UCF Center for Research in Computer Vision

CAP6412
Advanced Computer Vision

Mubarak Shah

shah@crcv.ucf.edu

Lecture-5-DeTR_TimeS_ViViT

UCF Center for Research in Computer Vision

Contents

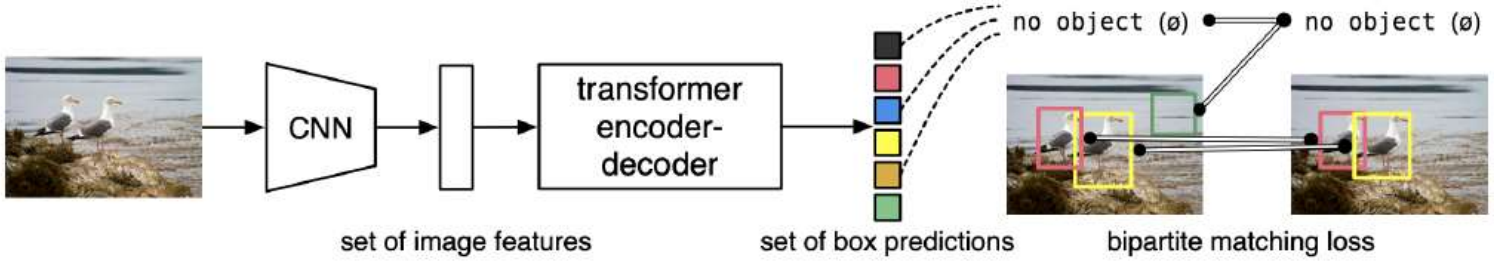
- DETr
- TimeS
- ViViT
- Transformers in Vision: A Survey

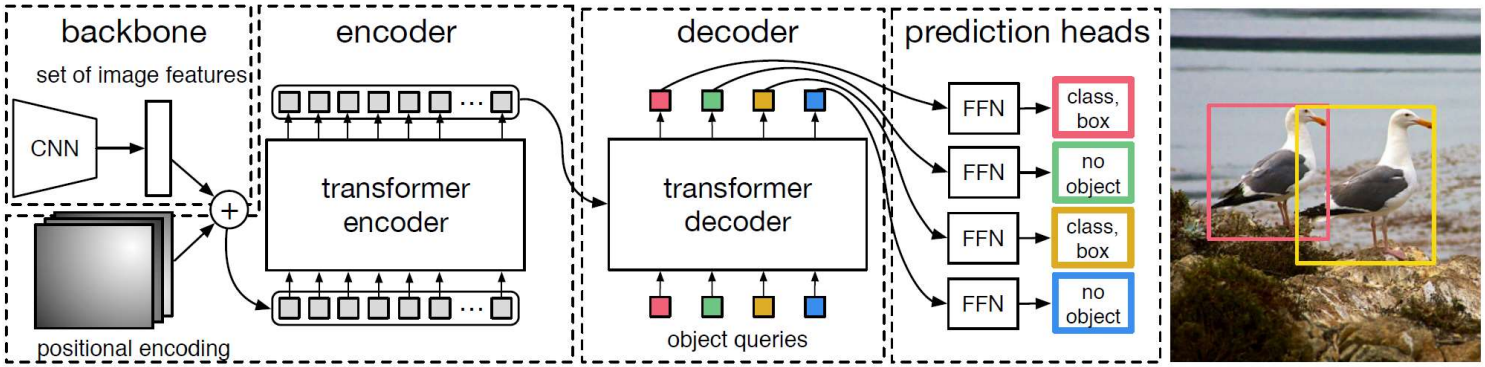
DETr (DEtection Transformer)

- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "*End-to-End Object Detection with Transformers*", ECCV-2020. (1421 Citations)

