On the Relationship between Self-Attention and Convolutional Layers

Jean-Baptiste Cordonnier, Andreas Loukas, Martin Jaggi
Ecole Polytechnique Federale de Lausanne (EPFL)

ICLR 2020

Presented by Eungyeup Kim

Vision Seminar
03 MAR 2020

Motivation

Attention mechanisms in vision is working well recently.

- *Bello et al.(2019)* replaced some convolutional layers with self-attention(SA) layers, leading to improvements on image classification and object detection tasks.

- *Ramachandran et al.(2019)* noticed that SOTA results are reached when attention and convolutional features are combined, or even when **only** SA only architecture is used.

- So, do SA layers process images in a similar manner to convolutional layers?

This work

1) Provides the constructive proof showing that multi-head SA layers with relative positional encoding can express any convolutional layers.

2) Provides experiments demonstrating the layers of attention-only architectures act like a convolutional layers, such as attending on grid-like pattern around each query pixel.

Backgrounds

Multi-head self-attention layer (MHSA)

- $\mathrm{Attention\ scores}$

$$A \coloneqq \mathrm{X}\mathrm{W}_{\mathrm{query}}\mathrm{W}_{\mathrm{key}}{}^{\top}\mathrm{X}^{\top}$$

- $\mathrm{Attention\ scores\ (with\ positional\ encoding)}$

$A \coloneqq (\mathrm{X}+\mathrm{P})\mathrm{W}_{\mathrm{query}}\mathrm{W}_{\mathrm{key}}{}^{\top}(\mathrm{X}+\mathrm{P})^{\top}$, where P indicates poisitional encoding

- $\mathrm{Self}-\mathrm{Attention}\ (X)_{\mathrm{q},:} = \sum_k softmax(A_{q,:})_k\, X_{k,:} W_{value}$, where $q = (i, j)$ indicating 2D coordinates.

- $\mathrm{MHSA}(X) \coloneqq \mathrm{concat}_{\mathrm{h}\in[N_h]}[\mathrm{Self}-\mathrm{Attention}_{\mathrm{h}}(\mathrm{X})]\ \mathrm{W}_{\mathrm{out}} + \mathrm{b}_{\mathrm{out}}$, where $\mathrm{W}_{\mathrm{out}} \in R^{N_h D_h \times D_{out}}$

Convolutional layer

- $Conv(X)_{i,j,:} \coloneqq \sum_{(\delta_1,\,\delta_2)\in\triangle_K} X_{i+\delta_1, j+\delta_2,:}\, W_{\delta_1,\delta_2,:,:} + b$, where $\triangle_K \coloneqq \left[-\left[\frac{K}{2}\right], \ldots, \left[\frac{K}{2}\right]\right] \times \left[-\left[\frac{K}{2}\right], \ldots, \left[\frac{K}{2}\right]\right]$ containing all possible shifts with a $K \times K$ kernel.

Backgrounds

Positional Encoding

- Absolute encoding : a vector $P_{p,:}$ is assigned to each pixel p.

$$A_{q,k}^{abs} \coloneqq (X_{q,:} + P_{q,:})^{\top} \mathrm{W}_{\mathrm{query}}{}^{\top} W_{key} (X_{k,:} + P_{k,:})$$

$$= X_{q,:}{}^{\top} \mathrm{W}_{\mathrm{query}}{}^{\top} W_{key} X_{k,:} + X_{q,:}{}^{\top} \mathrm{W}_{\mathrm{query}}{}^{\top} W_{key} P_{k,:} + \mathrm{P}_{\mathrm{q},:}{}^{\top} \mathrm{W}_{\mathrm{query}}{}^{\top} \mathrm{W}_{\mathrm{key}} X_{k,:} + \mathrm{P}_{\mathrm{q},:}{}^{\top} \mathrm{W}_{\mathrm{query}}{}^{\top} \mathrm{W}_{\mathrm{key}}{}^{\top} P_{k,:}$$

- Relative encoding : only consider the position difference between pixels. (Dai et al.)

$$A_{q,k}^{rel} = X_{q,:}{}^{\top} \mathrm{W}_{\mathrm{query}}{}^{\top} \mathrm{W}_{\mathrm{key}} X_{k,:} + X_{q,:}{}^{\top} \mathrm{W}_{\mathrm{query}}{}^{\top} \widehat{W}_{\mathrm{key}} r_{\delta} + u^{\top} \mathrm{W}_{\mathrm{key}} X_{k,:} + v^{\top} \widehat{W}_{\mathrm{key}} r_{\delta}$$

where the attention scores only depend on the shift $\delta \coloneqq k - q$.

