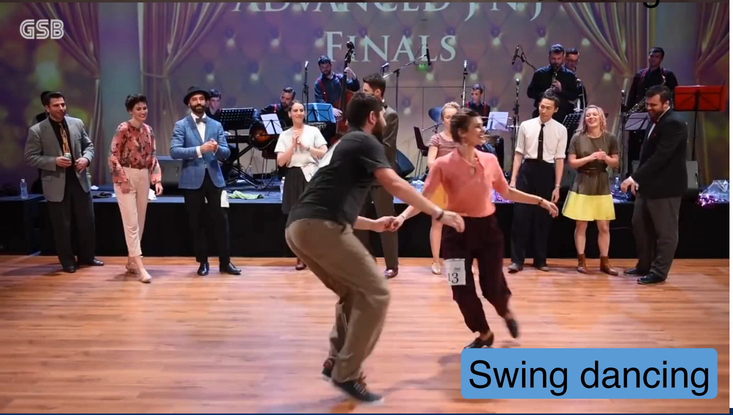
UNIVERSITY OF OXFORD

Keeping Your Eye on the Ball: Trajectory Attention for Video Transformers

Mandela Patrick[*1], Dylan Campbell[*2], Yuki M. Asano[*2]
Ishan Misra[1], Florian Metze[1], Christoph Feichtenhofer[1]
Andrea Vedaldi[1], João F. Henriques[2]

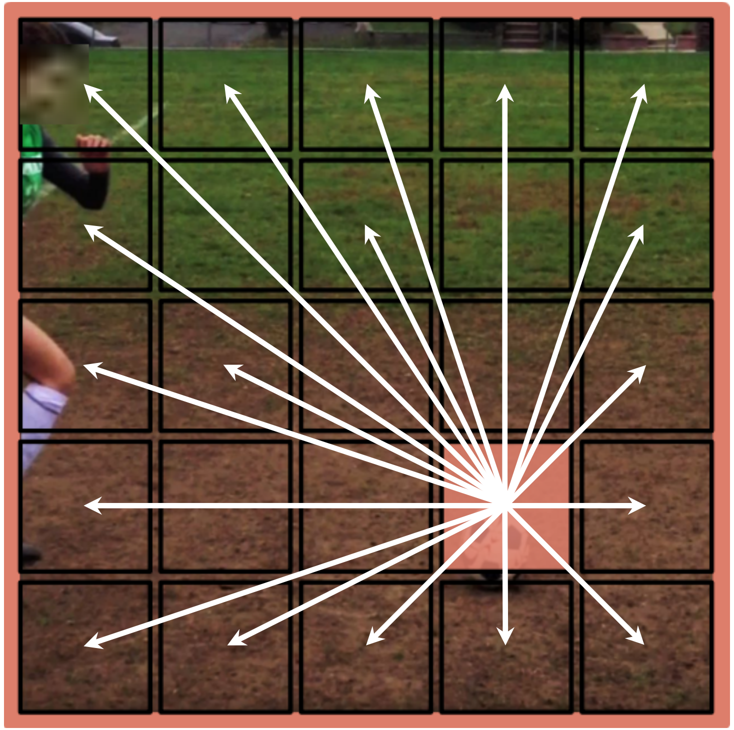
[1]Facebook AI
[2]Visual Geometry Group, Oxford
[*]Equal contribution

Recognising actions in video

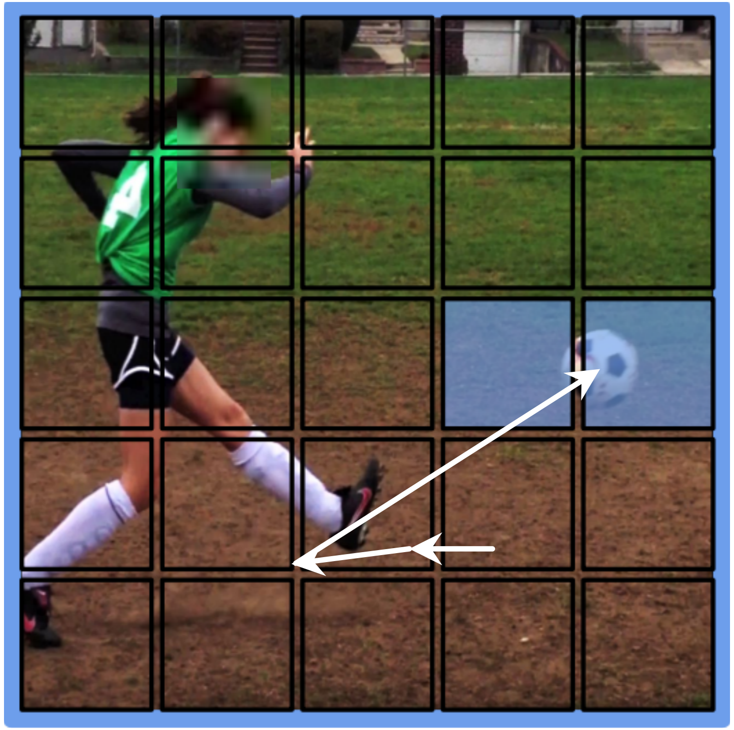


- **Proxy** for other video recognition tasks (≈ classification for images)
- Often requires **fine-grained** distinctions between subtle motions
- Often requires **long-range** associations
- *E.g.: swing dancing vs. salsa dancing;*
  dribbling basketball vs. dunking basketball;
  catching ball vs. throwing ball;
  ...