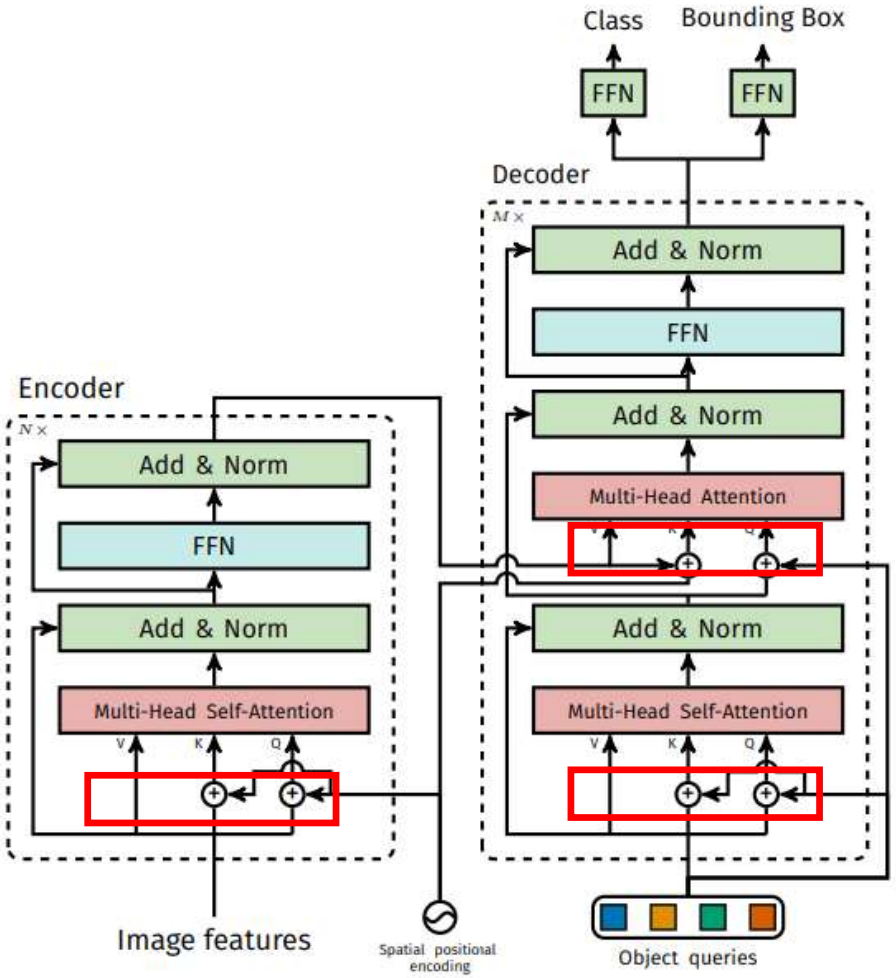
Detr

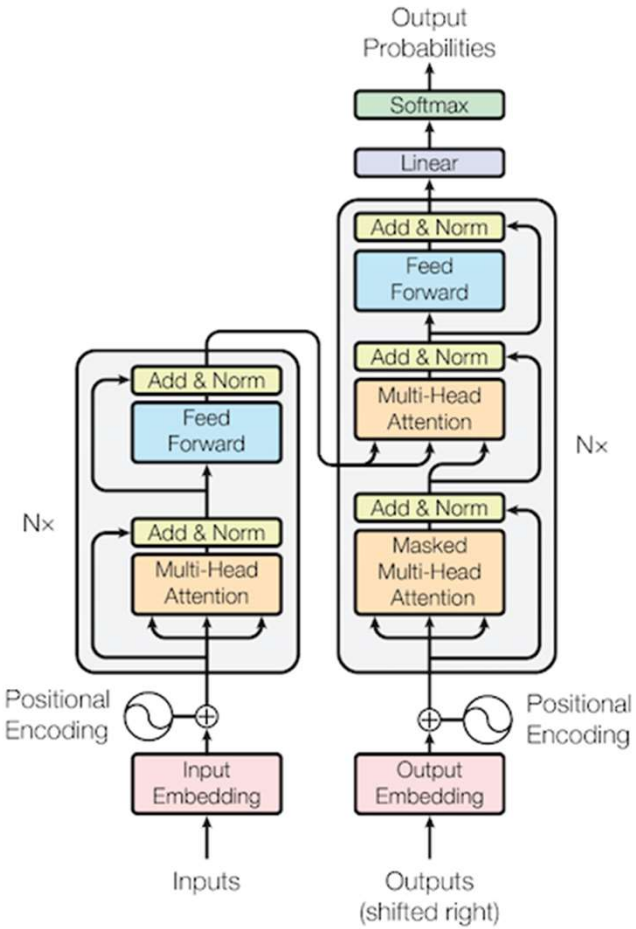
- A new method that views object detection as a set predictions
- Streamlines the detection pipeline, effectively removing the need for many hand-designed components
- Given a fixed small set of learned object queries, DETR outputs a set of predictions in parallel
- DETR demonstrates accuracy and run-time performance on par with Faster RCNN on COCO
- DETR can be generalized to produce panoptic segmentation


CNN
transformer encoder-decoder
no object (ø)
no object (ø)
set of image features
set of box predictions
bipartite matching loss


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

DETR


Class
Bounding Box
FFN
FFN
Decoder
M×
Add & Norm
FFN
Add & Norm
Multi-Head Attention
Add & Norm
Multi-Head Self-Attention
Encoder
N×
Add & Norm
FFN
Add & Norm
Multi-Head Self-Attention
Image features
Spatial positional encoding
Object queries


Output Probabilities
Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Nx
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding
Inputs
Outputs (shifted right)

DETR

y : ground truth set of objects

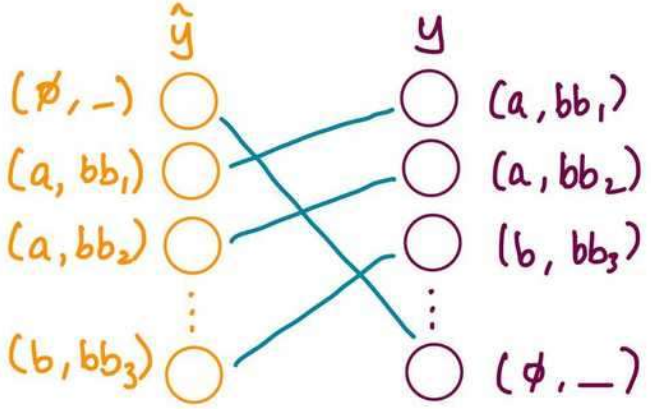
$\hat{y}\ (\{\hat{y}\}_{i=1}^{N})$: set of N predictions

$$L_{match}(y_i, \hat{y}_{\sigma(i)}) = -\mathbb{I}_{\{c_i \neq \emptyset\}} \hat{p}_{\sigma(i)}(c_i) + \mathbb{I}_{\{c_i \neq \emptyset\}} L_{box}(b_i, \hat{b}_{\sigma(i)})$$

