UNIVERSITÉ CÔTE D'AZUR

Inria

Lecture 7

Attention Mechanism

Rui Dai

rui.dai@inria.fr

About Me

Rui Dai

- Home page: https://dairui01.github.io/
- Ph.D. student at INRIA, STARS team.
- Research topic: “Action detection using Deep Learning”.

Outline

- Introduction to Attention Mechanism
- Attention Modules
- Self-Attention
- Transformer

Section 1

Introduction

What is Attention?

In psychology, attention is the cognitive process of selectively concentrating on one or a few things while ignoring others.

What is Attention?

In psychology, attention is the cognitive process of selectively concentrating on one or a few things while ignoring others.

What is Attention?

In psychology, attention is the cognitive process of selectively concentrating on one or a few things while ignoring others.

What is Attention?

In psychology, attention is the cognitive process of selectively concentrating on one or a few things while ignoring others.

What is Attention?

In psychology, attention is the cognitive process of selectively concentrating on one or a few things while ignoring others.

What is Attention?

In psychology, attention is the cognitive process of selectively concentrating on one or a few things while ignoring others.

Introduction

What object is it?

Introduction

What object is it?

What is Attention?

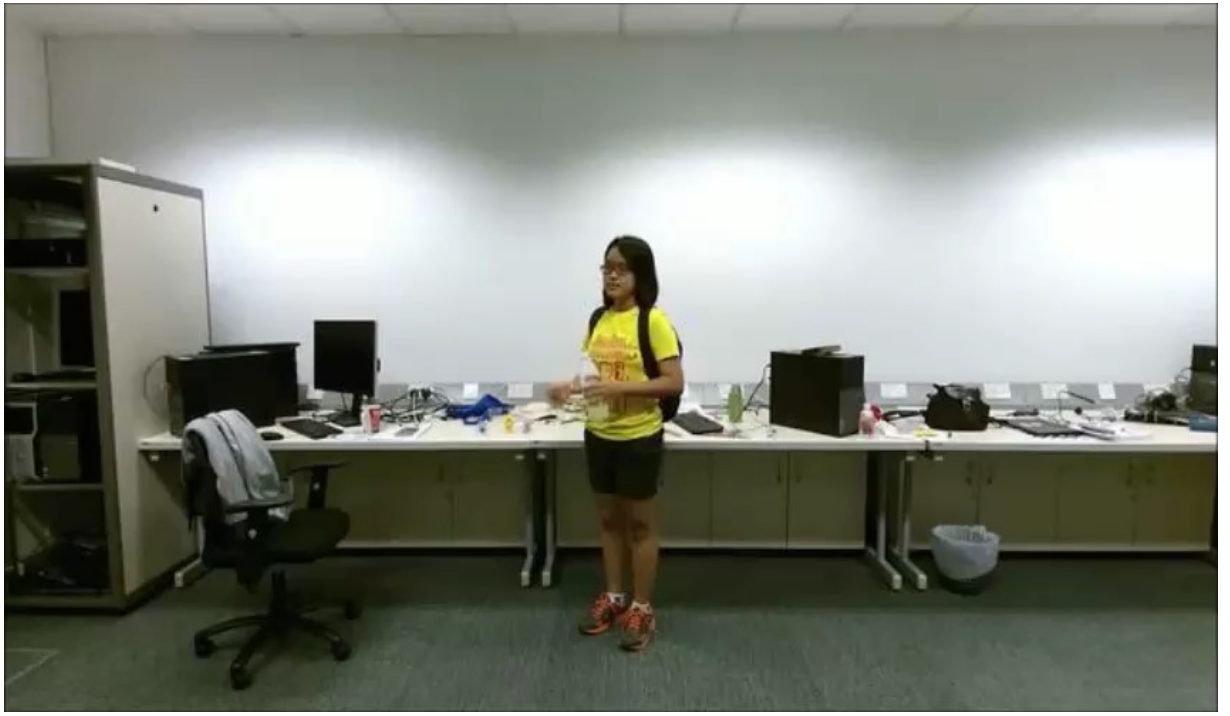
With Attention Mechanism:

Selectively concentrating on a few relevant things, while ignoring others in deep neural networks.

What object is it?

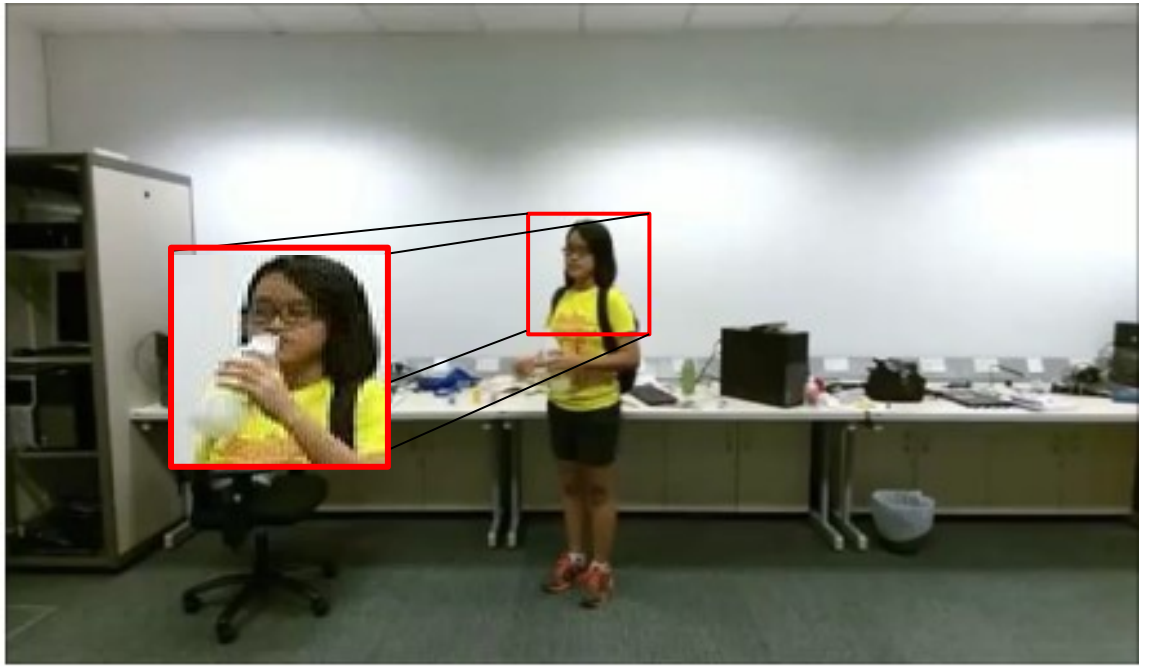
Example in Action Recognition

The girl is drinking water from a bottle



Do we really need the whole video to infer that?

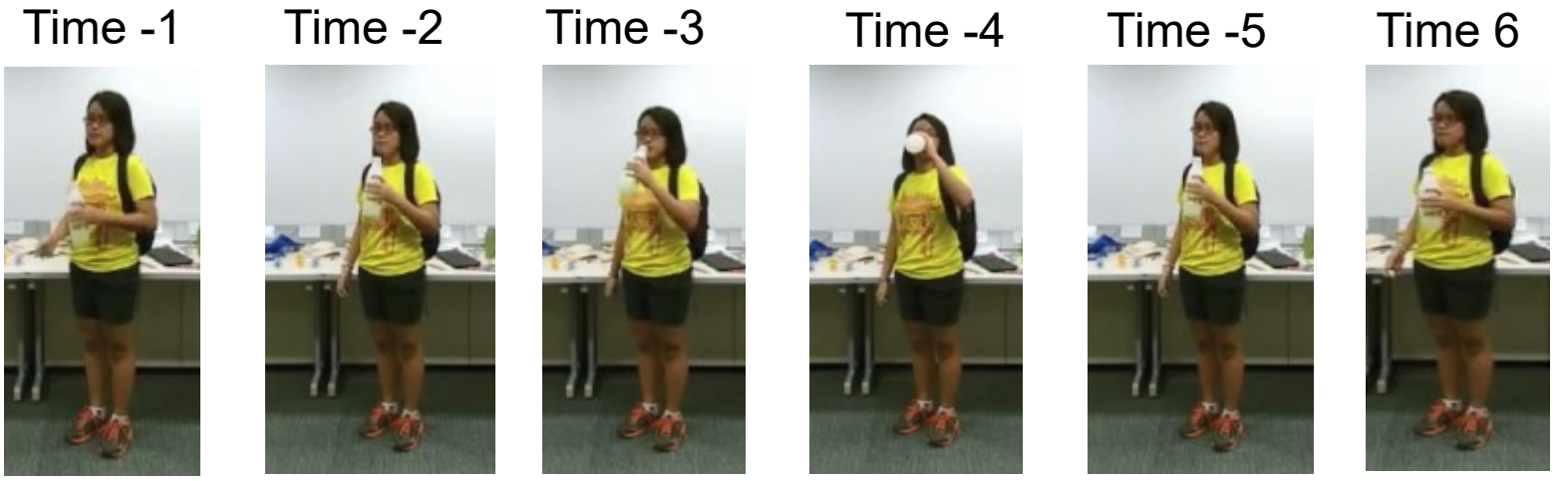
Example in Action Recognition



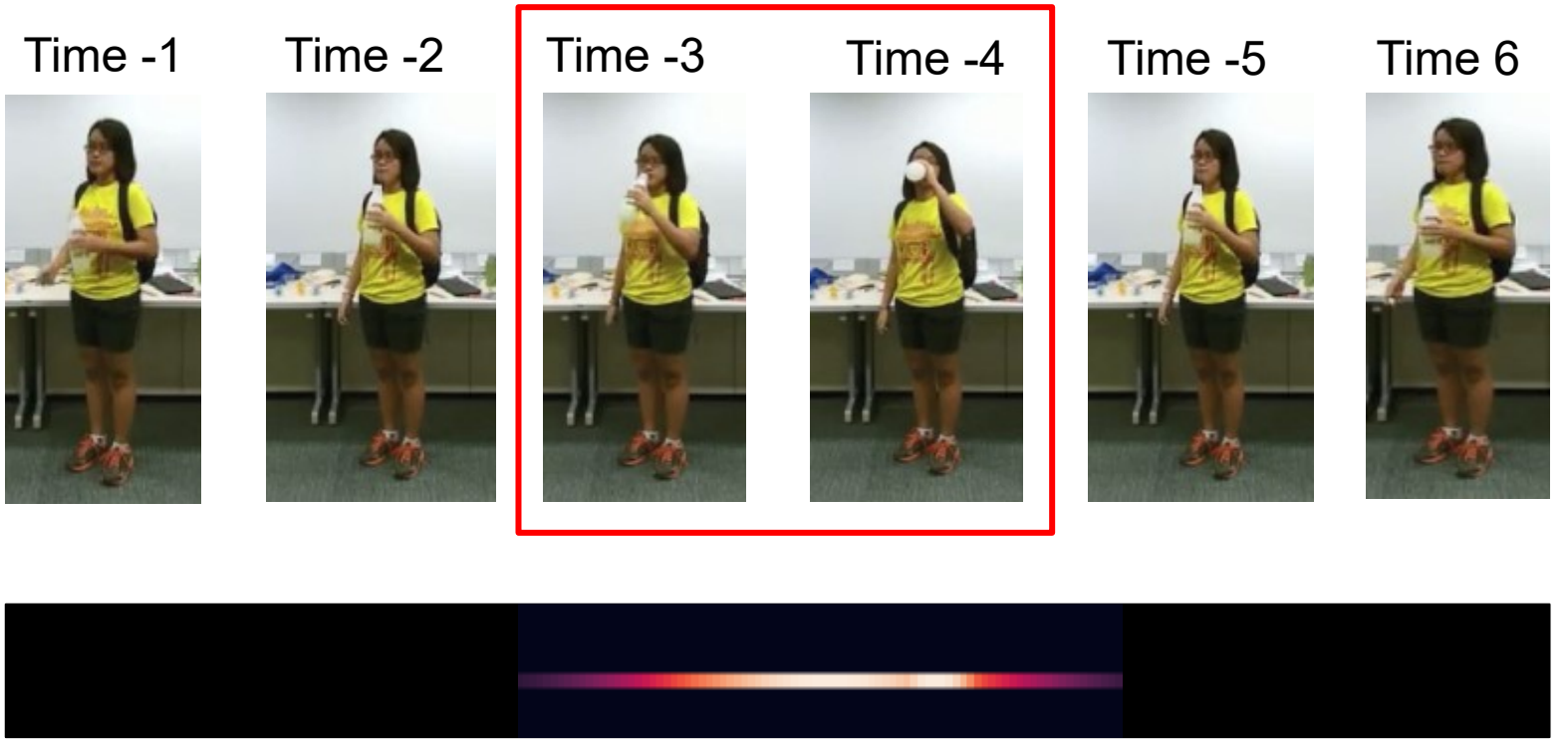
- Isn't this enough for an inference?

Focus in the **Spatial** space is required!