Recognising actions in video

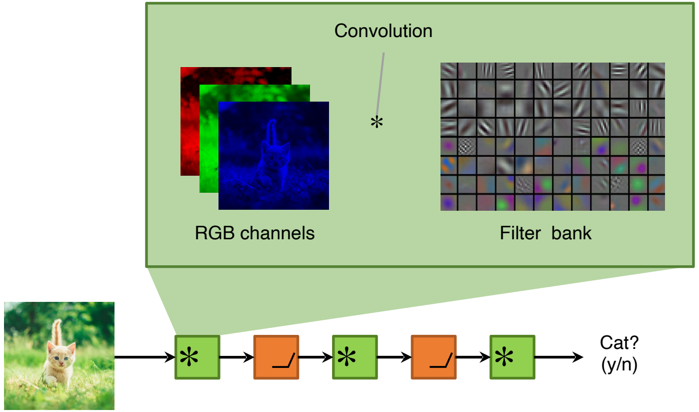


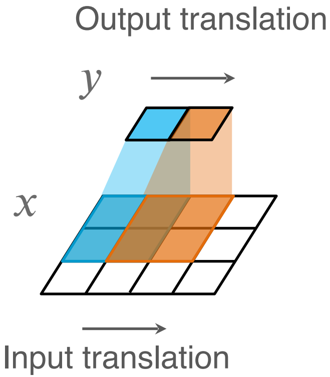
Recognising actions in video



- Camera motion
- Object motion

Background





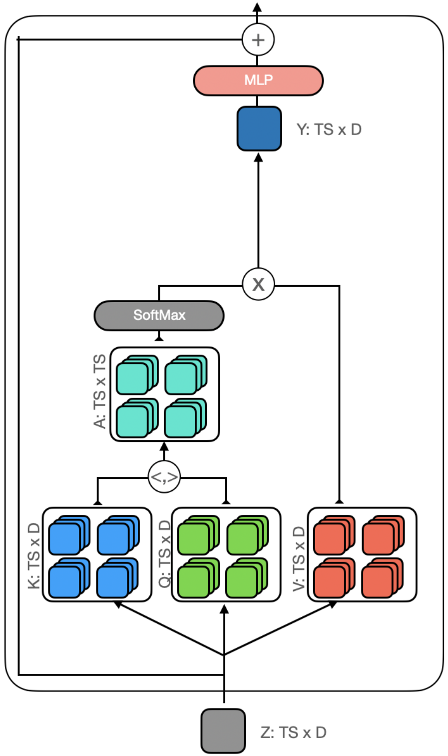
Convolutional networks

Convolutions limit the receptive field, both spatially and temporally

- Alleviated with *atrous* convolution
- Receptive field varies with resolution and framerate;
  can be difficult to tune

Tran et al., Learning spatiotemporal features with 3D convolutional networks. In ICCV, 2015.
Carreira & Zisserman, Quo vadis, action recognition? A new model and the Kinetics dataset. In CVPR, 2017.
Tranet et al., A closer look at spatiotemporal convolutions for action recognition. In CVPR, 2018.
Wang et et al., Non-local neural networks. In CVPR, 2018.

Background



Transformer networks

- Long-range associations / receptive field covers the **full input** at all stages
- Very little inductive bias compared to CNNs
  $\Rightarrow$ often harder to train, but more flexible
- Computation grows quadratically with input ($\mathcal{O}(S^2T^2)$ for input with T frames and S pixels)

Patrick et al., Support-set bottlenecks for video-text representation learning. In ICLR, 2021.

Dosovitskiy et al., An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.

Touvron et al., Training data-efficient image transformers & distillation through attention. In ICML, 2021.

Doersch et al., Crosstransformers: spatially-aware few-shot transfer. In NeurIPS, 2020.

Torresani et al., Is space-time attention all you need for video understanding? In ICML, 2021.

Inductive biases in video processing

Physical motivation:

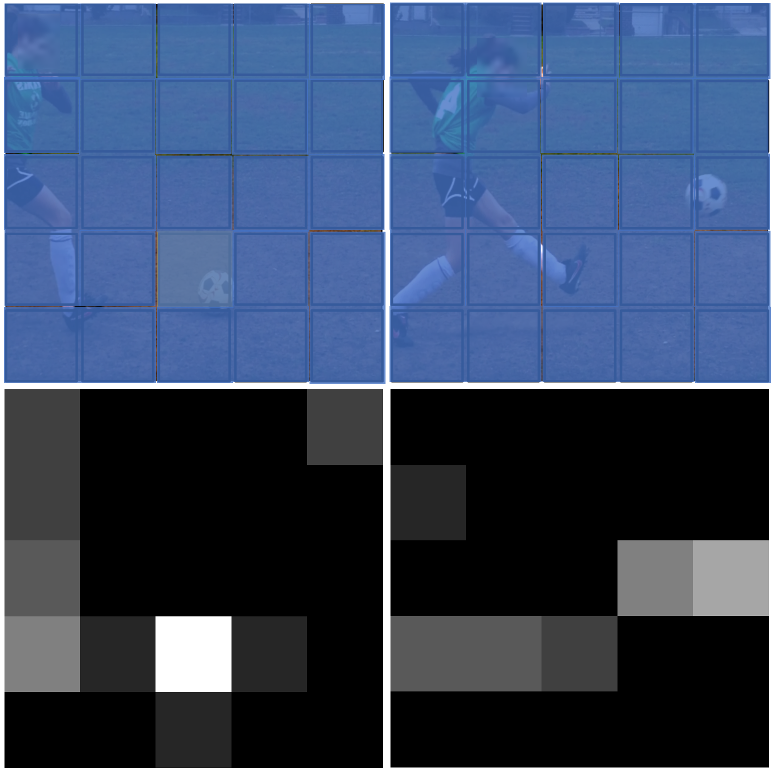
- Camera motion does not affect scene properties
- Motion path and appearance of an object can be disentangled
  - Translation equivariance is a subset of this desired behaviour

Advantages:

- Data efficiency
- Extrapolation beyond training set (generalization)
- Sometimes:
  - Improves computational efficiency
  - Reduces # of parameters and overfitting

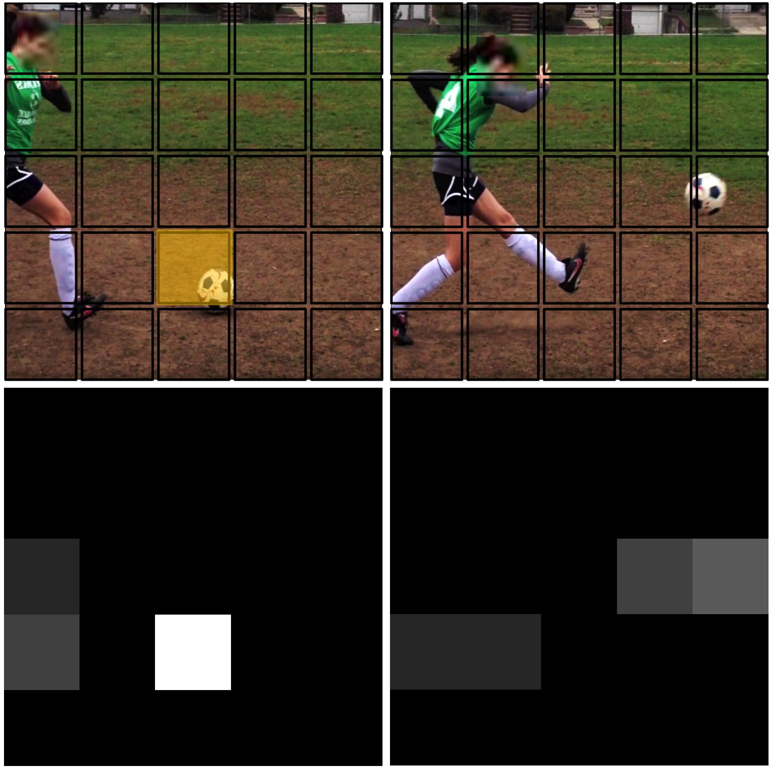
Video attention strategies: joint space-time