$y_i = (c_i, b_i)$

$b_i = box,\ c_i = class$

$\sigma(i) = index$



https://www.vivekkalyan.com/end-to-end-object-detection-with-transformers

$$\hat{\sigma} : argmin_{\sigma \in \mathfrak{S}_N} \sum_i^N L_{match}(y_i, \hat{y}_{\sigma(i)})$$

$\hat{p}_{\sigma(i)}(c_i) = prediction\ probability\ of\ c_i$

$$L_{Hungarian}(y, \hat{y}) = \sum_{i=1}^{N} \left[-log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbb{I}_{\{c_i \neq \emptyset\}} L_{box}(b_i, \hat{b}_{\hat{\sigma}(i)}) \right]$$

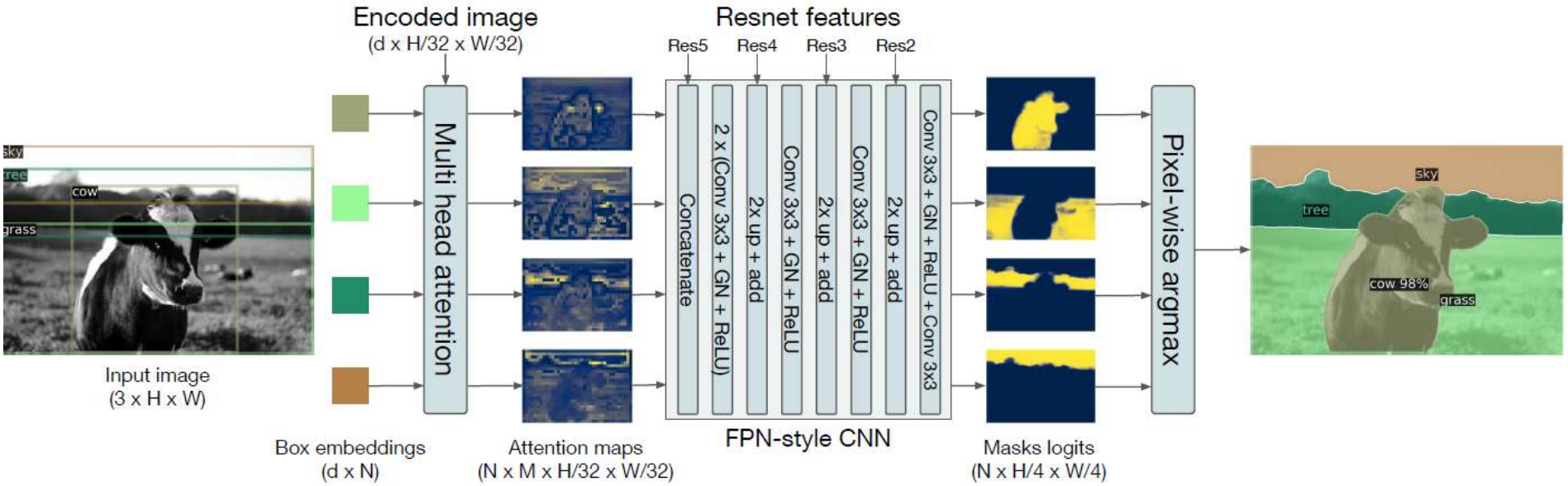
$$L_{box}(b_i, \hat{b}_{\hat{\sigma}(i)}) = \lambda_{iou} L_{iou}(b_i, \hat{b}_{\hat{\sigma}(i)}) + \lambda_{L1} ||b_i - \hat{b}_{\sigma(i)}||_1$$

COCO Results

Model	GFLOPS/FPS	#params	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Faster RCNN-DC5	320/16	166M	39.0	60.5	42.3	21.4	43.5	52.5
Faster RCNN-FPN	180/26	42M	40.2	61.0	43.8	24.2	43.5	52.0
Faster RCNN-R101-FPN	246/20	60M	42.0	62.5	45.9	25.2	45.6	54.6
Faster RCNN-DC5+	320/16	166M	41.1	61.4	44.3	22.9	45.9	55.0
Faster RCNN-FPN+	180/26	42M	42.0	62.1	45.5	26.6	45.4	53.4
Faster RCNN-R101-FPN+	246/20	60M	44.0	63.9	**47.8**	**27.2**	48.1	56.0
DETR	86/28	41M	42.0	62.4	44.2	20.5	45.8	61.1
DETR-DC5	187/12	41M	43.3	63.1	45.9	22.5	47.3	61.1
DETR-R101	152/20	60M	43.5	63.8	46.4	21.9	48.0	61.8
DETR-DC5-R101	253/10	60M	**44.9**	**64.7**	47.7	23.7	**49.5**	**62.3**

Activate Winc

Panoptic Segmentation



UCF Center for Research in Computer Vision

Results



DETr (DEtection Transformer)

- Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, "*End-to-End Object Detection with Transformers*", ECCV-2020. (1421 Citations)

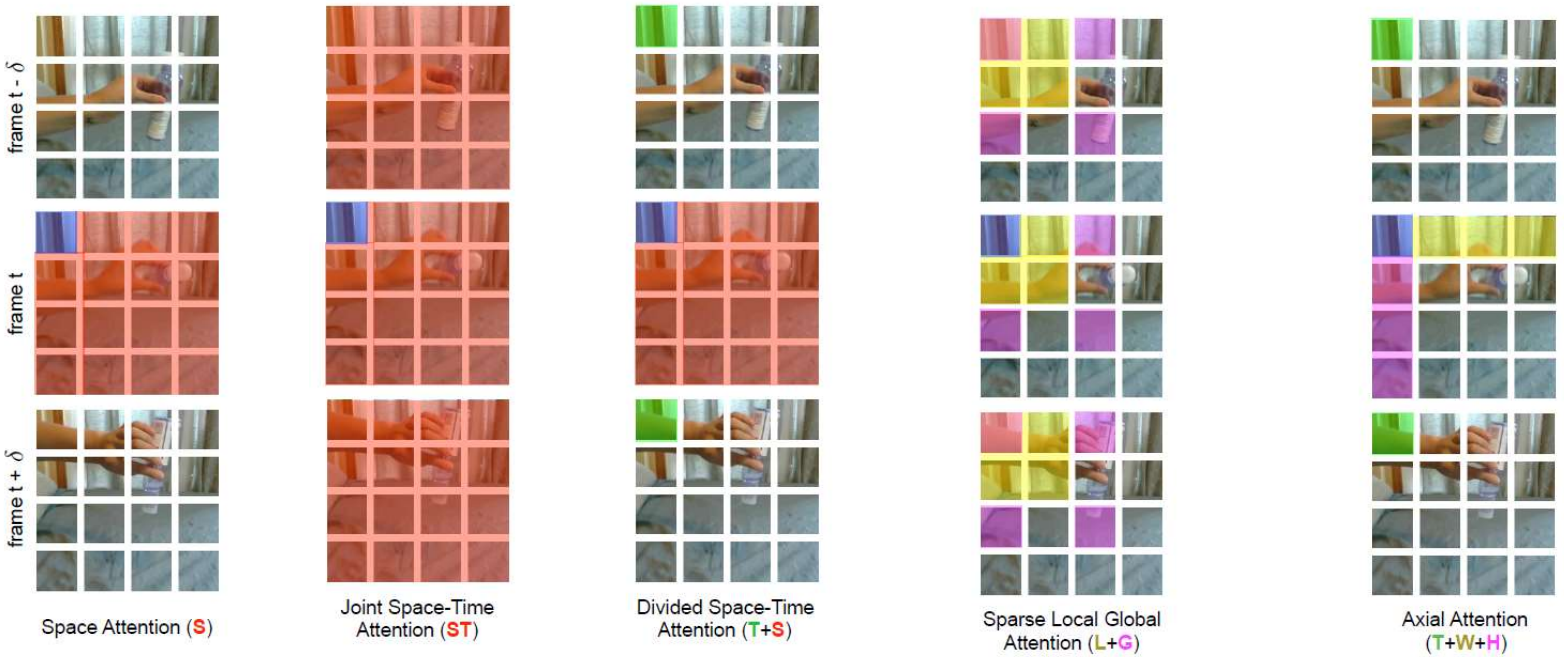
TimeSTransformer

- Gedas Bertasius Heng Wang Lorenzo, Torresanils, *" Space-Time Attention All You Need for Video Understanding?*", International Conference on Machine Learning, PMLR 139, 2021. (175 Citations)

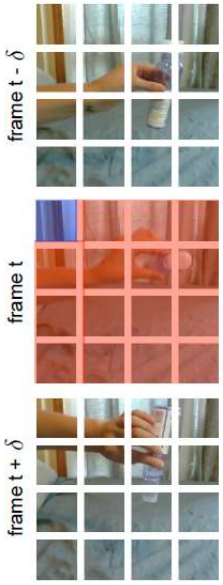
TimeSTransformer

- Adapts the standard Vision Transformer (ViT) architecture to video by enabling spatiotemporal feature learning
- “Divided attention” architecture, which separately applies temporal and spatial attention, works the best
- TimeSformer achieves state-of-the-art results on several action recognition benchmarks, including
  - Kinetics-400
  - Kinetics-600
- Compared to 3D convolutional networks, model is faster to train, it can achieve dramatically higher test efficiency

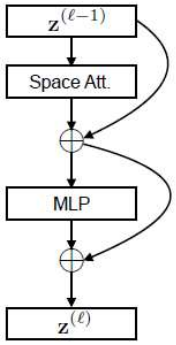
Self-Attention Schemes



Self-Attention Schemes



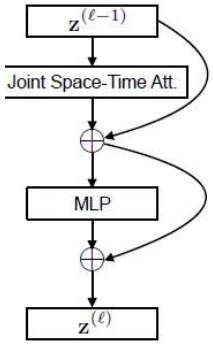
Space Attention (S)



Space Attention (S)



