On the Relationship Between Self-Attention and Convolutional Layers

Jean-Baptiste Cordonnier
with Andreas Loukas and Martin Jaggi

Swiss Machine Learning Days
November 13th, 2019

Conclusion

On the Relationship between
Self-Attention and Convolutional Layers
jbcordonnier.com

A Multi-Head Self-Attention Layer
can express any Convolutional Layer.

Building block of state of the art NLP models

Transformers
(Vaswani et al. 2017)

GPT2
(Radford et al. 2018)

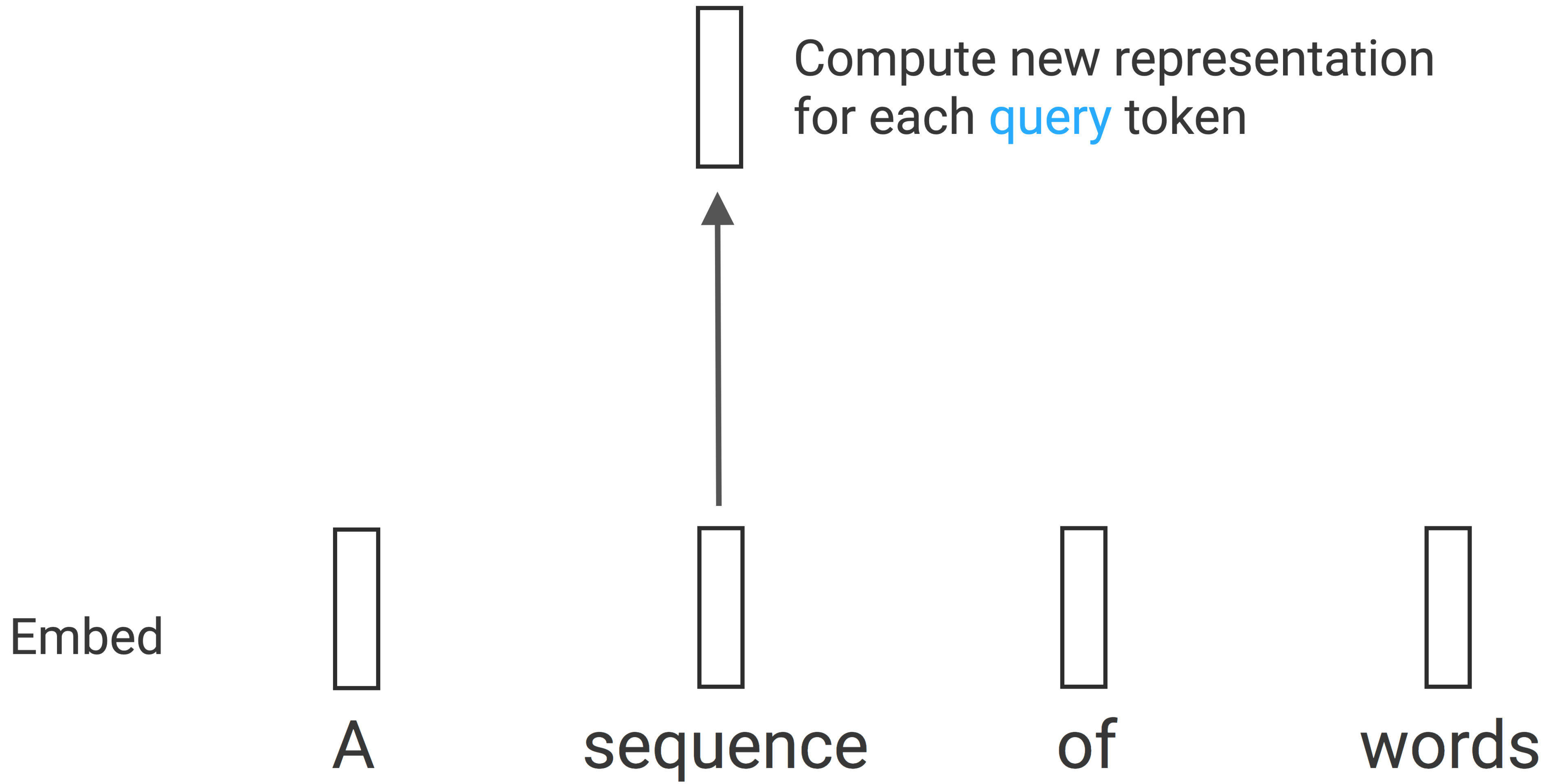
BERT
(Devlin et al. 2019)

are competitive with CNNs.

applied to vision task,
achieve same performance,
at same computation cost.

