Stand-Alone Self-Attention in Vision Models

Prajit Ramachandran, NikiParmar, Ashish Vaswani,
Irwan Bello, Anselm Levskaya, Jonathon Shlens

NeurIPS 2019

Presented Sanghyeon Lee

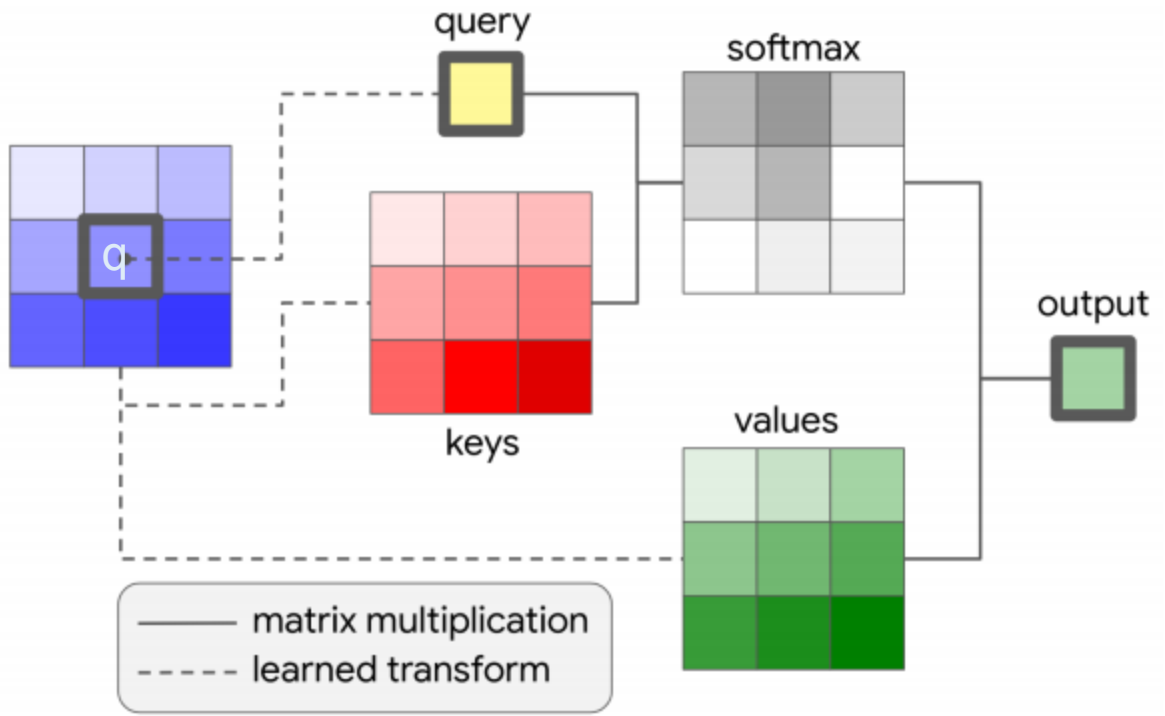
Motivation

- CNN has poor scaling properties with respect to large receptive fields
- In order to capture “long-range Dependencies”, recent approaches have argued for going beyond CNN
- Self - Attention is applied to a single context and can get more long-distance interaction
- Self-attention can be a stand-alone primitive for vision models instead of CNN?