- Trainable u and v are unique for each head, while non-trainable r_{δ} is shared by all layers and heads.
- We deliberately separate the $\mathrm{W}_{\mathrm{key}}$ and $\widehat{W}_{\mathrm{key}}$ for producing content-based key vectors and location-based key vectors respectively.
- Content-based addressing, content-dependent positional bias, global content bias, global positional bias

https://medium.com/@_init_/how-self-attention-with-relative-position-representations-works-28173b8c245a

Self-Attention vs Convolutional Layers

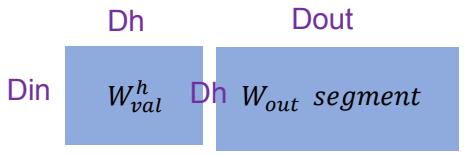
Theorem 1. A multi-head self-attention layer with N_h heads of dimension D_h, output dimension D_{out} and a relative positional encoding of dimension $D_p \geq 3$ can express any convolutional layer of kernel size $\sqrt{N_h} \times \sqrt{N_h}$ and min(D_h, D_{out}) output channels.

Lemma 1. Consider a multi-head self-attention layer consisting of $N_h = \mathrm{K}^2$ heads, $D_h \geq D_{out}$ and let $f: [N_h] \rightarrow \triangle_K$ be a bijective mapping of heads onto shifts. Further, suppose that for every head the following holds:

$$softmax\left(A^{(h)}_{q,:}\right)_k = \begin{cases} 1 & if\ f(h) = q - k \\ 0 & o/w \end{cases}$$

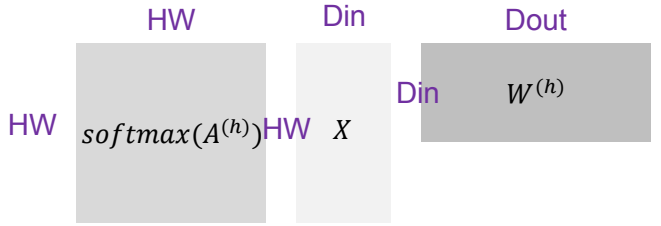
Then, for any convolutional layer with a $\mathrm{K} \times \mathrm{K}$ kernel and D_{out} output channels, there exists $\left\{W^{(h)}_{val}\right\}_{\mathrm{h} \in [N_h]}$ such that $MHSA(X) = Conv(X)$ for every $X \in \mathrm{R}^{\mathrm{W} \times \mathrm{H} \times D_{in}}$.

Self-Attention vs Convolutional Layers



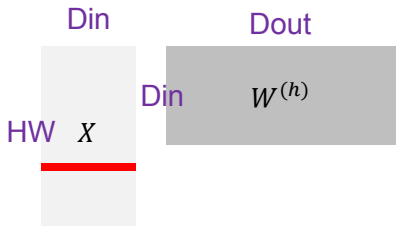
proof

$$MHSA(X) = b_{out} + \sum_{h\in[N_h]} softmax(A^{(h)})\, X\, W_{val}^{h}\, W_{out}[(h-1)D_h + 1 : hD_h + 1]$$



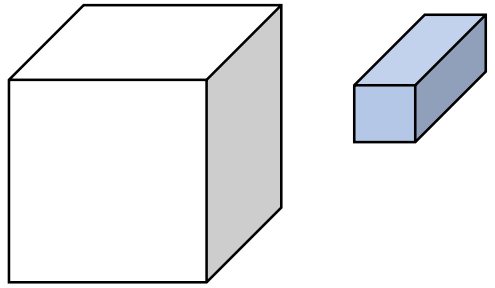
$$MHSA(X)_{q,:} = \sum_{h\in[N_h]}(\sum_{K} softmax\left(A_{q,:}^{(h)}\right)_{k} X_{k,:})W^{(h)} + b_{out}$$