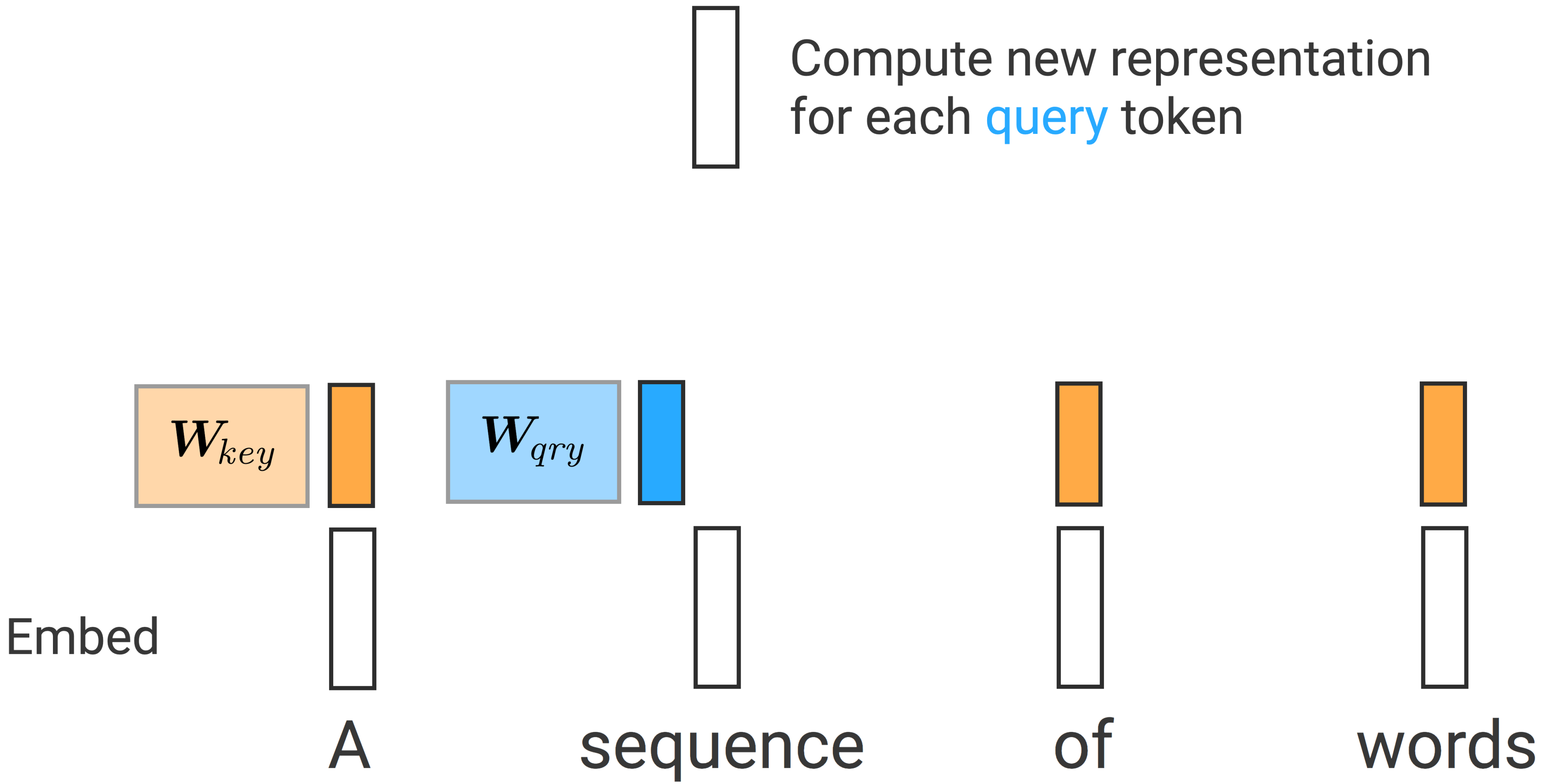
(Bello et al. 2019)
(Ramachandran et al. 2019)

Self-Attention (Vaswani et al. 2017)

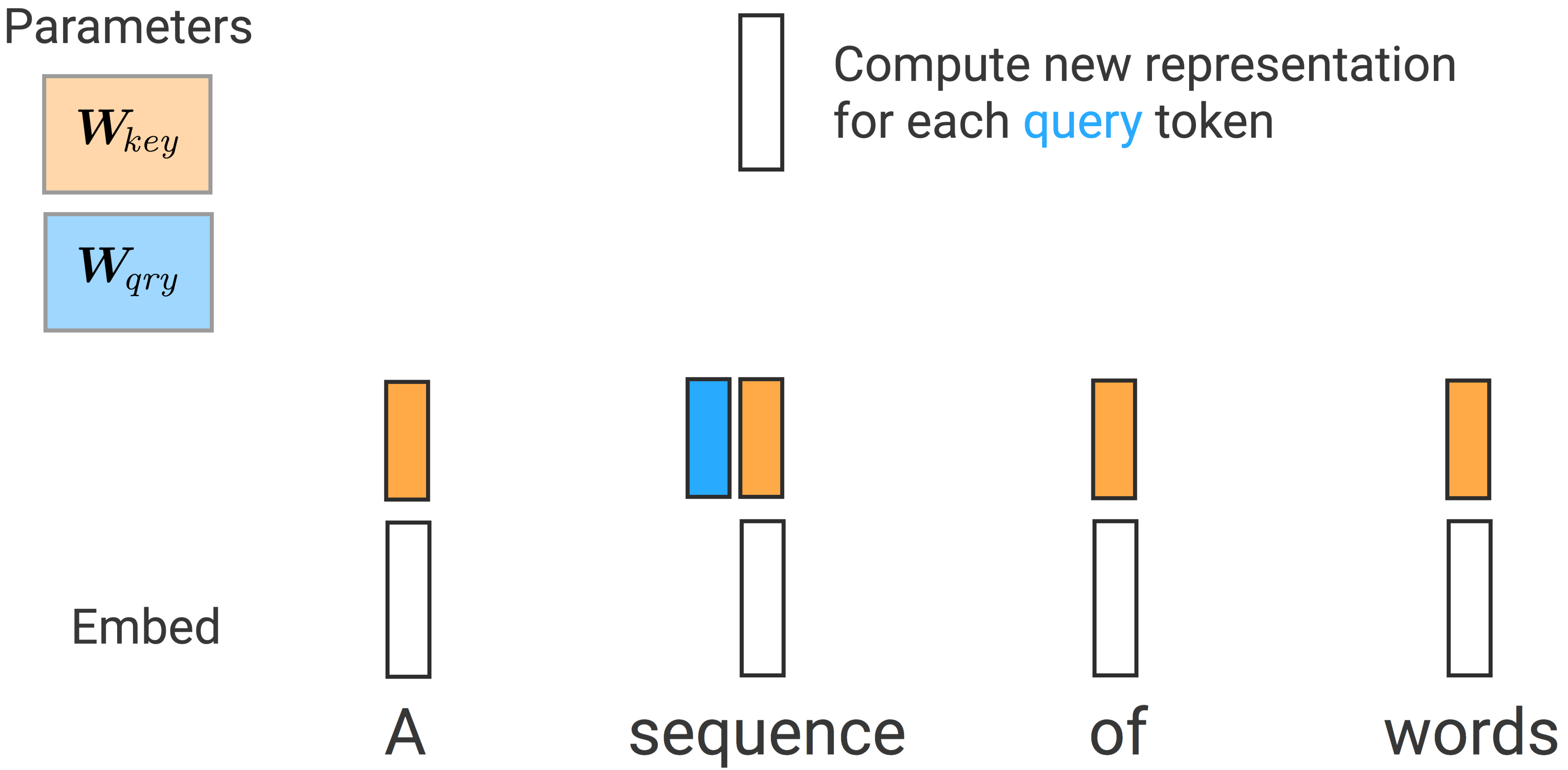


Self-Attention (Vaswani et al. 2017)