Example in Action Recognition



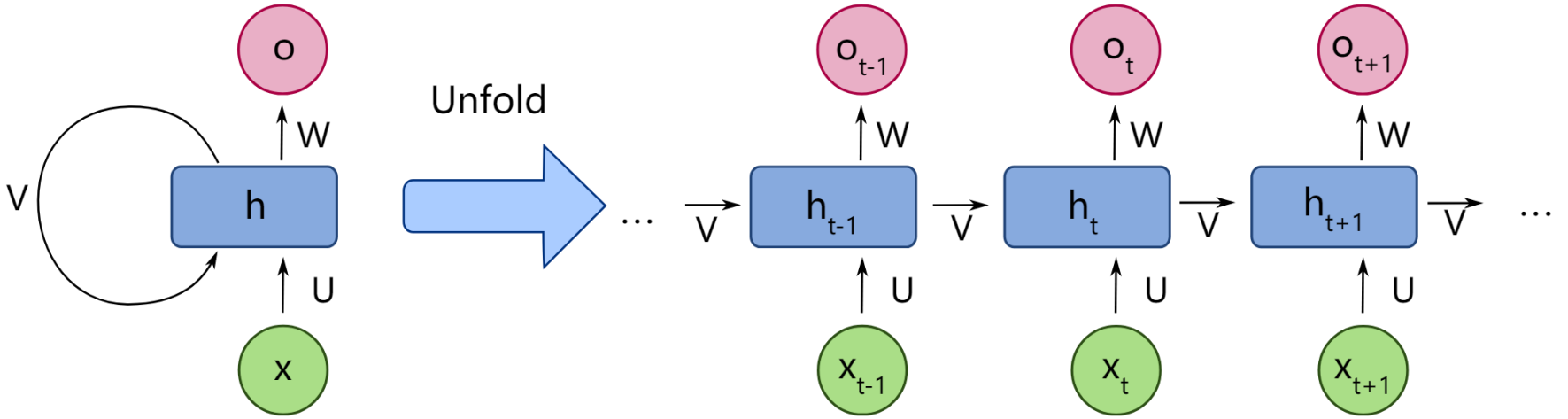
Example in Action Recognition



Section 2

Attention Modules

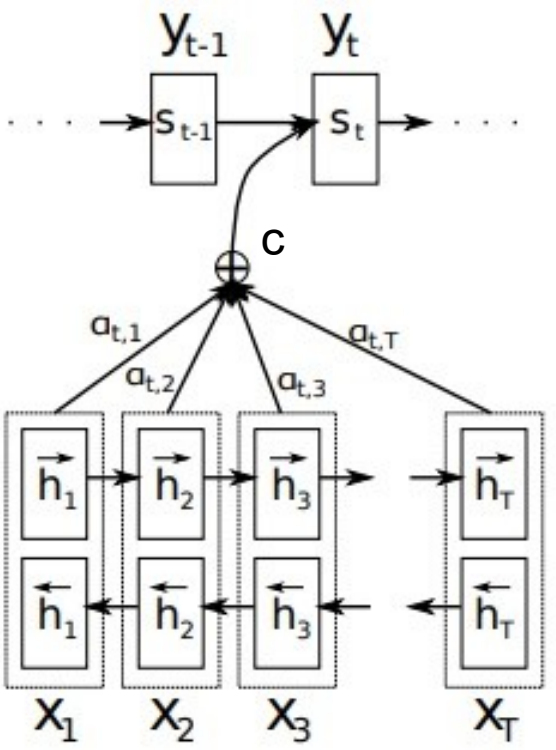
RNN

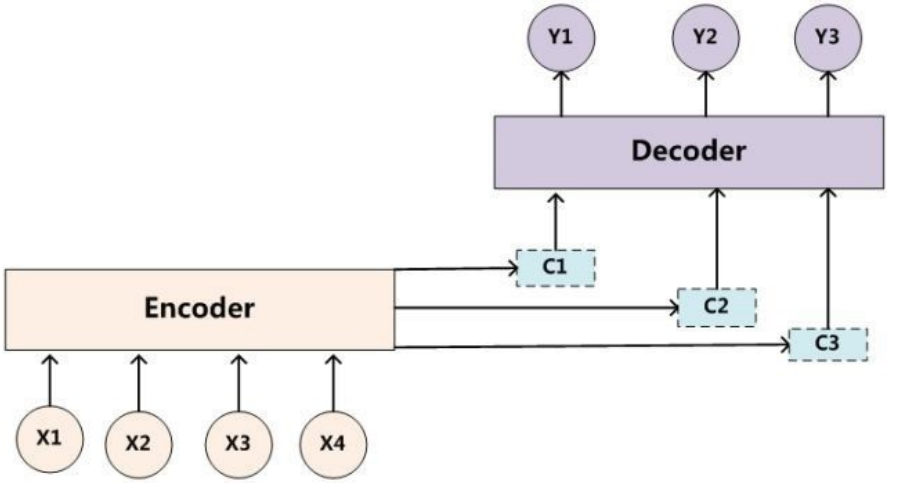


Drawbacks:

Can not learn long-term information.

Attention in RNN



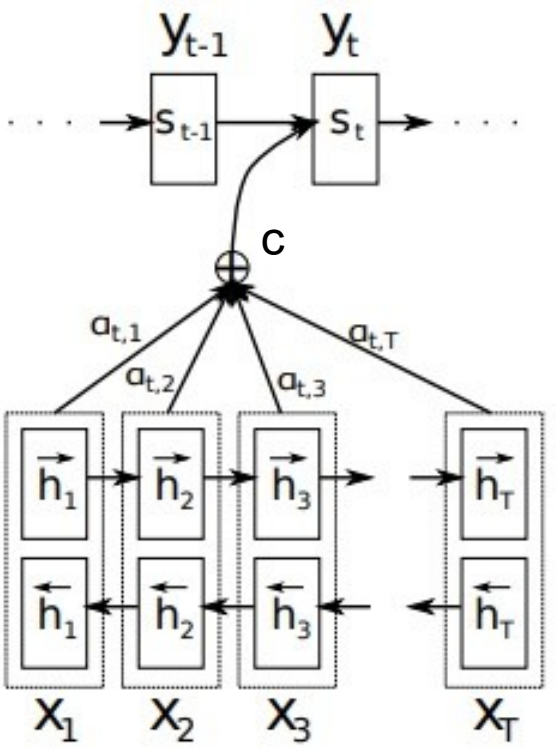


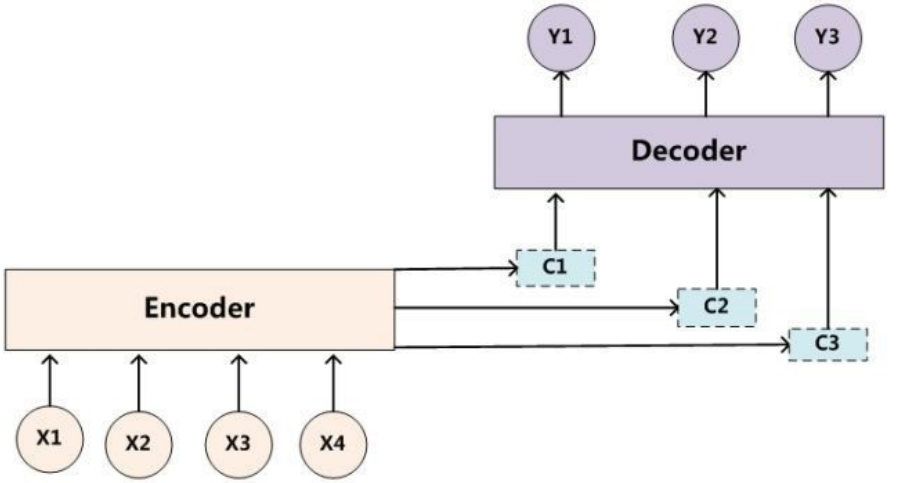
$$e_{ij} = a(s_{i-1}, h_j)$$
$$= v_a^T tanh(W_a s_{i-1} + U_a h_j)$$

