Rensselaer
why not change the world?®

Rensselaer
Radiation Measurement & Dosimetry Group

Attention Augmented Convolutional Networks

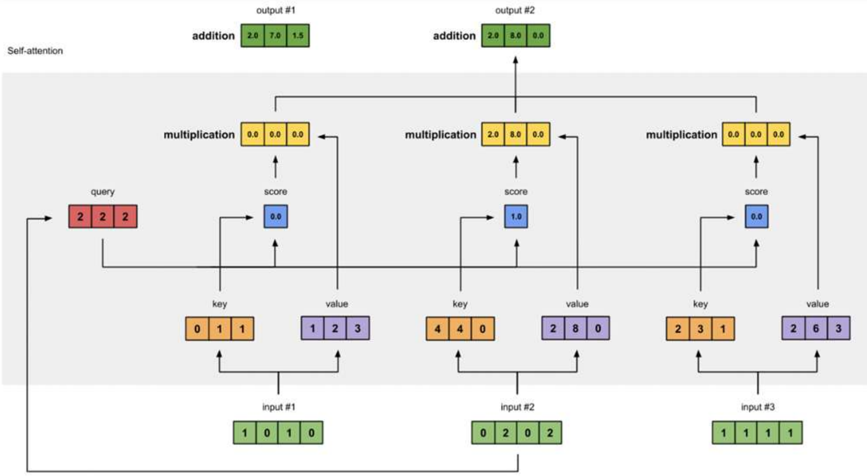
Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, Quoc V. Le
Google Brain

Presented by QIYUN CHENG | 02/19/2020

CNN and Self-Attention

CNN	Self-Attention
1. Widely used in Computer Vision Applications (Image classification) 2. Versatility and mature strategies for designing architectures 3. Lack of global information (long range interactions) due to the nature of the convolution kernel	1. Mostly been applied to sequence modeling and generative modeling tasks 2. Weights are produced dynamically (Ability to capture long range interactions without increasing the number of parameters) 3. Stand-alone computational primitive for image classification

Self-Attention Network



$$O_h = \text{Softmax}\left(\frac{(XW_q)(XW_k)^T}{\sqrt{d_k^h}}\right)(XW_v)$$

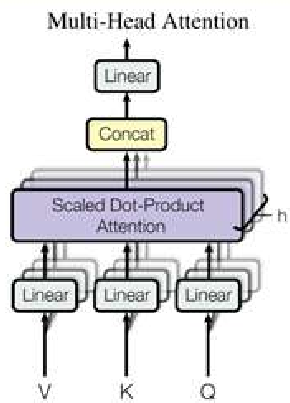
where $W_q, W_k \in \mathbb{R}^{F_{in} \times d_k^h}$ and $W_v \in \mathbb{R}^{F_{in} \times d_v^h}$ are learned linear transformations that map the input X to queries $Q = XW_q$, keys $K = XW_k$ and values $V = XW_v$.

Multi-Head Attention

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

- Multi-head attention allows the model to jointly attend to **information from different representation subspaces** at different positions.



$$\text{MHA}(X) = \text{Concat}\left[O_1, \ldots, O_{Nh}\right]W^O$$

where $W^O \in \mathbb{R}^{d_v \times d_v}$ is a learned linear transformation.

Two-dimensional Positional Encodings

Two requirements:

~~1. Permutation Equivariance~~

2. Translation Equivariance

$$\text{MHA}(\pi(X)) = \pi(\text{MHA}(X))$$

The attention logit for how much pixel i attends to pixel j:

$$l_{i,j} = \frac{q_i^T}{\sqrt{d_k^h}}(k_j + r^W_{j_x - i_x} + r^H_{j_y - i_y})$$

The output of head h becomes:

$$O_h = \text{Softmax}\left(\frac{QK^T + S_H^{rel} + S_W^{rel}}{\sqrt{d_k^h}}\right) V$$