Compute new representation for each query token



Self-Attention (Vaswani et al. 2017)



Parameters

$\boldsymbol{W}_{key}$

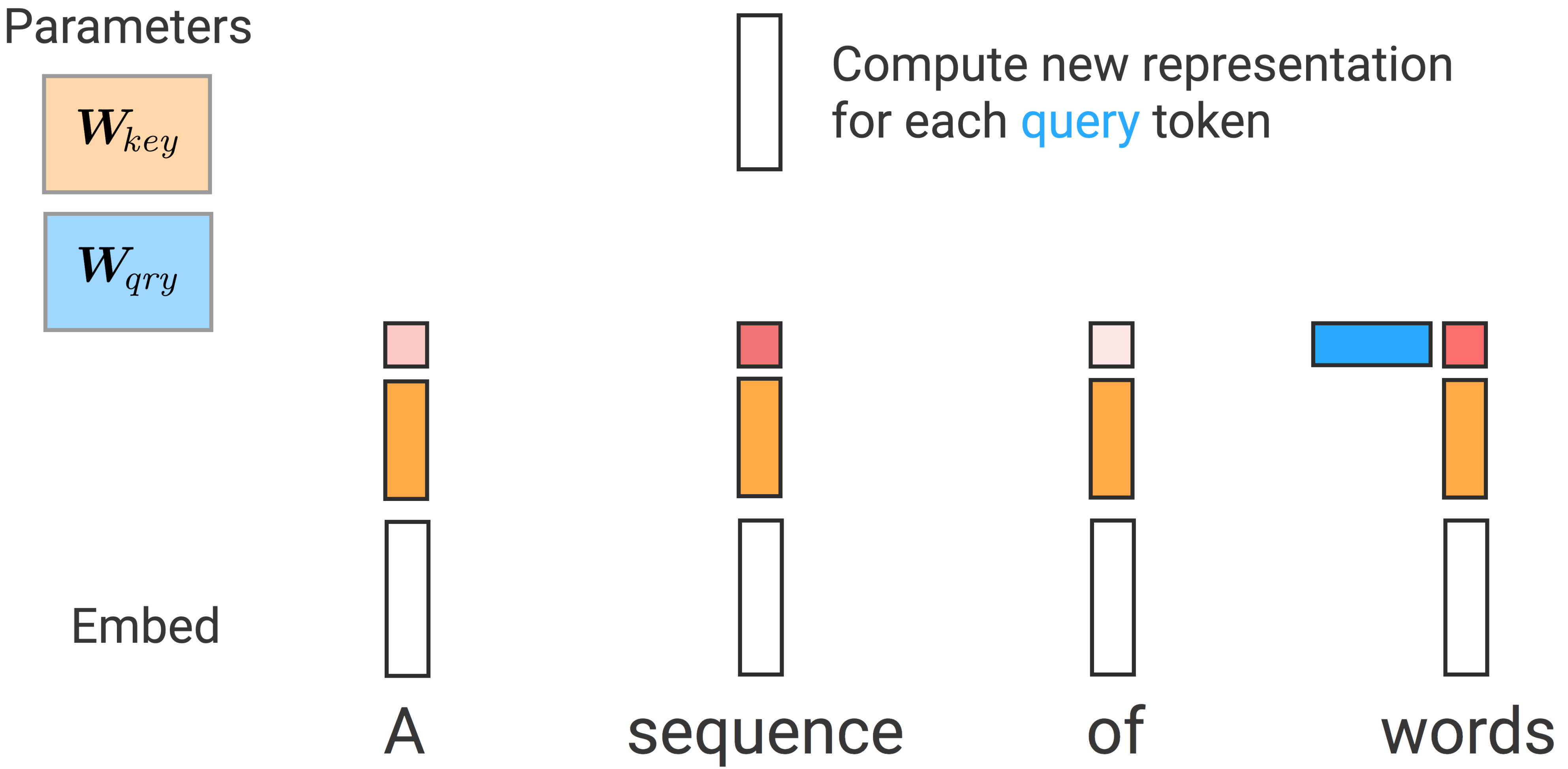
$\boldsymbol{W}_{qry}$

Compute new representation
for each query token

