ATTENTION AND DIFFUSION

Understanding Neural Network Architectures

Michał Karzyński

EuroPython 2023

THE **SPEAKER**

Michał Karzyński (@postrational)

Software Architect

ONNX Operators group

THE TALK: **MODELS**

Transformers	Diffusion
Natural Language	Images
GPT BERT T5	Stable Diffusion Midjourney DALL-E
Attention	Convolution & Attention

CHAT **GPT**

Model: GPT-4

Hi Chat, say "Hello" to people at EuroPython 2023, who are listening to a talk about you.

Hello EuroPython 2023 attendees! I'm thrilled that you're engaging in a discussion about me, ChatGPT. As an AI language model, my purpose is to assist, inspire, and educate. I look forward to learning about the insightful observations and exciting ideas that will emerge from this talk. Enjoy the conference, stay curious, and keep exploring the wonderful world of Python and AI!

DIFFUSION MODELS

Stable Diffusion

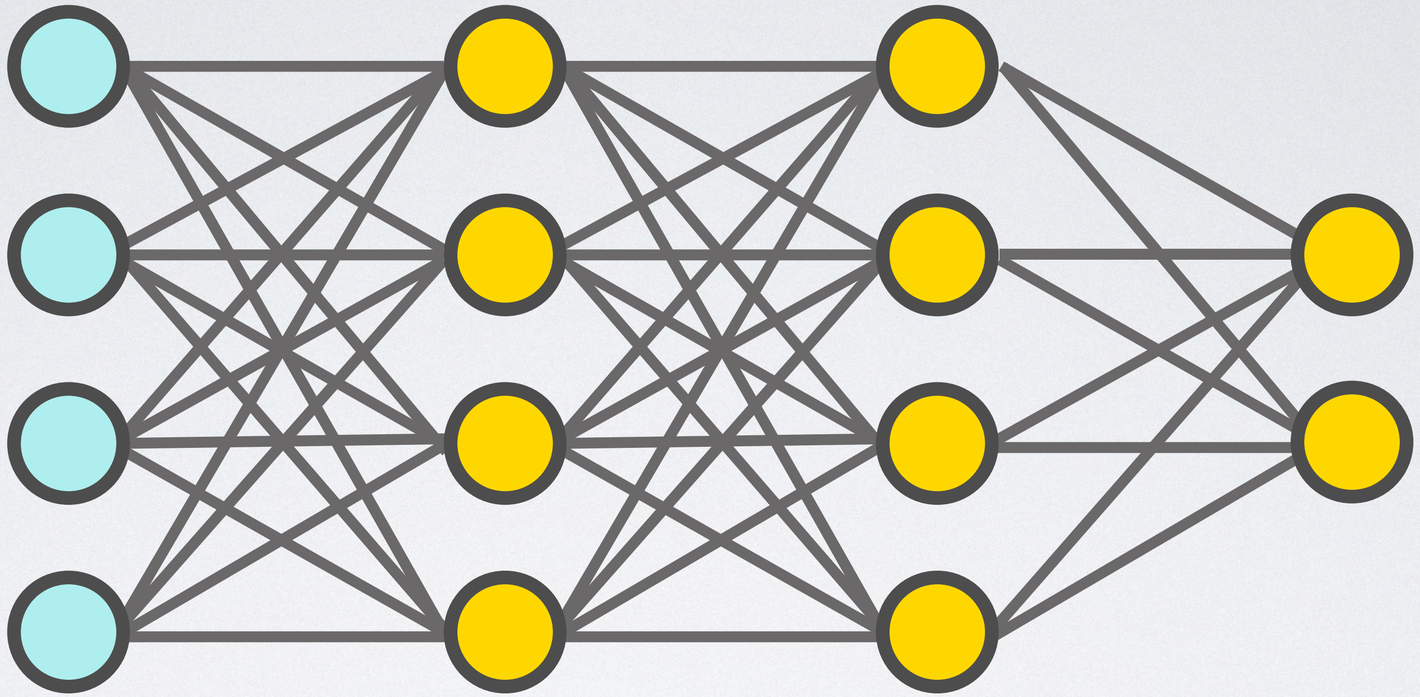
Midjourney

DALL-E

THE TALK: **OPERATIONS**

Linear	a.k.a.: Dense, Fully-connected $y = AX + b$
Convolution	Filter scan to produce feature maps
Attention	Key-value store lookup