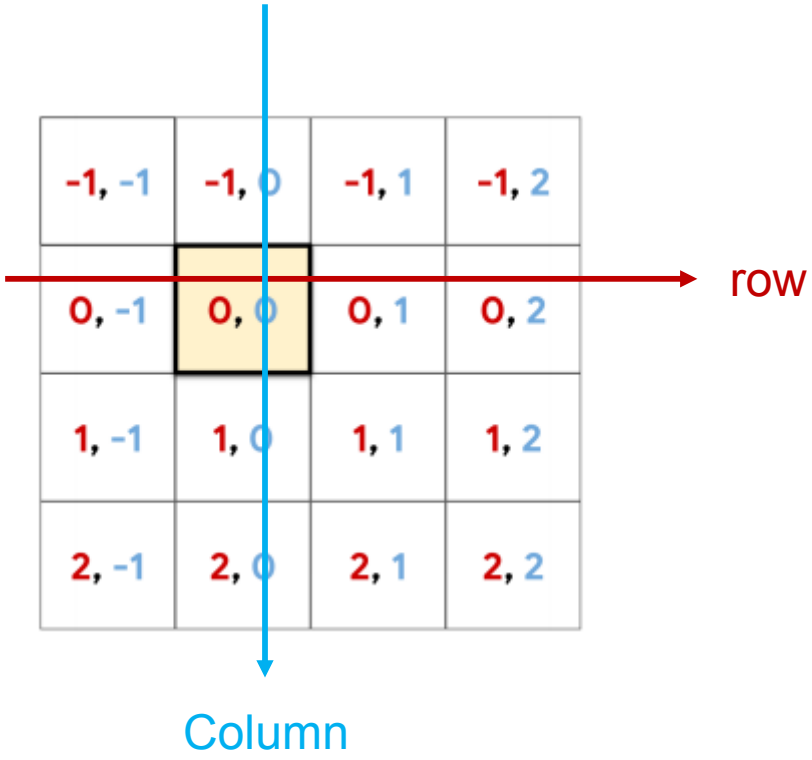
Self Attention Vision Model

Self Attention Vision Model

Self Attention Vision Model



Relative distance



Self Attention Vision Model

Self Attention Vision Model

- $\mathrm{At}tention\ scores$

$$A \coloneqq \mathrm{X_{ij}W_{query}W_{key}}^{\top}\mathrm{X_{ab}^{\top}}, \mathrm{X_{ij}} \in R^{D_{In}}, \mathrm{X}_{ab} \in R^{K*K\times D_{In}}, \mathrm{W}_{*} \in R^{D_{in}\times D_{out}}$$

K: filter size, D_{*}: channel, $\mathrm{X_{ij}}$: query pixel, X_{ab}: key pixel

- $\mathrm{At}tention\ scores\ (with\ relative\ positional\ encoding)$

$A \coloneqq \mathrm{X_{ij}W_{query}W_{key}}^{\top}\mathrm{X_{ab}^{\top}}, \mathrm{X_{ij}} + \mathrm{X_{ij}W_{query}r_{a\text{-}i,b\text{-}j}}$; r : embedded row and column offsets

- Self Attention output

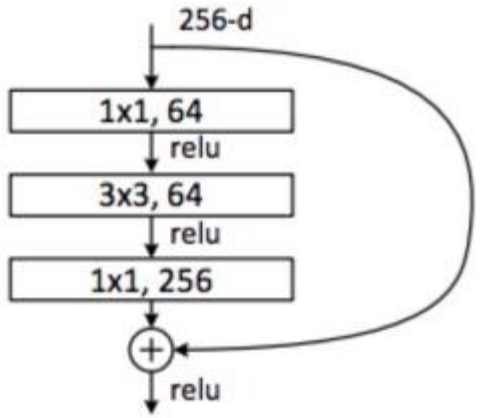
$\mathrm{y_{ij}} = \sum_{ab} softmax(A_{q_{ij}:})_{ab}\, X_{ab,:}W_{value}$; $\mathrm{x_{ij}}$ and $\mathrm{y_{ij}}$ are 1: 1

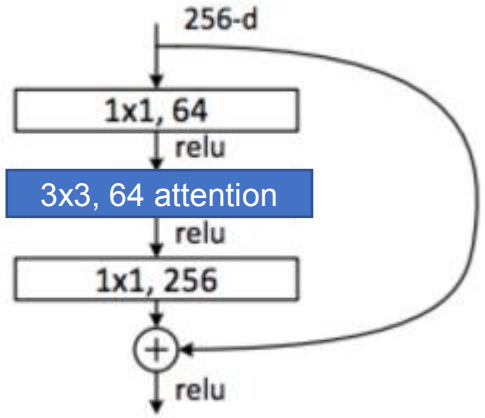
How to construct a fully attentional architecture?

