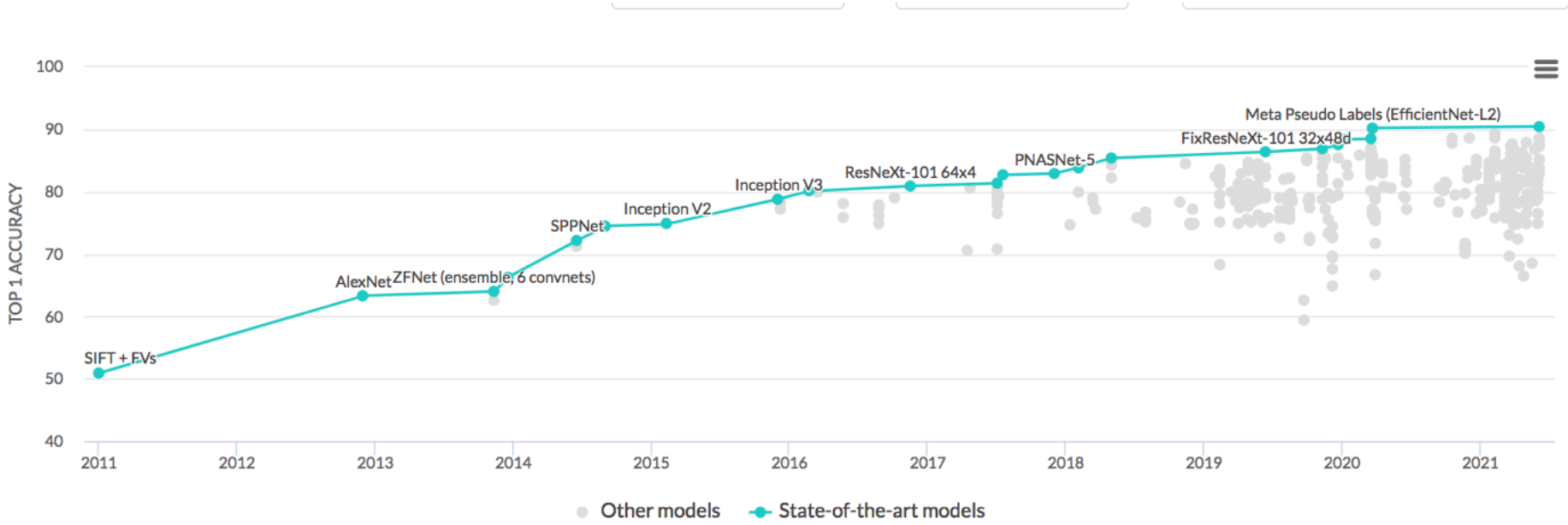
Video Understanding with Imperfect Data

Anurag Arnab

Google Research

Introduction

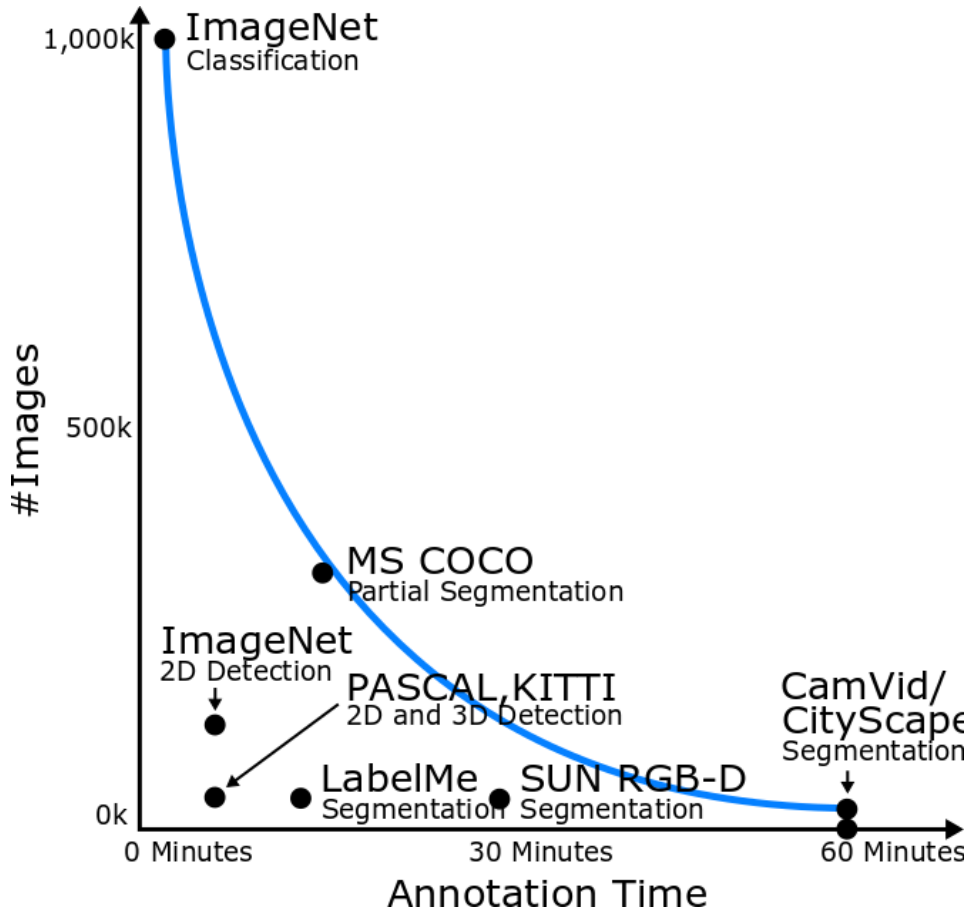
- Supervised learning works very well when we have a lot of data
- CNNs, and more recently Transformers, excel at learning from large supervised datasets.



ImageNet Top-1 Accuracy by year (from paperswithcode.com)

The Curse of Dataset Annotation

- Supervised learning works very well when we have a lot of data
