Searching for Better Spatio-temporal Alignment in Few-Shot Action Recognition

Yichao Cao[1*], **Xiu Su**[2*], **Qingfei Tang**[3], **Shan You**[4], **Xiaobo Lu**[1], **Chang Xu**[2†]
[1]School of Automation, Southeast University,
[2]School of Computer Science, Faculty of Engineering, The University of Sydney,
[3]Enbo Technology Co.,Ltd., China,
[4]SenseTime Research
caoyichao@seu.edu.cn, xisu5992@uni.sydney.edu.au, qingfeitang@gmail.com
youshan@sensetime.com, xblu@seu.edu.cn, c.xu@sydney.edu.au

36th Conference on Neural Information Processing Systems (NeurIPS 2022)

Method

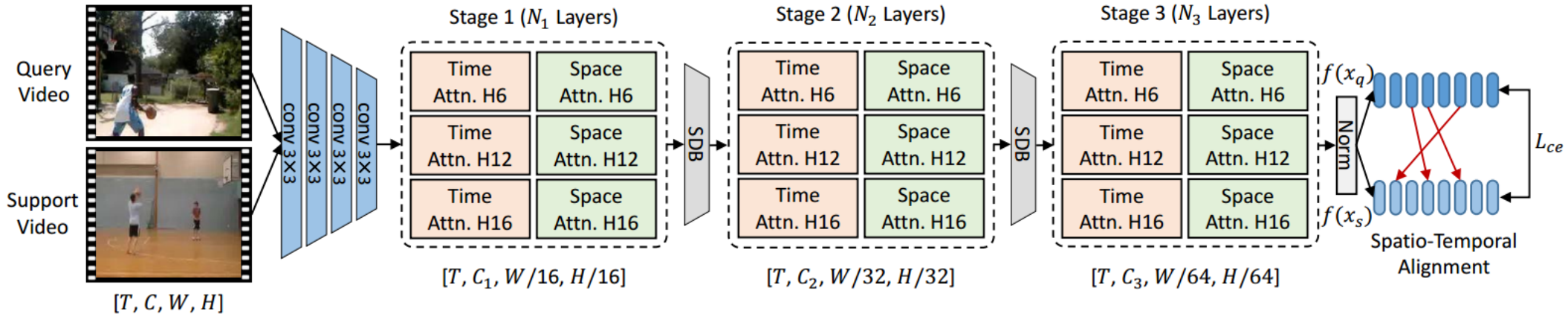


Figure 1: Overview of the neural module to be searched. The architecture of the network is determined by different operations per layers. There are three main stages in which space and time attention operations can be searched and selected.

Method

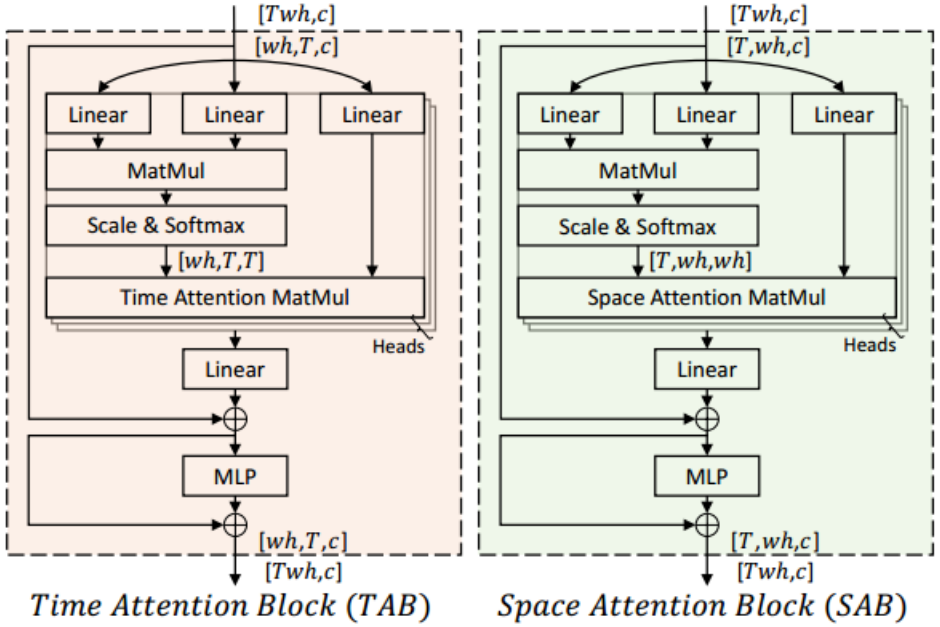


Figure 2: The essential operation of stack layer.