Step1. Fully Attentional Vision Models

1. Take an existing convolutional architecture
2. Replace every instance of a spatial convolution with an attention layer

* Spatial Down Sampling : Using 2x2 average polling with stride 2





How to construct a fully attentional architecture?

Step2. Replacing the Convolutional Stem

Stem: The initial layers of CNN
- It is important because of learning local features such as edges

Ex) Resnet: 7 x 7 convolution with stride 2 followed by 3 x 3 max pooling with stride 2

But at the stem layer, each RGB pixels are heavily spatially correlated and uninformative
➔It makes learning difficult for content-based mechanisms such as self-attention

Bridge the gap between

Inject distance based information in the pointwise 1 X 1 convolution
- $\mathrm{Attention\ scores}$

$$Value = \sum_m X_{ab} p(a,b,m)\, W^m_{value}, \mathrm{W}^m_* \in R^{D_{in} \times D_{out}/m}$$

$$p(a,b,m) = softmax_m\left(\left(emb_{row}(a) + emb_{col}(b)\right)^T \gamma^m\right)$$

γ^m: mixture embedding between a and b

Experiments

Image net classification

	ResNet-26			ResNet-38			ResNet-50		
	FLOPS (B)	**Params** (M)	**Acc.** (%)	**FLOPS** (B)	**Params** (M)	**Acc.** (%)	**FLOPS** (B)	**Params** (M)	**Acc.** (%)
Baseline	4.7	13.7	74.5	6.5	19.6	76.2	8.2	25.6	76.9
Conv-stem + Attention	4.5	10.3	**75.8**	5.7	14.1	**77.1**	7.0	18.0	**77.4**
Full Attention	4.7	10.3	74.8	6.0	14.1	76.9	7.2	18.0	**77.6**

Table 1: ImageNet classification results for a ResNet network with different depths. *Baseline* is a standard ResNet, *Conv-stem + Attention* uses spatial convolution in the stem and attention everywhere else, and *Full Attention* uses attention everywhere including the stem. The attention models outperform the baseline across all depths while having 12% fewer FLOPS and 29% fewer parameters.

