Stand-Alone Self-Attention in Vision Models

Prajit Ramachandran[*] **Niki Parmar**[*] **Ashish Vaswani**[*]

Irwan Bello **Anselm Levskaya**[†] **Jonathon Shlens**

Google Research, Brain Team
{prajit, nikip, avaswani}@google.com

Attention is
all you need!

Aim

Developing a fully self-attention model for object detection and recognition using local stand-alone self attention blocks.

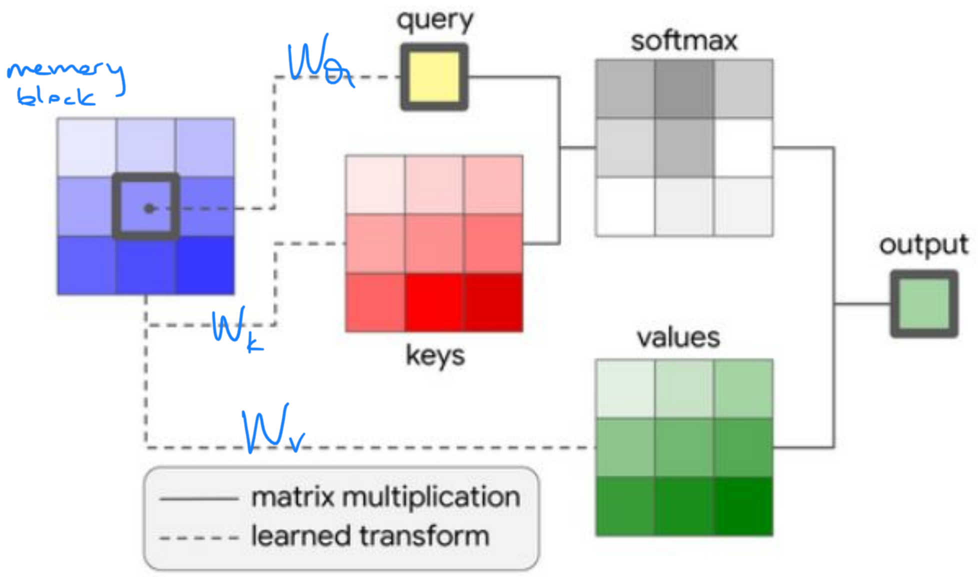
have successful examples in language modelling and generation.

- Fewer parameters? (yes)
- Faster runtime? (yes)
- Better accuracy? (yes)

under certain circumstances

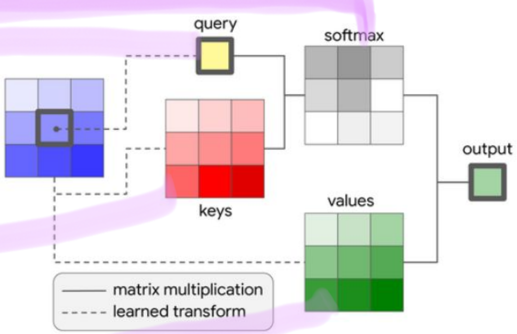
Contribution: Spatial Self Attention Block

A stand-alone self-attention layer that can be used to replace spatial convolutions and building a fully attentional model.

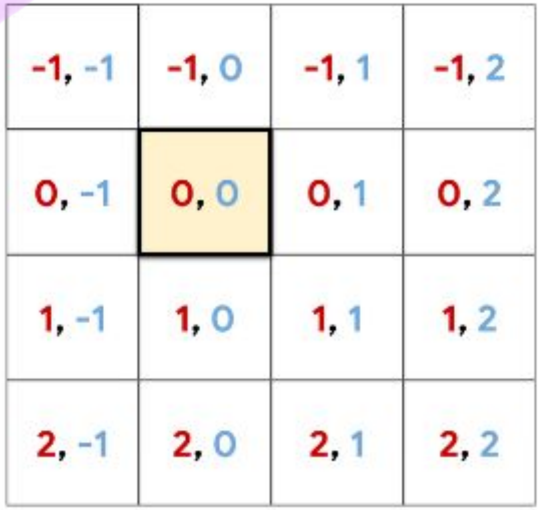


Relative Distance Computation

Using the relative distance to query pixel, spatial-relative attention is:



$$out = \sum softmax\left(q \times keys + q \times \underset{pos.}{relative} \right) \times values$$



-1, -1	-1, 0	-1, 1	-1, 2
0, -1	0, 0	0, 1	0, 2
1, -1	1, 0	1, 1	1, 2
2, -1	2, 0	2, 1	2, 2

Fully Attentional Vision Models

ResNet: – core mechanism:

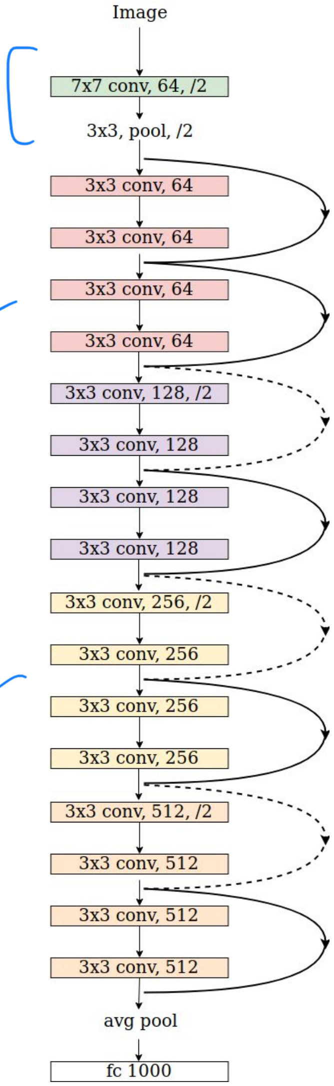
1x1 conv → downsample

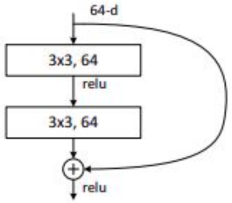
3x3 spatial conv

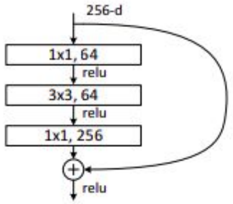
1x1 conv → upsample

convolutional stem

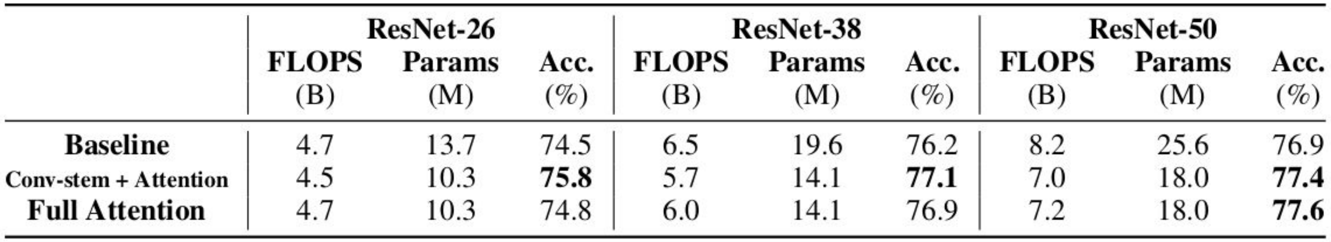
changing to attention decreases performance significantly.



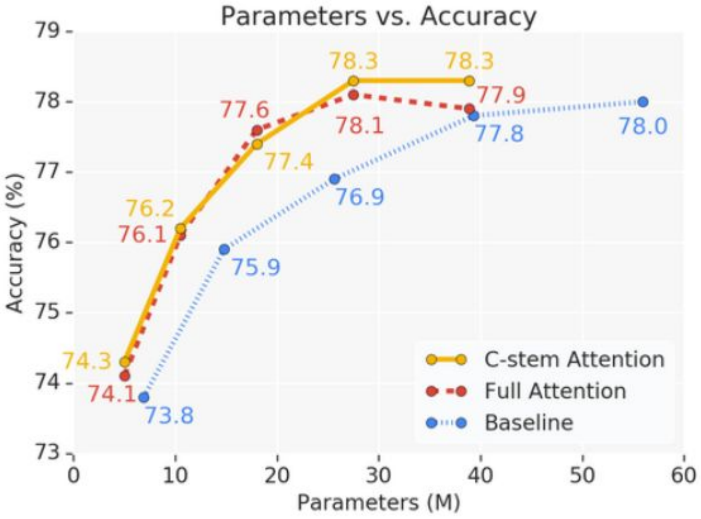


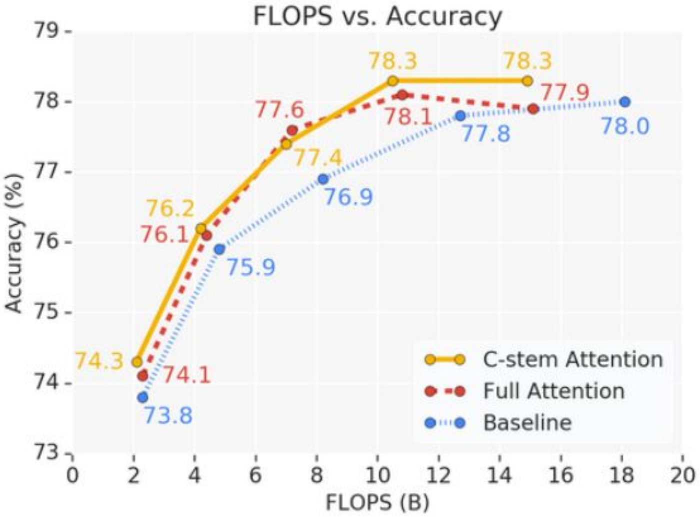


Experiments: ResNet on ImageNet



	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**





Which components are important in attention?

All these info is based on models with convolutional stem

- Spatial extent, k is 11 (improvement 3 -------> 11)
- Relative position encodings perform 2% better than absolute encodings.
-