Table 1: Macro Transformer space of our SST model. Given an input video with dimension $T \times C \times W \times H$, we search over operations and the number of heads at 3 stage. Within each stage, "Choice Block" indicates the search from blocks of "TAB" and "SAB"."Spatial Downsample" represents the spatial downsampling block (SDB).

Stage	# Layers	Operations	# Heads	Output Size
Patch Embedding	4	Convolution	-	$[T, C_1, W/16, H/16]$
Choice Block	8	$\{TAB, SAB\}$	6, 12, 16	$[T, C_1, W/16, H/16]$
Spatial Downsample	1	SDB	12	$[T, C_2, W/32, H/32]$
Choice Block	8	$\{TAB, SAB\}$	6, 12, 16	$[T, C_2, W/32, H/32]$
Spatial Downsample	1	SDB	12	$[T, C_3, W/64, H/64]$
Choice Block	8	$\{TAB, SAB\}$	6, 12, 16	$[T, C_3, W/64, H/64]$
Output	1	Norm	-	$[T, C_o]$

Method

Algorithm 1: Training supernet with Transformer space shrinking

Input: Supernet $\mathcal{N}$ with weight $\mathcal{W}$ and Transformer space $\mathcal{A}$, maximum training epochs $\mathcal{T}$, warm up epochs $\mathcal{T}_{\boldsymbol{w}}$, shrink epochs $\mathcal{P}$, score threshold Thr and shrink percentage $\mathcal{K}$.

Init $\tau = 0$;

while $\tau \leq \mathcal{T}$ **do**

  randomly sample subnets from supernet $\mathcal{N}$;

  train one-shot supernet with Transformer space $\mathcal{A}$;

  if $\tau \geq \mathcal{T}_{\boldsymbol{w}}$ **then**

    record the loss and FLOPs of the subnet $\boldsymbol{a}_{i,j}^{\boldsymbol{o}}$ for each operation $\boldsymbol{o}_{i,j}$;