$$a_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T} exp(e_{ik})}$$

a compute the correlation between the target s and input h

Attention in RNN





$$c_i = \sum_{j=1}^{T} a_{ij} h_j$$

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

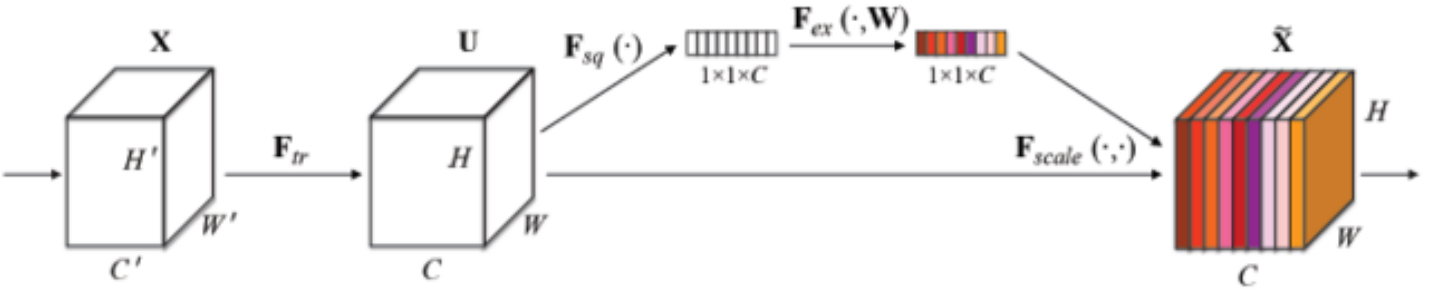
$$y_i = g(y_{i-1}, s_i, c_i)$$

Squeeze-and-Excitation Attention

Attention on Channels

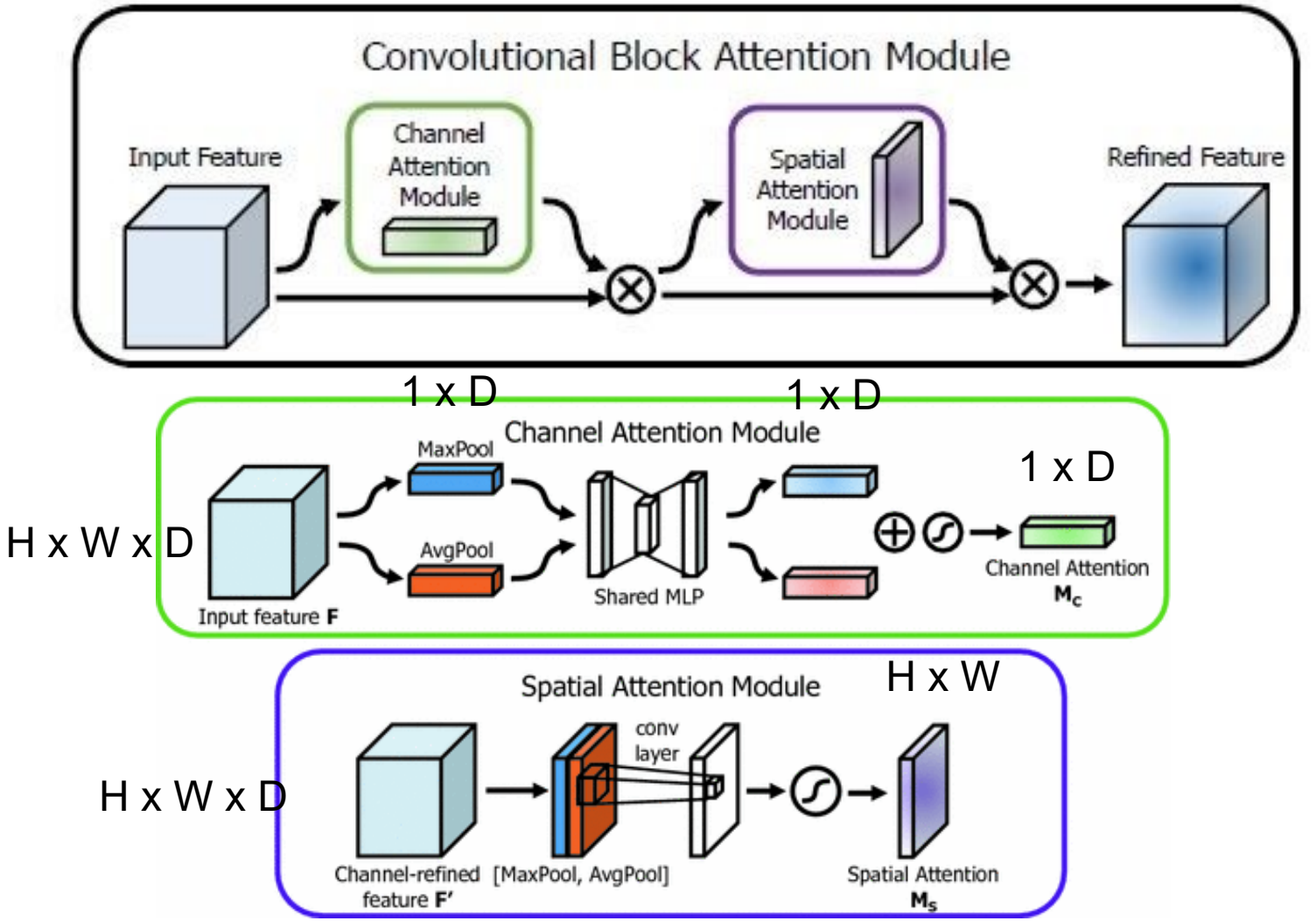
Squeeze: Average Pool

Excitation: FC + Sigmoid

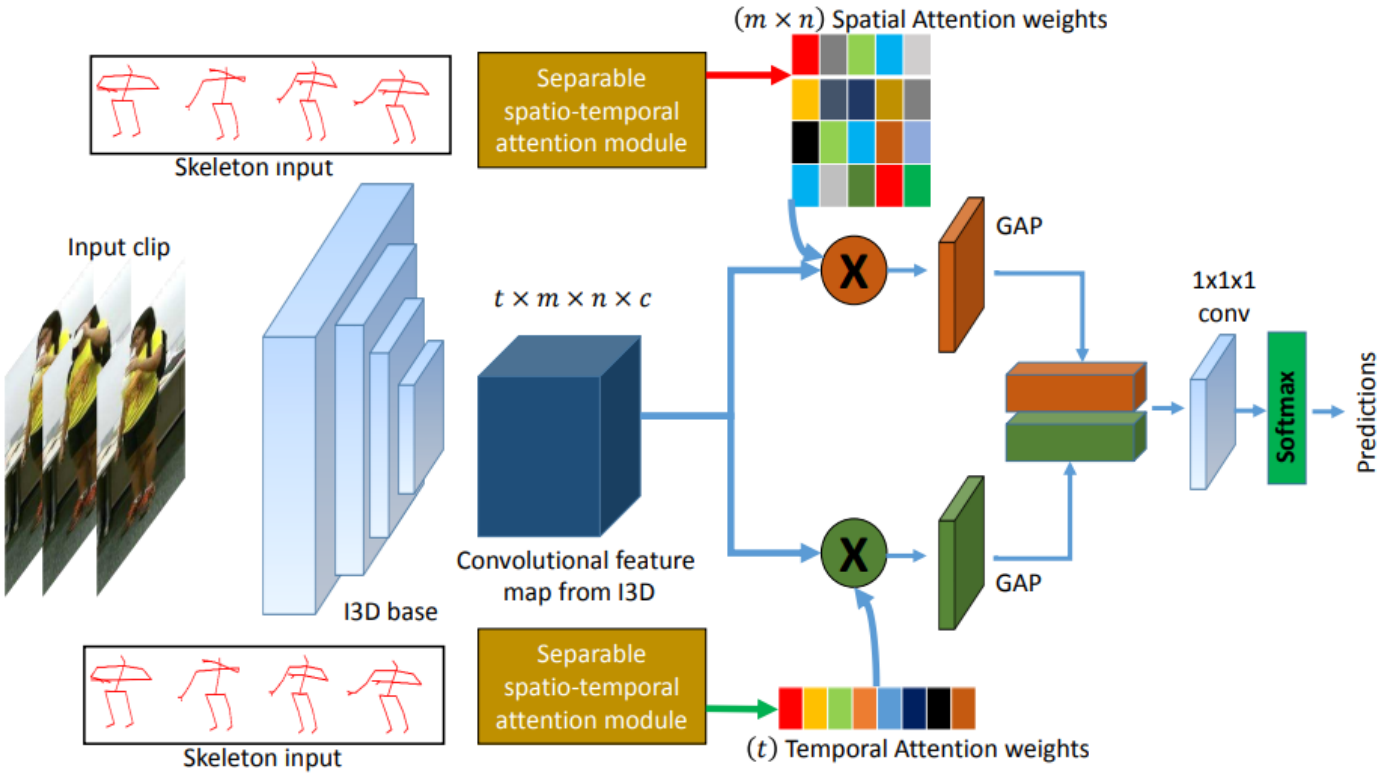


Convolutional Block Attention Module

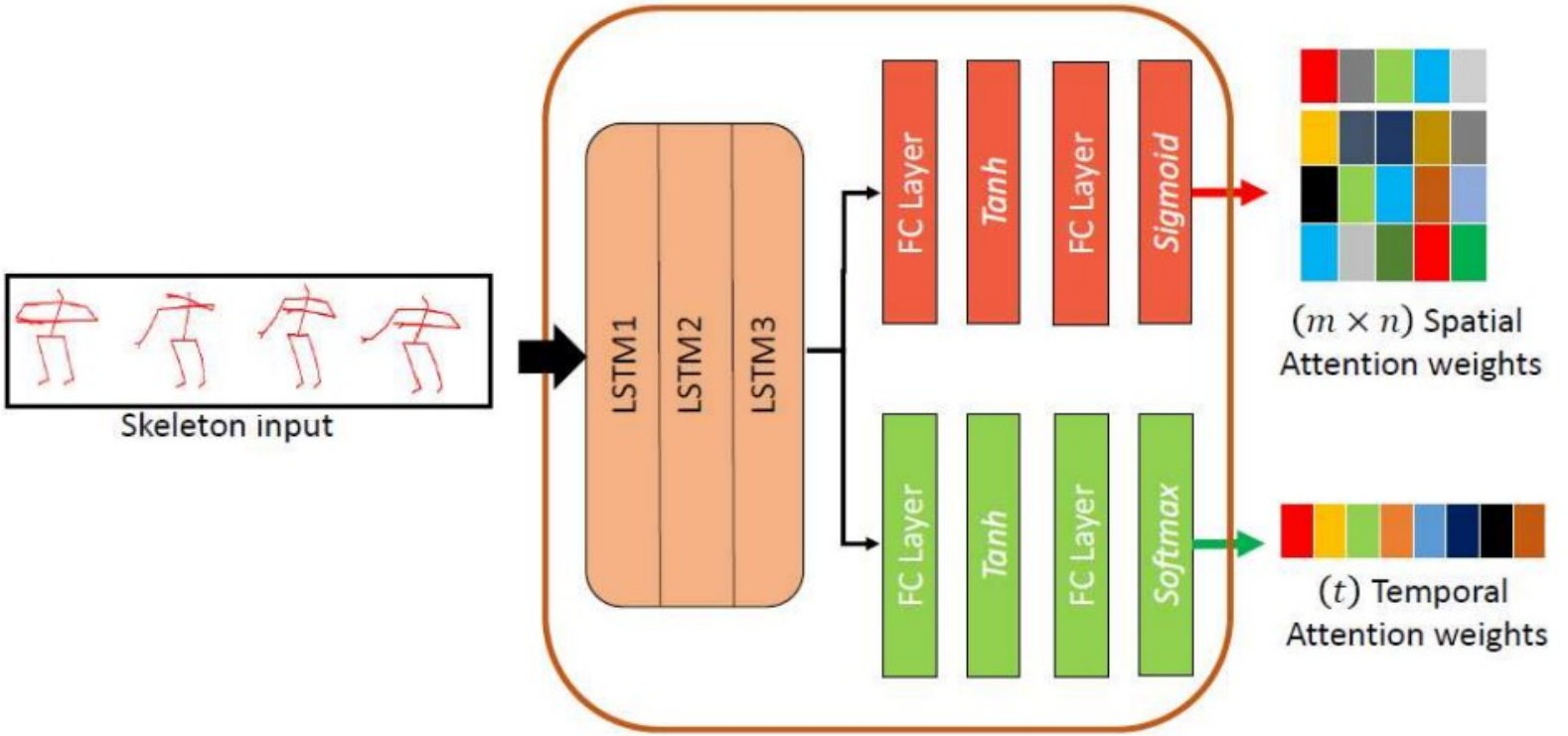
Channel + Spatial



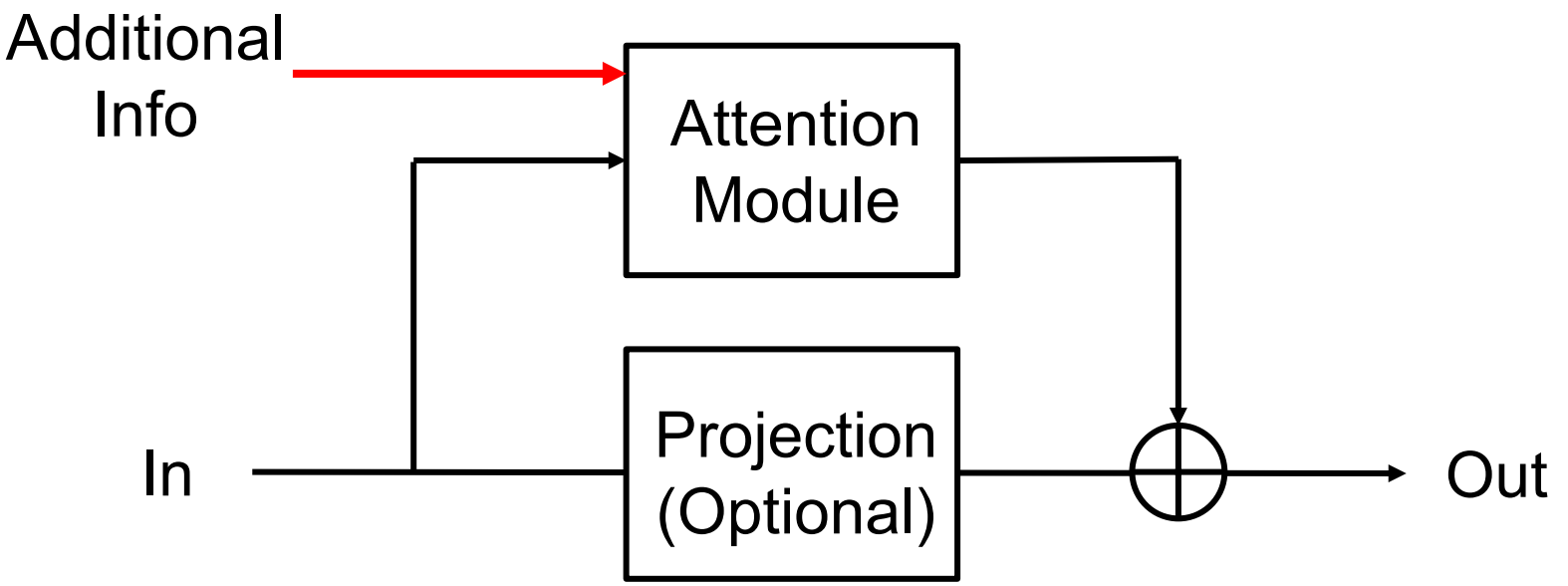
Spatial-Temporal Attention Network



Spatial-Temporal Attention Network



Attention Module Structure

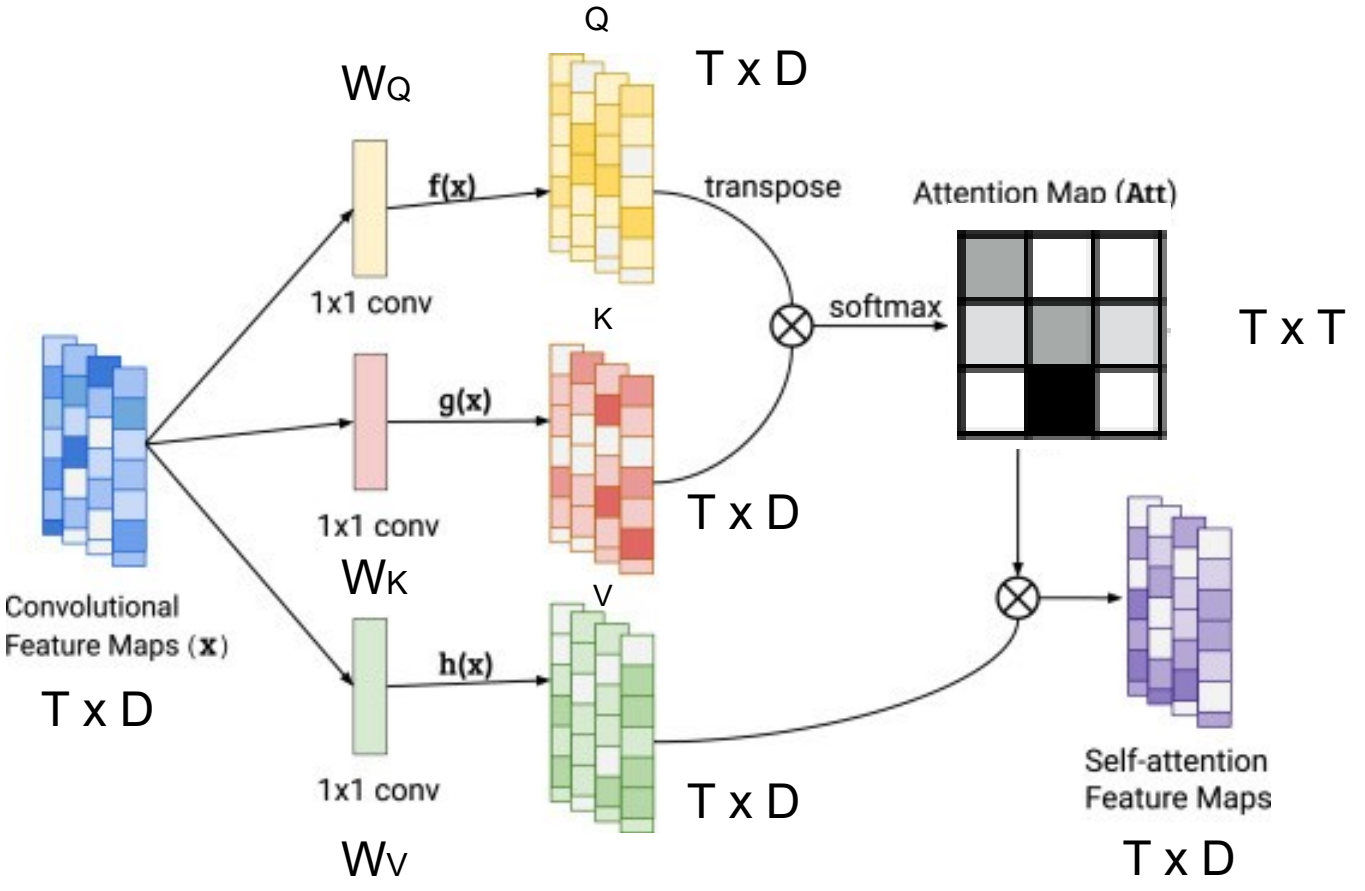


Section 3

Self-Attention

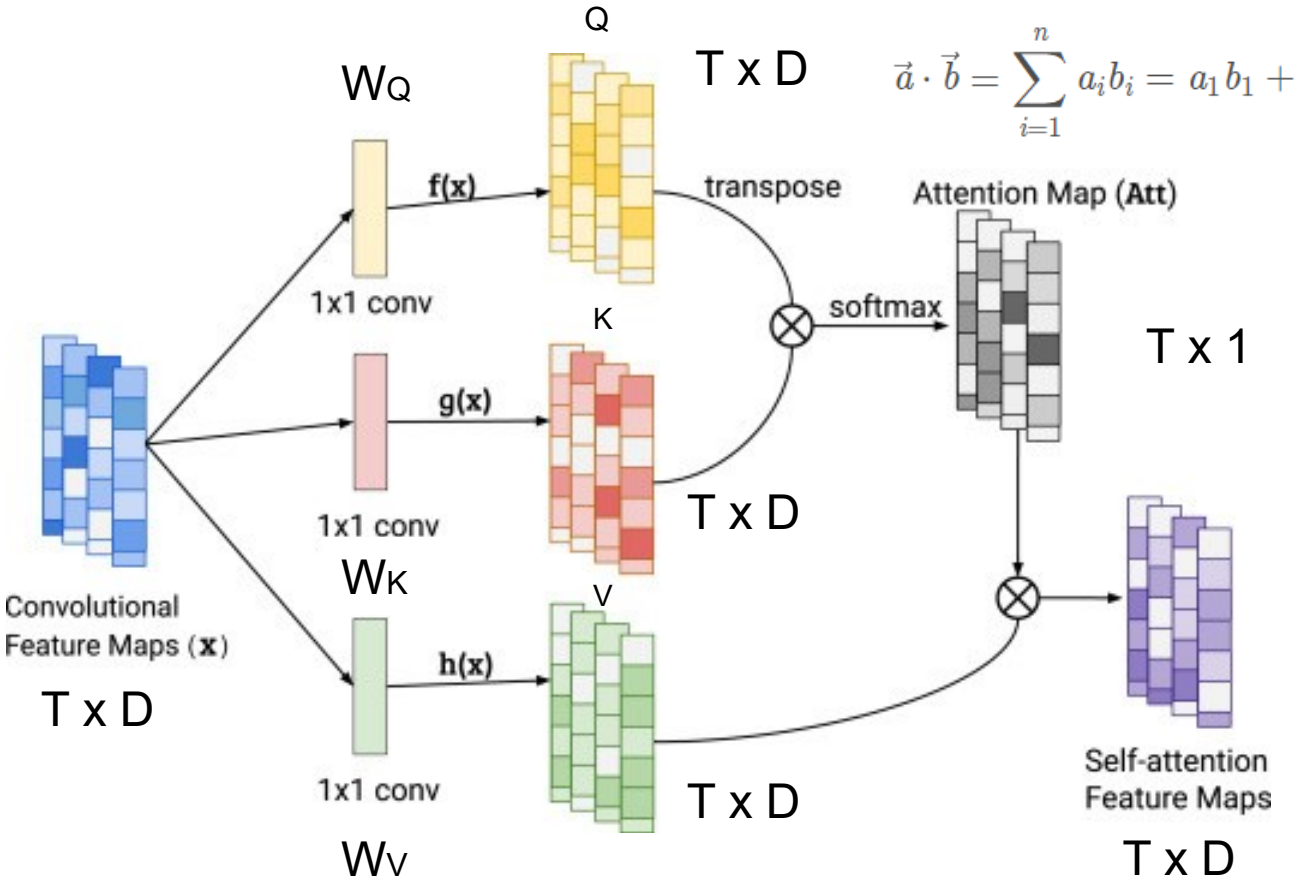
Self-Attention (1D)