MULTI-LAYER **PERCEPTRON**



Multi-layer Perceptron

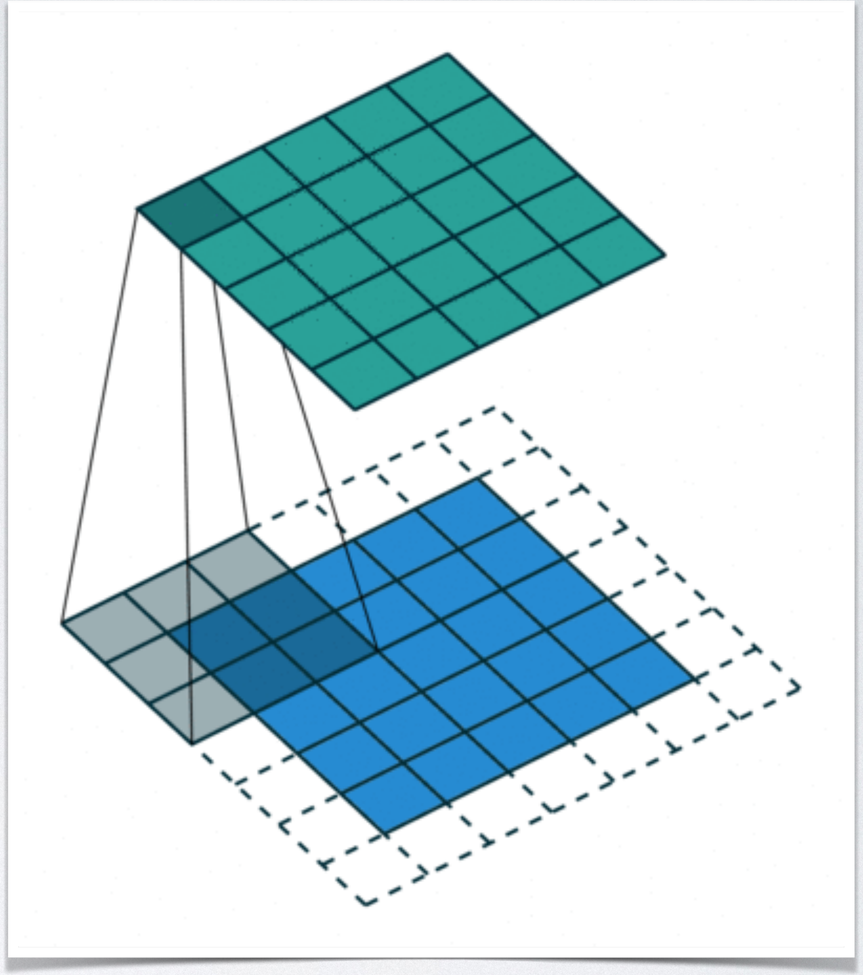
linear input

Multi-layer Perceptron

linear

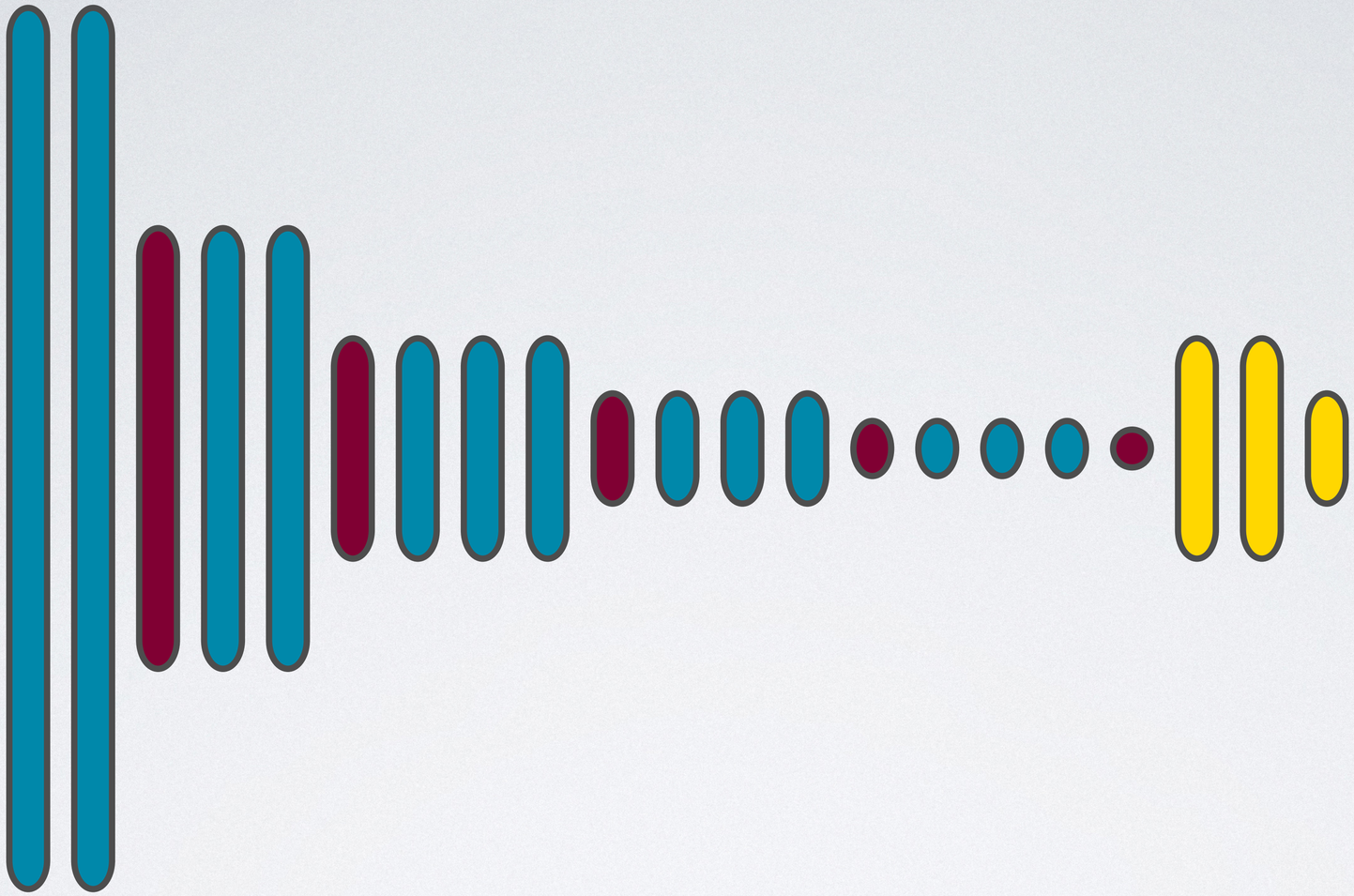
OPERATION: **CONVOLUTION**

OPERATION:
CONVOLUTION



$$map(x, y) = \sum_{fx=-a}^{a} \sum_{fy=-b}^{b} filter(fx, fy) * input(x - fx, y - fy)$$