Video attention strategies: joint space-time

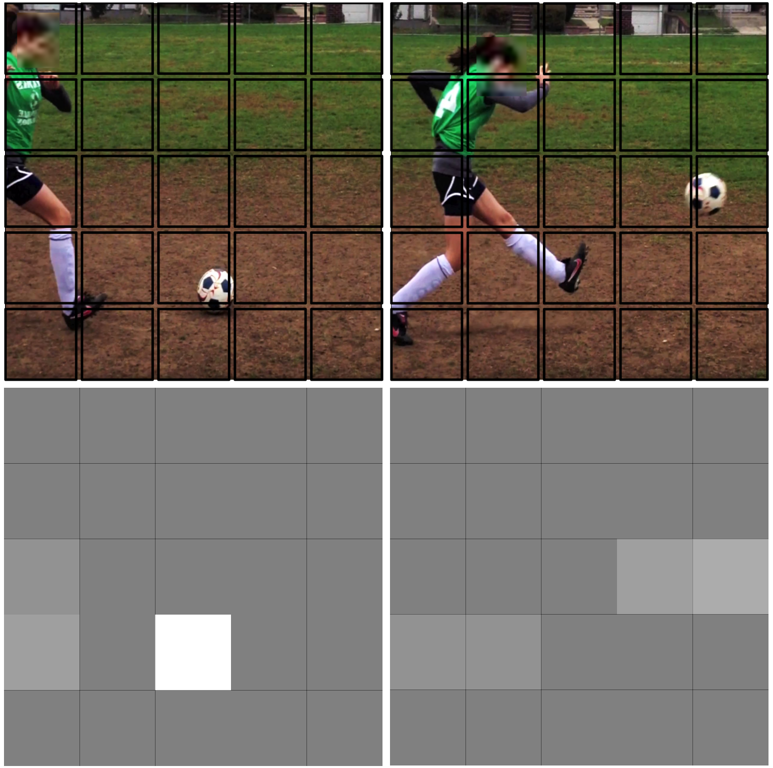


Video attention strategies: joint space-time

Softmax normalization
across volume

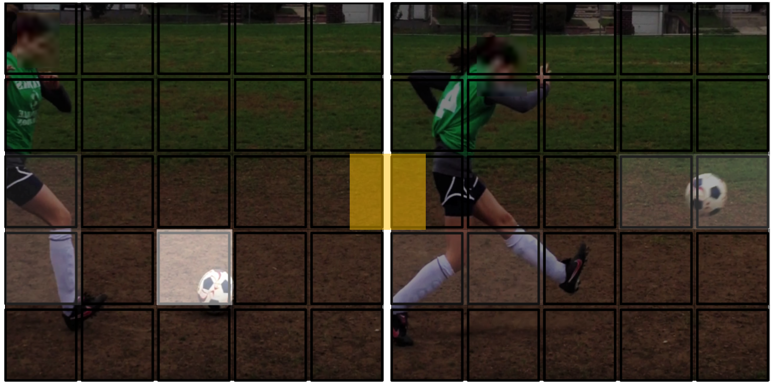


Video attention strategies: joint space-time



Video attention strategies: joint space-time

UNIVERSITY OF OXFORD



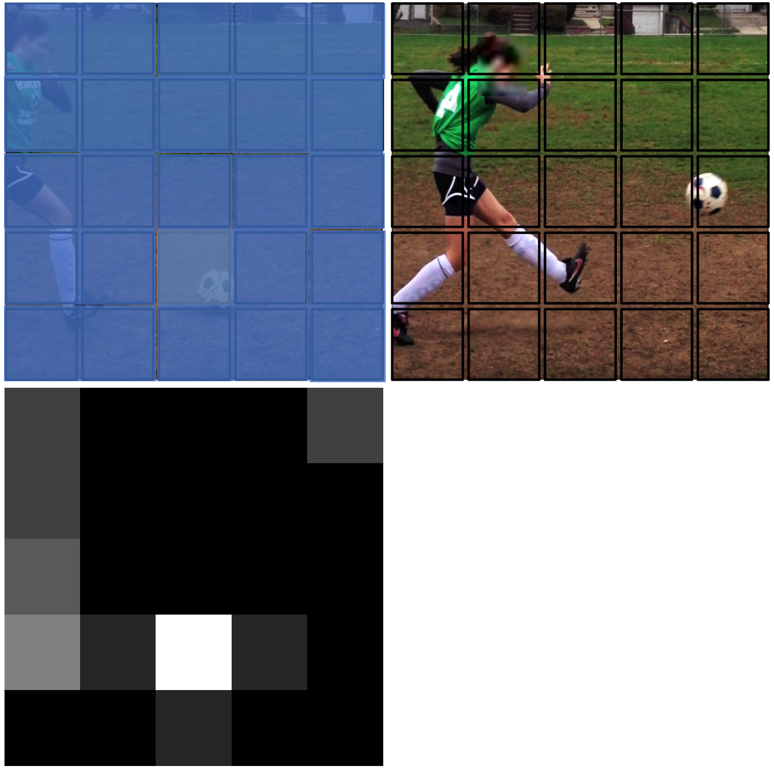
- Computational complexity: $\mathcal{O}(S^2T^2)$
- Infeasible for long and high-res videos
- Can we get closer to $\mathcal{O}(ST)$?

Bertasius et al., Is space-time attention all you need for video understanding? In ICML, 2021.