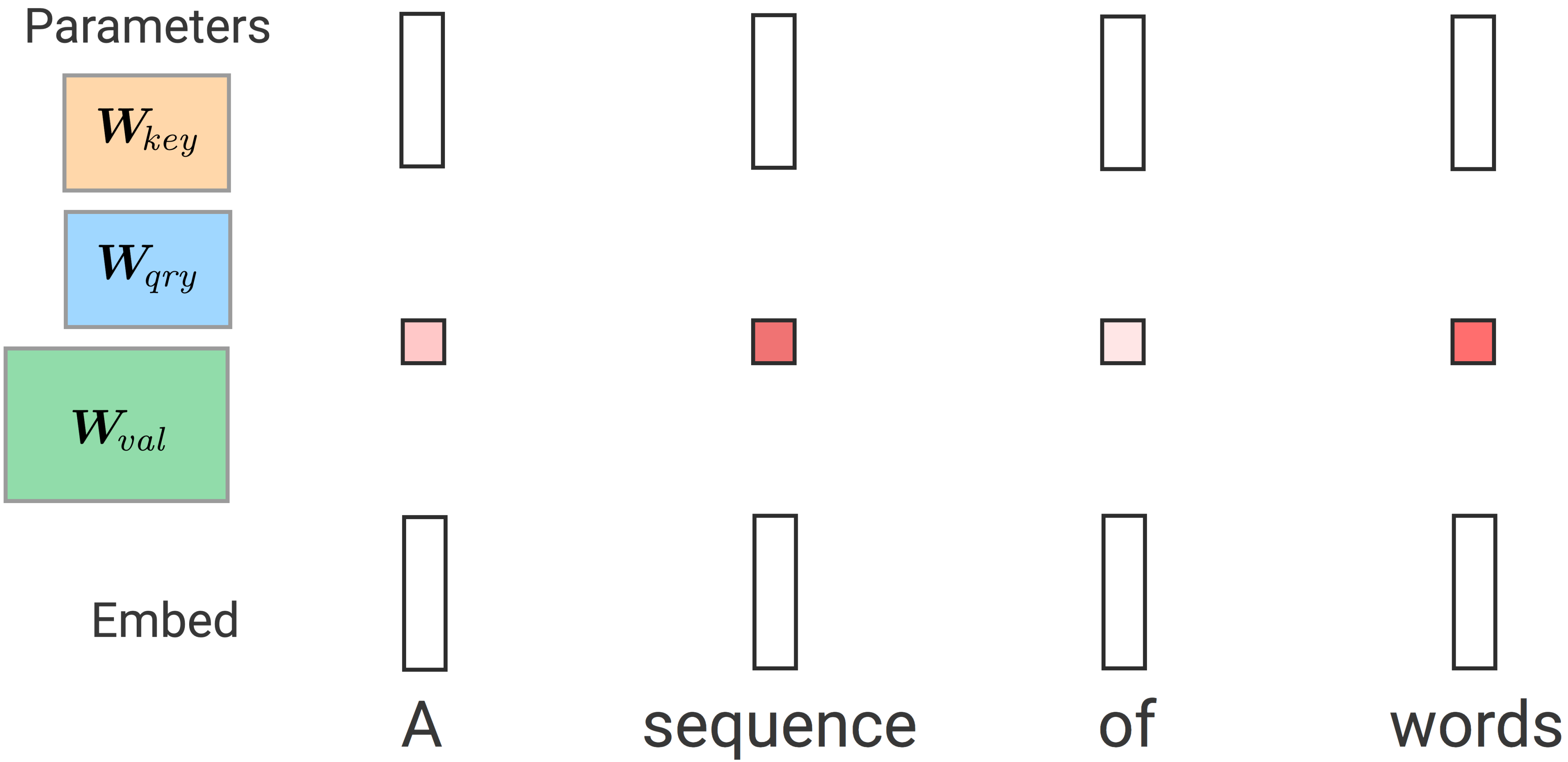
Embed

A sequence of words

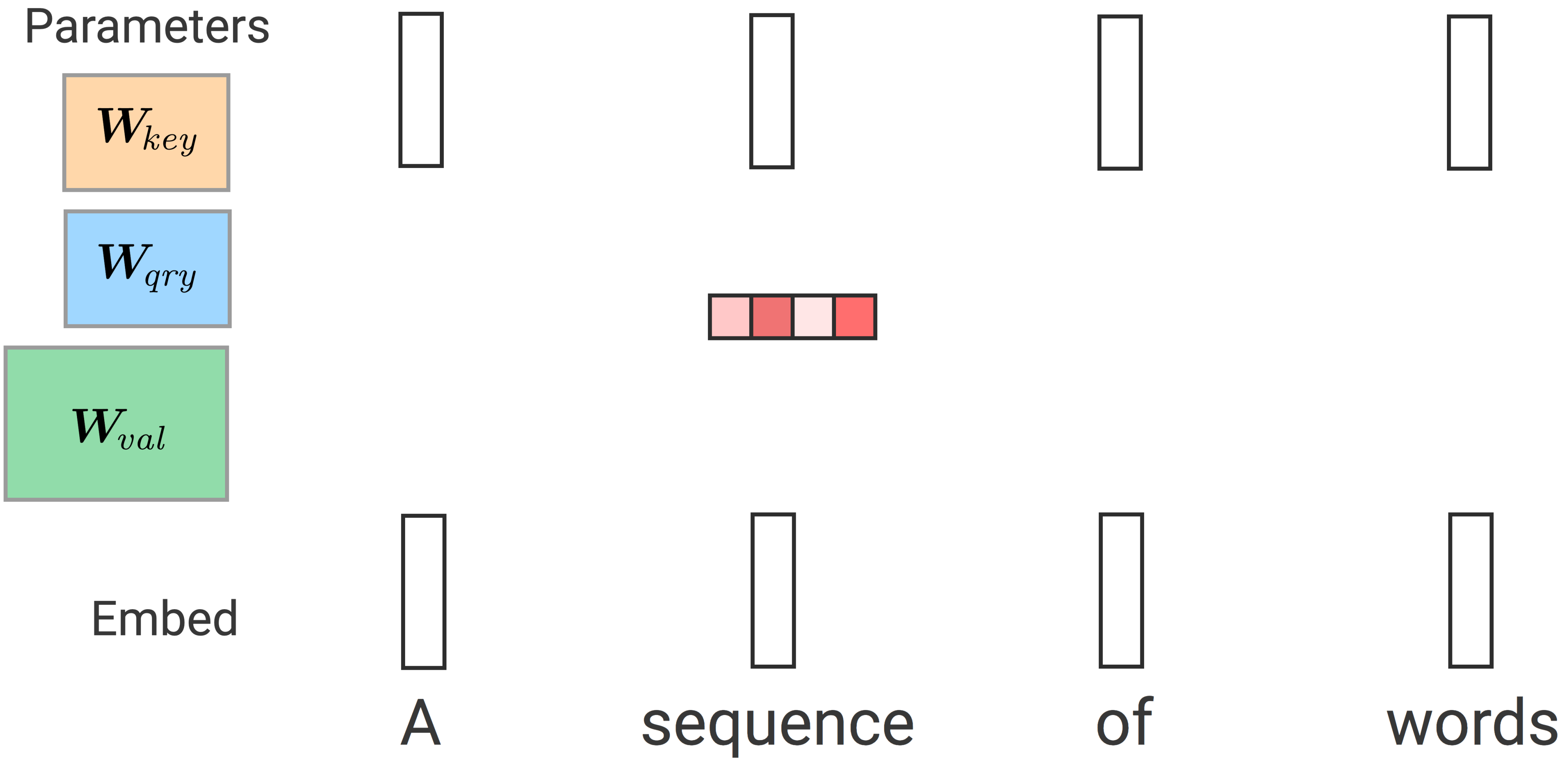
Self-Attention (Vaswani et al. 2017)



Self-Attention (Vaswani et al. 2017)



Self-Attention (Vaswani et al. 2017)