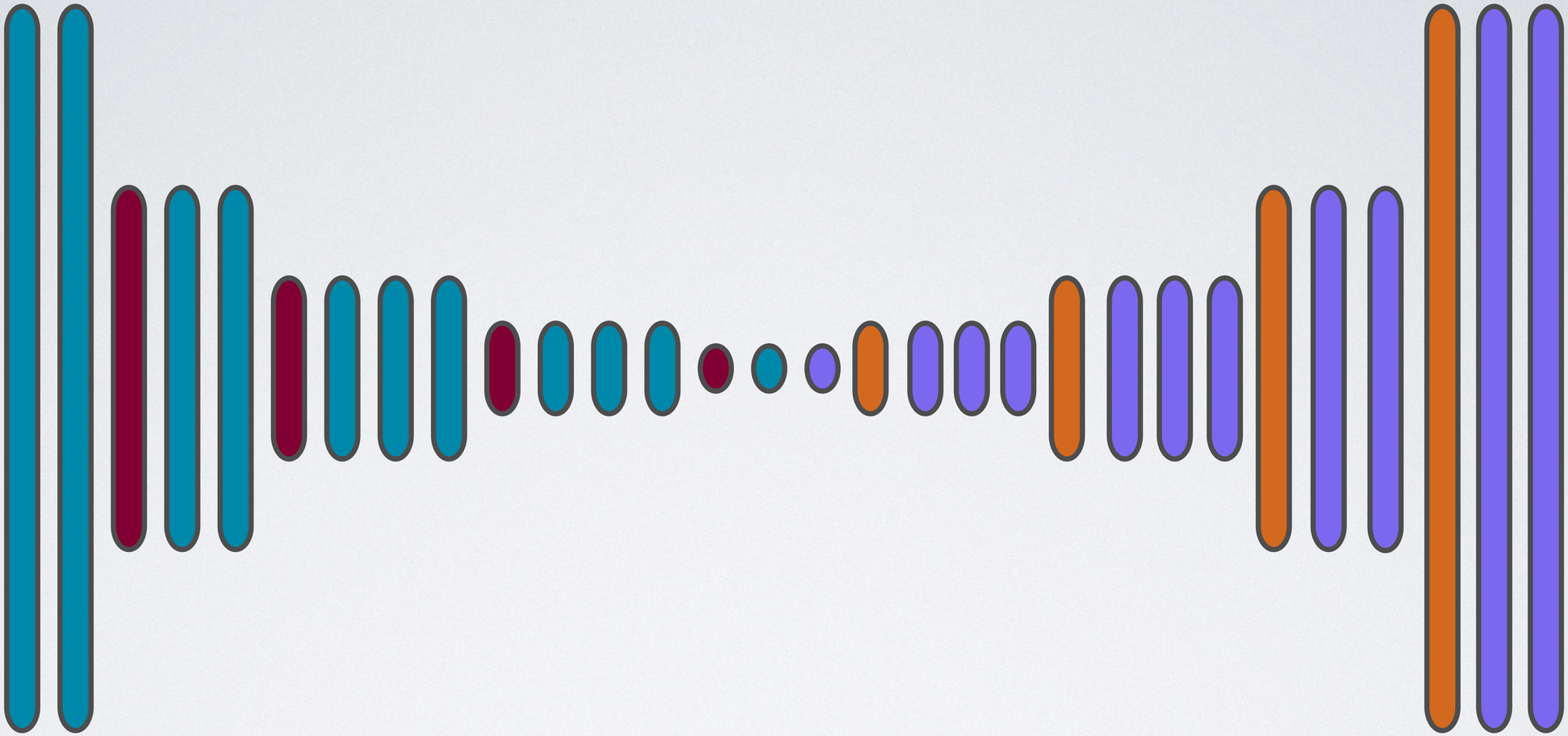
CONVOLUTIONAL NETWORKS



VGG-16

convolution max pool linear