Experiments

Image net classification

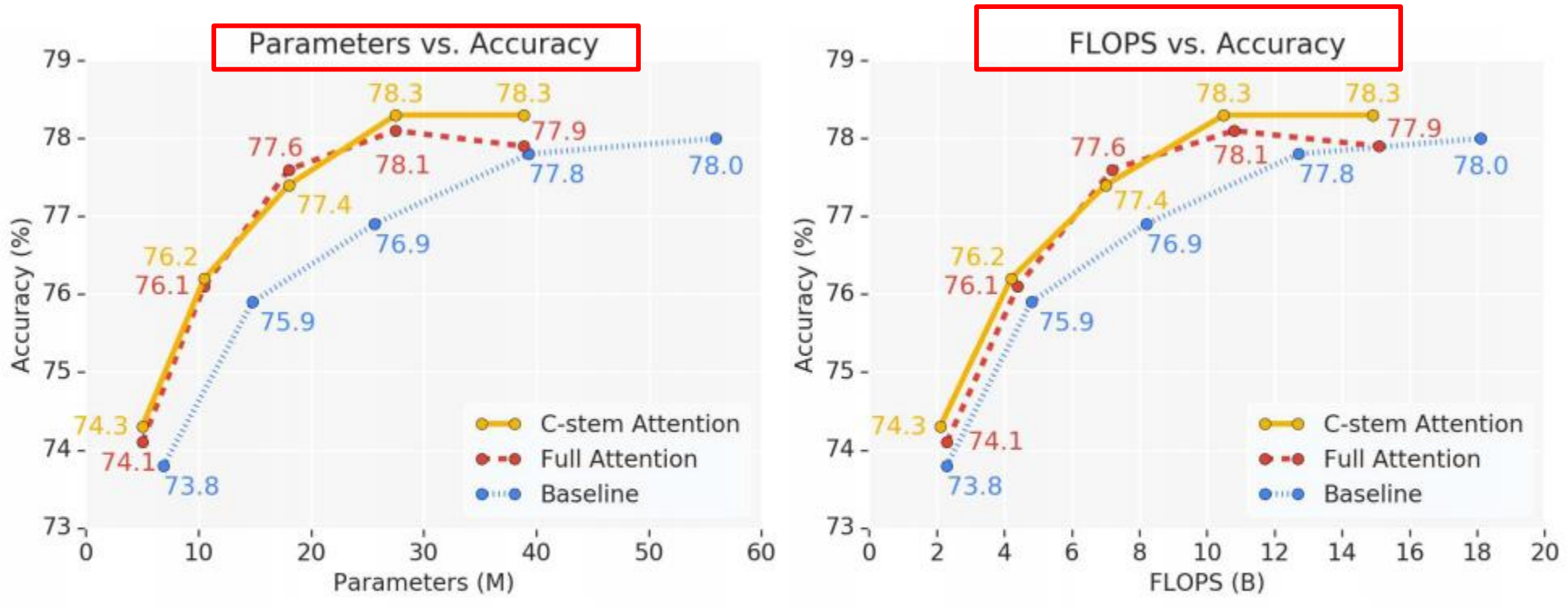


Figure 5: Comparing parameters and FLOPS against accuracy on ImageNet classification across a range of network widths for ResNet-50. Attention models have fewer parameters and FLOPS while improving upon the accuracy of the baseline.

Experiments

Where is stand-alone attention most useful?

Full Network **Stem**

Conv Groups	Attention Groups	FLOPS (B)	Params (M)	Top-1 Acc. (%)
-	1, 2, 3, 4	7.0	18.0	80.2
1	2, 3, 4	7.3	18.1	**80.7**
1, 2	3, 4	7.5	18.5	**80.7**
1, 2, 3	4	8.0	20.8	80.2
1, 2, 3, 4	-	8.2	25.6	79.5
2, 3, 4	1	7.9	25.5	79.7
3, 4	1, 2	7.8	25.0	79.6
4	1, 2, 3	7.2	22.7	79.9

Table 3: Modifying which layer groups use which primitive. Accuracies computed on validation set. The best performing models use convolutions for early groups and attention for later groups.

Experiments

Which components are important in attention?

1. Spatial extent

Spatial Extent ($k \times k$)	FLOPS (B)	Top-1 Acc. (%)
3×3	6.6	76.4
5×5	6.7	77.2
7×7	7.0	77.4
9×9	7.3	77.7
11×11	7.7	77.6

Table 4: Varying the spatial extent k. Parameter count is constant across all variations. Small k perform poorly, but the improvements of larger k plateaus off.

2. Positional Encoding Type

Positional Encoding Type	FLOPS (B)	Params (M)	Top-1 Acc. (%)
none	6.9	18.0	77.6
absolute	6.9	18.0	78.2
relative	7.0	18.0	80.2

Table 5: The effect of changing the positional encoding type for attention. Accuracies computed on the validation set. Relative encodings significantly outperform other strategies.

3. Attention Type

Attention Type	FLOPS (B)	Params (M)	Top-1 Acc. (%)
$q^\top r$	6.1	16.7	76.9
$q^\top k + q^\top r$	7.0	18.0	77.4

Table 6: The effect of removing the $q^\top k$ interactions in attention. Using just $q^\top r$ interactions only drops accuracy by 0.5%.

4. Attention Stem Type

Attention Stem Type	FLOPS (B)	Top-1 Acc. (%)
stand-alone	7.1	76.2
spatial convolution for values	7.4	77.2
spatially aware values	7.2	**77.6**

Table 7: Ablating the form of the attention stem. Spatially-aware value attention outperforms both stand-alone attention and values generated by a spatial convolution.

Thank you