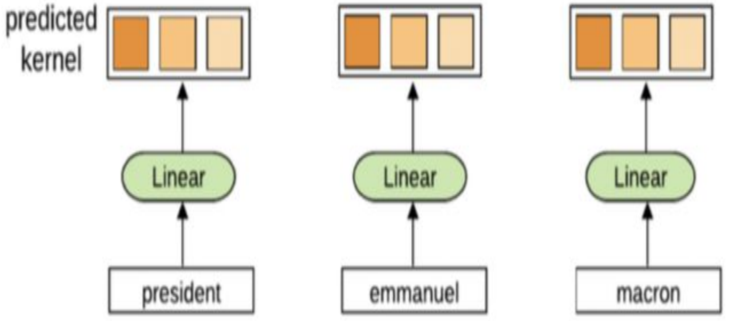


Depth Wise Convolutions

Self-attention and Dynamic Convolution



(a) Self-attention



(b) Dynamic convolution

Overview:

- Lightweight convolutions perform competitively/ on par to self attention
- Dynamic Convolutions- Simpler and Efficient than self-attention
- Predict convolution kernels based only on current time step to determine importance of context elements [1]

Depth-wise convolutions over self-attention?

- Does self-attention really model long-range dependencies? [3]
- Self attention is Computationally challenging -- due to quadratic complexity in input length
- Long sequences require hierarchies [4]

Method

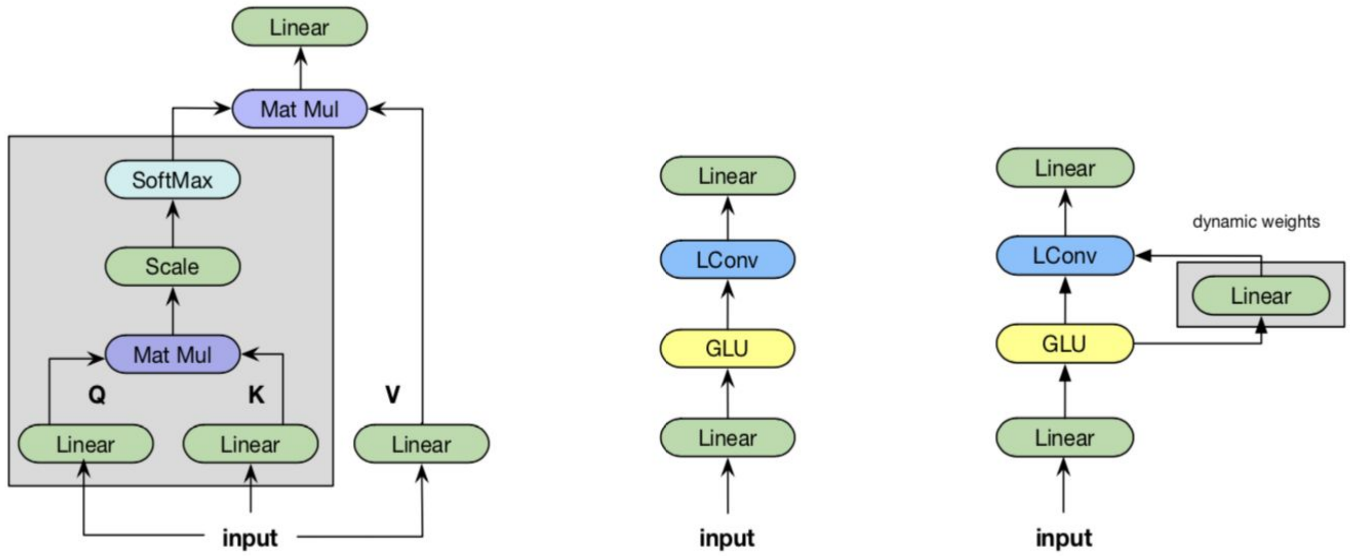
Lightweight convolutions:

- Depth wise separable
- Less weights compared to self-attention
- Weights are reused

Dynamic convolutions:

- Built on Lightweight convolutions
- Depth wise separable
- Predicting different convolutional kernel at every time step
- Weights are dynamically generated

Model Comparison



(a) Self-attention (b) Lightweight convolution (c) Dynamic convolution

Figure 2: Illustration of self-attention, lightweight convolutions and dynamic convolutions