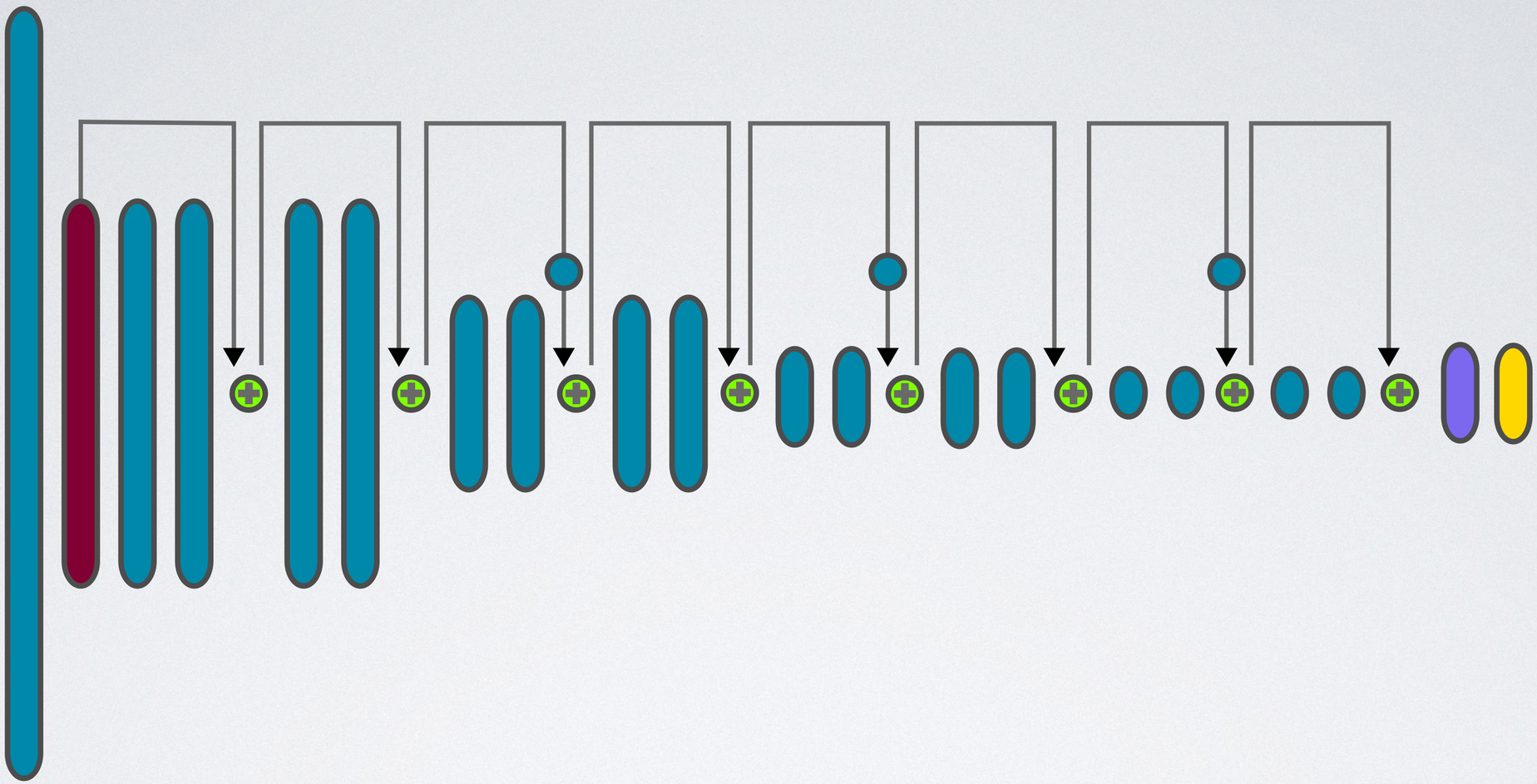
ENCODER-DECODER ARCHITECTURE



Convolutional Autoencoder

convolution max pool