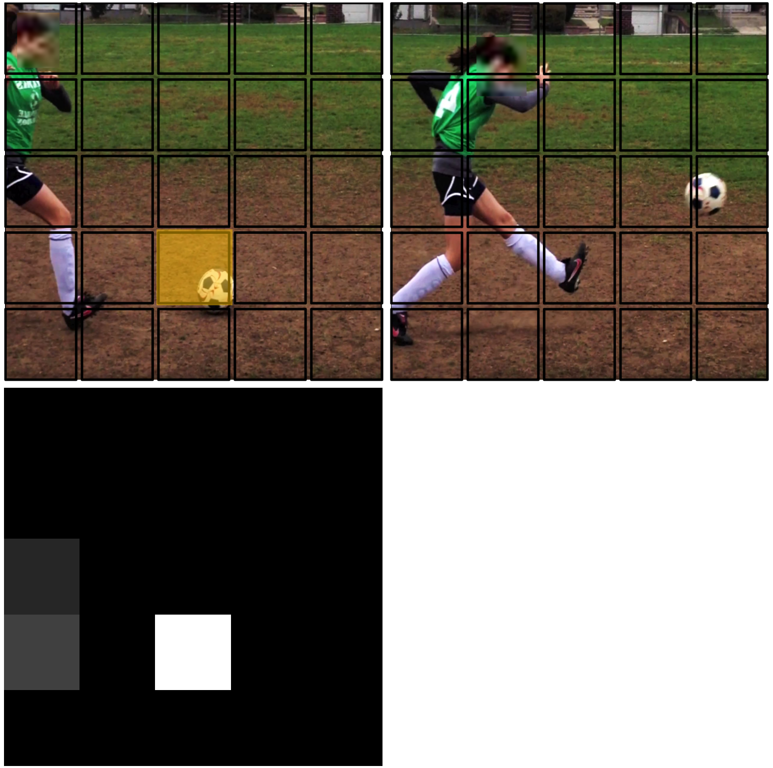
Arnab et al., Vivit: A video vision transformer, 2021.

Video attention strategies: divided space-time

Video attention strategies: divided space-time

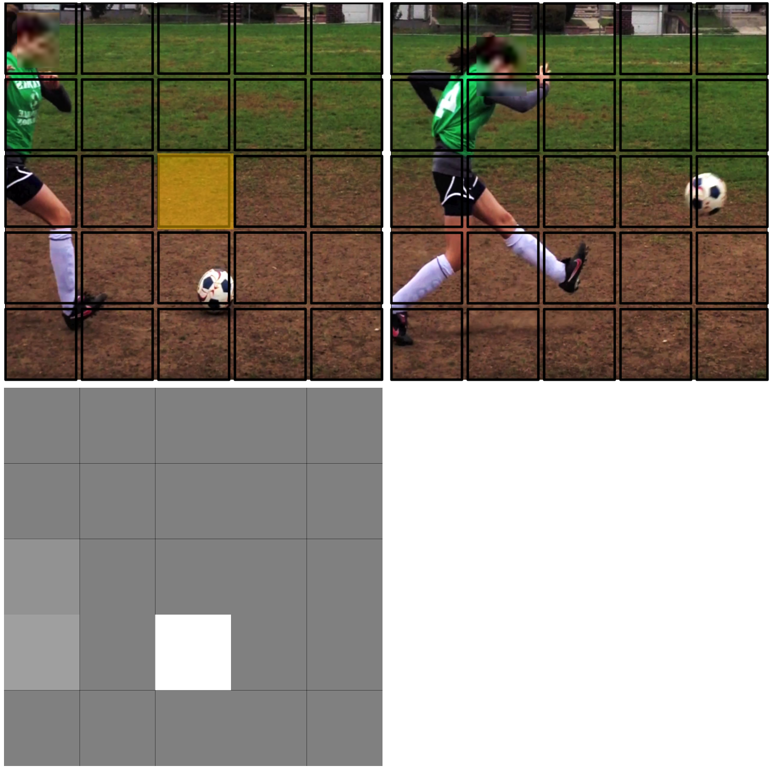


Video attention strategies: divided space-time

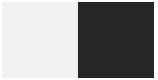


Softmax normalization

Video attention strategies: divided space-time



Video attention strategies: divided space-time

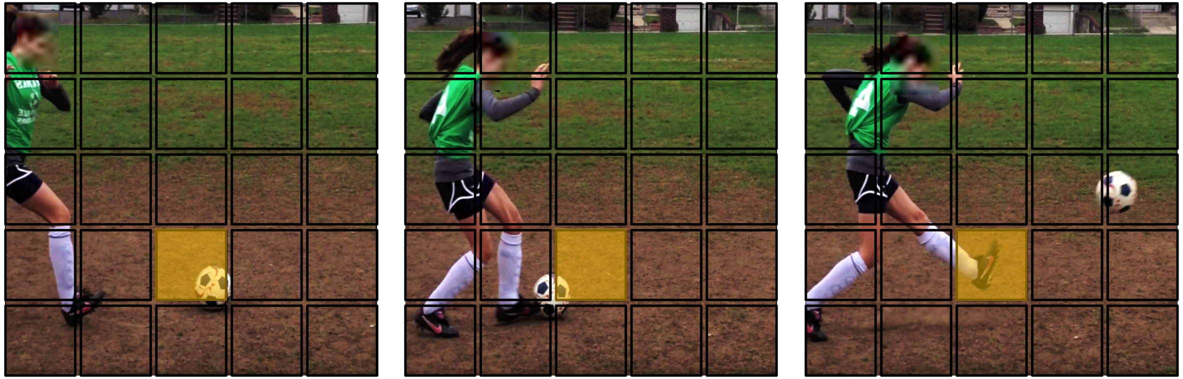


Video attention strategies: divided space-time

Video attention strategies: divided space-time

- Significant computation/memory gains: spatial attn $\mathcal{O}(S^2T)$, temporal attn $\mathcal{O}(ST^2)$
- Still has a quadratic bottleneck in each dimension
- Axis-aligned pooling is artificial