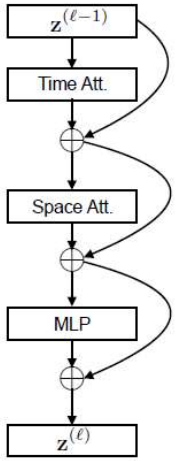
Joint Space-Time Attention (ST)



Joint Space-Time Attention (ST)



Divided Space-Time Attention (T+S)

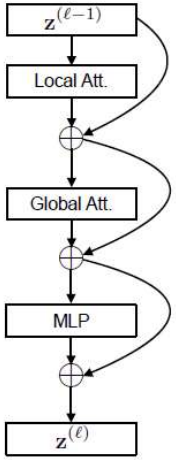


Divided Space-Time Attention (T+S)

Self-Attention Schemes



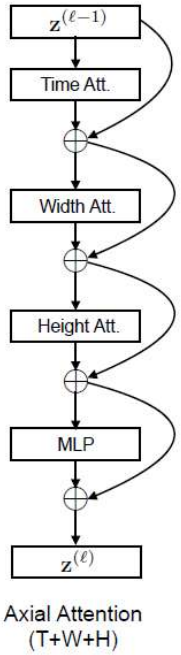
Sparse Local Global Attention (L+G)



Sparse Local Global Attention (L+G)



Axial Attention (T+W+H)



Axial Attention (T+W+H)

Performance

Attention	Params	K400	SSv2
Space	85.9M	76.9	36.6
Joint Space-Time	85.9M	77.4	58.5
Divided Space-Time	121.4M	**78.0**	**59.5**
Sparse Local Global	121.4M	75.9	56.3
Axial	156.8M	73.5	56.2

Performance Comparison

Model	Pretrain	K400 Training Time (hours)	K400 Acc.	Inference TFLOPs	Params
I3D 8x8 R50	ImageNet-1K	444	71.0	1.11	28.0M
I3D 8x8 R50	ImageNet-1K	1440	73.4	1.11	28.0M
SlowFast R50	ImageNet-1K	448	70.0	1.97	34.6M
SlowFast R50	ImageNet-1K	3840	75.6	1.97	34.6M
SlowFast R50	N/A	6336	76.4	1.97	34.6M
TimeSformer	ImageNet-1K	**416**	75.8	**0.59**	121.4M
TimeSformer	ImageNet-21K	**416**	**78.0**	**0.59**	121.4M

TimeSTransformer

- Gedas Bertasius Heng Wang Lorenzo, Torresanils, *" Space-Time Attention All You Need for Video Understanding?*", International Conference on Machine Learning, PMLR 139, 2021. (175 Citations)

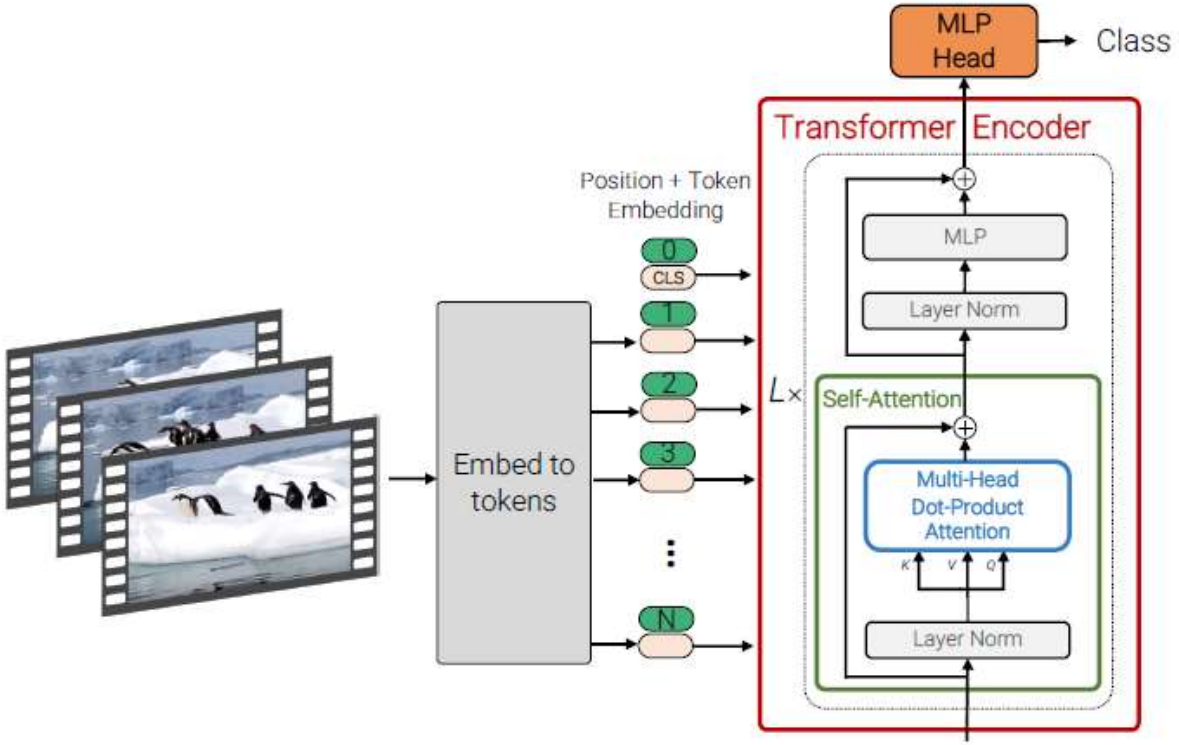
ViViT

- Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun Mario, Lu˘ci´cy, Cordelia Schmidy, "*ViViT: A Video Vision Transformer*", ICCV-2021. (124 Citations)