Due to the conditions of Lemma 1, for the h^{th} attention head the attention probability is one when $k = q - f(h)$ and zero otherwise. The layer's output at pixel q is thus equal to



$$MHSA(X)_q = \sum_{h\in[N_h]} X_{q-f(h),:} W^{(h)} + b_{out}$$

- For $K = \sqrt{N_h}$, the above can be seen to be equivalent to a convolutional layer.
- There is a one to one mapping (implied by f) between the matrices $W^{(h)}$ and the matrices $W_{k1,k2,:,:}$ for all $(k1, k2) \in [K]^2$



Self-Attention vs Convolutional Layers

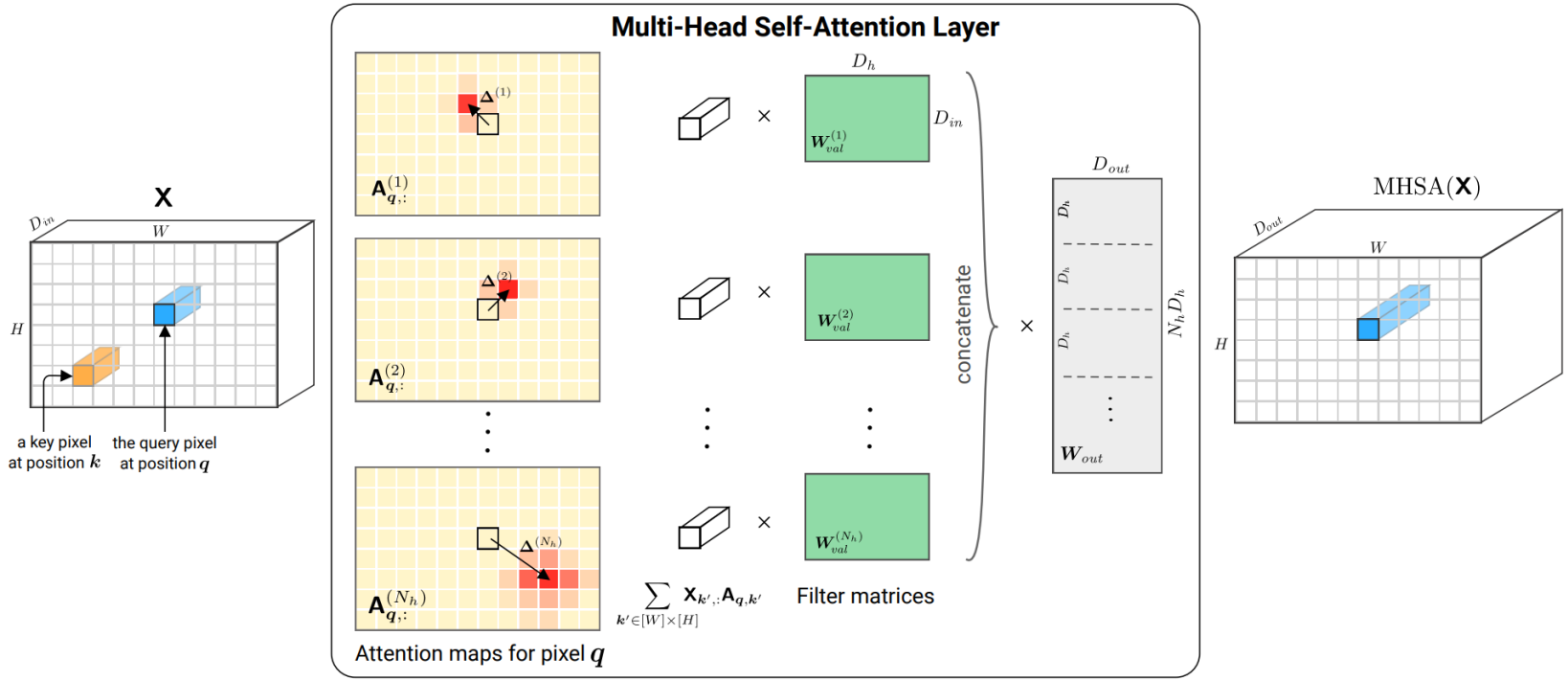


Figure 1: Illustration of a Multi-Head Self-Attention layer applied to a tensor image $\mathbf{X}$. Each head h attends pixel values around shift $\boldsymbol{\Delta}^{(h)}$ and learn a filter matrix $\boldsymbol{W}_{val}^{(h)}$. We show attention maps computed for a query pixel at position $\boldsymbol{q}$.