- But what if we cannot get this data?

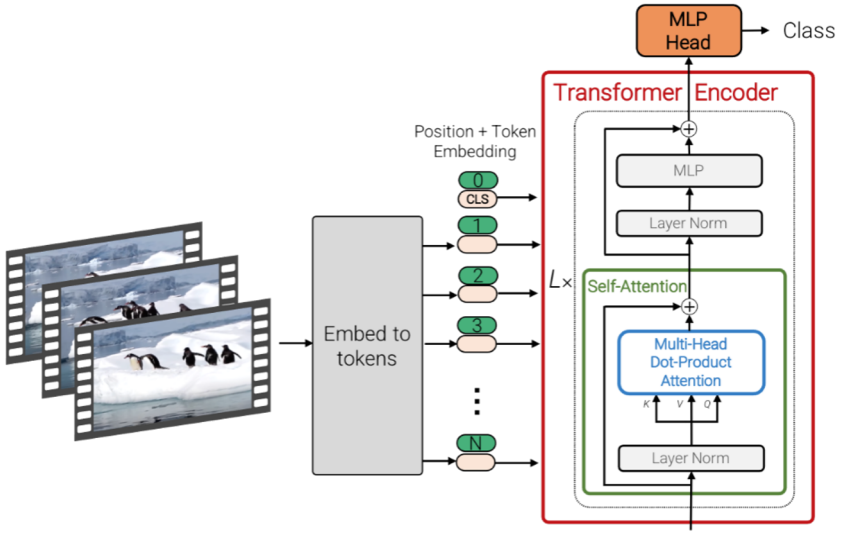


Xie *et al.* CVPR 2016

The Curse of Dataset Annotation: Video

- Simply too expensive to obtain labels for complex tasks
  - Requires labelling every frame of the video
  - Spatio-temporal action recognition, video segmentation
- Detailed annotations are ambiguous
  - Ambiguous when an action starts and ends
- Datasets as large-scale as those used for image classification or NLP don’t exist.