deconvolution unpooling

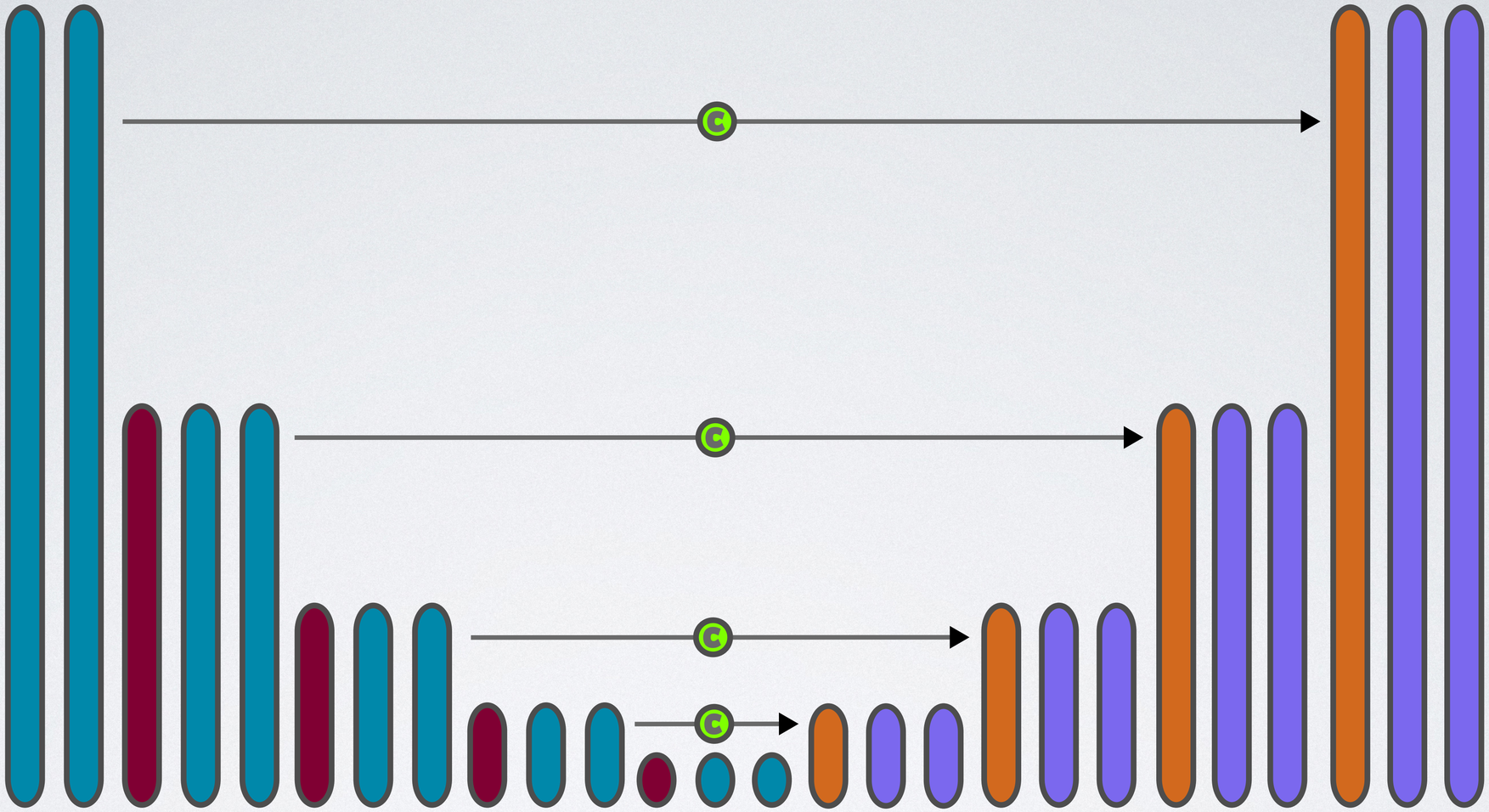
SKIP CONNECTIONS AND **RESIDUAL** NETWORKS