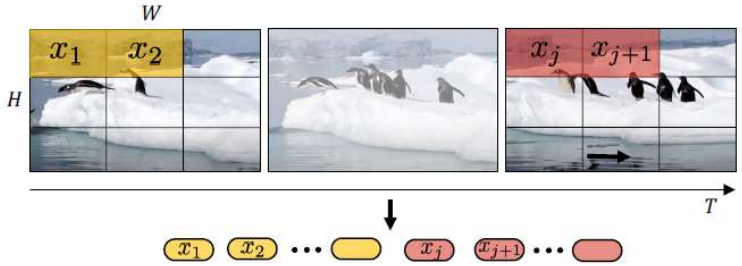
ViViT

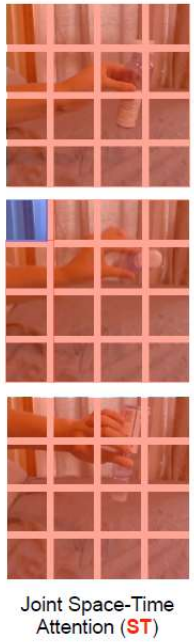
- Model extracts spatiotemporal tokens from the input video, which are then encoded by a series of transformer layers
- Efficient variants of model which factorize the spatial- and temporal-dimensions of the input
- Video models are initialized from pretrained ImageNet-21K or the larger JFT dataset
- Achieve state of-the-art results on multiple video classification benchmarks
  - Kinetics 400 and 600
  - Epic Kitchens 100
  - Something-Something v2
  - Moments in Time