Compute the Correlation in a sequence



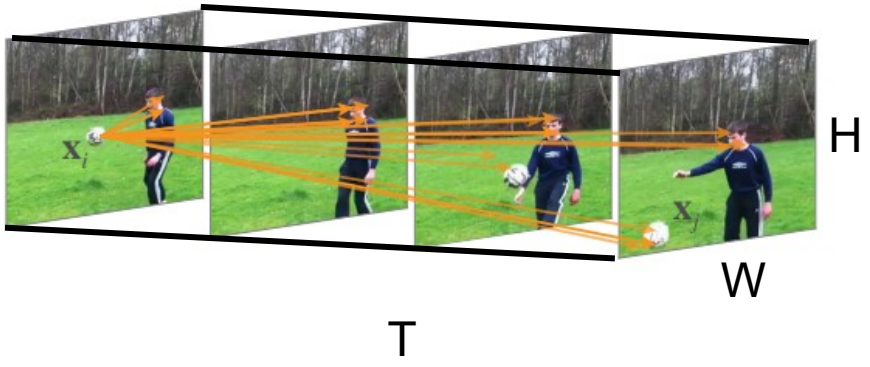
Self-Attention (1D)

Compute the Correlation in a sequence

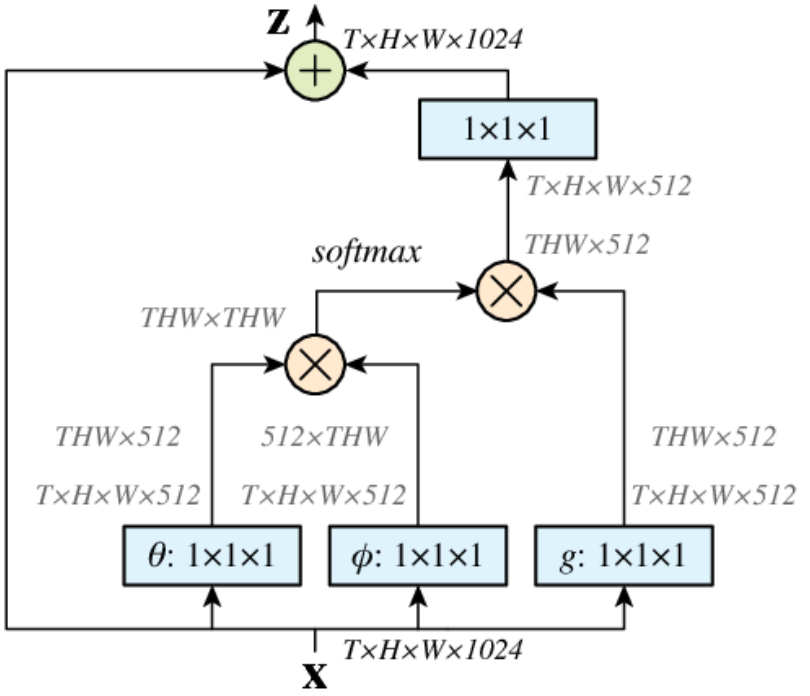


Dot product version

Application- Non local block (3D)



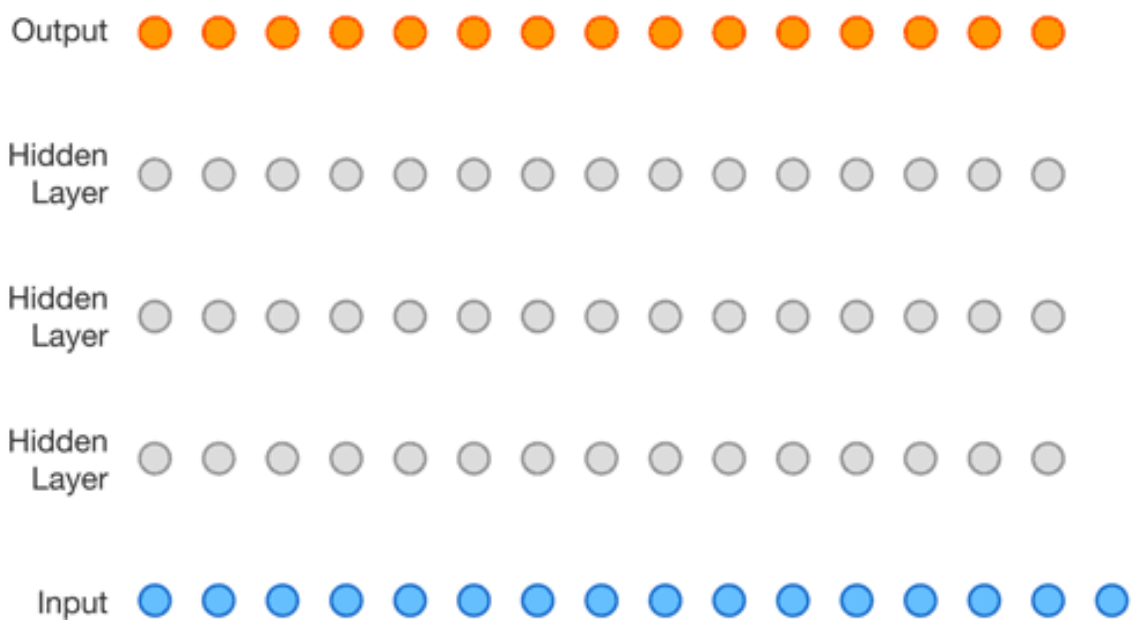
Statistic the Global information
as complementary to the local embeddings