Outline

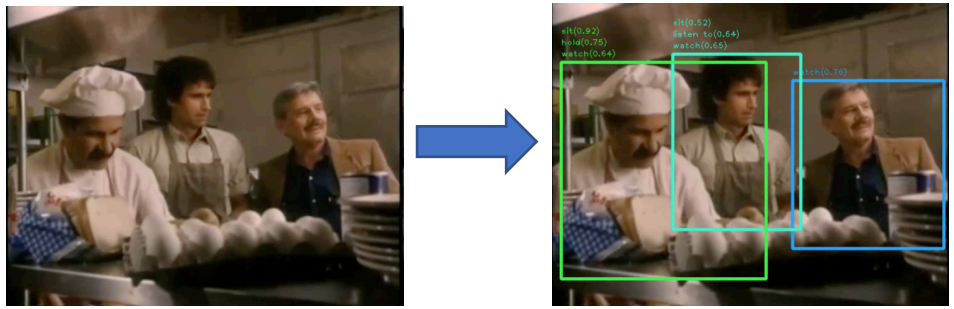


ViViT: A Video Vision Transformer (arxiv 2021)

Training large pure-transformer video models without large datasets

Annotation

Prediction



Labels: *sit, hold, watch, talk to, listen to*

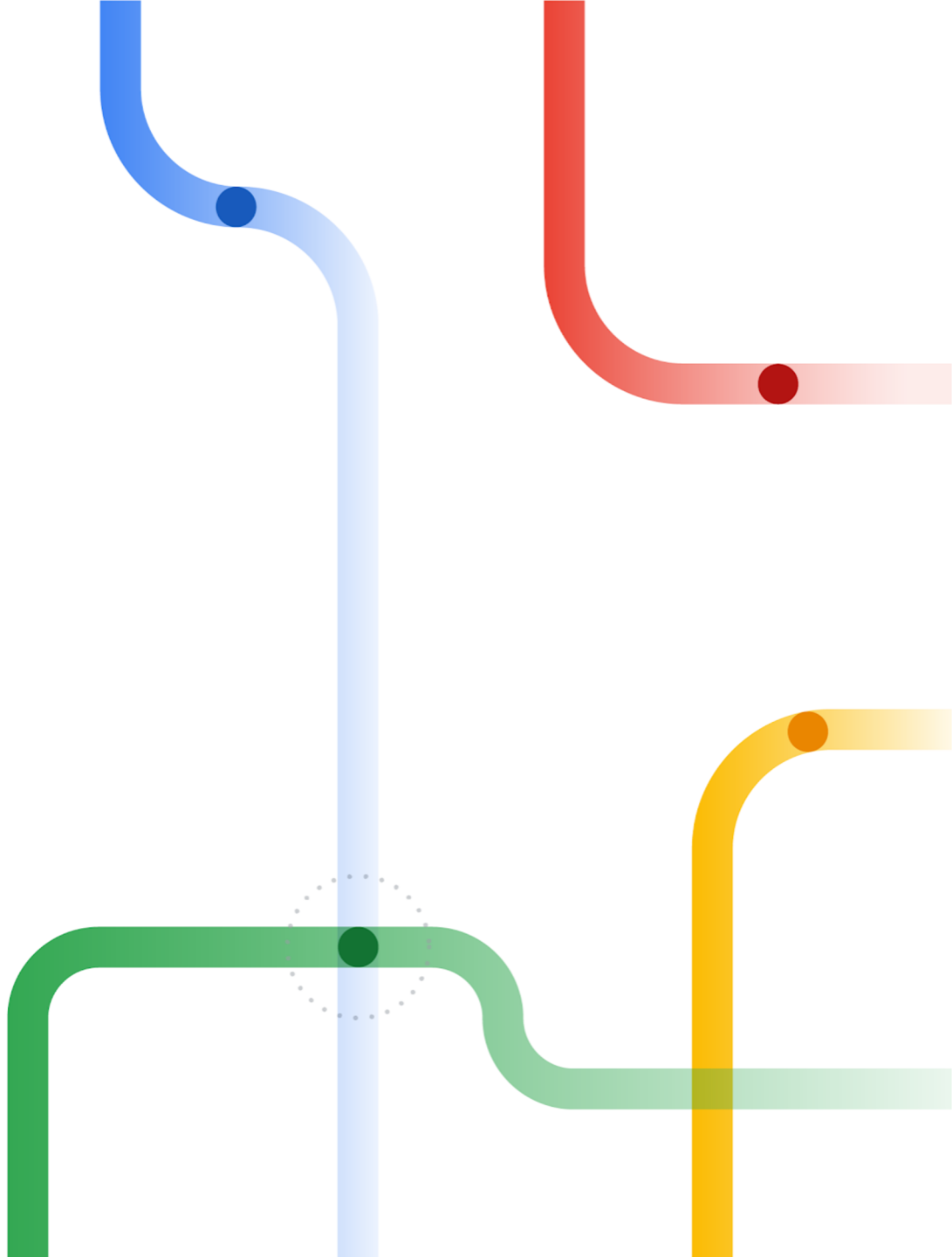
Uncertainty-Aware Weakly Supervised Action Detection from Untrimmed Videos (ECCV 2020)