ViViT


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

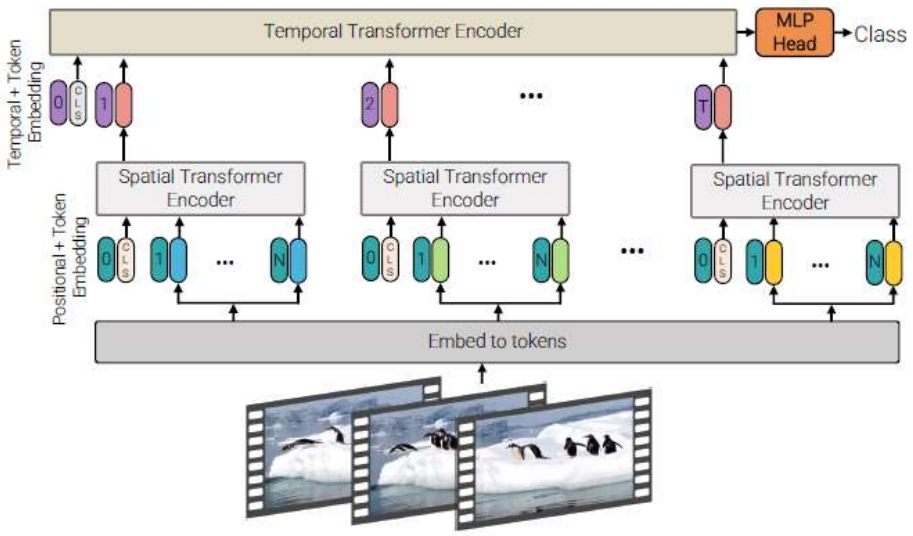
Model-1 Spatiotemporal Attention

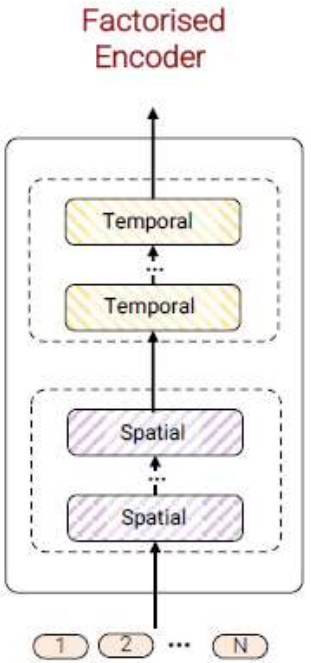




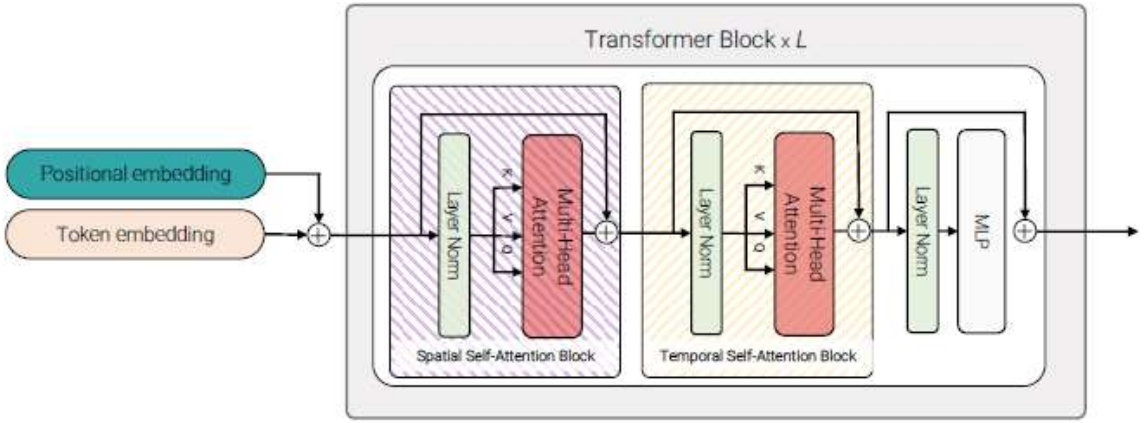
Joint Space-Time Attention (ST)

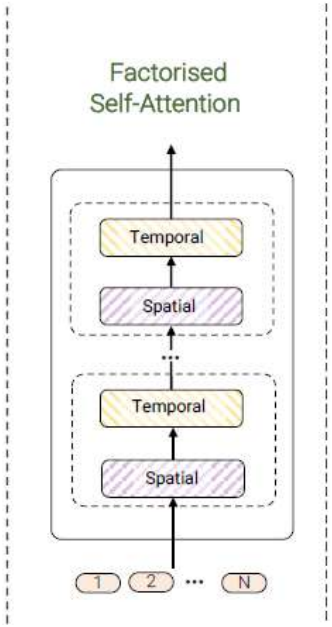
Model-2 Factorized Encoder





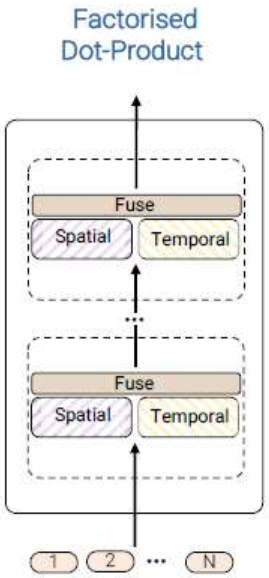
Model-3 Factorized Self-Attention





Model 4: Factorised dot-product attention

Attention weights for each token are computed separately over the spatial and temporal-dimensions using different heads



Results

	K400	EK	FLOPs ($\times 10^9$)	Params ($\times 10^6$)	Runtime (ms)
Model 1: Spatio-temporal	80.0	43.1	455.2	88.9	58.9
Model 2: Fact. encoder	78.8	43.7	284.4	115.1	17.4
Model 3: Fact. self-attention	77.4	39.1	372.3	117.3	31.7
Model 4: Fact. dot product	76.3	39.5	277.1	88.9	22.9
Model 2: Ave. pool baseline	75.8	38.8	283.9	86.7	17.3

ViViT

- Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun Mario, Lu˘ci´cy, Cordelia Schmidy, "*ViViT: A Video Vision Transformer*", ICCV-2021. (124 Citations)

Transformers in Vision: A Survey

- Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah, "*Transformers in Vision: A Survey*" ACM Computing Surveys, 2021. (182 citations)

Key Design Choices

Task	Method	Design Highlights (focus on differences with the standard form)	Input Data Type	Label Type	Loss
Image Classification	ViT [11]	Directly adopted NLP Transformer Encoder for images, Mechanism to linearly embed image patches with positional embedding suitable for the Encoder.	2D Image	Class labels	Cross-entropy
	DeiT [12]	Transformer as s student while CNN as a teacher, Distillation tokens to produce estimated labels from teacher, Attention between class and distillation tokens.	2D Image	Class labels	Cross-entropy, Distillation loss based on KL-divergence
	CLIP [195]	Jointly train image and text encoders on image-text pairs, to maximize similarity of valid pairs and minimize otherwise	2D Images & texts	Image-text pairs	Symmetric cross-entropy
Object Detection	DETR [13]	Linear projection layer to reduce CNN feature dimension, Spatial positional embedding added to each multi-head self-attention layer of both encoder and decoder. Object queries (output positional encoding) added to each multi-head self-attention laver of decoder.	2D Image	Class labels	Hungarian loss based on bipartite matching between predicted and ground truths
	D-DETR [14]	Deformable Transformer consists of deformable attention layers to introduce sparse priors in Transformers, Multi-scale attention module.	2D Image	Class labels	Hungarian loss