Self-Attention vs Convolutional Layers

On which condition can $softmax\left(A_{q,:}^{(h)}\right)_k = \begin{cases} 1 & if\ f(h) = q - k \\ 0 & o/w \end{cases}$ **be derived?**

Lemma 2. There exists a relative encoding scheme $\{r_\delta \in R^{Dp}\}_{\delta \in Z^2}$ with $D_p \geq 3$ and parameter $\mathrm{W}_{\mathrm{query}}$, $\mathrm{W}_{\mathrm{key}}$, u with $D_p \leq D_k$ such that, for every $\triangle \in \triangle_K$ there exists some vector v yielding $softmax\left(A_{q,:}\right)_k = 1$ if $k - q = \triangle$ and zero, otherwise.

proof

$$A_{q,k} = {X_{q,:}}^\top {\mathrm{W}_{\mathrm{query}}}^\top \mathrm{W}_{\mathrm{key}} X_{k,:} + {X_{q,:}}^\top {\mathrm{W}_{\mathrm{query}}}^\top \widehat{W}_{\mathrm{key}} r_\delta + u^\top \mathrm{W}_{\mathrm{key}} X_{k,:} + v^\top \widehat{W}_{\mathrm{key}} r_\delta$$

As the attention probabilities are independent of the input tensor X, we set $\mathrm{W}_{\mathrm{query}} = \mathrm{W}_{\mathrm{key}} = 0$. Thus

$$A_{q,k} = v^\top \widehat{W}_{\mathrm{key}} r_\delta = v^\top r_\delta \ (\widehat{W}_{\mathrm{key}}\ is\ set\ to\ be\ identity)$$

where $r_\delta \coloneqq k - q$

Suppose we could write

$$A_{q,k} = -\alpha(\|\delta - \triangle\|^2 + c)$$

for some constant c.

Self-Attention vs Convolutional Layers

In this way, we have

$$\lim_{\alpha\to\infty} softmax\left(A_{q,:}\right)_k = \lim_{\alpha\to\infty} \frac{e^{-\alpha\left(\|\delta-\triangle\|^2+c\right)}}{\sum_{k'} e^{-\alpha(\|(k-q)-\triangle\|^2+c)}} = \lim_{\alpha\to\infty} \frac{e^{-\alpha\left(\|\delta-\triangle\|^2\right)}}{\sum_{k'} e^{-\alpha(\|(k-q)-\triangle\|^2)}}$$

For $\delta = \triangle$,

$$\lim_{\alpha\to\infty} softmax\left(A_{q,:}\right)_k = \lim_{\alpha\to\infty} \frac{e^{-\alpha\left(\|\delta-\triangle\|^2\right)}}{\sum_{k'} e^{-\alpha(\|(k-q)-\triangle\|^2)}} = \frac{1}{1+\lim_{\alpha\to\infty}\sum_{k'\neq k} e^{-\alpha(\|(k-q)-\triangle\|^2)}} = 1$$

For $\delta \neq \triangle$,

$$\lim_{\alpha\to\infty} softmax\left(A_{q,:}\right)_k = \lim_{\alpha\to\infty} \frac{e^{-\alpha\left(\|\delta-\triangle\|^2\right)}}{\sum_{k'} e^{-\alpha(\|(k-q)-\triangle\|^2)}} = 0$$

So, which v and $\{r_\delta\}_{\delta\in Z^2}$ satisfy $A_{q,k} = -\alpha(\|\delta - \triangle\|^2 + c)$?

If we set

$$v = -\alpha(1, -\triangle_1, -\triangle_2), \qquad r_\delta = (\|\delta\|, \delta_1, \delta_2)$$

Then, $A_{q,k} = v^\top r_\delta = -\alpha(\|\delta - \triangle\|^2 + c)$

Self-Attention vs Convolutional Layers

Quadratic Encoding

$$v = -\alpha(1, -\triangle_1, -\triangle_2), \qquad r_\delta = (\|\delta\|, \delta_1, \delta_2)$$