Training spatio-temporal action detection models from only video-level tags.

ViViT: A Video Vision Transformer

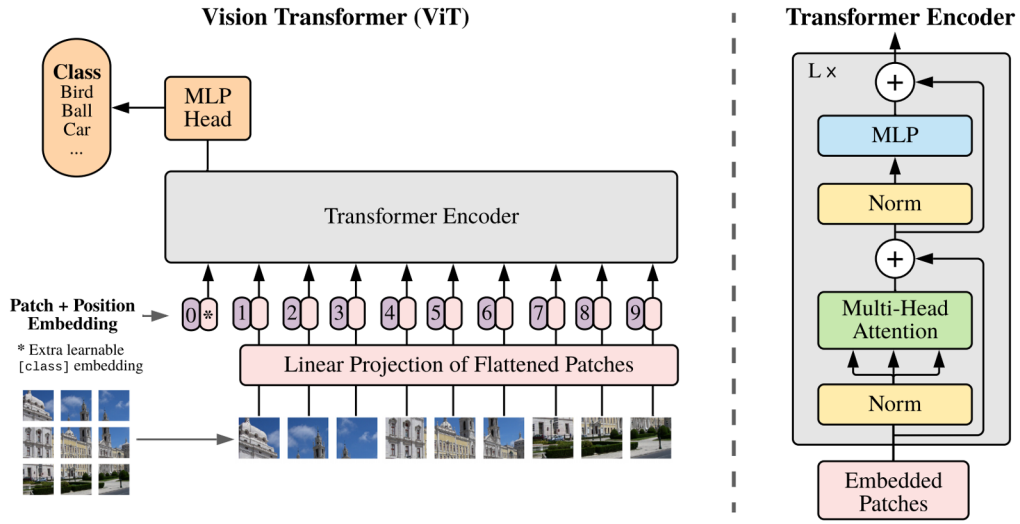
Anurag Arnab, Mostafa Dehghani,
Georg Heigold, Chen Sun,
Mario Lucic, Cordelia Schmid