ResNet-18

convolution max pool linear average pool add

U-NET



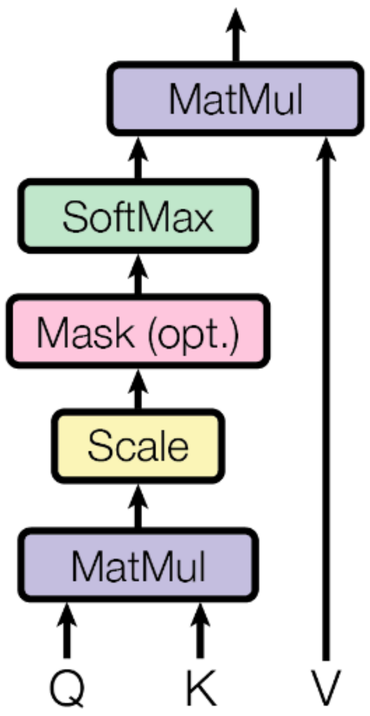
Convolutional U-Net

convolution max pool deconvolution unpooling concatenate

OPERATION: **ATTENTION**

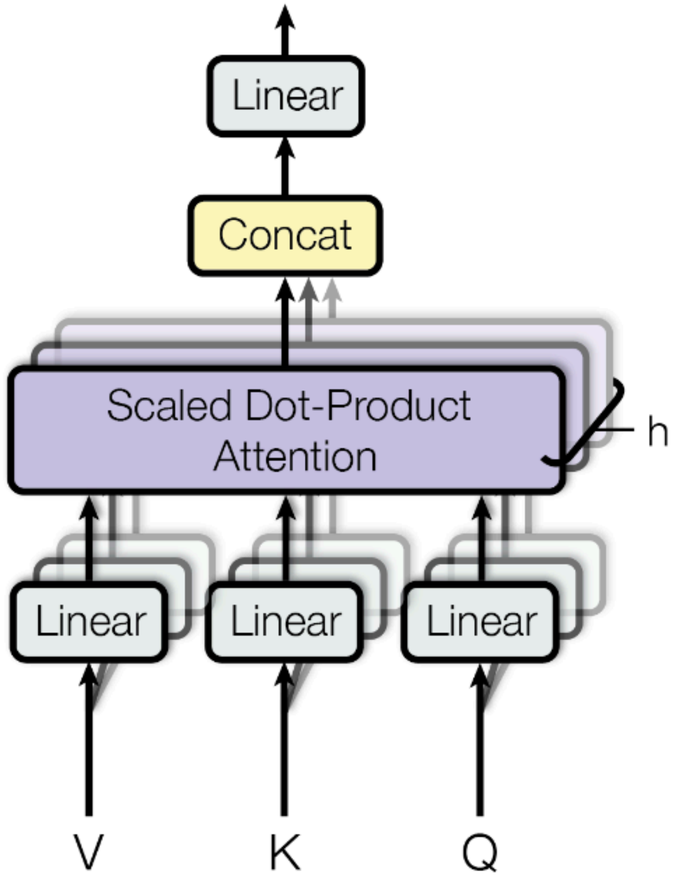
```
store = {
    'key1': 'value1',
    'key2': 'value2',
    'key3': 'value3',
}

query = 'key1'
value = store[query]
```