Advantages and Limitations of different Transformers based Methods

Task	Method	Metric	Dataset	Performance	Highlights	Limitations
Image Classification	ViT [11] ICLR'21	Top-1 Acc.	ImageNet	88.55	**a)** First application of Transformer (global self-attention) directly on image patches, **b)** Convolution-free network architecture, **c)** Outperforms CNN models such as ResNet.	**a)** Requires training on large-scale data *e.g.*, 300-Million images, **b)** Requires careful transfer learning to the new task, **c)** Requires large model with 632-Million parameters to achieve SOTA results.
	DeiT [12] arXiv'20	Top-1 Acc.	ImageNet	83.10	**a)** Successfully trains Transformer on ImageNet only, **b)** Introduces attention-based distillation method. **c)** Produces competitive performance with small (86-Million pa-	**a)** Requires access to pretrained CNN based teacher model thus performance depends on the quality of the teacher model.
	Swin-T [36] arXiv'21	Top-1 Acc.	ImageNet	84.5	**a)** Provides a general purpose backbone for different vision tasks e.g., classification, detection and segmentation **b)** A hierarchical design	**a)** Hard to train from scratch on smaller datasets **b)** Quadratic compute complexity inherent to the self-attention operation.
Object Detection	DETR [13] ECCV'20	AP	COCO	44.9	**a)** Use of Transformer allows end-to-end training pipeline for object detection, **b)** Removes the need for hand-crafted post-processing steps.	**a)** Performs poorly on small objects, **b)** Requires long training time to converge.
	D-DETR [14] ICLR'21	AP	COCO	43.8	**a)** Achieves better performance on small objects than DETR [13], **b)** Faster convergence than DETR [13]	Obtain SOTA results with 52.3 AP but with two stage detector design and test time augmentations.

Comparison of Transformer and CNN Models

Method	#Param (M)	GFLOPs	Top-1 Acc (%)
ResNet18 [67]⋆	11.7	1.8	69.8
EfficientNet-B3 [87]⋆	12.0	1.8	81.6
DeiT-T [12]	5.7	1.3	72.2
T2T-ViT$_t$-7 [35]	5.0	1.3	71.7
LocalViT-T [107]	5.9	1.3	74.8
CrossViT-T [104]	6.9	1.6	73.4
PVTv1-T [93]	13.2	1.9	75.1
ResT-Lite [110]	10.5	1.4	77.2
CaiT-XXX-24 [243]	12.0	2.5	77.6
PVTv2-B1 [97]	13.1	2.1	78.7
Lv-ViT-T [89]	8.5	–	79.1
RegionViT-T [100]	13.8	2.4	80.4
ResNet50 [67]⋆	25.6	4.1	76.1
ResNeXt50-32x4d [244]⋆	25.0	4.3	77.6
RegNetY-4G [86]⋆	21.0	4.0	80.0
EfficientNet-B4 [87]⋆	19.0	4.2	82.9
DeiT-S [12]	22.1	4.6	79.9
PVTv1-S [93]	24.5	3.8	79.8
LocalViT-S [107]	22.4	4.6	80.8
CrossViT-S [104]	26.7	5.6	81.0
TNT-S [88]	23.8	5.2	81.3
Swin-T [36]	29.0	4.5	81.3
NesT-T [111]	17.0	5.8	81.5
T2T-ViT$_t$-14 [35]	21.5	5.2	81.5
CvT-13 [96]	20.0	4.5	81.6
ResT-B [110]	30.3	4.3	81.6
Twins-SVT-S [37]	24.0	2.8	81.7
PVTv2-B2-Li [97]	22.6	3.9	82.1
RegionViT-S [100]	30.6	5.6	82.5
Lv-ViT-S [89]	26.0	6.6	83.3

Method	#Param (M)	GFLOPs	Top-1 Acc (%)
ResNet101 [67] ⋆	44.7	7.9	77.4
ResNeXt101-32x4d [244]⋆	44.2	8.0	78.8
RegNetY-8G [86]⋆	39.0	8.0	81.7
EfficientNet-B5 [87] ⋆	30.0	9.9	83.6
CvT-21 [96]	32.0	7.1	82.5
CaiT-S-24 [243]	32.2	9.4	82.7
T2T-ViT$_t$-19 [35]	39.0	9.8	81.4
PVTv1-M [93]	44.2	6.7	81.2
PVTv2-B3 [97]	45.2	6.9	83.2
NesT-S [111]	38.0	10.4	83.3
ResNet152 [67] ⋆	60.2	11.6	78.3
CaiT-S-36 [243]	48.0	13.9	83.3
T2T-ViT$_t$-24 [35]	64.0	15.0	82.2
PVTv1-L [93]	61.4	9.8	81.7
TNT-B [88]	66.0	14.1	82.8
Swin-S [36]	50.0	8.7	83.0
Twins-SVT-B [37]	56.0	8.3	83.2
RegionViT-B [100]	72.7	13.0	83.3
PVTv2-B4 [97]	62.6	10.1	83.6
ResNeXt101-64x4d [244] ⋆	83.5	15.6	79.6
RegNetY-16G [86] ⋆	84.0	16.0	82.9
EfficientNet-B6 [87] ⋆	43.0	19.0	84.0
NesT-B [111]	68.0	17.9	83.8
ViT-B/16 [11]	86.6	17.6	79.8
DeiT-B/16 [12]	86.6	17.6	81.8
Swin-B [36]	88.0	15.4	83.3
Twins-SVT-L [37]	99.2	14.8	83.7
PVTv2-B5 [97]	82.0	11.8	83.8
Lv-ViT-M [89]	56.0	16.0	84.1

UCF Center for Research in Computer Vision

CAP6412
Advanced Computer Vision

Mubarak Shah

shah@crcv.ucf.edu

Lecture-5-DeTR_TimeS_ViViT