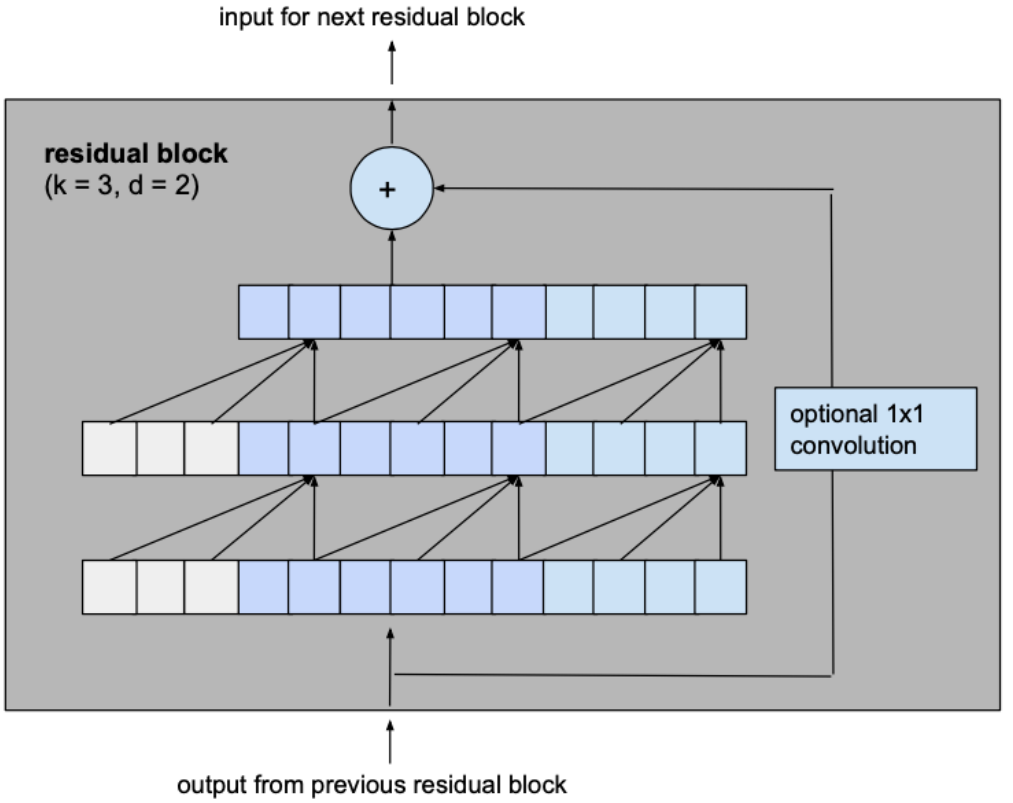
I3D Network

Video

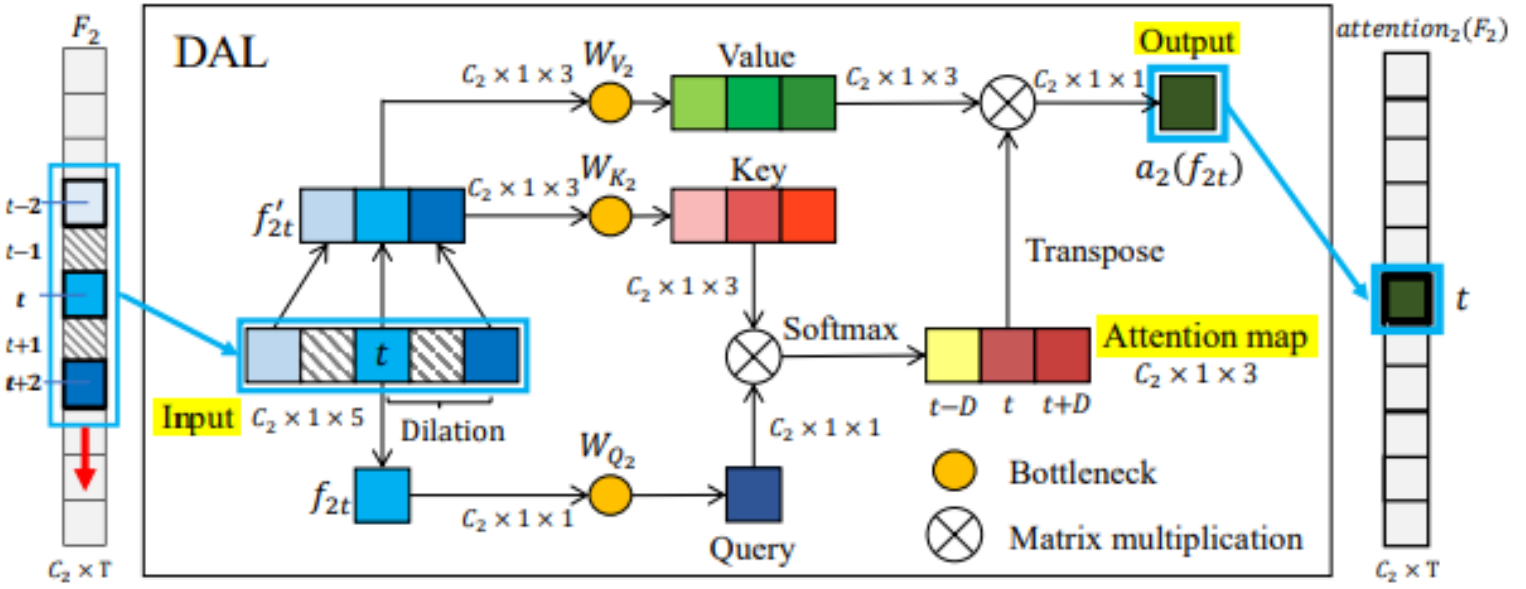
1 Dimensional Convolution



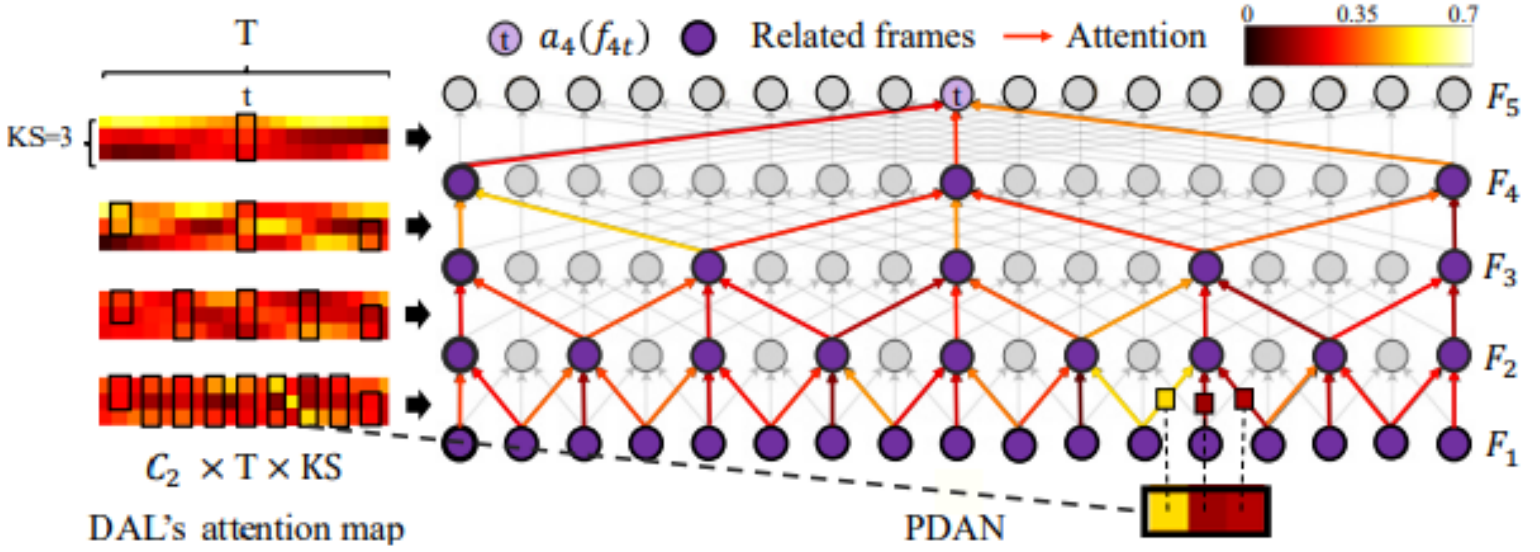
1 Dimensional Convolution



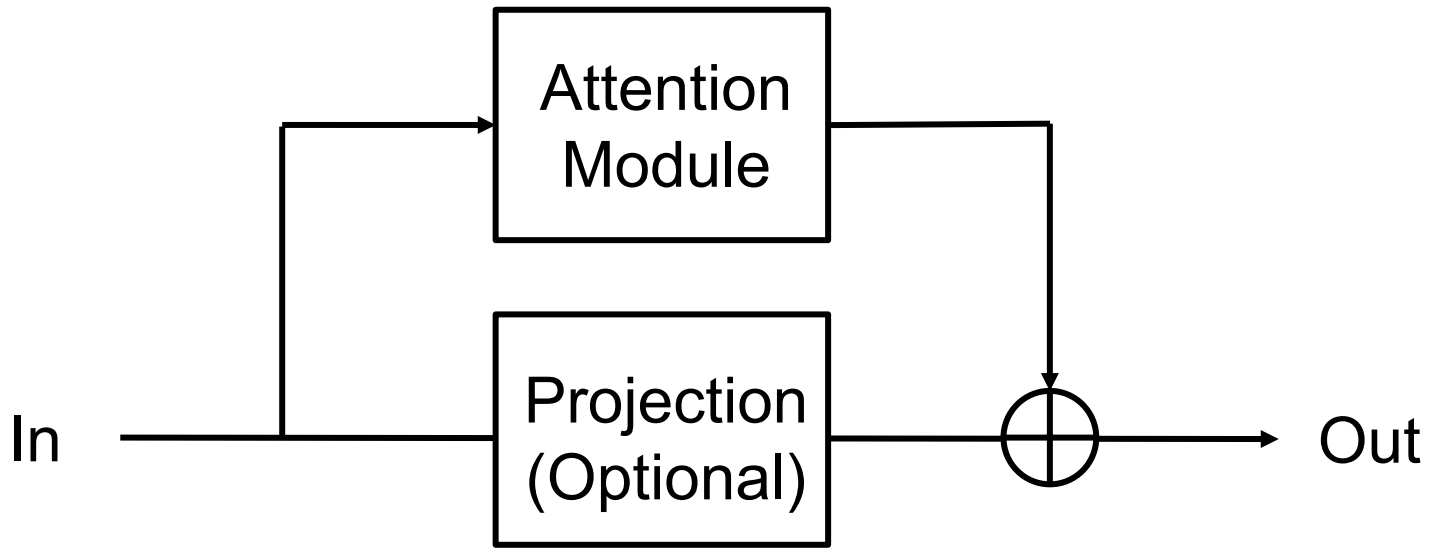
Application- PDAN (1D)



Application- PDAN (1D)



Attention Mechanism Structure



Section 4

Transformer

History

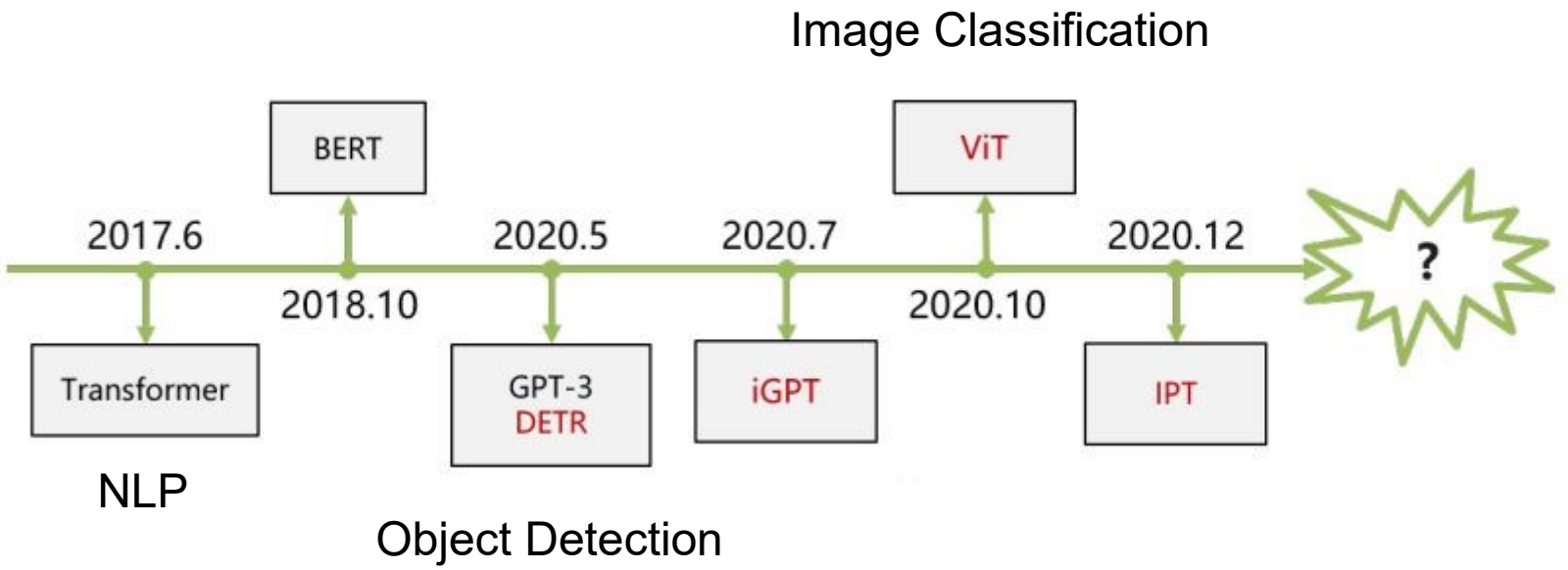
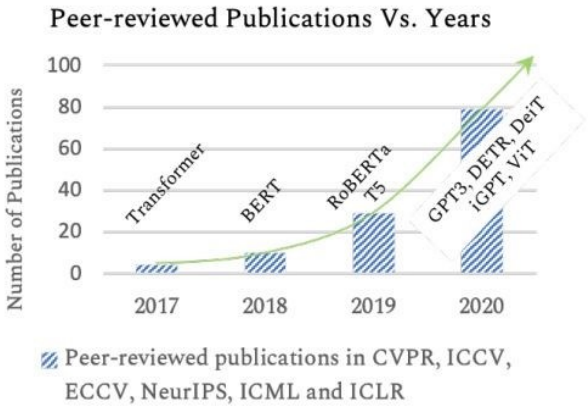
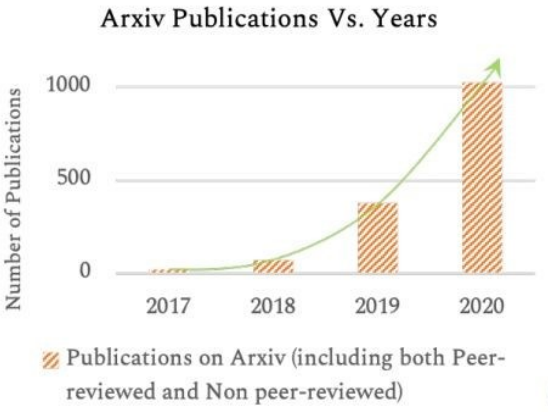
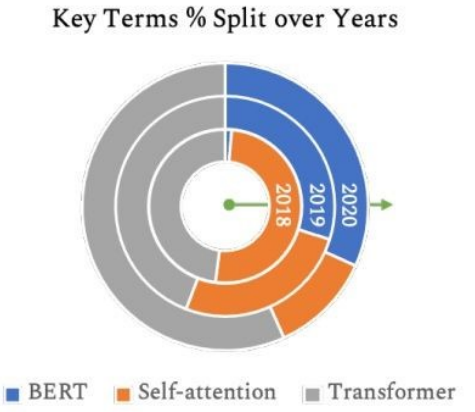


Image Classification
BERT
ViT
2017.6
2020.5
2020.7
2020.12
?
2018.10
2020.10
Transformer
GPT-3
DETR
iGPT
IPT
NLP
Object Detection

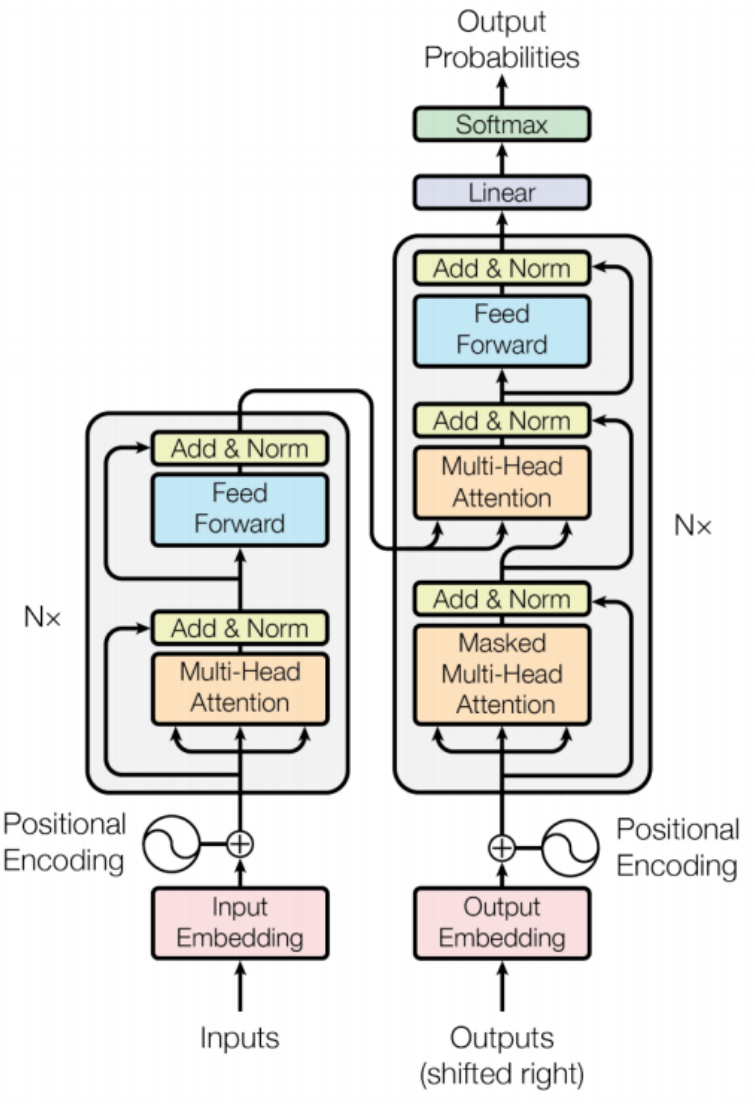

Peer-reviewed Publications Vs. Years
Number of Publications
100
80
60
40
20
0
2017
2018
2019
2020
Transformer
BERT
RoBERTa
T5
GPT3, DETR, DeiT
iGPT, ViT
Peer-reviewed publications in CVPR, ICCV, ECCV, NeurIPS, ICML and ICLR


Arxiv Publications Vs. Years
Number of Publications
1000
500
0
2017
2018
2019
2020
Publications on Arxiv (including both Peer-reviewed and Non peer-reviewed)