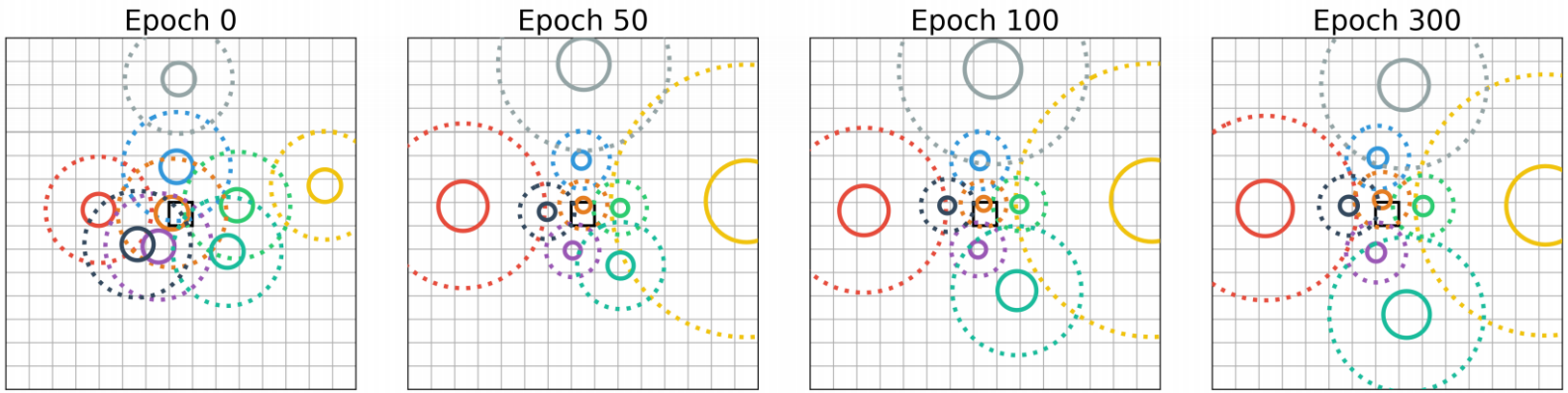
It is important that above encoding is not the only one for which the conditions of Lemma 1 are satisfied. The relative encoding learned by the neural network also matched the conditions of the lemma.

Nevertheless, the encoding defined above is very efficient in terms of size, as only $D_p = 3$ dimensions suffice to encode the relative position of pixels, while also reaching similar or better empirical performance than the learned one.

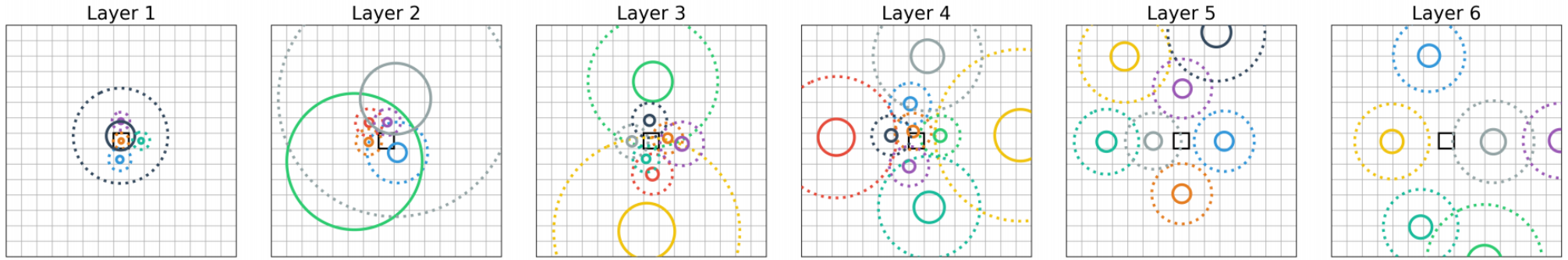
Experiments

A fully attentional model consisting of six multi-head self-attention layers is trained on Cifar-100 and ImageNet for image classification.

Using quadratic encoding : $\triangle^{h} \sim \mathrm{N}(\mathrm{mu}, \mathrm{sigma})$



[This figure shows how the initial positions of the heads at **layer 4** changed during training.]