Moving camera + moving objects



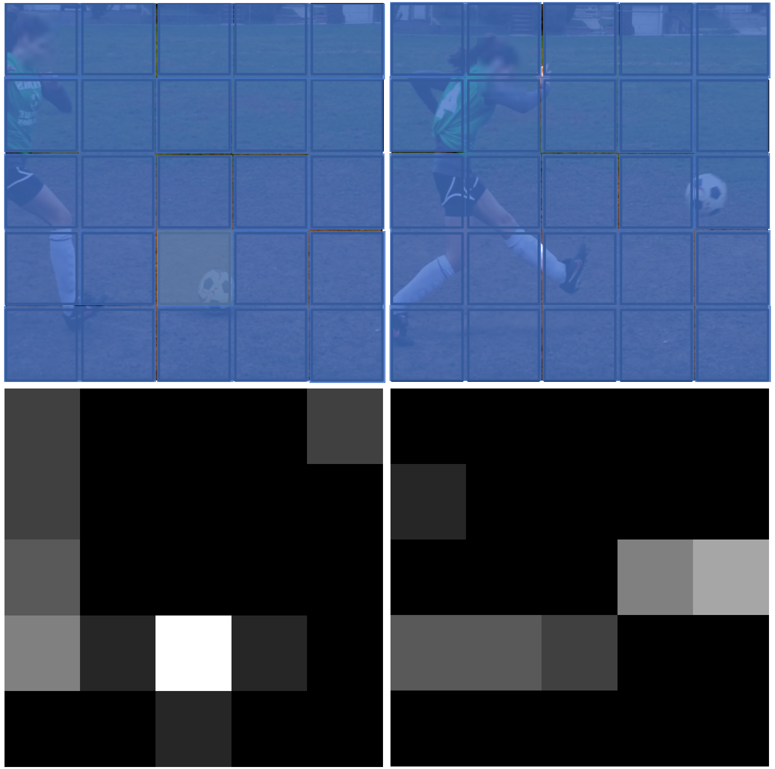
Bertasius et al., Is space-time attention all you need for video understanding? In ICML, 2021.

Arnab et al., Vivit: A video vision transformer, 2021.

Trajectory attention: motivation

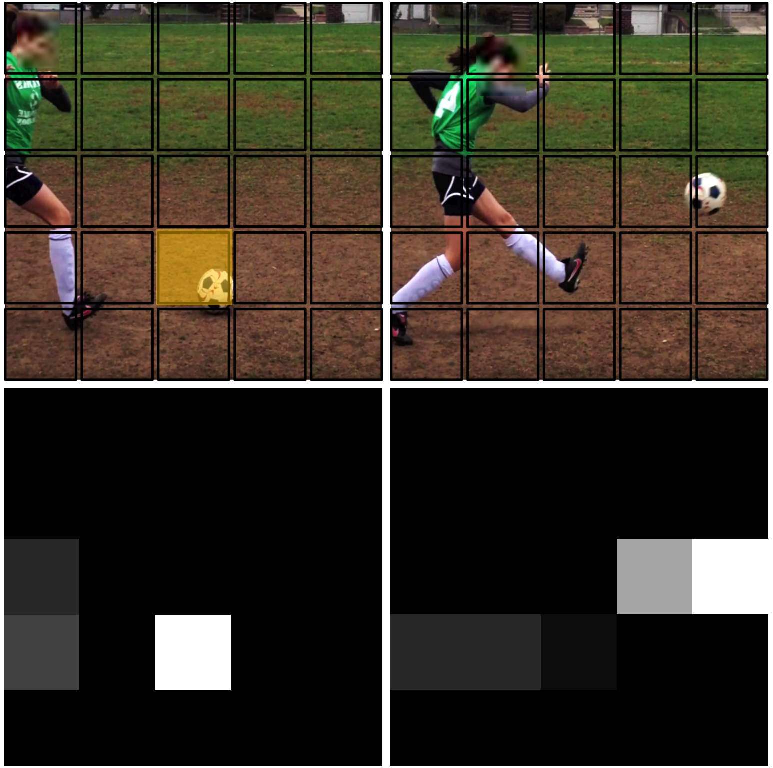
- **Aim:** *find other patches that contain the ball and aggregate their information into a single output*
- **Why?**
  - To leverage **multiple views** of the same object to better understand its properties
  - To reason about the **motion** of the object
- **How?**
  - **Attention:** computes feature similarities across space-time and pools information

Trajectory attention



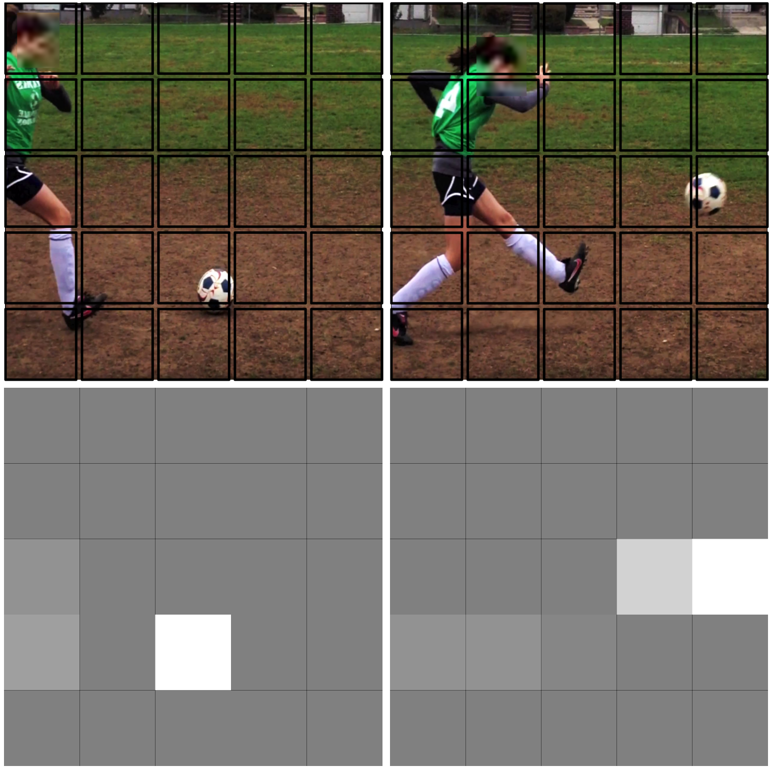
Trajectory attention

UNIVERSITY OF OXFORD



Softmax normalization
per frame

Trajectory attention



Trajectory attention

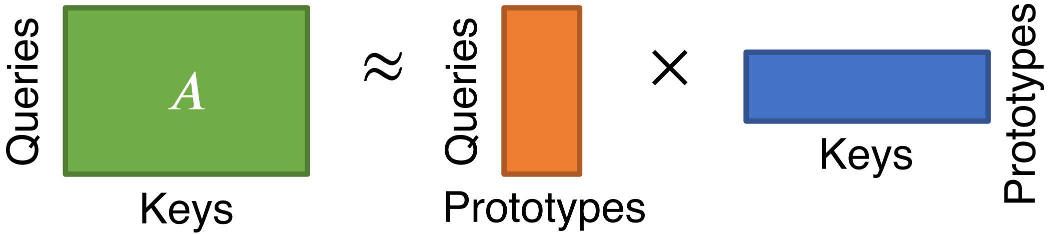
Trajectory attention

Trajectory attention

- Overall complexity: $\mathcal{O}(S^2T^2)$
  - No better than before
  - Needs to be improved by other means

Computational efficiency

Idea: Take inspiration from **matrix factorization** methods / low-rank decomposition



Cost of multiplying this matrix by an arbitrary vector: $\mathcal{O}(S^2T^2) \rightarrow \mathcal{O}(STP)$

- Not just multiplication, in general: $A \approx f(\quad , \quad)$
- Due to the softmax, attention matrices usually have high rank

  $\Rightarrow$ Poorly approximated by PCA/low-rank decompositions
- Prototypes must be few, and representative of all keys/queries

Xiong et al., Nyströmformer: A Nyström-based algorithm for approximating self-attention. In AAAI, 2021.