Parameters

$\boldsymbol{W}_{key}$

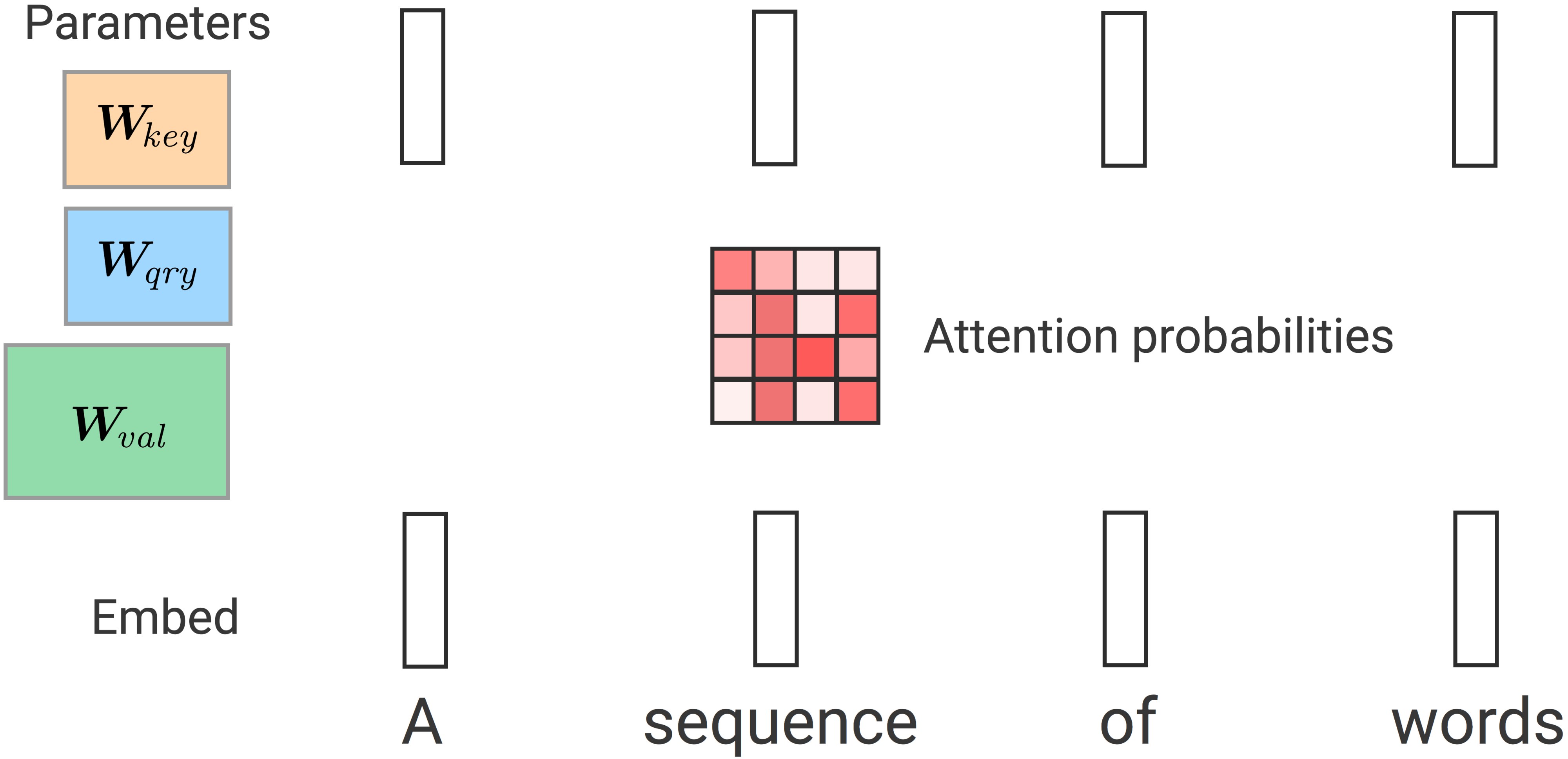
$\boldsymbol{W}_{qry}$

$\boldsymbol{W}_{val}$

Embed

A sequence of words

Self-Attention (Vaswani et al. 2017)



