Pay Less Attention with Lightweight and Dynamic Convolutions

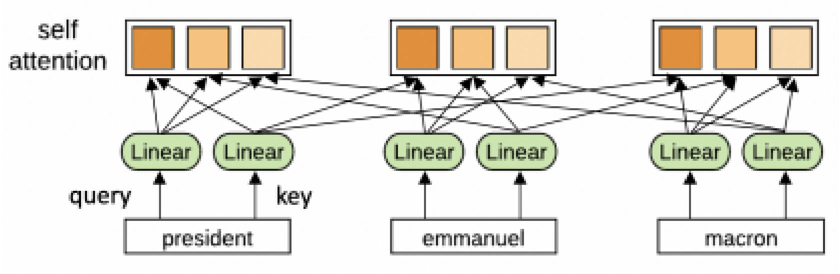
Felix Wu, Angela Fan, Alexei Baevski, Yann Dauphin, Michael Auli

Motivation

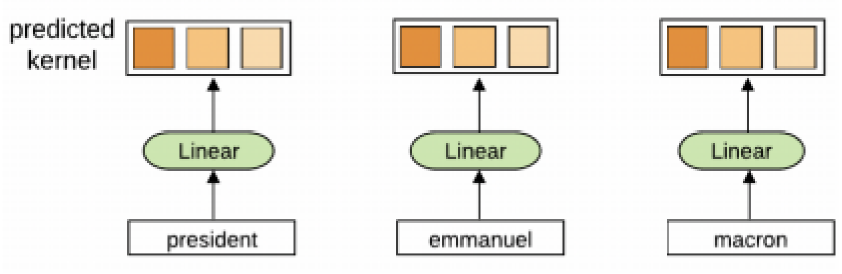
- Self-attention determines the importance of context elements by comparing each element to the current time step.
- Is the self-attention the most important component in the structure of transformer?
- The number of operations required by self-attention scales quadratic in the input length.
- Is there any way to reduce this to linear complexity?

Solution

• Convolution Network



Self-Attention

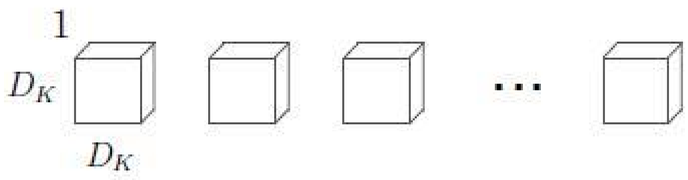


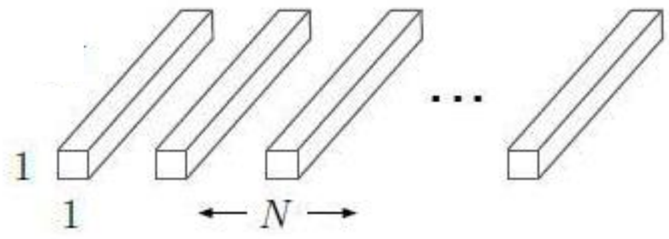
Convolution

Depthwise convolutions

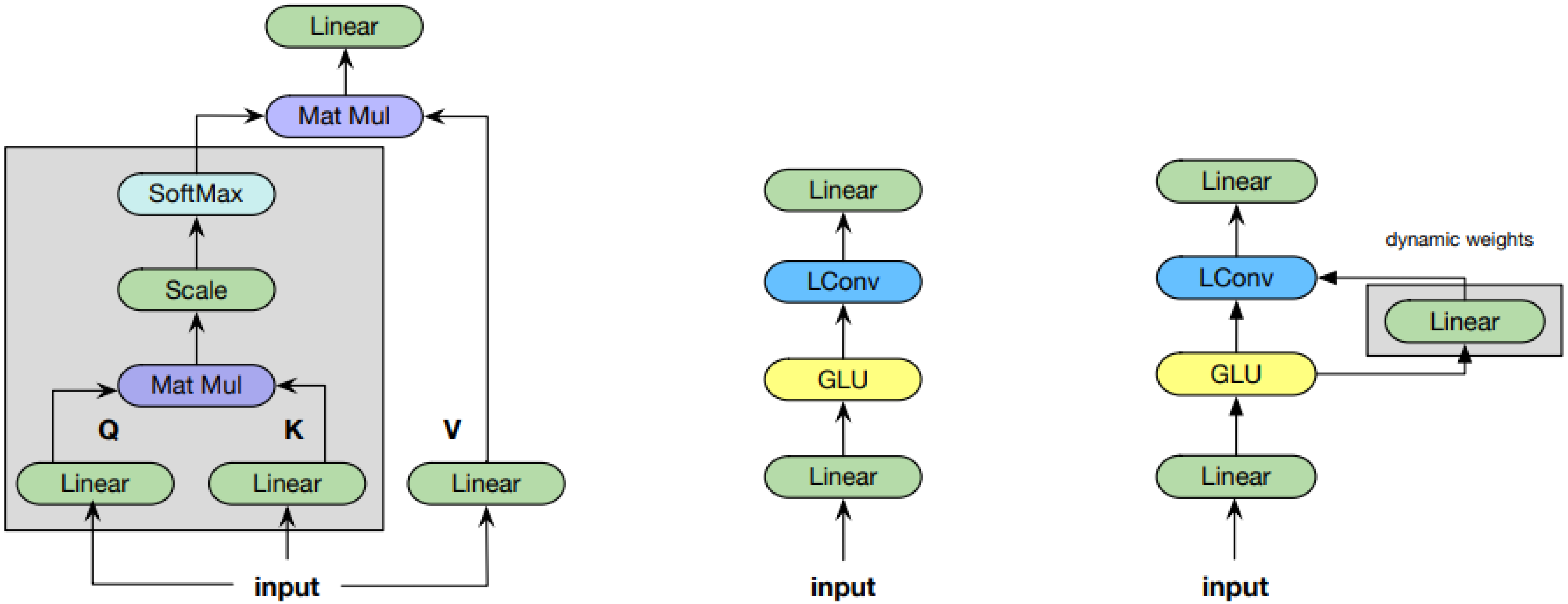
- Come from Xception – extreme Iception

- Fundamental hypothesis: cross-channel correlations and spatial correlations can be entirely decoupled.