Beltagy et al., Longformer: The long-document transformer, 2020.

Choromanski et al., Rethinking attention with performers. In ICLR, 2021.

Computational efficiency

UNIVERSITY OF OXFORD

Formulate attention probabilistically

- Attention operator defines a **parametric model** of the probability of event A_{ij} (assignment of key j to query i), with a multinomial logistic function:

$$P(A_{i:}) = \underbrace{\mathcal{S}}_{\text{Softmax}}(\underbrace{\mathbf{q}_i^\top}_{\text{Query vector}}\underbrace{\mathbf{K}}_{\text{Key vectors (concatenated)}})$$

- Introduce **latent variables** U_{jl} (assignment of key j to prototype l)

- Then (without approximation):

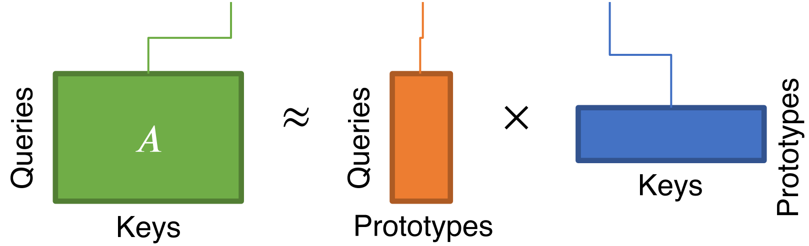
$$P(A_{ij}) = \sum_\ell P(A_{ij} \mid U_{\ell j}) P(U_{\ell j})$$

- But, $P(A \mid U)$ is intractable ➔ approximate with a similar parametric model

$$\mathcal{O}(S^2T^2) \rightarrow \mathcal{O}(STP)$$

- All together:

$$\tilde{P}(A)\mathbf{V} = \mathcal{S}(\mathbf{Q}^\top\mathbf{P})\,(\mathcal{S}(\mathbf{P}^\top\mathbf{K})\mathbf{V})$$



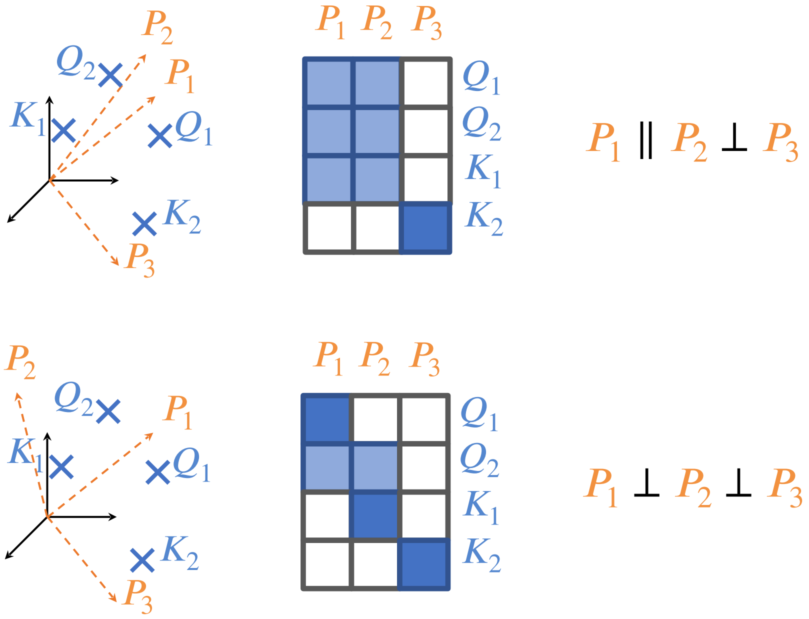
Selecting prototypes

Priorities:

- **Dynamically** adjust to keys/queries to ensure their region is reconstructed well
- Minimize **redundancy** between prototypes

Some suboptimal choices:

- Trainable vectors *(not adaptive)*
- Random sampling from keys/queries *(often selects collinear vectors)*
- Clustering keys/queries online *(expensive)*

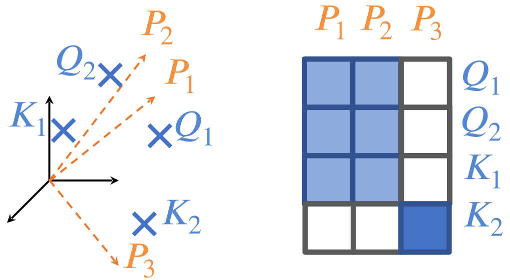


Selecting prototypes

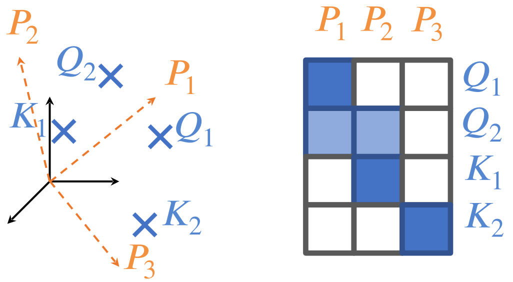
Objective:
Select the **most orthogonal** subset of keys/queries

A greedy algorithm:

$X \leftarrow$ random subset of $K \cup Q$
For $l \in \{1, \ldots, |P|\}$:
$$i^* \leftarrow \underset{i}{\text{argmin}} \sum_{j=1}^{l-1} \left| \left\langle X_i, P_j \right\rangle \right|$$
$$P_l \leftarrow X_{i^*}$$



$P_1 \parallel P_2 \perp P_3$



$P_1 \perp P_2 \perp P_3$

Experiments: approximating attention

UNIVERSITY OF OXFORD

Comparison to state-of-the-art on the Long Range Arena benchmark

Model	ListOps	Text	Retrieval	Image	Pathfinder	Avg↑	GFLOPS↓	Mem.↓
Exact [76]	36.69	63.09	78.22	31.47	66.35	55.16	1.21	4579
Performer-256 [14]	36.69	63.22	**78.98**	29.39	**66.55**	54.97	0.49	885
Nyströmformer-128 [85]	**36.90**	64.17	78.67	**36.16**	52.32	53.64	0.62	745
Orthoformer-64	33.87	**64.42**	78.36	33.26	66.41	**55.26**	**0.24**	**344**

- Best overall results with far fewer prototypes (64) than other methods
- About **half** the memory and GFLOPS of the best approximations
- **No loss** of performance on average (unlike the other approximations)

Experiments: approximating attention

Comparison on action recognition datasets (Kinetics-400, Something-Something)

(a) Orthoformer is competitive with Nyström.