Parameters

$\boldsymbol{W}_{key}$

$\boldsymbol{W}_{qry}$

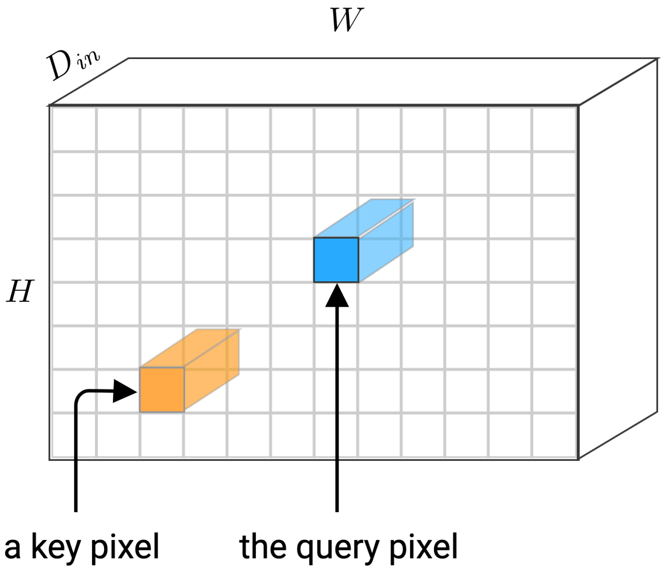
$\boldsymbol{W}_{val}$

Attention probabilities

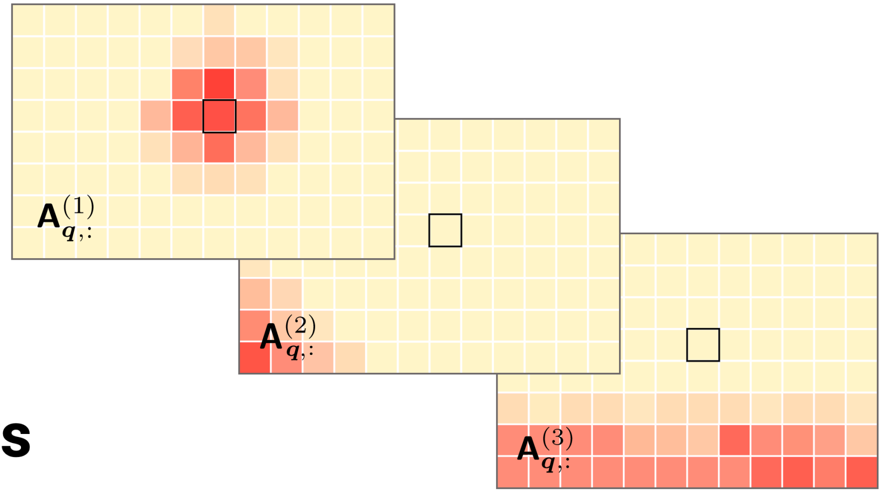
Embed

A sequence of words

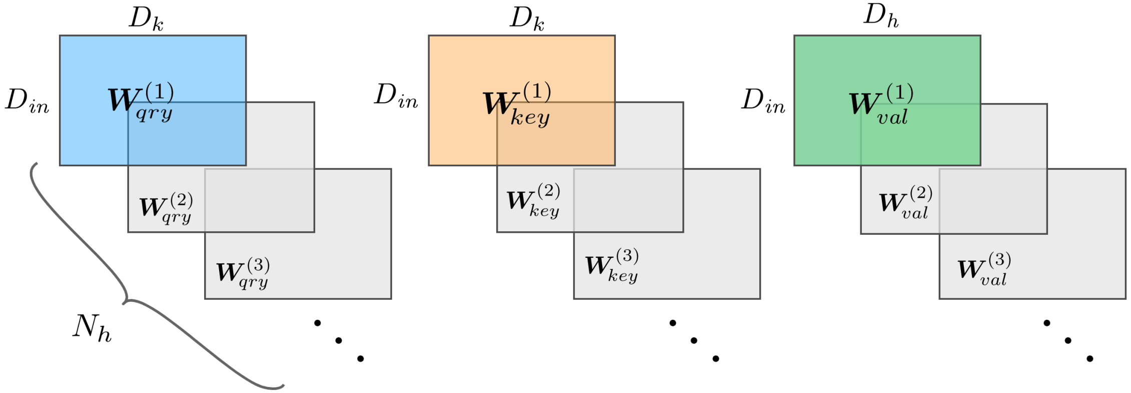
Self-Attention on Images



Attention probabilities

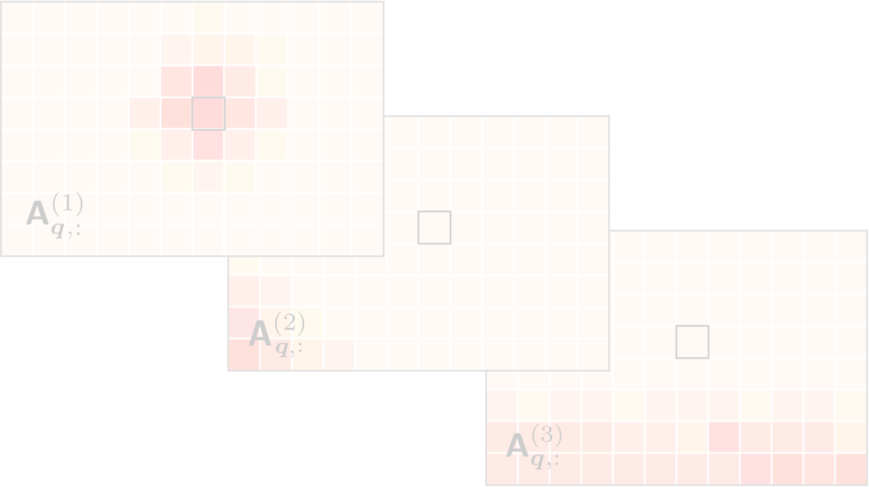

$\mathbf{A}^{(1)}_{\boldsymbol{q},:}$
$\mathbf{A}^{(2)}_{\boldsymbol{q},:}$
$\mathbf{A}^{(3)}_{\boldsymbol{q},:}$

Parameters

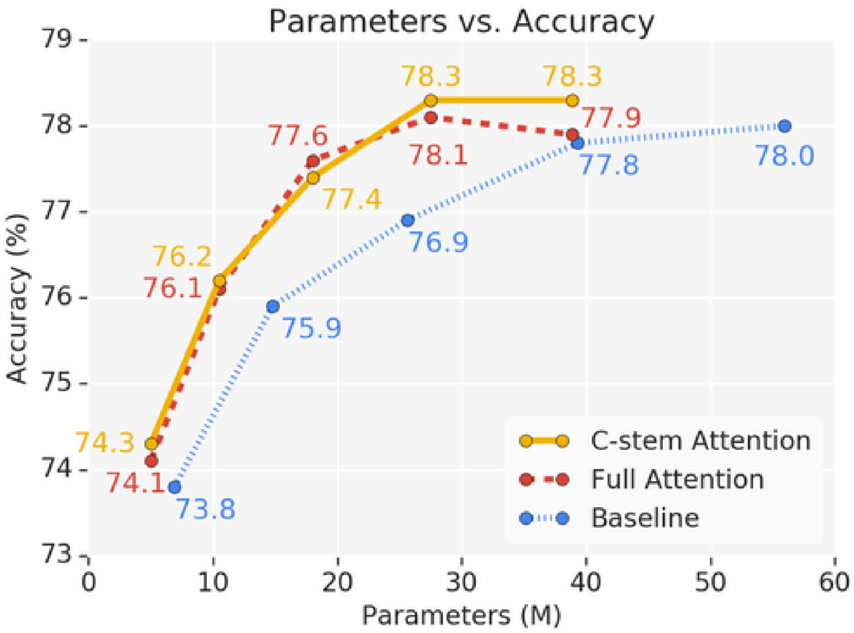

D_k
D_{in}
$\boldsymbol{W}^{(1)}_{qry}$
$\boldsymbol{W}^{(2)}_{qry}$
$\boldsymbol{W}^{(3)}_{qry}$
N_h
D_k
D_{in}
$\boldsymbol{W}^{(1)}_{key}$
$\boldsymbol{W}^{(2)}_{key}$
$\boldsymbol{W}^{(3)}_{key}$
D_h
D_{in}
$\boldsymbol{W}^{(1)}_{val}$
$\boldsymbol{W}^{(2)}_{val}$
$\boldsymbol{W}^{(3)}_{val}$