Google Research



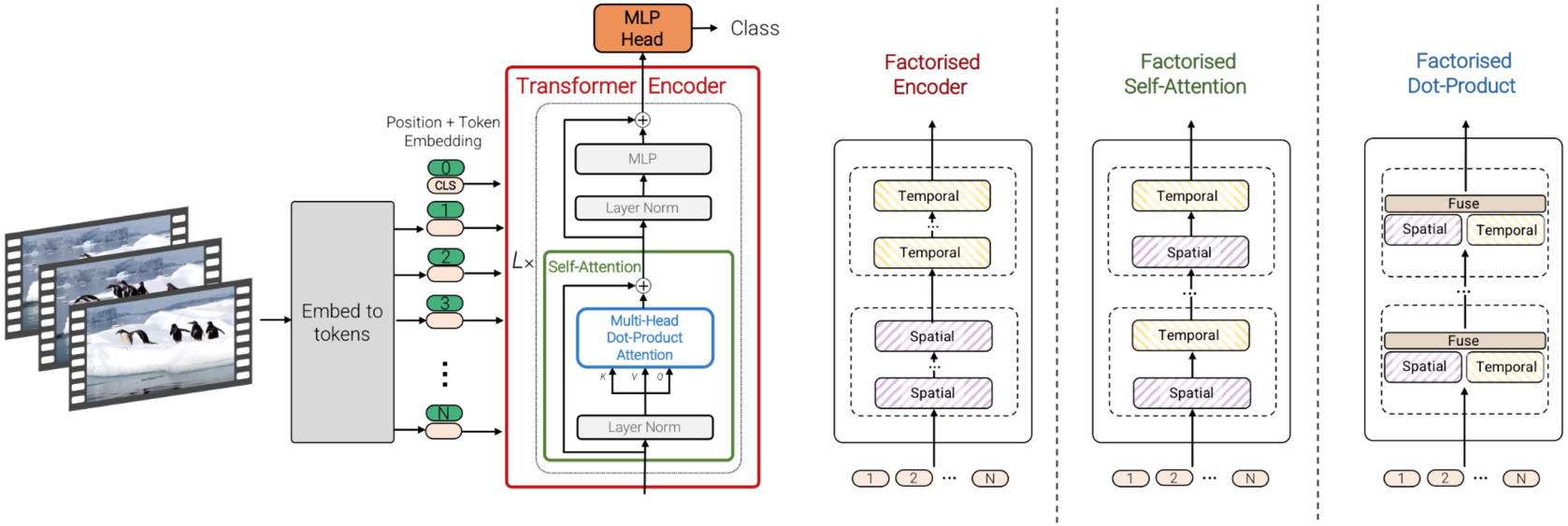
Introduction

- CNNs are architecture of choice in Vision ; Transformers are architecture of choice in Natural Language
- Vision Transformers: recent pure-transformer architecture for images
- Benefits of such architectures realised at large scale



ViViT: Video Vision Transformers

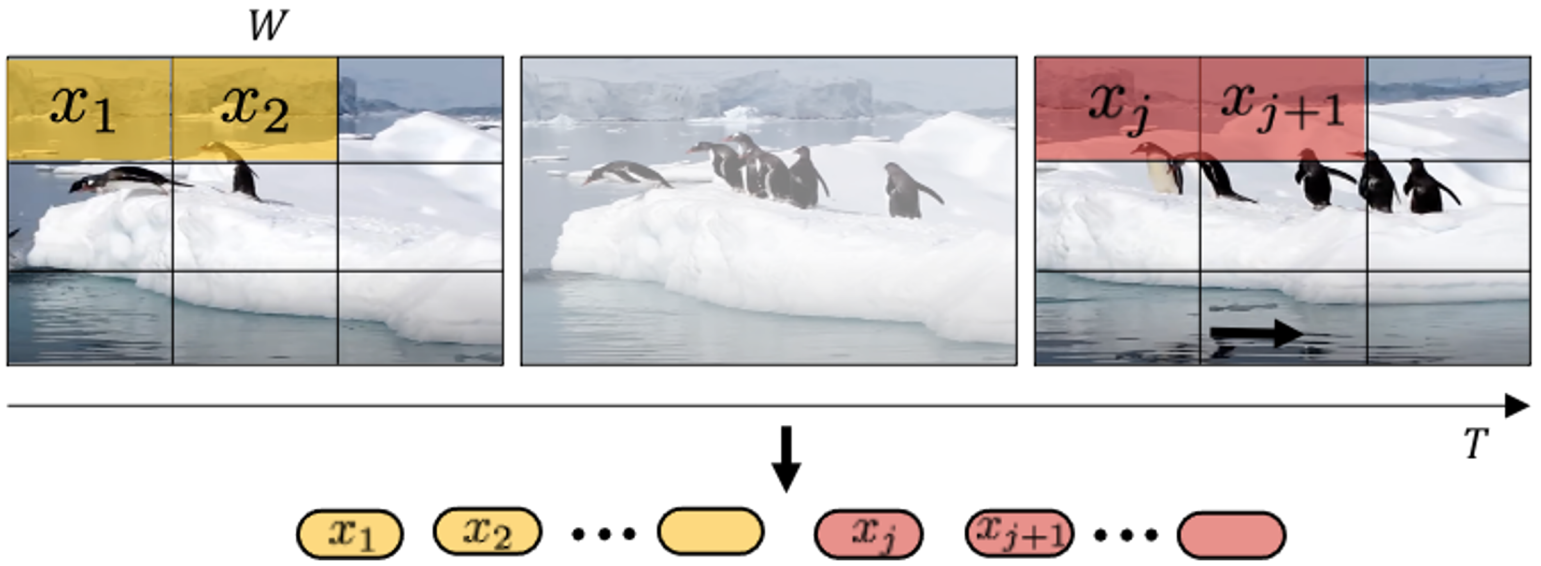
- Extend idea of ViT (static images) to videos
- To handle large number of tokens, explore more efficient factorised attention variants.
- Regularisation to train on comparatively small video datasets.