    if $\tau \% \mathcal{P} = 0$ **then**

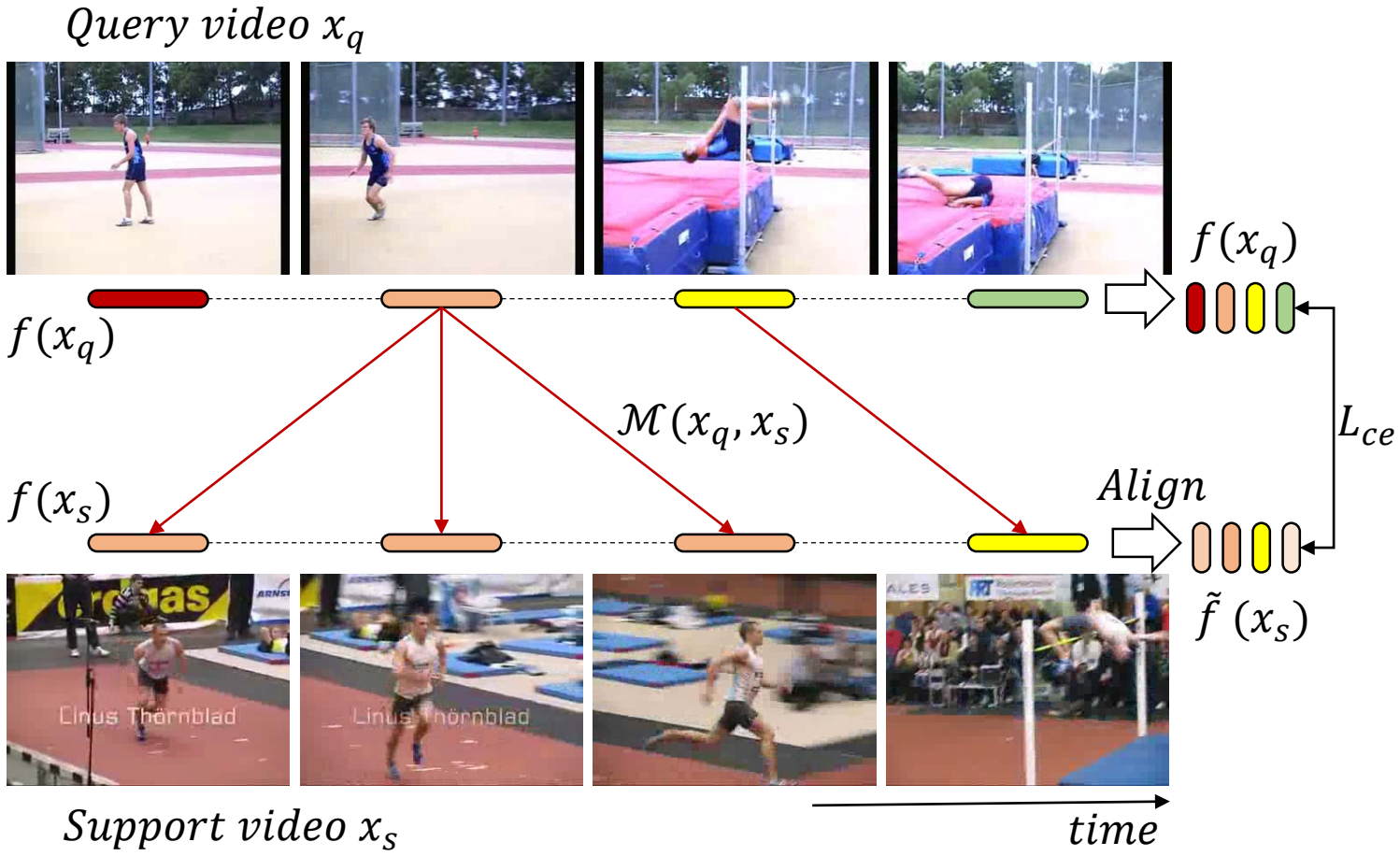
      calculate the $\mathcal{S}(i, j)$ for each operation with Eq. (6) and Eq. (7);

      $\mathcal{A} \leftarrow \mathcal{A}.\text{Shrink}(\mathcal{K}, Thr)$

end

Output: The optimized weight $\mathcal{W}_{\mathcal{A}}^{*}$ for supernet $\mathcal{N}$, and the shrunk Transformer space $\mathcal{A}$.

Method


Query video x_q
$f(x_q)$
$f(x_q)$
$\mathcal{M}(x_q, x_s)$
$f(x_s)$
Align
L_{ce}
$\tilde{f}(x_s)$
Linus Thörnblad
Linus Thörnblad
Support video x_s
time

Experiments

Table 2: Few-shot action classification results on HMDB51 [21].

Method	Frames	5-way 1-shot			5-way 5-shot		
		Acc	Params	FLOPs	Acc	Params	FLOPs
TimeSformer [2]	8	33.2	40.7M	73.35G	41.7	40.7M	73.35G
TRX [27]	8	29.1	25.6M	41.43G	46.4	25.6M	41.43G
ARN [47]	20	45.2	-	-	60.6	-	-
Ours	4	39.2	8.54M	6.83G	57.1	8.53M	6.81G
	8	51.1	8.89M	13.64G	60.4	8.91M	13.65G
	12	52.4	8.87M	20.49G	62.2	8.86M	20.48G

Table 3: Few-shot action classification results on UCF101 [31].

Method	Frames	5-way 1-shot			5-way 5-shot		
		Acc	Params	FLOPs	Acc	Params	FLOPs
TimeSformer [2]	8	42.0	40.7M	73.35G	63.0	40.7M	73.35G
TRX [27]	8	46.7	25.6M	41.43G	67.0	25.6M	41.43G
Ours	4	60.1	8.61M	6.79G	68.2	8.63M	6.83G
	8	63.8	8.87M	13.72G	69.7	8.84M	13.67G
	12	65.4	8.76M	20.34G	70.4	8.87M	20.45G

Thanks