Pay Less Attention with Lightweight and Dynamic Convolutions

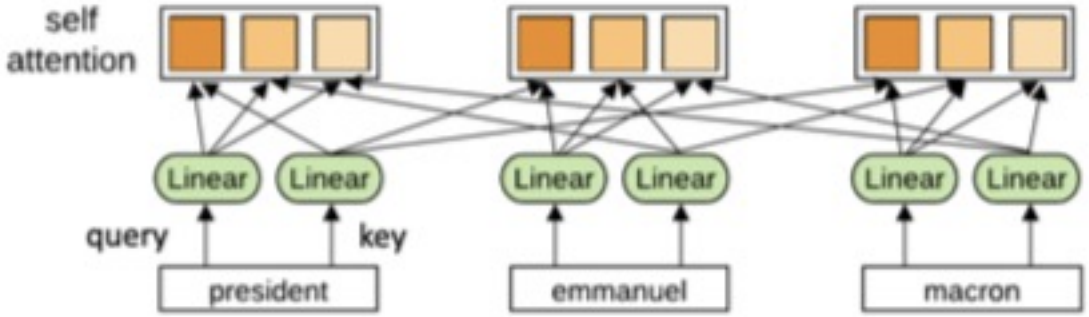
Presenter: Baosong Yang

Motivation

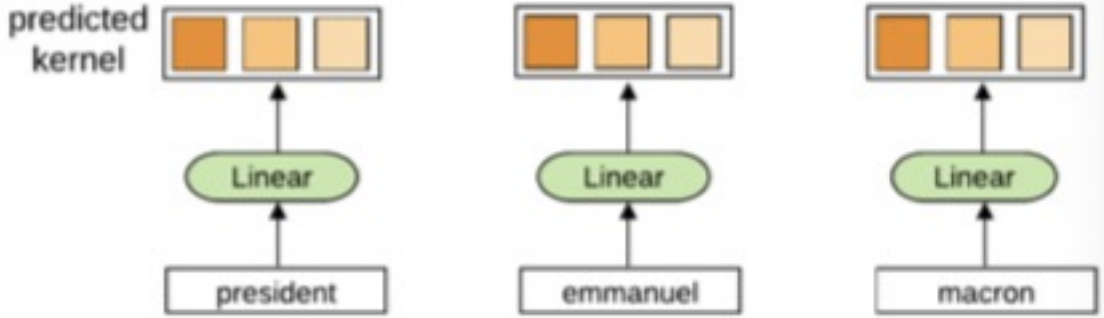
- SAN
- More Parameter
- Complexity
- (Is it necessary to build word dependencies?)

Model

- **Lightweight Convolution:** Reuse the same weights for context elements, regardless of the current time-step
- **Dynamic Convolution:** build on lightweight convolutions by predicting a different convolution kernel at every time-step.



(a) Self-attention



(b) Dynamic convolution

Lightweight

$$O_{i,c} = \text{DepthwiseConv}(X, W_{c,:}, i, c) = \sum_{j=1}^{k} W_{c,j} \cdot X_{(i+j-\lceil \frac{k+1}{2} \rceil),c}$$

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

- Channel wise
- Share weights => Multi-head
- Regardless of context elements

Dynamic Convolution

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

$$\sum_{c=1}^{d} W^{Q}_{h,j,c} X_{i,c}$$

- Changes the weights assigned to current element
- Do not depends on entire context
- Similar to source2token self-attention

Experiments

- 7 layers; K=3,5,15,31*4; H=16

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

Experiments

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**

Conclusion

- Local scope is benefit for feature extraction
- Maybe it is unnecessary to model dependencies between Q and K
- Maybe a new widely applied model: faster, simplistic and effectiveness