Self-Attention on Images

Attention probabilities

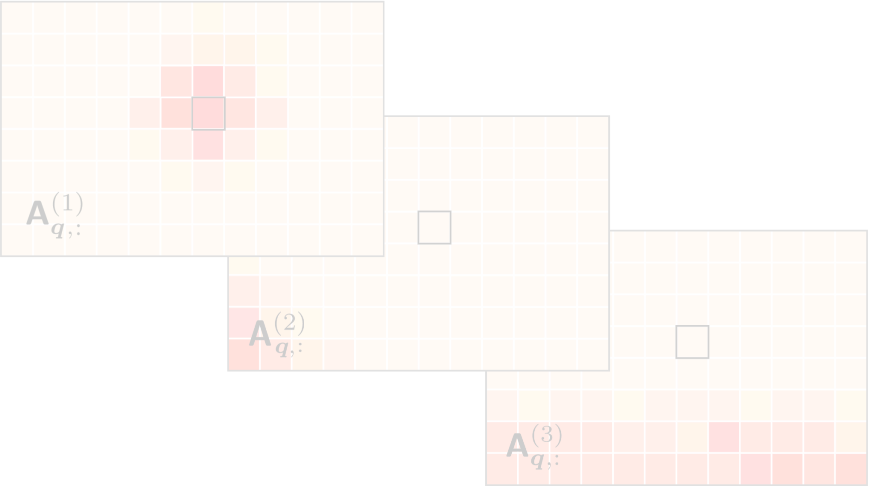


Same performance as ResNet on ImageNet (Ramachandran, 2019)

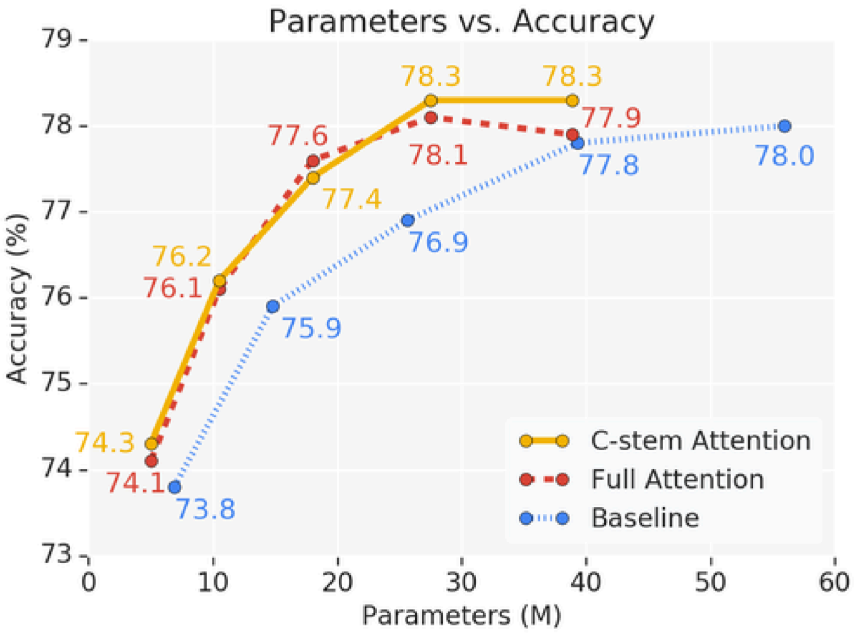


Self-Attention on Images

Attention probabilities

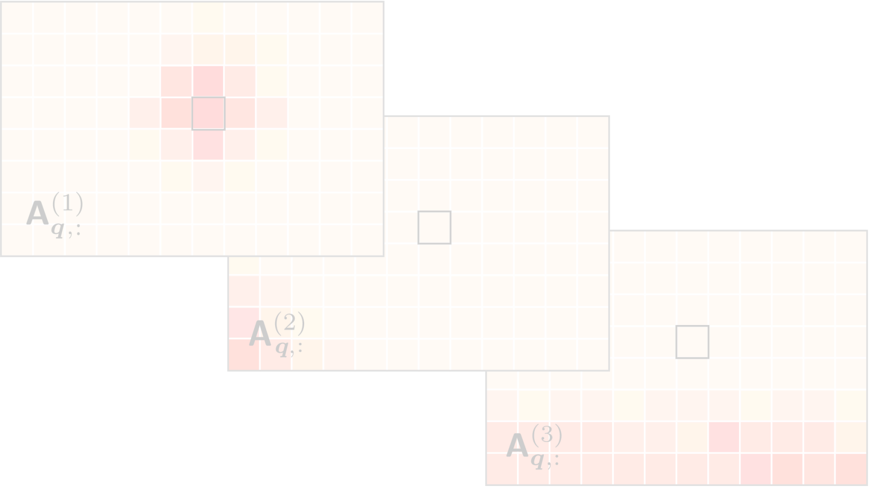


Same performance as ResNet on ImageNet (Ramachandran, 2019)