Attention	Approx.	Mem.	K-400	SSv2
Trajectory (E)	N/A	7.4	**79.7**	**66.5**
Trajectory (A)	Performer	5.1	72.9	52.7
	Nyströmformer	3.8	**77.5**	**64.0**
	Orthoformer	3.6	**77.5**	63.8

(b) Selecting orthogonal prototypes is the best strategy.

Attention	Selection	Mem.	K-400	SSv2
Trajectory (E)	N/A	7.4	**79.7**	**66.5**
Trajectory (A)	Seg-Means	3.6	75.8	60.3
	Random	3.6	76.5	62.5
	Orthogonal	3.6	**77.5**	**63.8**

Experiments: setup

UNIVERSITY OF OXFORD

Application: **action recognition**

- Use ViT [1] as the base model (12 layers / 12 attention heads / embeddings size 768)
- Separate space and time positional encodings (TimeSformer [2])
- Cubic image tokenization (ViViT [3])
- Adding our **Trajectory Attention**

Datasets:

- Kinetics-400/600 *(appearance cues are more dominant)*
- Something-Something V2 (*motion cues are more dominant*)
- Epic Kitchens 100

Keeps objects **consistent** across different action classes

[1] Dosovitskiy et al., An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 20

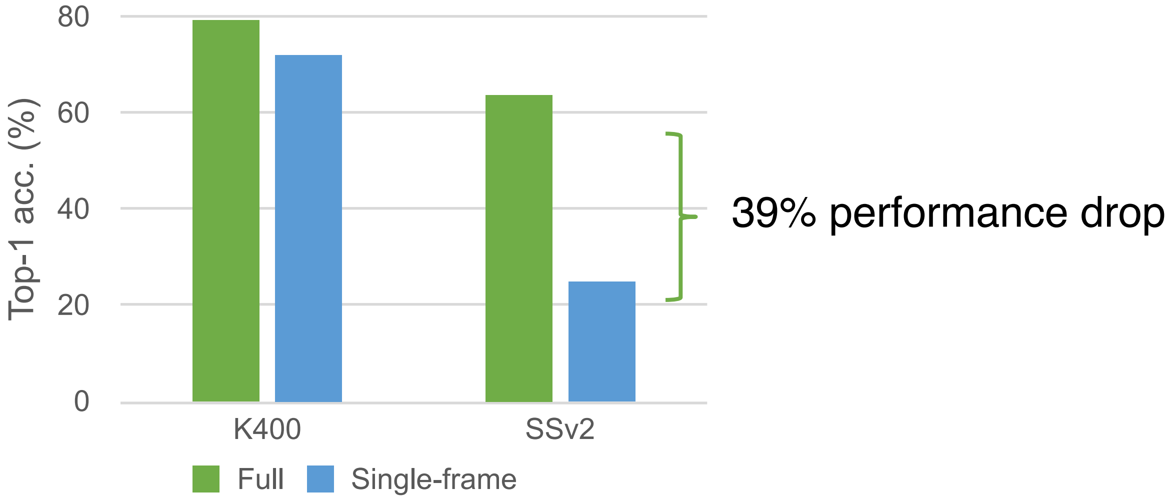
[2] Bertasius et al., Is space-time attention all you need f video understanding? In ICML, 2021.

[3] Arnab et al., Vivit: A video vision transformer, 2021.

Experiments: setup

UNIVERSITY OF OXFORD

Train model on **single frames only** and assess drop in performance



Experiments: attention ablation

UNIVERSITY OF OXFORD

Comparison of attention mechanisms

Attention	K-400	SSv2
Joint Space-Time	79.2	64.0
Divided Space-Time	78.5	64.2
Trajectory	**79.7**	**66.5**

Experiments: benchmark results

UNIVERSITY OF OXFORD

(a) **Something–Something V2**

Model	Pretrain	Top-1	Top-5	GFLOPs ×views
SlowFast [25]	K-400	61.7	-	65.7×3×1
TSM [46]	K-400	63.4	88.5	62.4×3×2
STM [33]	IN-1K	64.2	89.8	66.5×3×10
MSNet [40]	IN-1K	64.7	89.4	67×1×1
TEA [45]	IN-1K	65.1	-	70×3×10
bLVNet [23]	IN-1K	65.2	90.3	128.6×3×10
VidTr-L [44]	IN-21K+K-400	60.2	-	351×3×10
Tformer-L [7]	IN-21K	62.5	-	1703×3×1
ViViT-L [2]	IN-21K+K-400	65.4	89.8	3992×4×3
MViT-B [22]	K-400	67.1	90.8	170×3×1
Mformer	IN-21K+K-400	66.5	90.1	369.5×3×1
Mformer-L	IN-21K+K-400	**68.1**	**91.2**	1185.1×3×1
Mformer-HR	IN-21K+K-400	67.1	90.6	958.8×3×1

(b) **Kinetics-400**

Method	Pretrain	Top-1	Top-5	GFLOPs×views
I3D [10]	IN-1K	72.1	89.3	108×N/A
R(2+1)D [75]	-	72.0	90.0	152×5×23
S3D-G [84]	IN-1K	74.7	93.4	142.8×N/A
X3D-XL [24]	-	79.1	93.9	48.4×3×10
SlowFast [25]	-	79.8	93.9	234×3×10
VTN [51]	IN-21K	78.6	93.7	4218×1×1
VidTr-L [44]	IN-21K	79.1	93.9	392×3×10
Tformer-L [7]	IN-21K	80.7	94.7	2380×3×1
MViT-B [22]	-	81.2	95.1	455×3×3
ViViT-L [2]	IN-21K	**81.3**	94.7	3992×3×4
Mformer	IN-21K	79.7	94.2	369.5×3×10
Mformer-L	IN-21K	80.2	94.8	1185.1×3×10
Mformer-HR	IN-21K	81.1	**95.2**	958.8×3×10