Key Terms % Split over Years
2018
2019
2020
BERT
Self-attention
Transformer

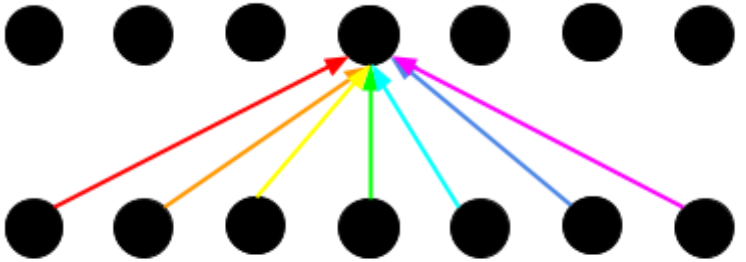
Transformer

- Positional Encoding
- Multi-head attention
- Feed Forward
- Outputs

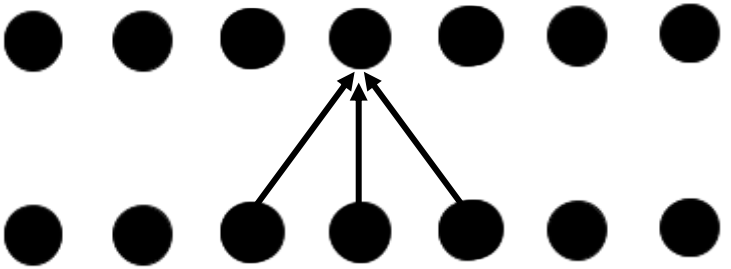


Transformer and Convolution

Transformer layer



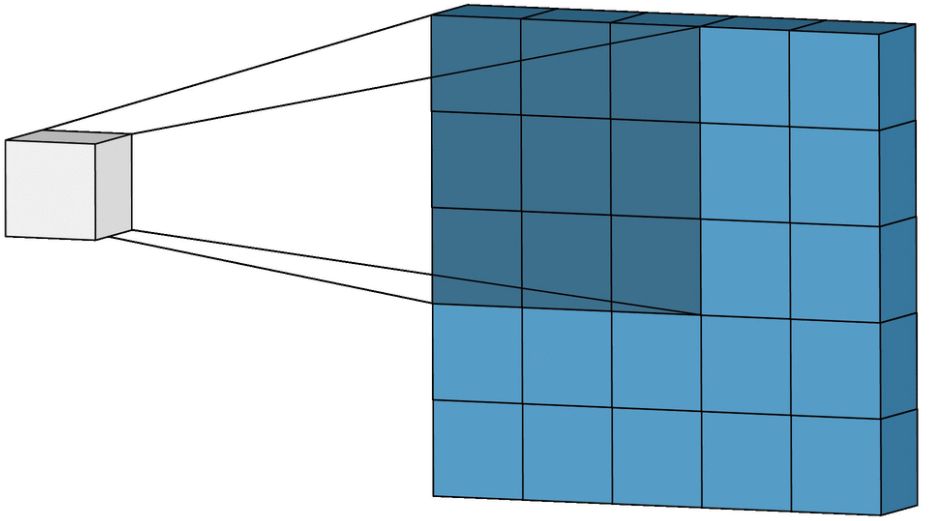
Convolution layer



Pros:
Global relation
Attention enhanced

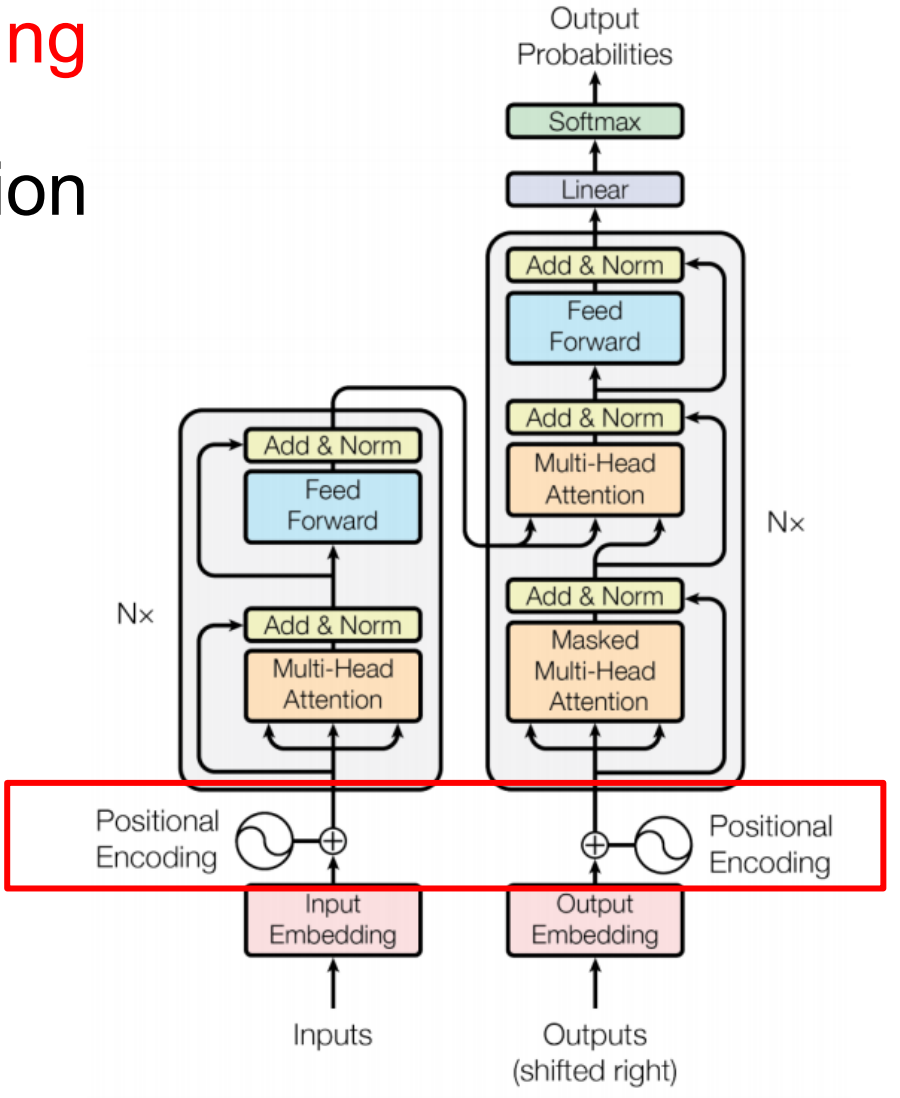
Cons:
More Flops
Loss location info

Missing location info



Transformer

- Positional Encoding
- Multi-head attention
- Feed Forward
- Outputs



Positional Encoding

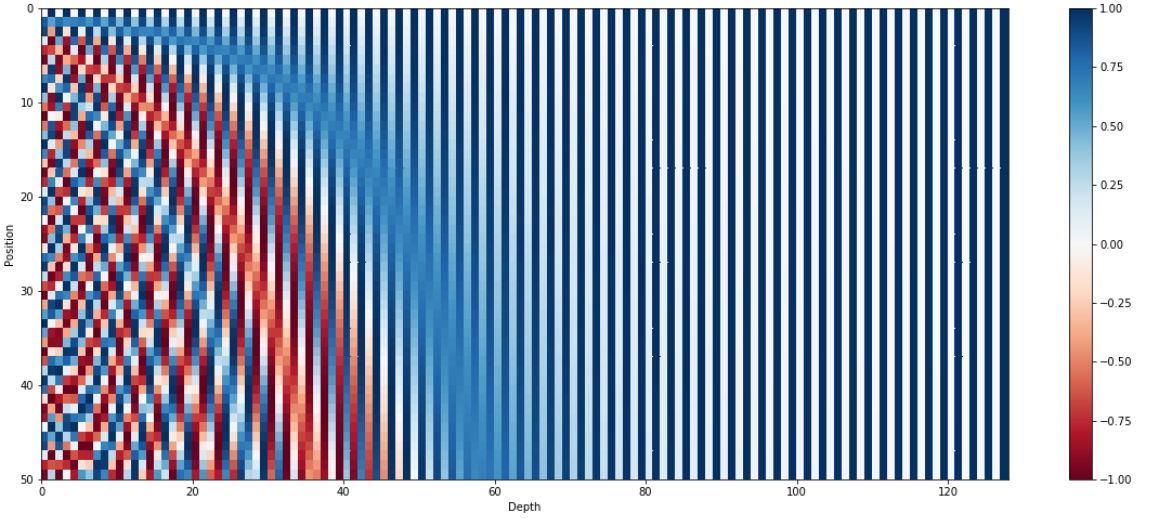
Indicating the positional information

0:	0	0	0	0	8:	1	0	0	0
1:	0	0	0	1	9:	1	0	0	1
2:	0	0	1	0	10:	1	0	1	0
3:	0	0	1	1	11:	1	0	1	1
4:	0	1	0	0	12:	1	1	0	0
5:	0	1	0	1	13:	1	1	0	1
6:	0	1	1	0	14:	1	1	1	0
7:	0	1	1	1	15:	1	1	1	1

Positional Encoding

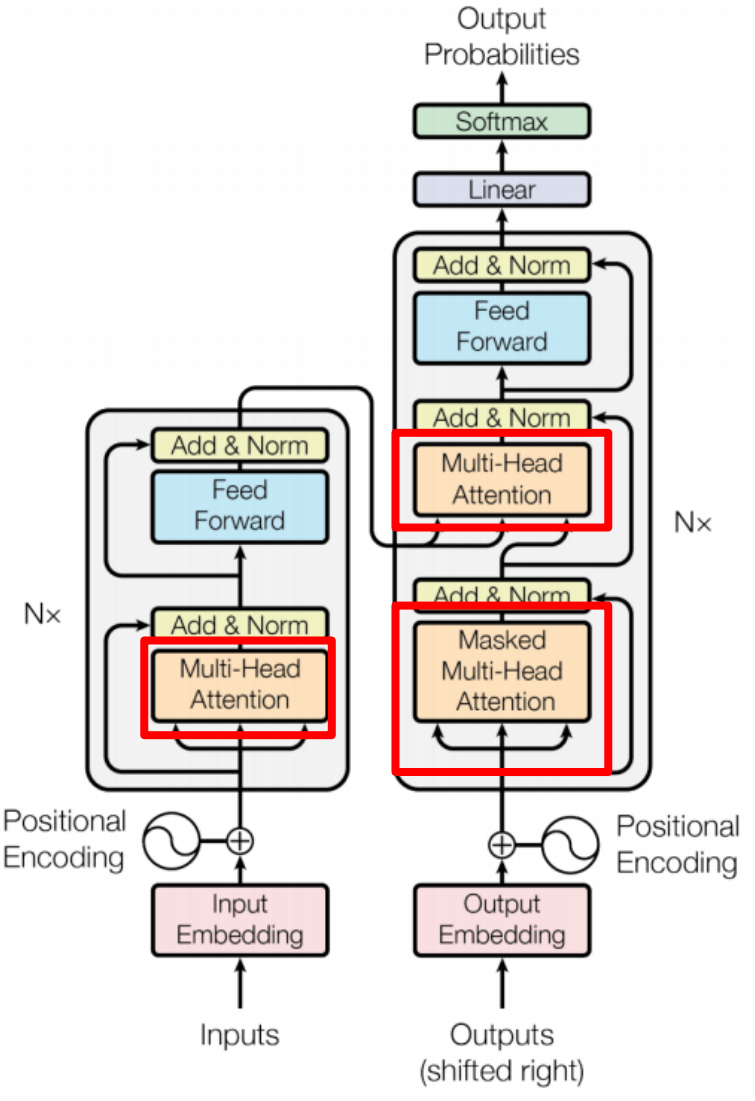
$$\overrightarrow{p_t}^{(i)} = f(t)^{(i)} := \begin{cases} \sin(\omega_k . t), & \text{if } i = 2k \\ \cos(\omega_k . t), & \text{if } i = 2k+1 \end{cases} \qquad \omega_k = \frac{1}{10000^{2k/d}}$$

$$\overrightarrow{p_t} = \begin{bmatrix} \sin(\omega_1 . t) \\ \cos(\omega_1 . t) \\ \\ \sin(\omega_2 . t) \\ \cos(\omega_2 . t) \\ \\ \vdots \\ \\ \sin(\omega_{d/2} . t) \\ \cos(\omega_{d/2} . t) \end{bmatrix}_{d \times 1}$$