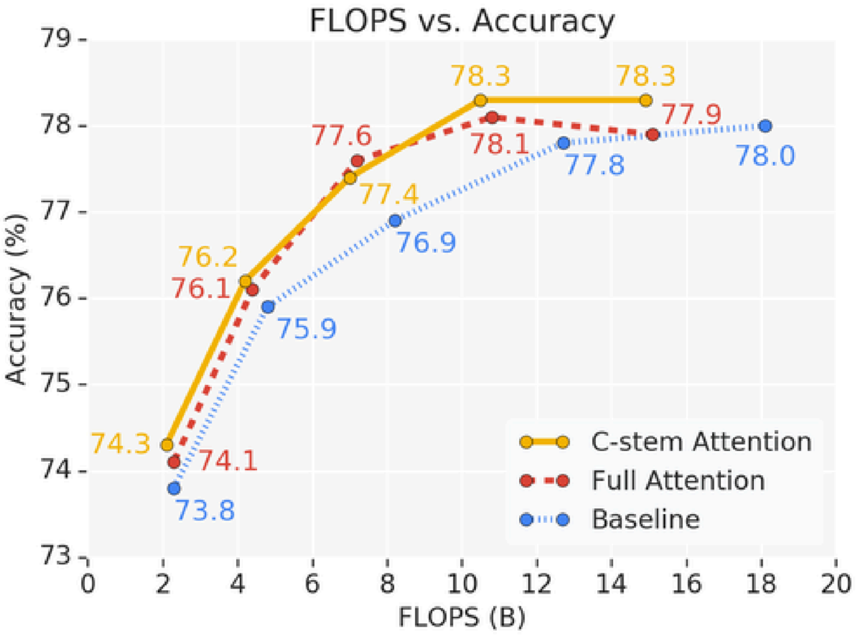


Self-Attention on Images

Attention probabilities

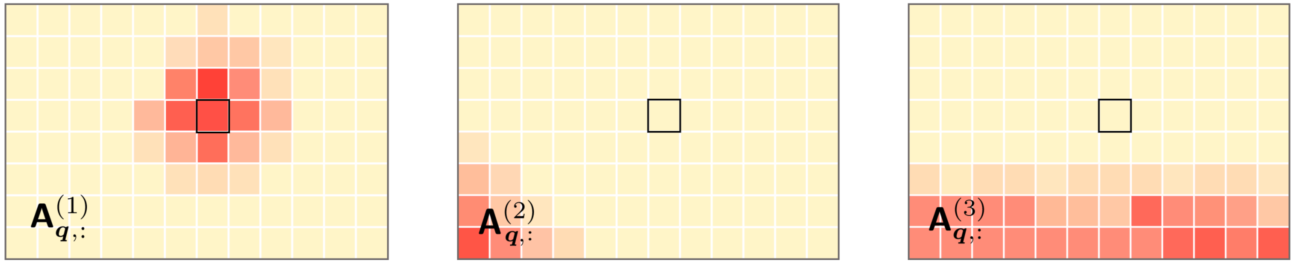


Same performance as ResNet on ImageNet (Ramachandran, 2019)

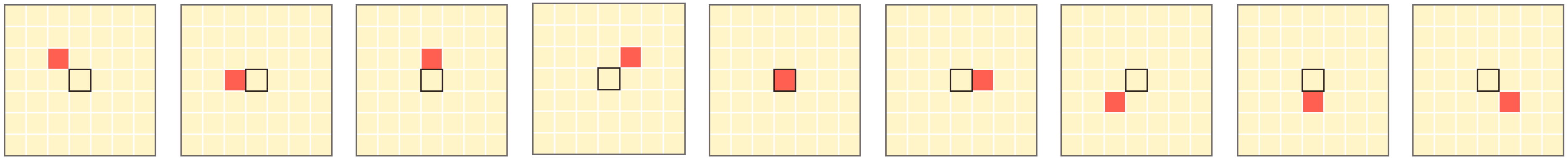


Self-Attention on Images

Attention probabilities

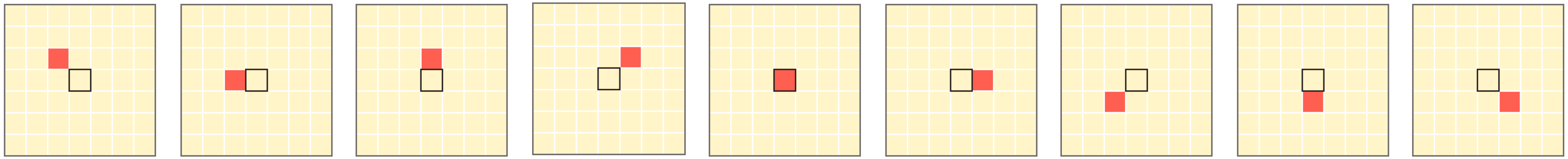

$\mathbf{A}^{(1)}_{\boldsymbol{q},:}$
$\mathbf{A}^{(2)}_{\boldsymbol{q},:}$
$\mathbf{A}^{(3)}_{\boldsymbol{q},:}$

What if attention probabilities could look like this?



Self-Attention on Images

What if attention probabilities could look like this?

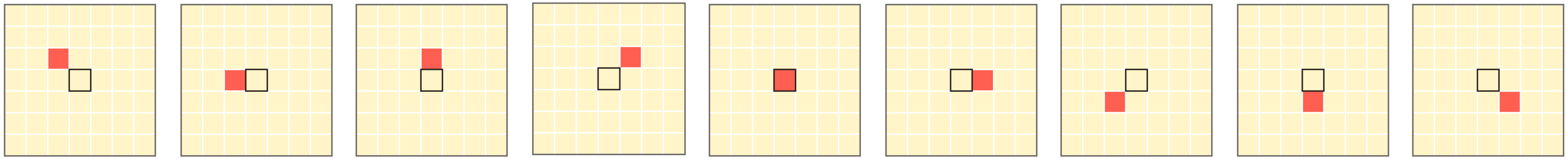


Then the multi-head self-attention could express any 3x3 convolution

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.*

Self-Attention on Images

What if attention probabilities could look like this?

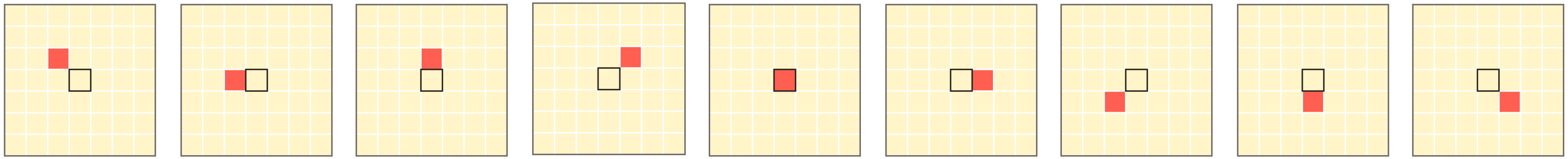


Then the multi-head self-attention could express any 3x3 convolution

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.*

Self-Attention on Images

What if attention probabilities could look like this?



Then the multi-head self-attention could express any 3x3 convolution

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.*

Conclusion

On the Relationship between
Self-Attention and Convolutional Layers
jbcordonnier.com

A Multi-Head Self-Attention Layer
can express any Convolutional Layer.

Building block of state of the art NLP models

Transformers
(Vaswani et al. 2017)

GPT2
(Radford et al. 2018)

BERT
(Devlin et al. 2019)

are competitive with CNNs.

applied to vision task,
achieve same performance,
at same computation cost.

(Bello et al. 2019)
(Ramachandran et al. 2019)