Scaled Dot-Product Attention



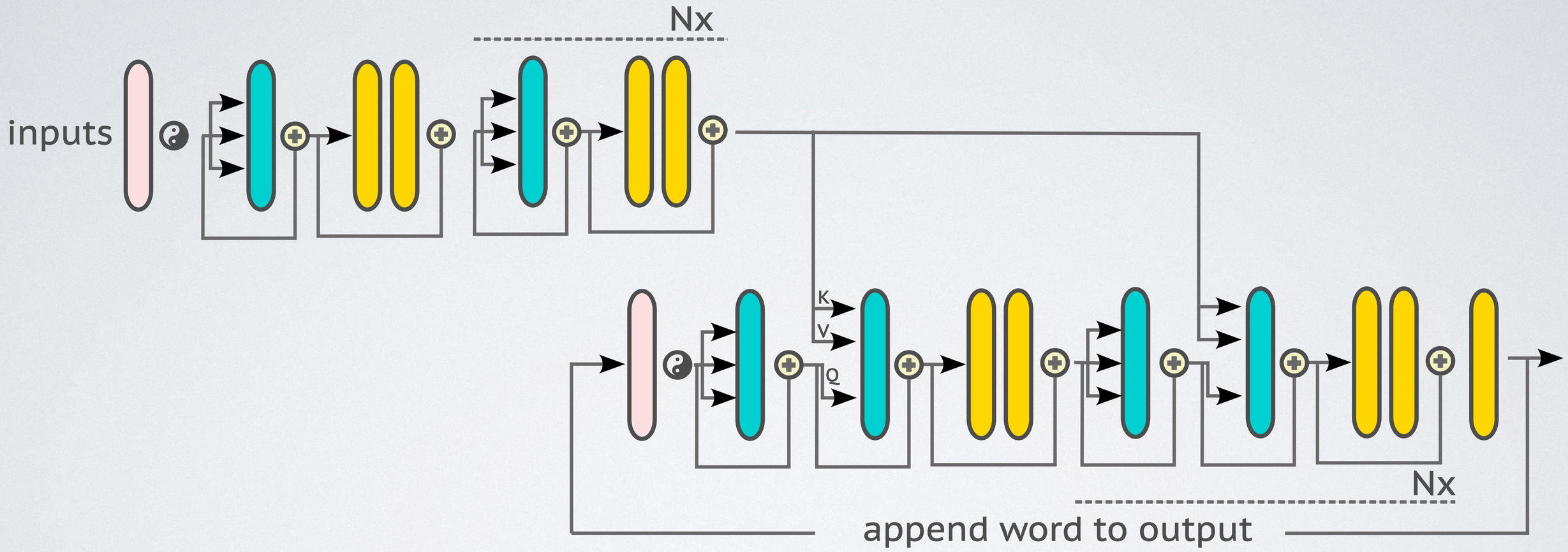
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Multi-Head Attention



Vaswani, Ashish, et al. "Attention Is All You Need." arXiv:1706.03762v5 [cs.CL], 2017