Google Research

Input Encoding 1: Uniform Frame Sampling

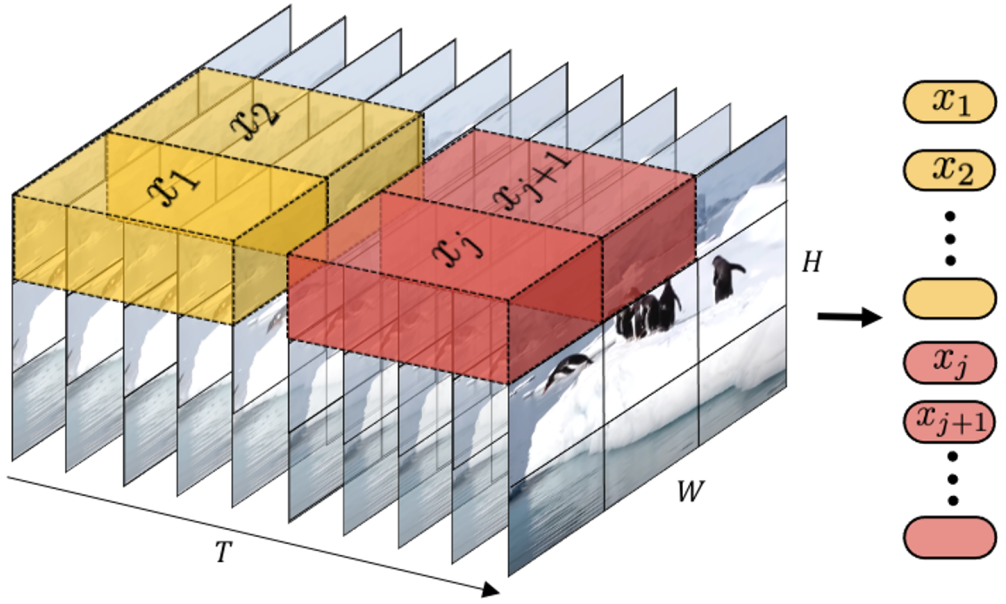
- Sample frames, extract 2D patches and linearly project (as in ViT)
- Effectively consider a video as a “big image”