Structure Comparation



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

Lightweight Convolutions

- Depthwise convolution
- Weights are normalized across the temporal dimension using a softmax
- Weights are shared within different output channels

$$\mathrm{LightConv}(X, W_{\lceil \frac{cH}{d} \rceil,:}, i, c) = \mathrm{DepthwiseConv}(X, \mathrm{softmax}(W_{\lceil \frac{cH}{d} \rceil,:}), i, c)$$

- Example : a regular convolution requires 7,340,032 ($d^2 \times k$) weights for d = 1024 and k = 7, a depthwise separable convolution has 7,168 weights ($d \times k$), and with weight sharing, H = 16, we have only 112 ($H \times k$) weights

Dynamic Convolutions

- Takes the same form as LightConv but uses a time-step dependent kernel that is computed using a function $\boldsymbol{f}: \boldsymbol{R}^{\boldsymbol{d}} \rightarrow \boldsymbol{R}^{\boldsymbol{H \times k}}$

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

- Here $\boldsymbol{f}$ is simple linear module with learned weight $W^{Q} \in R^{H \times k \times d}$:

$$f(X_i) = \sum_{c=1}^{d} W_{h,j,c}^{Q} X_{i,c}$$

Experiment

- Setting
  - Use same setting as “Attention is all you need”
  - Replace the self-attention module for lightweight and dynamic convolutions
  - The encoder and decoder’s kernel sizes to 3, 7, 15, 31x4 for each block respectively
- Tasks
  - Machine Translation – WMT Zh-En; WMT En-De; WMT En-Fr; IWSLT Zh-En
  - Language Modeling - Billion word dataset
  - Summarization - CNN-DailyMail

Result

- Machine Translation

Model	Param (En-De)	WMT En-De	WMT En-Fr
Gehring et al. (2017)	216M	25.2	40.5
Vaswani et al. (2017)	213M	28.4	41.0
Ahmed et al. (2017)	213M	28.9	41.4
Chen et al. (2018)	379M	28.5	41.0
Shaw et al. (2018)	-	29.2	41.5
Ott et al. (2018)	210M	29.3	43.2
LightConv	202M	28.9	43.1
DynamicConv	213M	**29.7**	**43.2**

Model	Param (Zh-En)	IWSLT	WMT Zh-En
Deng et al. (2018)	-	33.1	-
Hassan et al. (2018)	-	-	24.2
Self-attention baseline	292M	34.4	23.8
LightConv	285M	34.8	24.3
DynamicConv	296M	**35.2**	**24.4**

Result

- Machine Translation

Model	Param	BLEU	Sent/sec
Vaswani et al. (2017)	213M	26.4	-
Self-attention baseline (k=inf, H=16)	210M	26.9 ± 0.1	52.1 ± 0.1
Self-attention baseline (k=3,7,15,31x3, H=16)	210M	26.9 ± 0.3	54.9 ± 0.2
CNN (k=3)	208M	25.9 ± 0.2	68.1 ± 0.3
CNN Depthwise (k=3, H=1024)	195M	26.1 ± 0.2	67.1 ± 1.0
+ Increasing kernel (k=3,7,15,31x4, H=1024)	195M	26.4 ± 0.2	63.3 ± 0.1
+ DropConnect (H=1024)	195M	26.5 ± 0.2	63.3 ± 0.1
+ Weight sharing (H=16)	195M	26.5 ± 0.1	63.7 ± 0.4
+ Softmax-normalized weights [LightConv] (H=16)	195M	26.6 ± 0.2	63.6 ± 0.1
+ Dynamic weights [DynamicConv] (H=16)	200M	26.9 ± 0.2	62.6 ± 0.4
Note: DynamicConv(H=16) w/o softmax-normalization	200M	diverges	
AAN decoder + self-attn encoder	260M	26.8 ± 0.1	59.5 ± 0.1
AAN decoder + AAN encoder	310M	22.5 ± 0.1	59.2 ± 2.1

Result

- Language Modeling

Model	Param	Valid	Test
2-layer LSTM-8192-1024 (Józefowicz et al., 2016)	–	–	30.6
Gated Convolutional Model (Dauphin et al., 2017)	428M	–	31.9
Mixture of Experts (Shazeer et al., 2017)	4371M †	–	28.0
Self-attention baseline	331M	26.67	26.73
DynamicConv	339M	26.60	**26.67**

Result

- Summarization

Model	Param	Rouge-1	Rouge-2	Rouge-l
LSTM (Paulus et al., 2017)	-	38.30	14.81	35.49
CNN (Fan et al., 2017)	-	39.06	15.38	35.77
Self-attention baseline	90M	39.26	15.98	36.35
LightConv	86M	39.52	15.97	36.51
DynamicConv	87M	**39.84**	**16.25**	**36.73**
Bottom-Up (Gehrmann et al., 2018)	-	41.22	18.68	**38.34**
RL (Celikyilmaz et al., 2018)	-	**41.69**	**19.47**	37.92

Conclusion

- Demonstrates that self-attention is not critical to achieve good accuracy on the language tasks.
- Both lightweight convolution and dynamic convolution are 20% faster at runtime than self-attention.
- Get comparable or better results in all tasks to self-attention