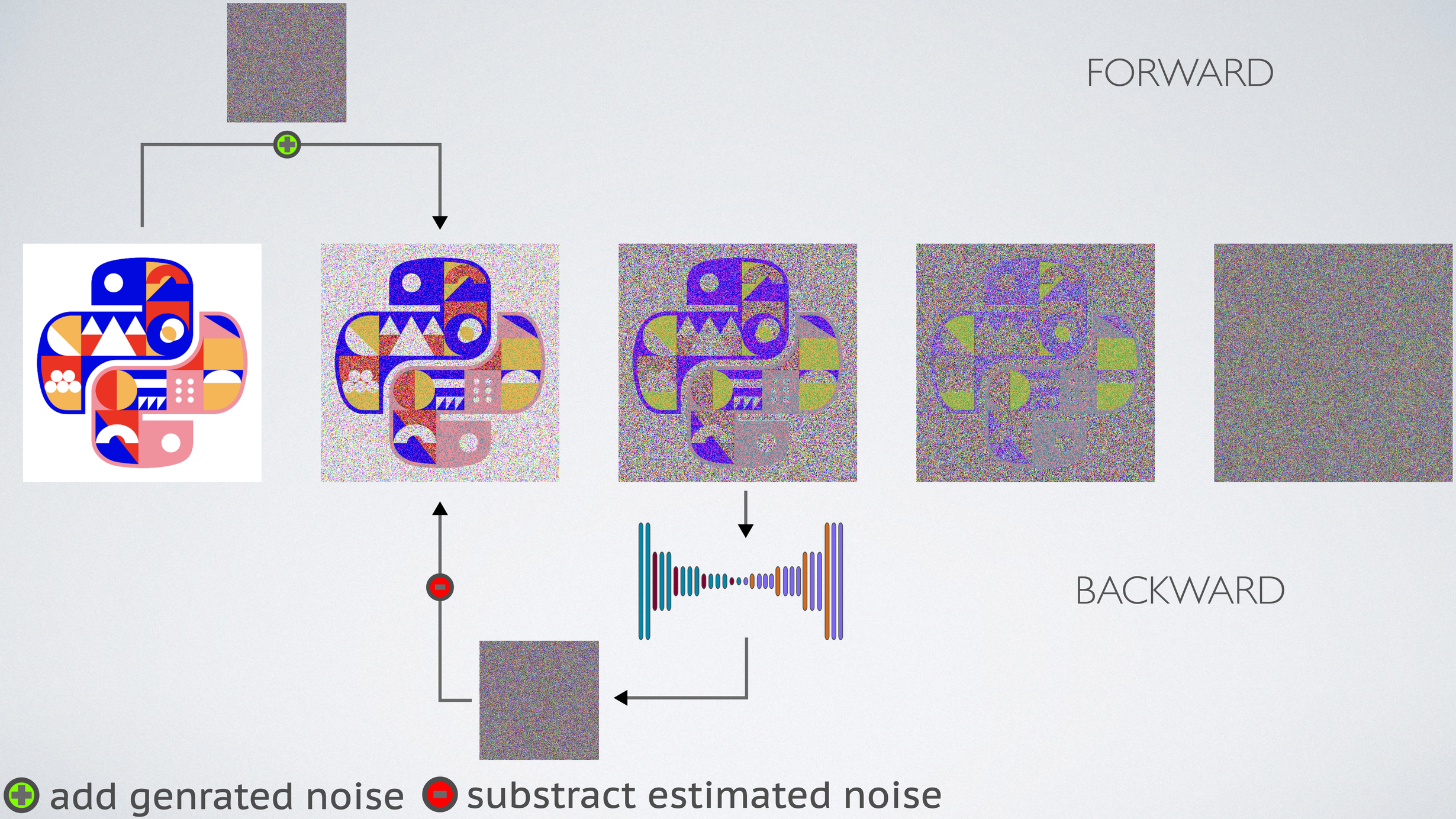
TRANSFORMER ARCHITECTURE



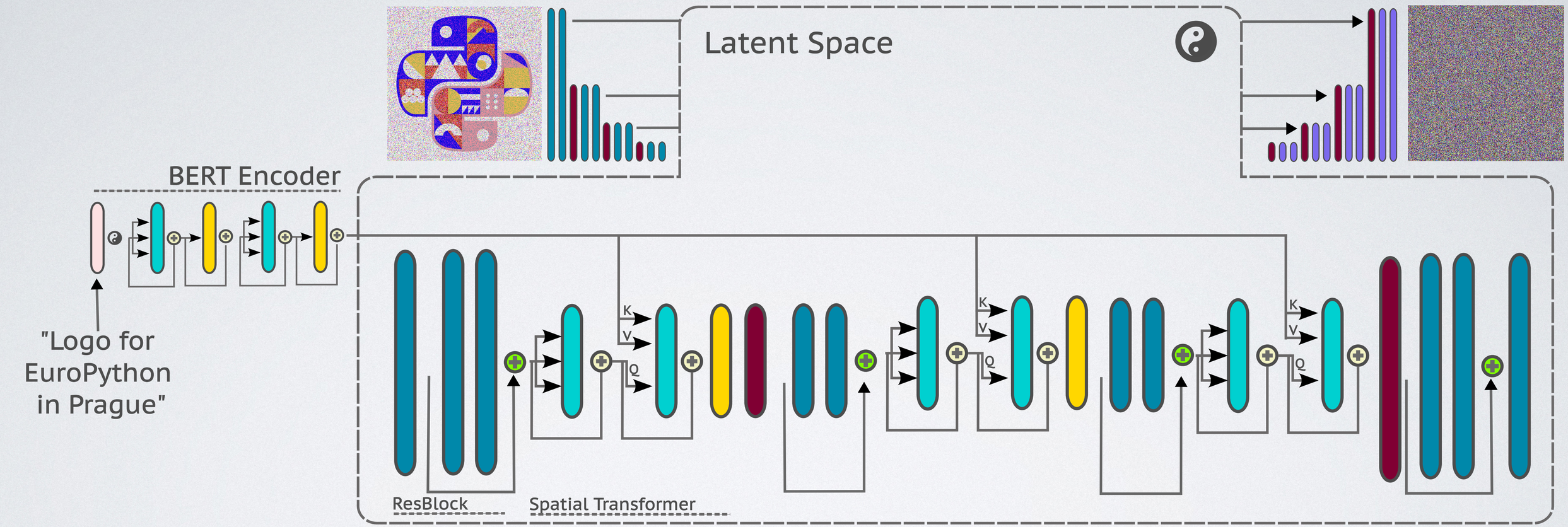
Transformer

multi-head attention linear embedding positional encoding add

FORWARD AND BACKWARD
DIFFUSION


FORWARD
BACKWARD
add genrated noise
substract estimated noise

LATENT DIFFUSION



Latent Diffusion

multi-head attention · linear · embedding · positional encoding
convolution · deconvolution · up/down sample

CLOSING REMARKS

THANK **YOU**