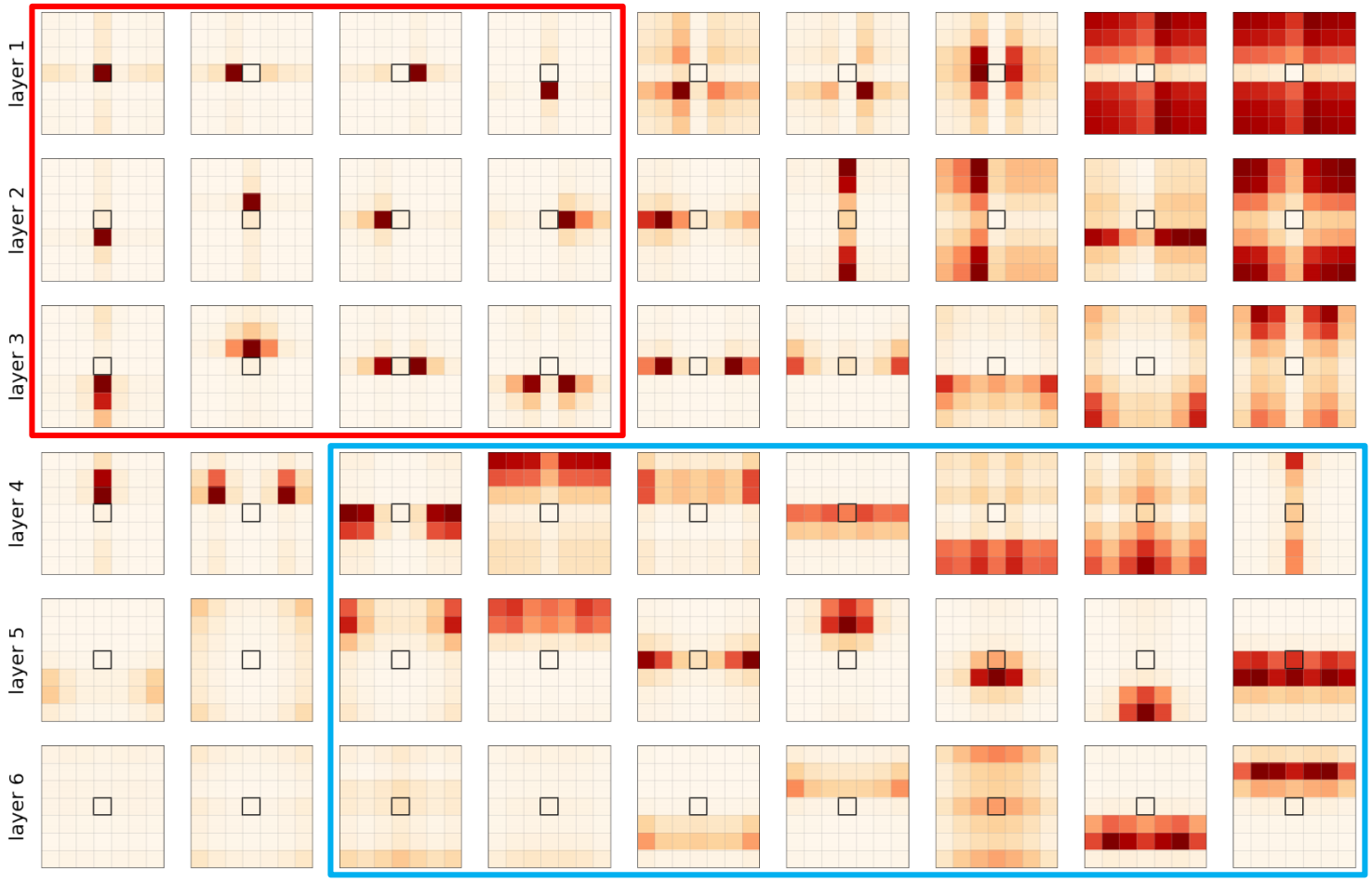


[This figure shows that the first few layers tend to focus on local patterns, while deeper layers also attend to larger patterns by positioning the center of attention further from the query pixel.]

Experiments

Using learned encoding(without content-based attention) :

Satisfying lemma 1, thus theorem as well. (local pattern)



Capturing the horizontally-symmetric but not localized patterns. (long-rage pixel inter-dependency)

Figure 5: Attention probabilities of each head (*column*) at each layer (*row*) using learned relative positional encoding without content-based attention. The central black square is the query pixel. We reordered the heads for visualization and zoomed on the 7x7 pixels around the query pixel.