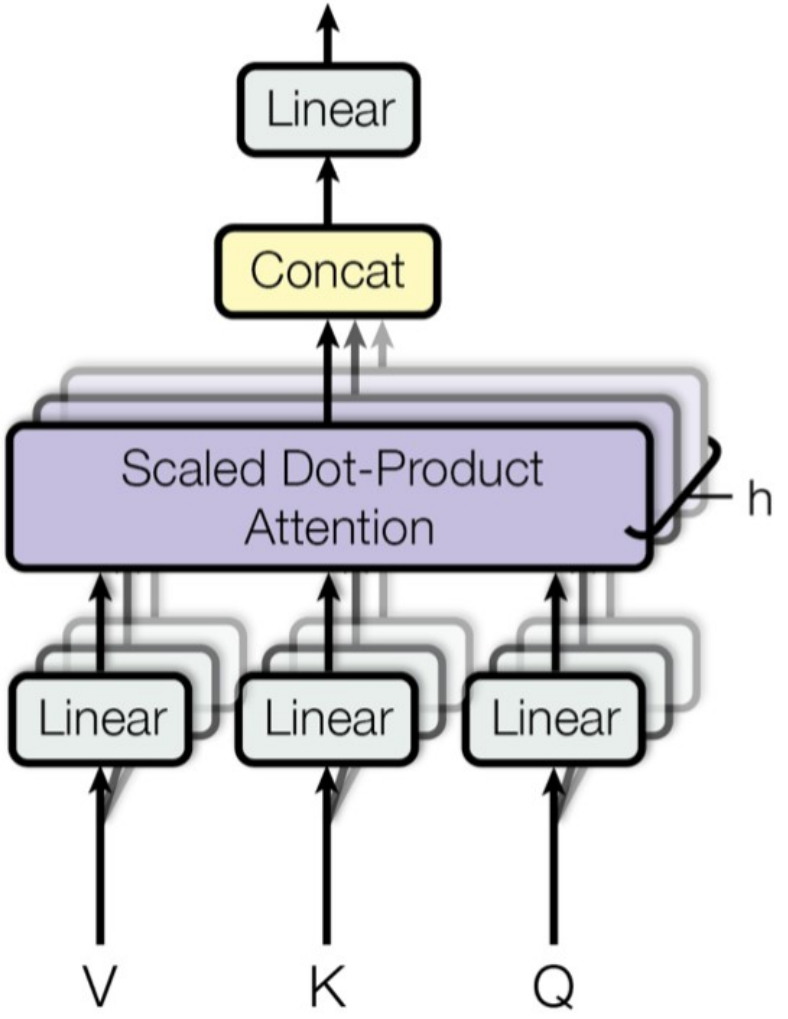


Transformer

- Positional Encoding
- Multi-head attention
- Feed Forward
- Outputs

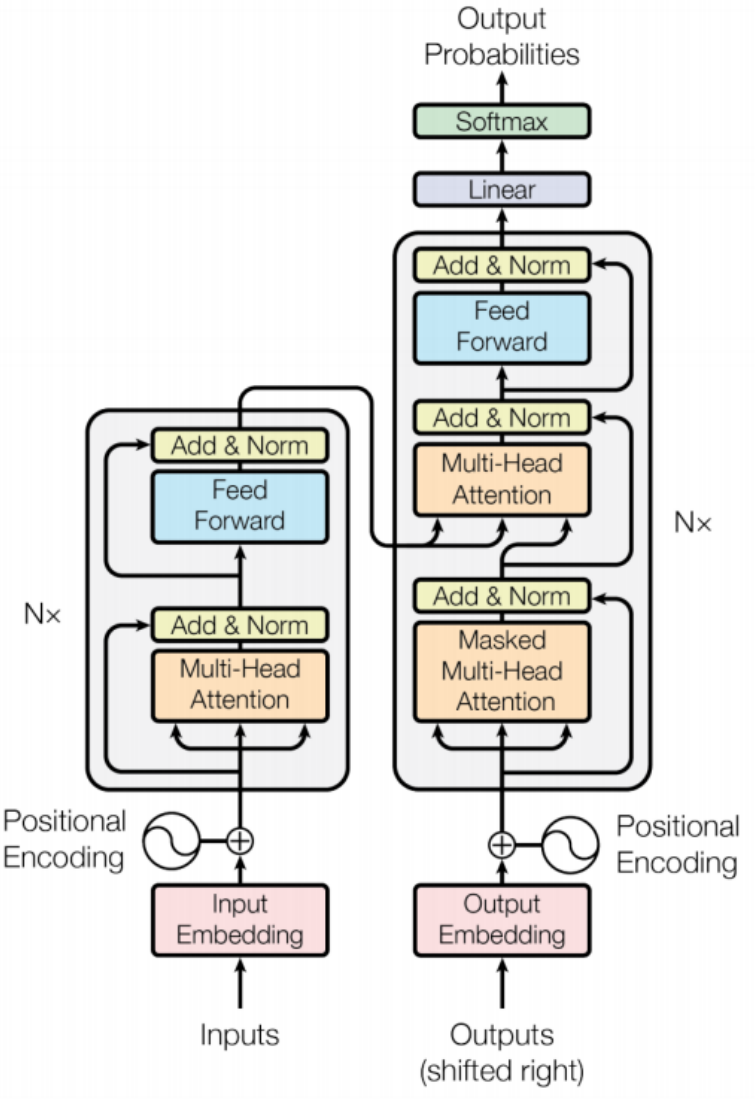


Multi-head Attention



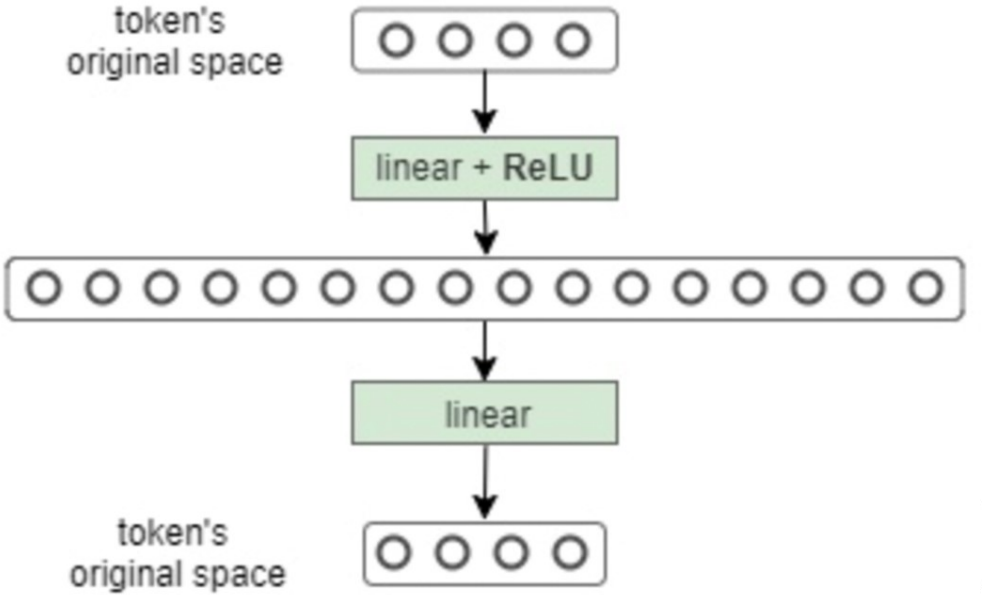
Transformer

- Positional Encoding
- Multi-head attention
- Feed Forward
- Outputs



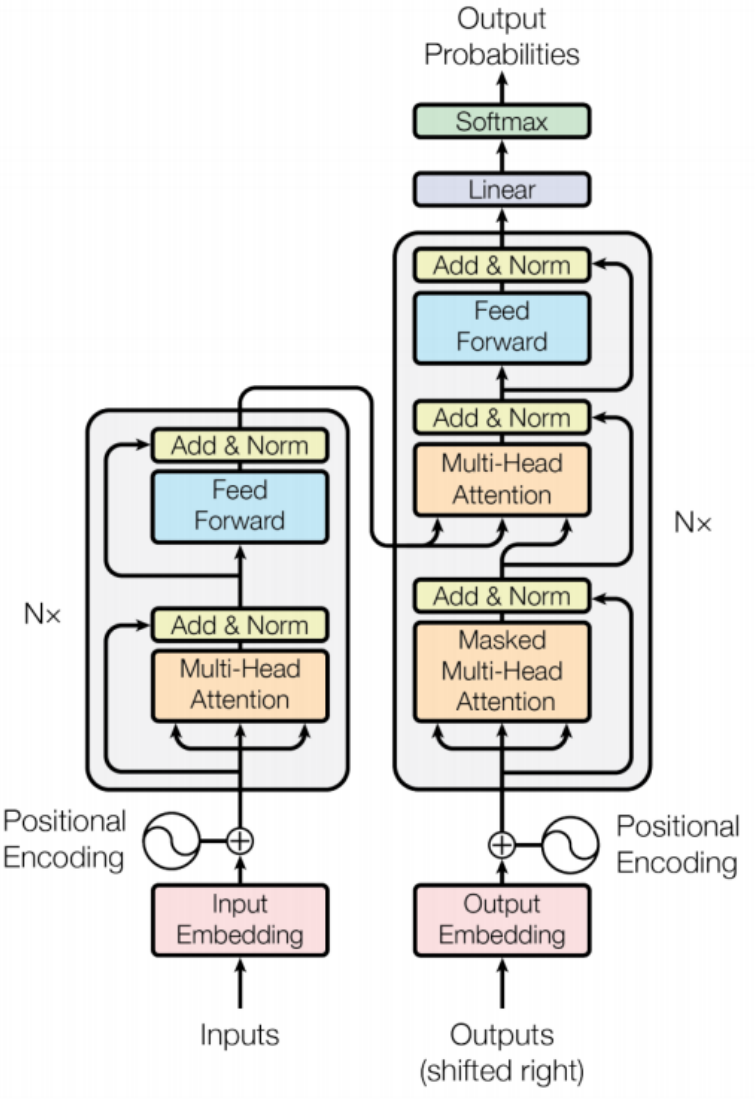
Transformer

Project the feature into a larger space, for extracting the features easier (Similar to SVM)

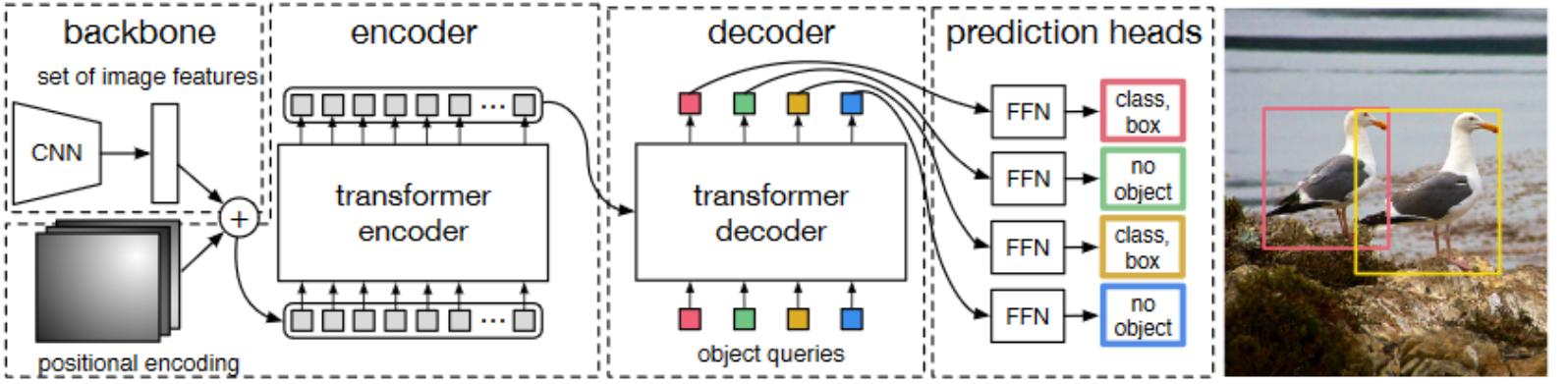


Transformer

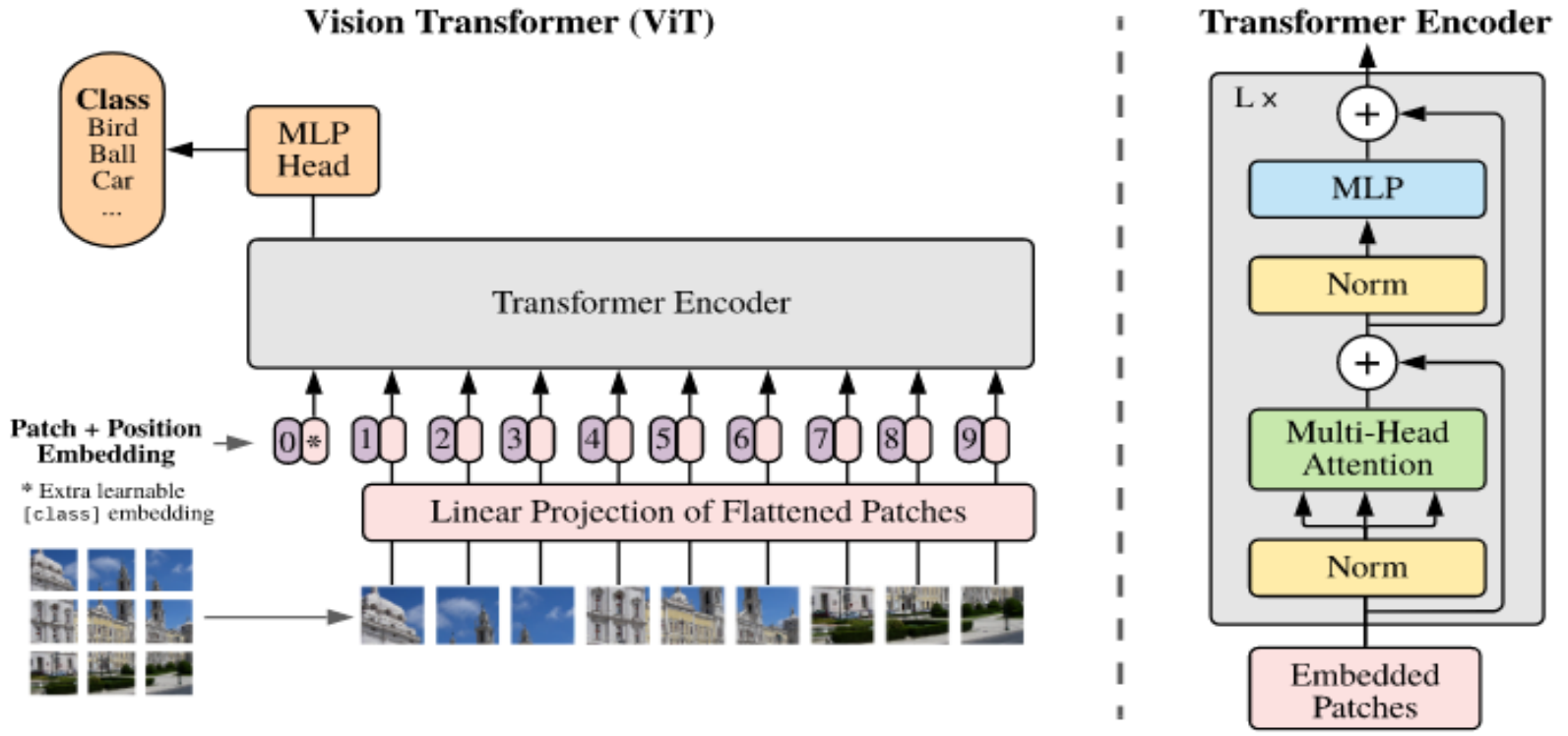
- Positional Encoding
- Multi-head attention
- Feed Forward
- Outputs



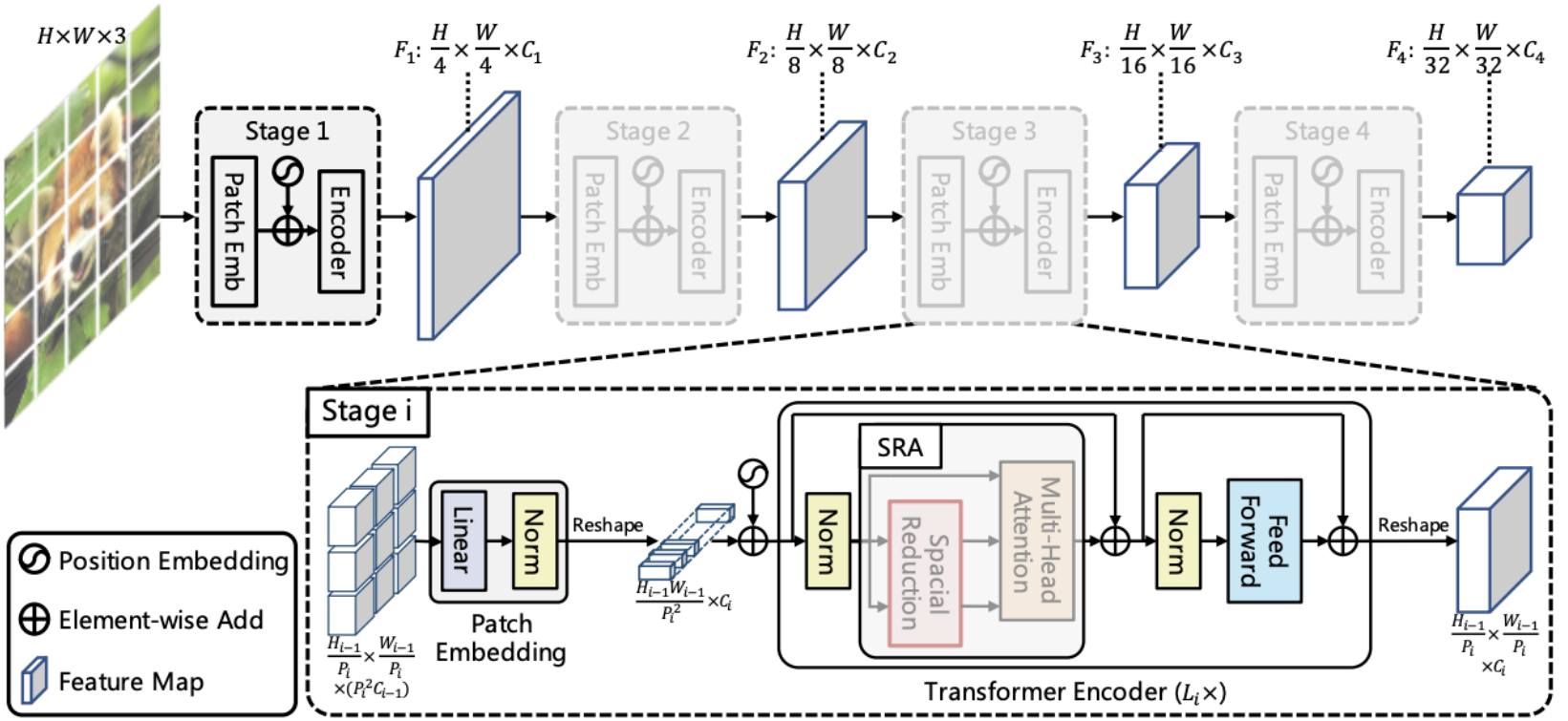
DETR


backbone
set of image features
CNN
positional encoding
encoder
transformer encoder
decoder
transformer decoder
object queries
prediction heads
FFN
FFN
FFN
FFN
class, box
no object
class, box
no object