Google Research

Input Encoding 2: Tubelet embedding

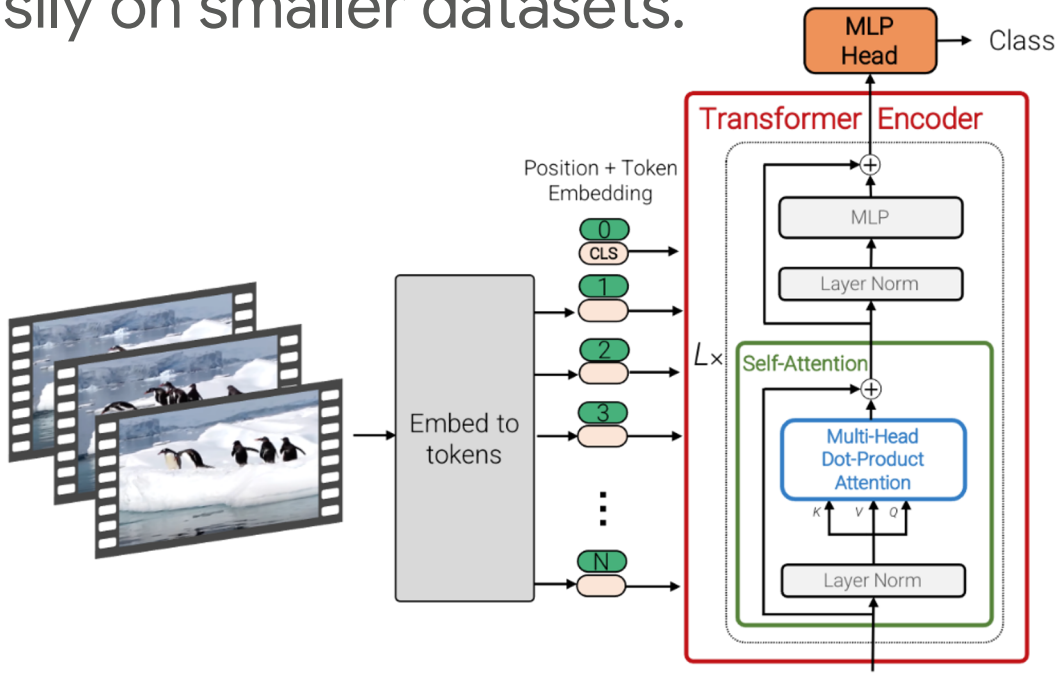
- Extract 3D tubelets to encode spatio-temporal “tubes” into tokens
- Temporal information included from the initial tokenisation stage.
- Works better when initialised appropriately.



Google Research

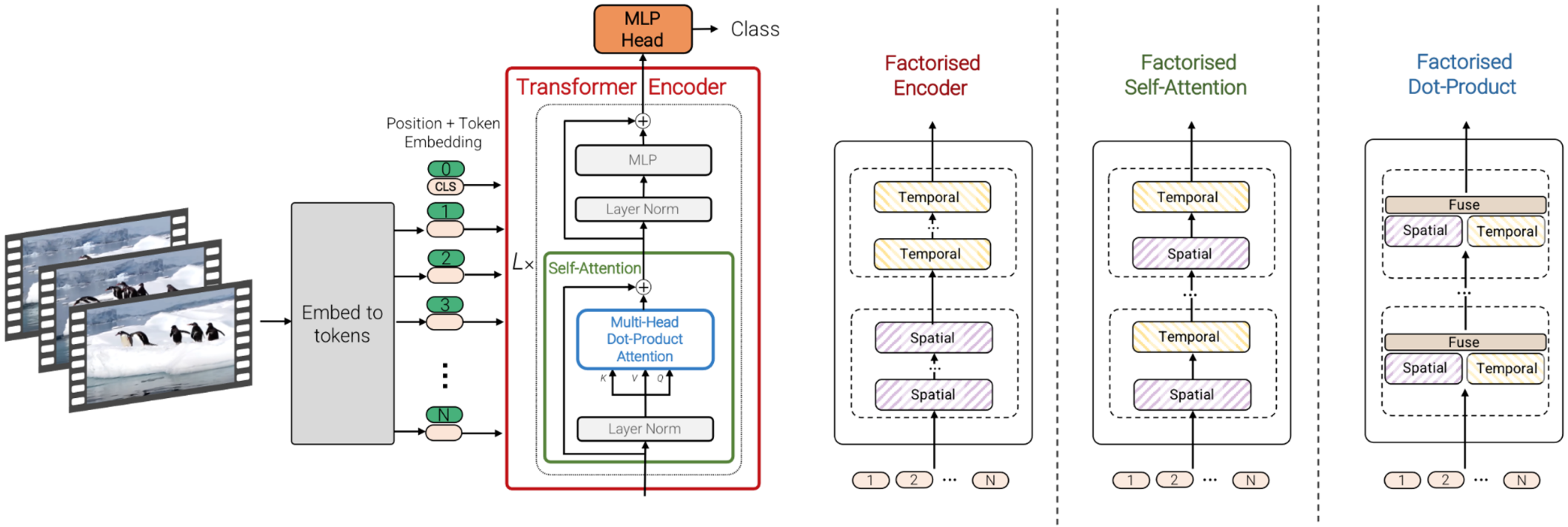
ViViT: Joint Spatio-Temporal Attention

- Simply forward many spatio-temporal tokens through multiple transformer layers.
- Requires a lot of computation, and high-capacity means it can overfit easily on smaller datasets.