- **SOTA** on SSv2 (+1%), which is more reliant on motion cues
- Competitive with the much larger ViViT-L model on K400

Experiments: benchmark results

UNIVERSITY OF OXFORD

(c) **Epic-Kitchens**

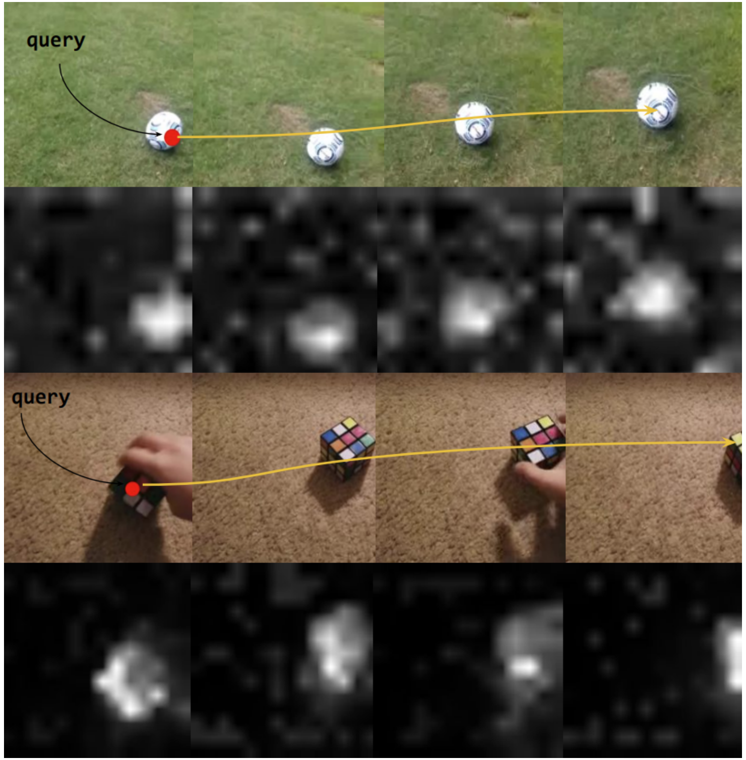
Method	Pretrain	A	V	N
TSN [78]	IN-1K	33.2	60.2	46.0
TRN [86]	IN-1K	35.3	65.9	45.4
TBN [36]	IN-1K	36.7	66.0	47.2
TSM [46]	IN-1K	38.3	**67.9**	49.0
SlowFast [25]	K-400	38.5	65.6	50.0
ViViT-L [2]	IN-21K+K-400	44.0	66.4	56.8
Mformer	IN-21K+K-400	43.1	66.7	56.5
Mformer-L	IN-21K+K-400	44.1	67.1	57.6
Mformer-HR	IN-21K+K-400	**44.5**	<u>67.0</u>	**58.5**

(d) **Kinetics-600**

Model	Pretrain	Top-1	Top-5	GFLOPs $\times$ views
AttnNAS [81]	-	79.8	94.4	-
LGD-3D [56]	IN-1K	81.5	95.6	-
SlowFast [25]	-	81.8	95.1	$234 \times 3 \times 10$
X3D-XL [24]	-	81.9	95.5	$48.4 \times 3 \times 10$
Tformer-HR [7]	IN-21K	82.4	96.0	$1703 \times 3 \times 1$
ViViT-L [2]	IN-21K	83.0	95.7	$3992 \times 3 \times 4$
MViT-B-24 [22]	-	**83.8**	**96.3**	$236 \times 1 \times 5$
Mformer	IN-21K	81.6	95.6	$369.5 \times 3 \times 10$
Mformer-L	IN-21K	82.2	96.0	$1185.1 \times 3 \times 10$
Mformer-HR	IN-21K	<u>82.7</u>	96.1	$958.8 \times 3 \times 10$

- **SOTA** on Epic-Kitchens Nouns (+2.3%), which is more reliant on motion cues
- Competitive performance on K600

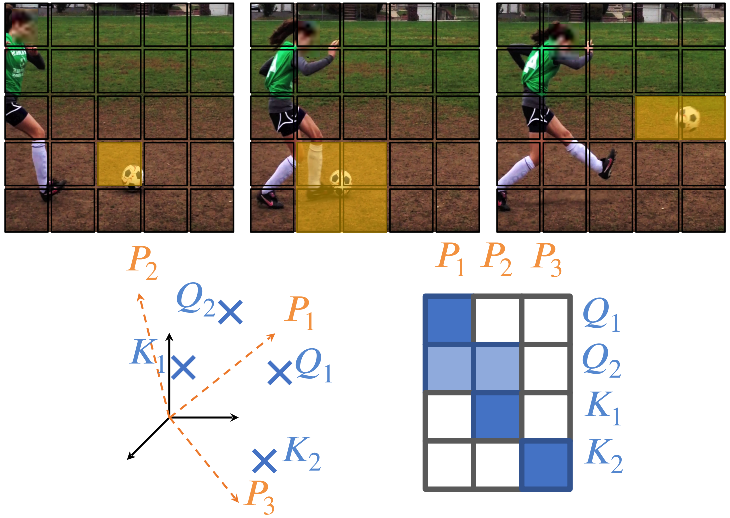
Experiments: attention maps



Conclusions

UNIVERSITY OF OXFORD

- ✓ Aggregating information along implicit **motion trajectories** can inject a helpful inductive bias into video transformers
- ✓ **Quadratic dependency** on input size can be reduced to **linear**
- ✓ **Orthogonality** is the most effective prototype selection criteria
- ✓ SOTA results on **motion-focused** datasets



Algorithm 1 Orthoformer (proposed) attention

1: $\mathbf{P} \leftarrow \text{MostOrthogonalSubset}(\mathbf{Q}, \mathbf{K}, R)$
2: $\mathbf{\Omega}_1 = \mathcal{S}(\mathbf{Q}^\top \mathbf{P}/\sqrt{D})$
3: $\mathbf{\Omega}_2 = \mathcal{S}(\mathbf{P}^\top \mathbf{K}/\sqrt{D})$
4: $\mathbf{Y} = \mathbf{\Omega}_1(\mathbf{\Omega}_2 \mathbf{V})$

Thank you

Mandela Patrick[*] Dylan Campbell[*] Yuki M. Asano[*]

Ishan Misra Florian Metze Christoph Feichtenhofer

Andrea Vedaldi João F. Henriques

University of Oxford / Facebook AI Research

[*]Equal Contribution