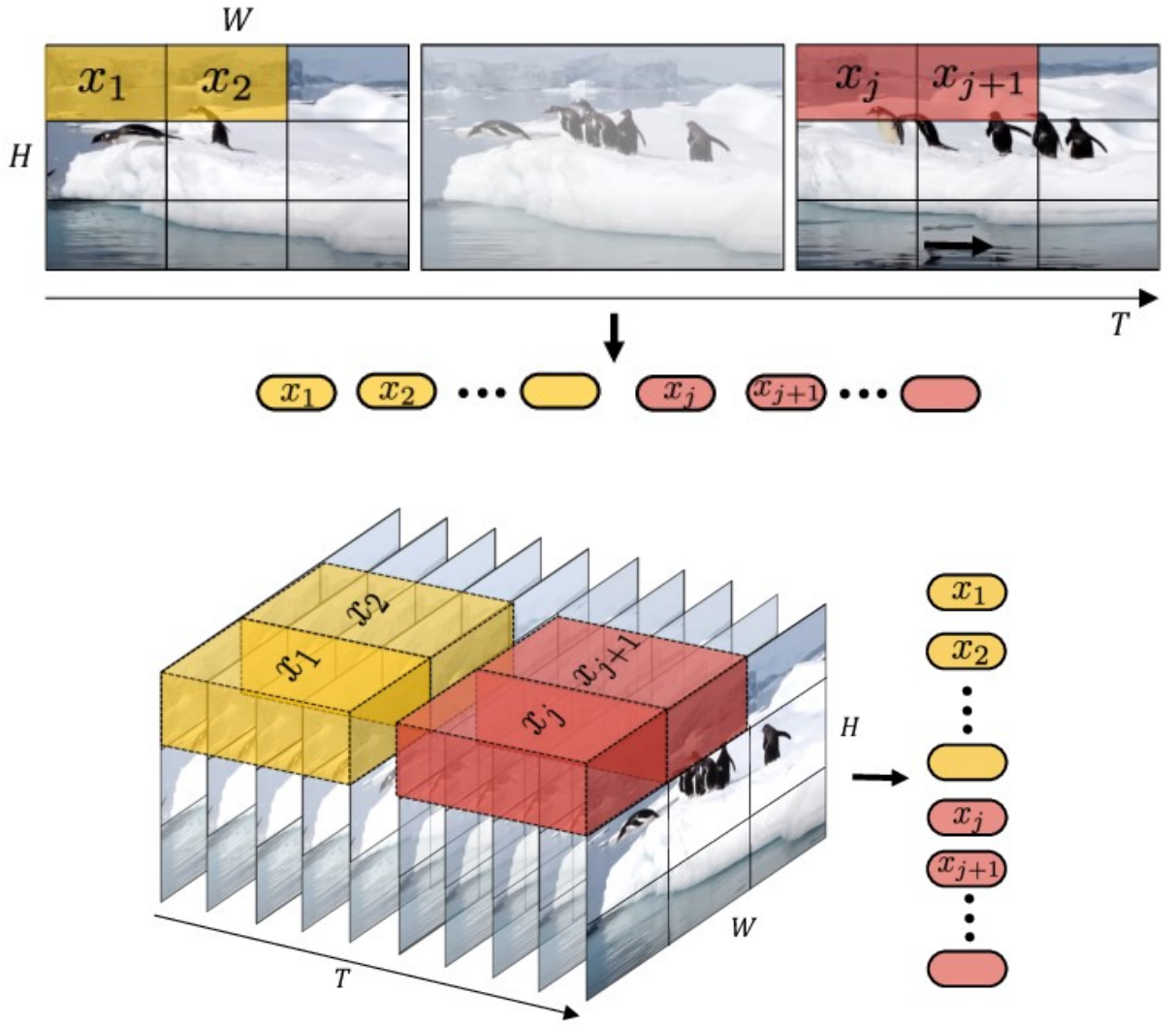
Vision Transformer



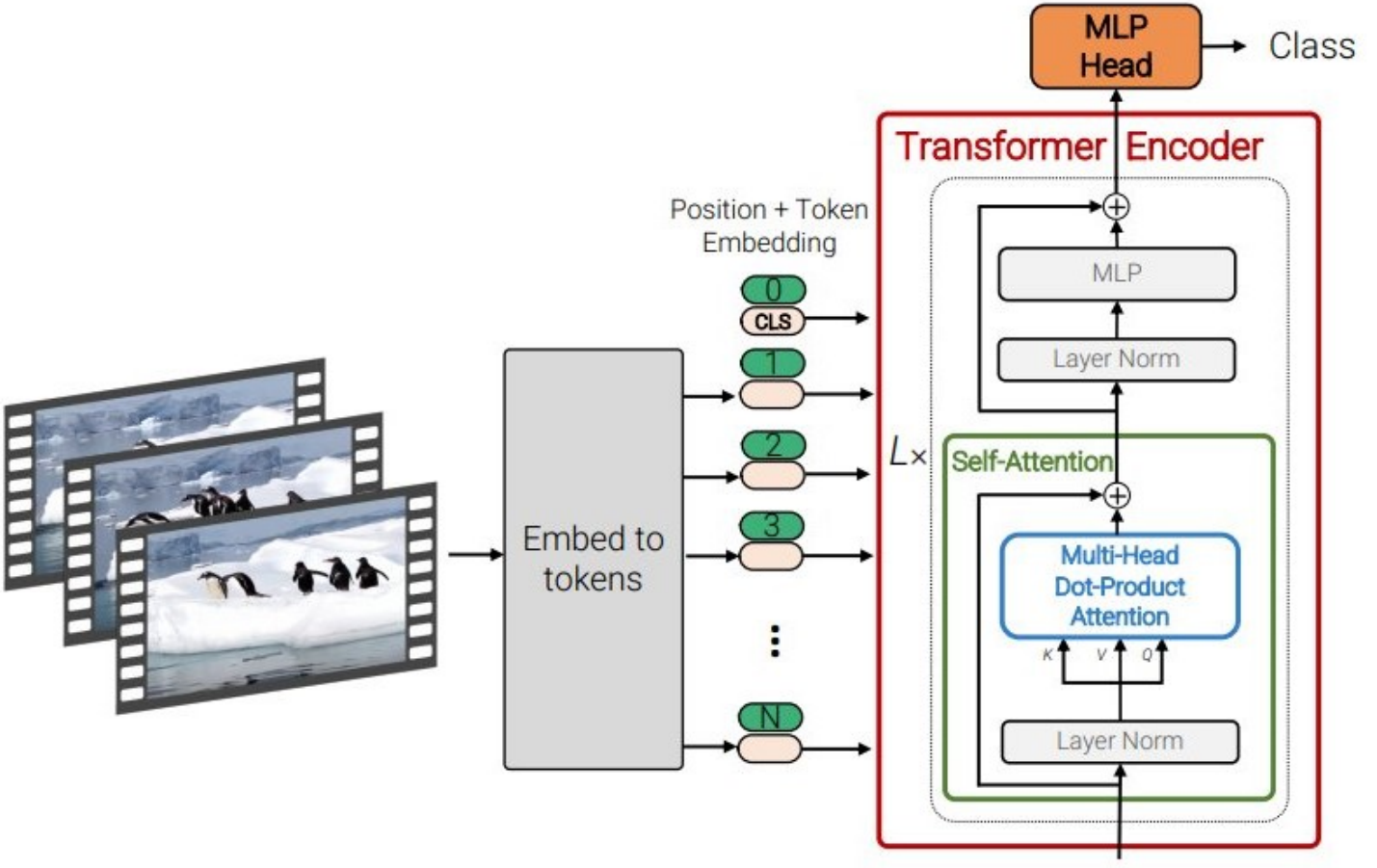
Pyramid Vision Transformer



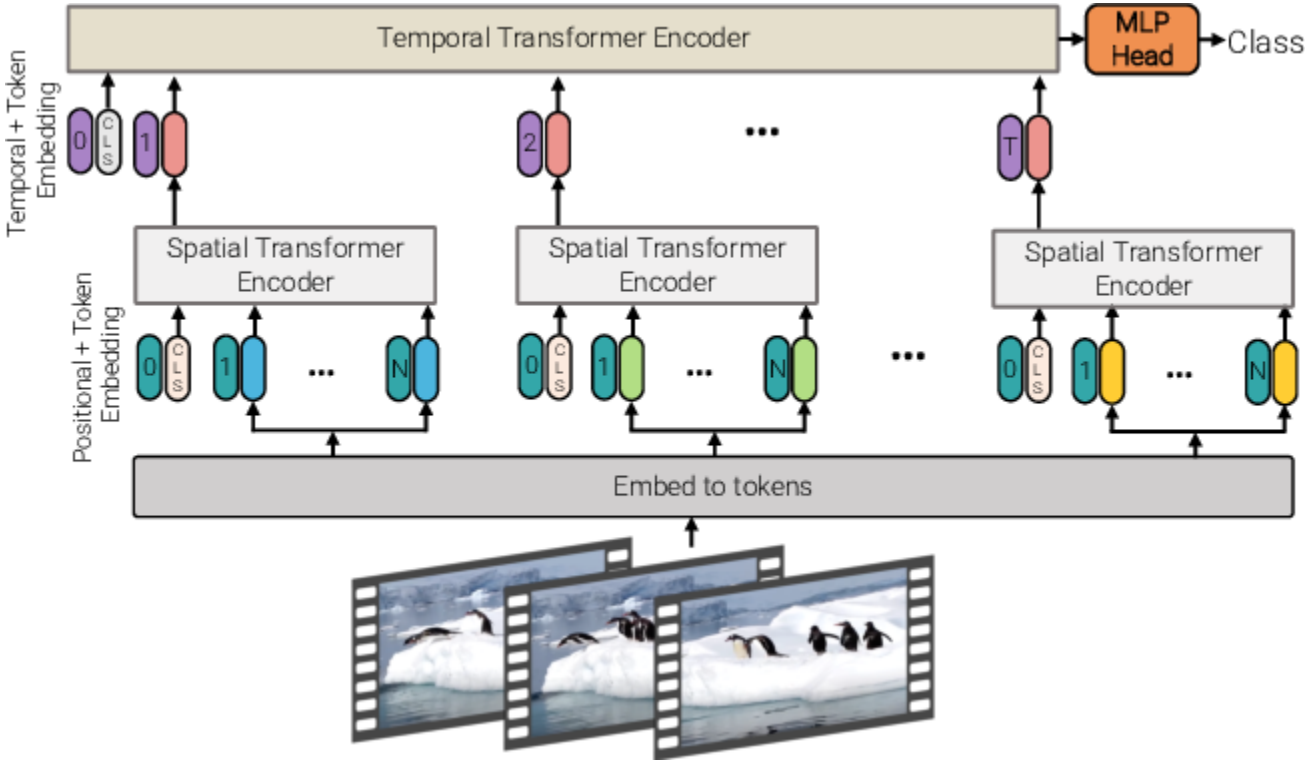
ViVit - patch



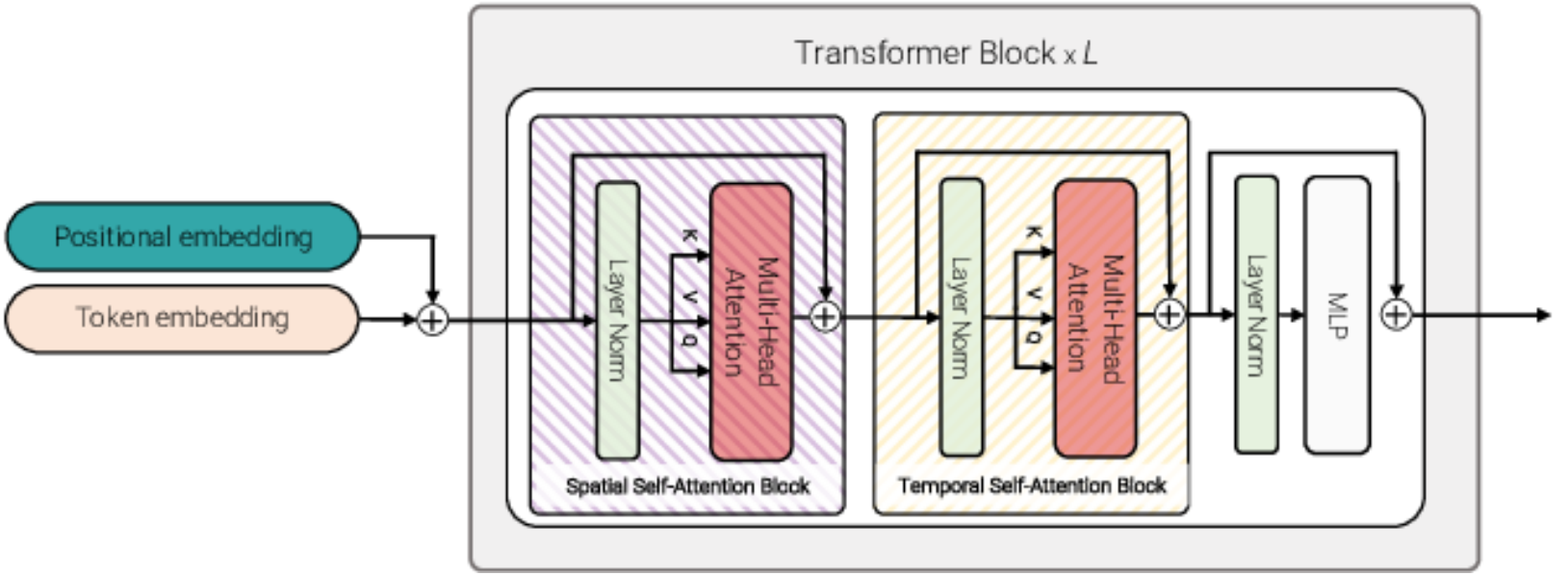
ViVit –M1


MLP Head
Class
Transformer Encoder
Position + Token Embedding
0
CLS
1
2
3
N
Embed to tokens
L×
MLP
Layer Norm
Self-Attention
Multi-Head Dot-Product Attention
K
V
Q
Layer Norm

ViVit – M2


Temporal Transformer Encoder
MLP Head
Class
Temporal + Token Embedding
Spatial Transformer Encoder
Spatial Transformer Encoder
Spatial Transformer Encoder
Positional + Token Embedding
Embed to tokens

ViVit – M3


Transformer Block x L
Positional embedding
Token embedding
Layer Norm
K
V
Q
Multi-Head Attention
Spatial Self-Attention Block
Layer Norm
K
V
Q
Multi-Head Attention
Temporal Self-Attention Block
Layer Norm
MLP

Travaux Pratiques

Practice

- Positional encoding: https://colab.research.google.com/drive/1ibERwNZ_QcDXqb52Ac3f61v1wBfvqvpn?usp=sharing
- Self-attention: https://colab.research.google.com/drive/19M9W5fx6yx7LZ275ccNy7aQKcM3cVwSi?usp=sharing
- Transformer for Text task: https://colab.research.google.com/drive/18LTQ5FgDJKSQiU1f0nAZ-K3eyKA5ZPH4?usp=sharing

Thanks!

E-mail: rui.dai@inria.fr