Gated Linear Units (GLU)

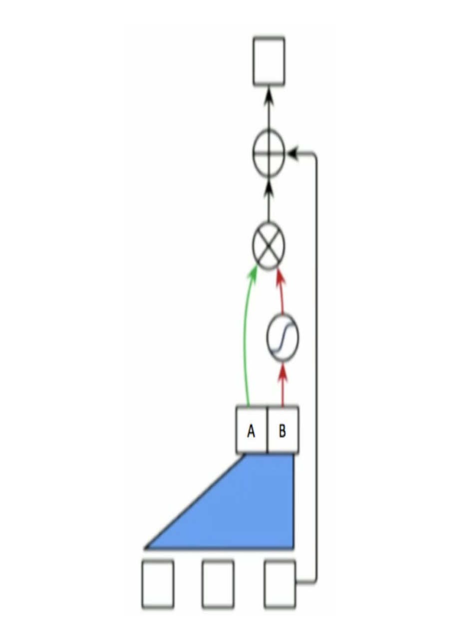
- Uses half of the inputs as gates using sigmoid function
- Pointwise product with other inputs [5]

Advantages:

- Multiplicative skip connection avoiding gradient flow

■ Gated linear units (GLU)

$$h_l(\mathbf{E}) = (\mathbf{E} * \mathbf{W} + b) \otimes \sigma(\mathbf{E} * \mathbf{V} + c)$$



Gated Linear Unit (GLU), with residual skip connection. A convolutional block with window k=3 produces two convolutional outputs, A and B. A is element-wise multiplied with sigmoid(B), and the residual is added to the output. Image

Dynamic Convolutions:

- Uses a timestep dependent kernel
- They change weights over time
- Diff b/w self-attention:
  - The weights are dependent only on the current time step
  - Self-attention- High computational cost (Quadratic operations)
  - Dynamic Convs- Less computational cost (Scales linearly in the sequence)

$$\mathrm{DynamicConv}(X, i, c) = \mathrm{LightConv}(X, f(X_i)_{h,:}, i, c)$$

Model Architecture

- Encoder (contains 2 blocks):
  - First Block:
    - LightConv or DynamicConv module
  - Second Block:
    - Feed-forward module with Relu Activation
- Decoder
  - Identical.
  - Additional source target attention sub-block
  - Source target attention= Self- attention [6]

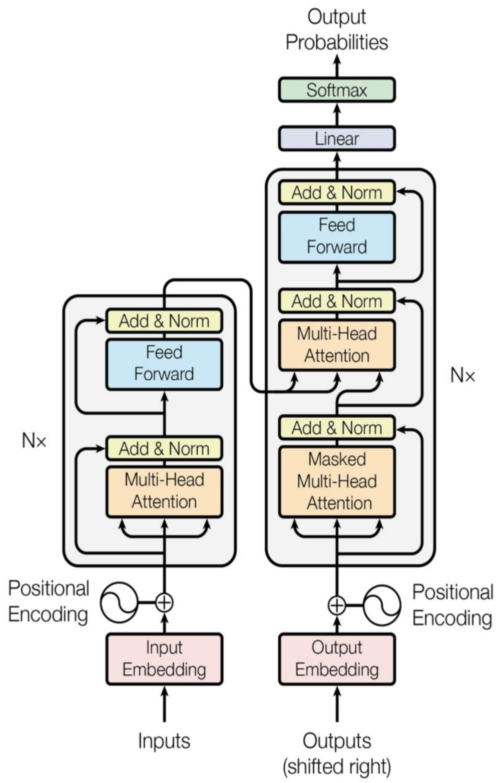


Figure 1: The Transformer - model architecture.

Model Comparison

- Lightweight convolutions perform competitively with self-attention models
- Dynamic convolutions outperform with self-attention in various tasks
- 20% faster run time than self-attention

Results and Evaluation

Three task evaluation

- Machine translation
- Language Modeling
- Abstractive Summarization

References

1. https://openreview.net/pdf?id=SkVhlh09tX
2. https://www.geeksforgeeks.org/depth-wise-separable-convolutional-neural-networks/
3. https://arxiv.org/pdf/1808.08946.pdf
4. https://arxiv.org/pdf/1801.10198.pdf
5. https://vimeo.com/238222385
6. https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Thank you!