Experiments

Using learned encoding(with content-based attention) :

Satisfying lemma 1, thus theorem as well. (local pattern)

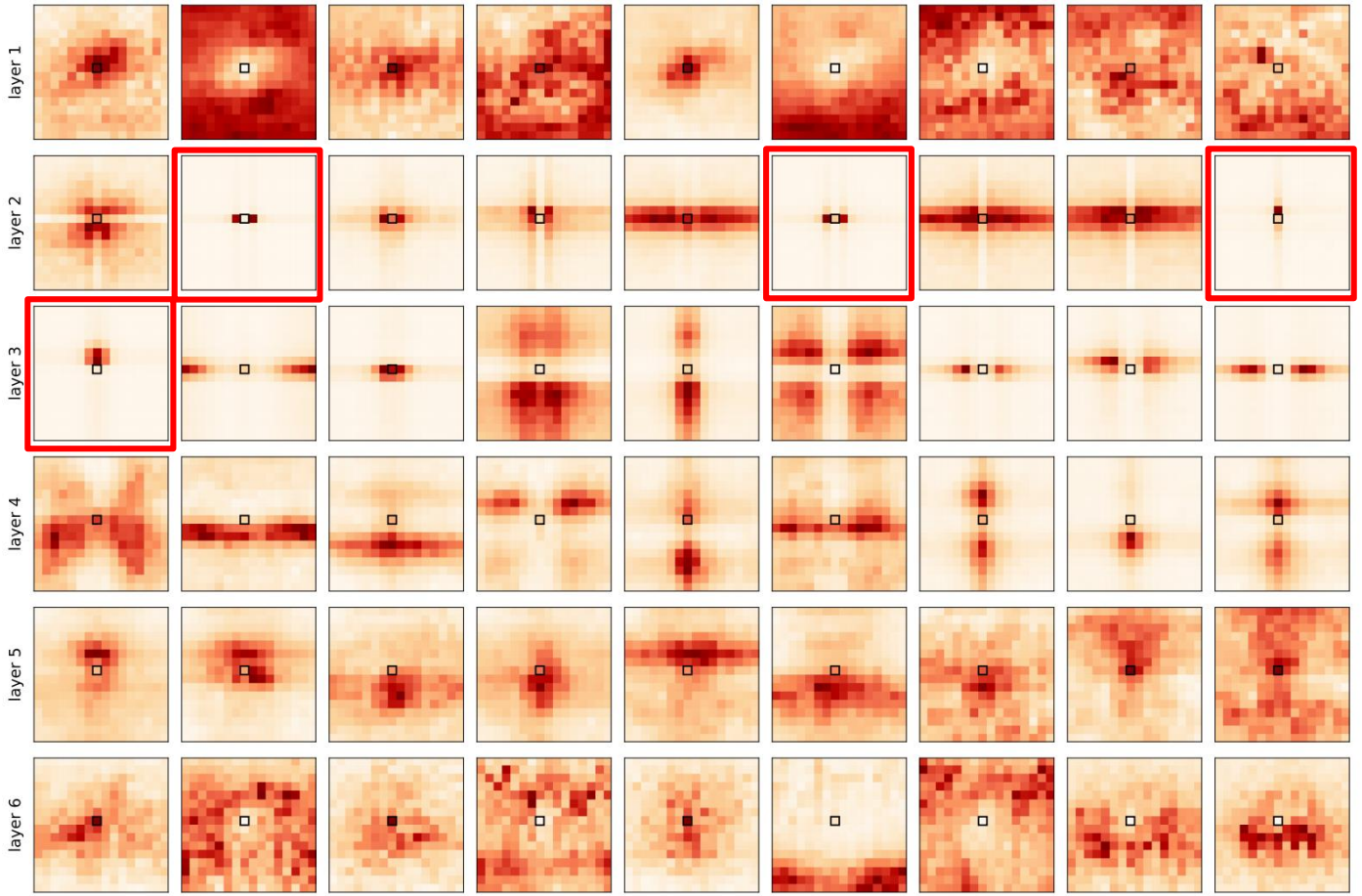
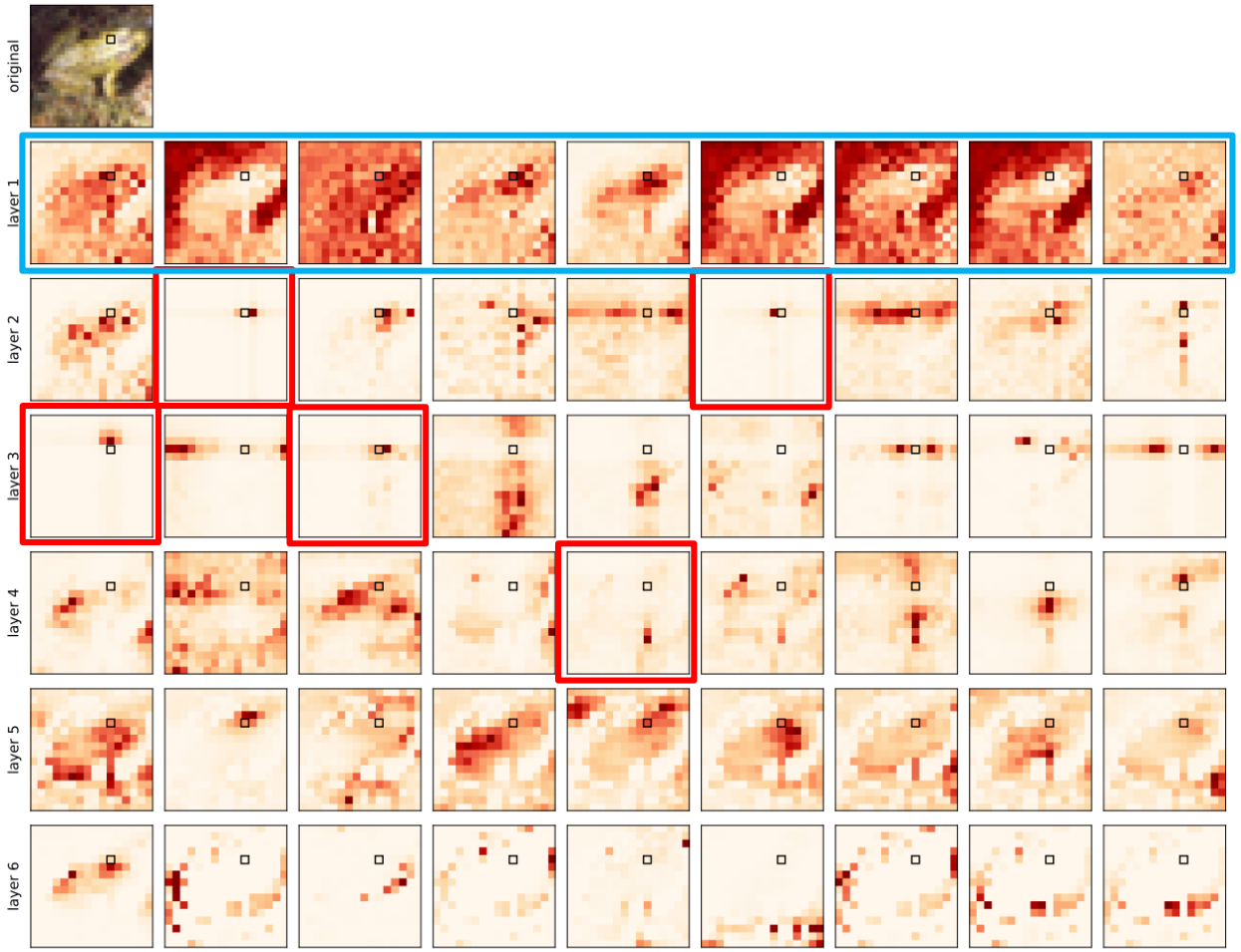


Figure 6: Attention probabilities for a model with 6 layers (*rows*) and 9 heads (*columns*) using learned relative positional encoding and content-content based attention. Attention maps are averaged over 100 test images to display head behavior and remove the dependence on the input content. The black square is the query pixel. More examples are presented in Appendix A.

- Some heads use more **content-based attention** heads, leveraging the advantage of self-attention over CNN. This effectiveness was shown by *Bello et al.(2019)* that combining CNN and self-attention features outperforms each taken separately.

Experiments

Using learned encoding(with content-based attention) :



Capturing highly content-dependent patterns.

Satisfying lemma 1, thus theorem as well. (local pattern)

Figure 8: Attention probabilities for a model with 6 layers (*rows*) and 9 heads (*columns*) using learned relative positional encoding and content-content based attention. The query pixel (black square) is on the frog head.

Thank you