Attention Augmented Convolution

$$\text{AAConv}(X) = \text{Concat}\Big[\text{Conv}(X), \text{MHA}(X)\Big]$$

1. Equivariant to translation
2. Readily operate on inputs of different spatial dimensions

Attention-Augmented Convolution Network Structure

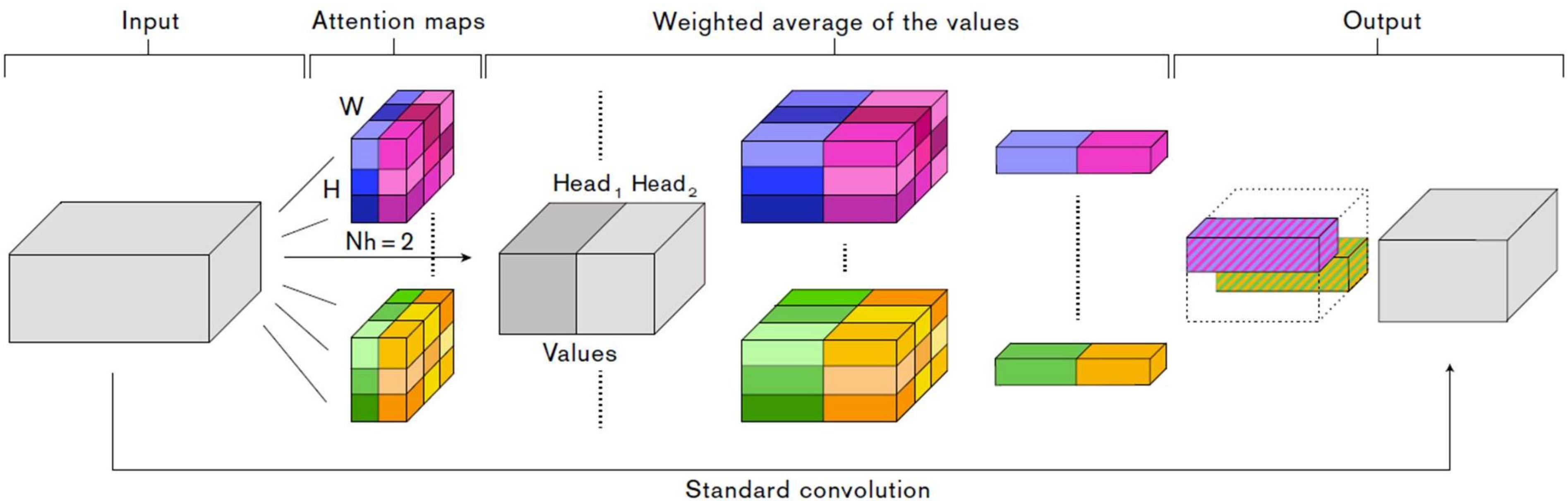


Figure 2. **Attention-augmented convolution**: For each spatial location (h, w), N_h attention maps over the image are computed from queries and keys. These attention maps are used to compute N_h weighted averages of the values V. The results are then concatenated, reshaped to match the original volume's spatial dimensions and mixed with a pointwise convolution. Multi-head attention is applied in parallel to a standard convolution operation and the outputs are concatenated.

Efficiency

Change of Parameters: $\Delta_{params} \sim F_{in}F_{out}(2\kappa + (1-k^2)\upsilon + \frac{F_{out}}{F_{in}}\upsilon^2)$

$\kappa = \frac{d_k}{F_{out}}$

$\upsilon = \frac{d_v}{F_{out}}$

A slight decrease in parameters when replacing 3x3 convolutions and a slight increase in parameters when replacing 1x1 convolutions.

Memory Cost: $O((N_h(HW)^2)$

we augment convolutions with attention starting from the last layer (with smallest spatial dimension) until we hit memory constraints. To reduce the memory footprint of augmented networks, we typically resort to a smaller batch size and sometimes additionally down sample the inputs to self-attention in the layers with the largest spatial dimensions where it is applied.

Ablation Study

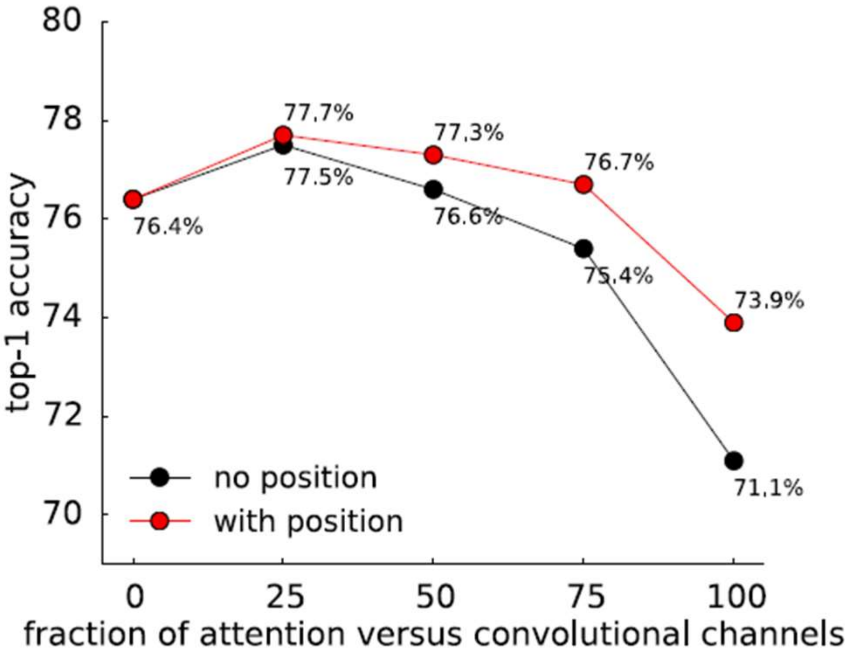
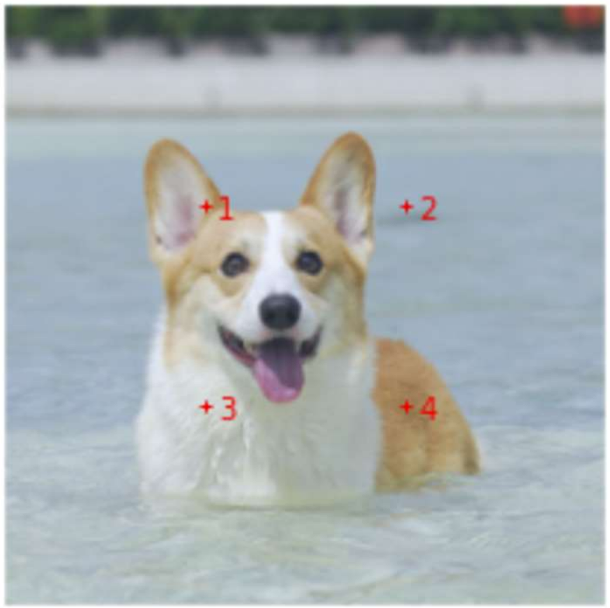


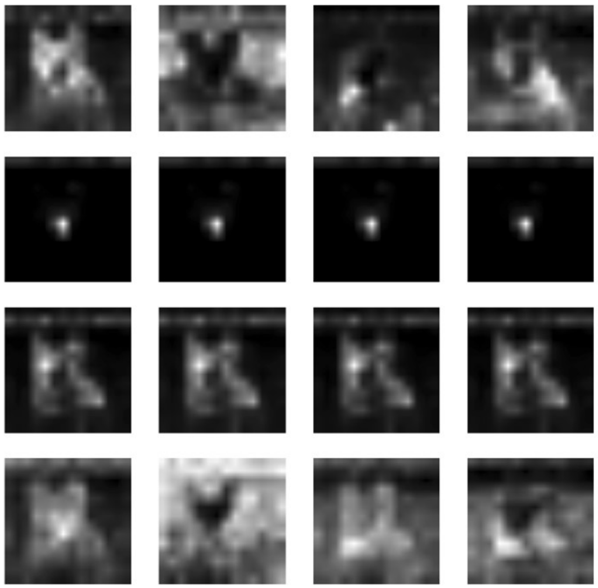
Figure 4. Effect of relative position embeddings as the ratio of attentional channels increases on our Attention-Augmented ResNet50.

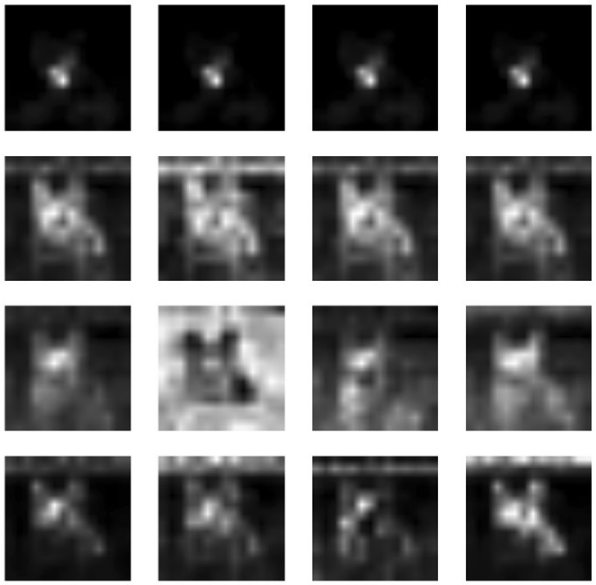
Architecture	GFlops	Params	top-1	top-5
ResNet-34 [14]	7.4	21.8M	73.6	91.5
ResNet-50 [14]	8.2	25.6M	76.4	93.1
$\kappa = \upsilon = 0.25$	7.9	24.3M	77.7	93.8
$\kappa = \upsilon = 0.5$	7.3	22.3M	77.3	93.6
$\kappa = \upsilon = 0.75$	6.8	20.7M	76.7	93.2
$\kappa = \upsilon = 1.0$	6.3	19.4M	73.9	91.5

Table 6. Attention Augmented ResNet-50 with varying ratios of attentional channels.

Visualization of attention maps







Rensselaer

why not change the world?®