Google Research

ViViT: Space/Time Factorisations



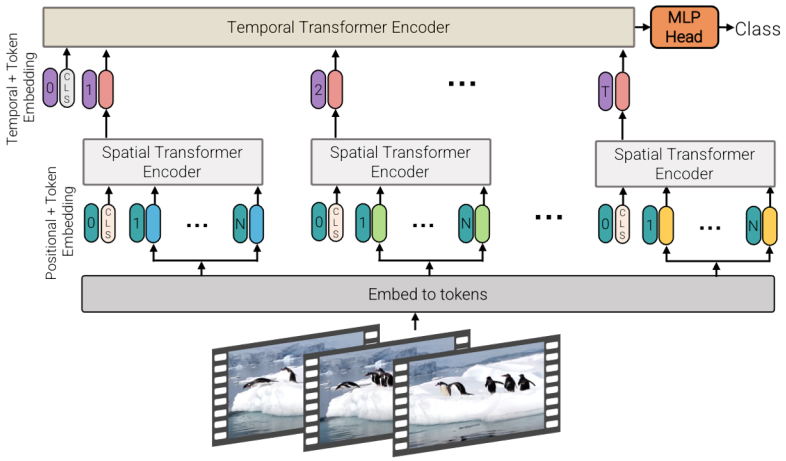
Alternative ways of mixing the temporal and spatial information

Reduces complexity from $O((w * h)^2 + t^2)$ instead of $O((w*h*t)^2)$

ViViT Factorisations

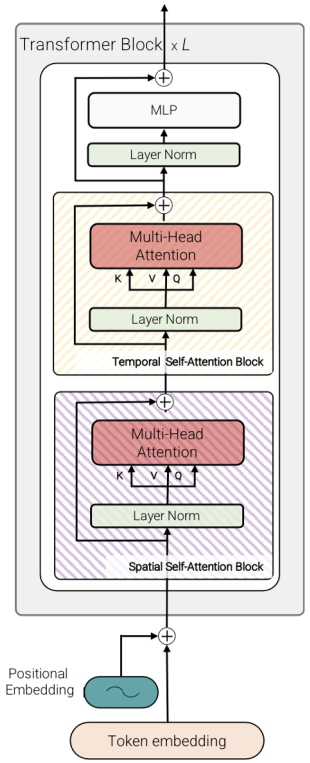
Factorised encoder

- “Late fusion” of spatial and temporal information



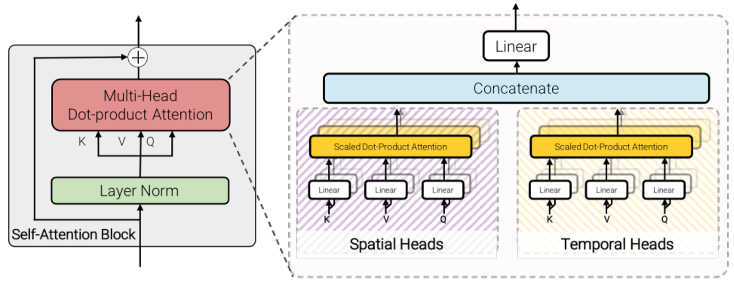
Factorised self-attention

- Perform self-attention separately over space and time



Factorised dot-product

- Attention heads separated over space and time dimensions.



Google Research

Input Encoding

- Tubelet embedding works better if 3D filter is initialised appropriately.
  - Filter inflation [1, 2]: $\mathbf{E} = \frac{1}{t}[\mathbf{E}_{\text{image}}, \ldots, \mathbf{E}_{\text{image}}, \ldots, \mathbf{E}_{\text{image}}]$.
  - Central frame initialiser: $\mathbf{E} = [\mathbf{0}, \ldots, \mathbf{E}_{\text{image}}, \ldots, \mathbf{0}]$.
    - Initialise to “select” central frame using 2D filter weights.

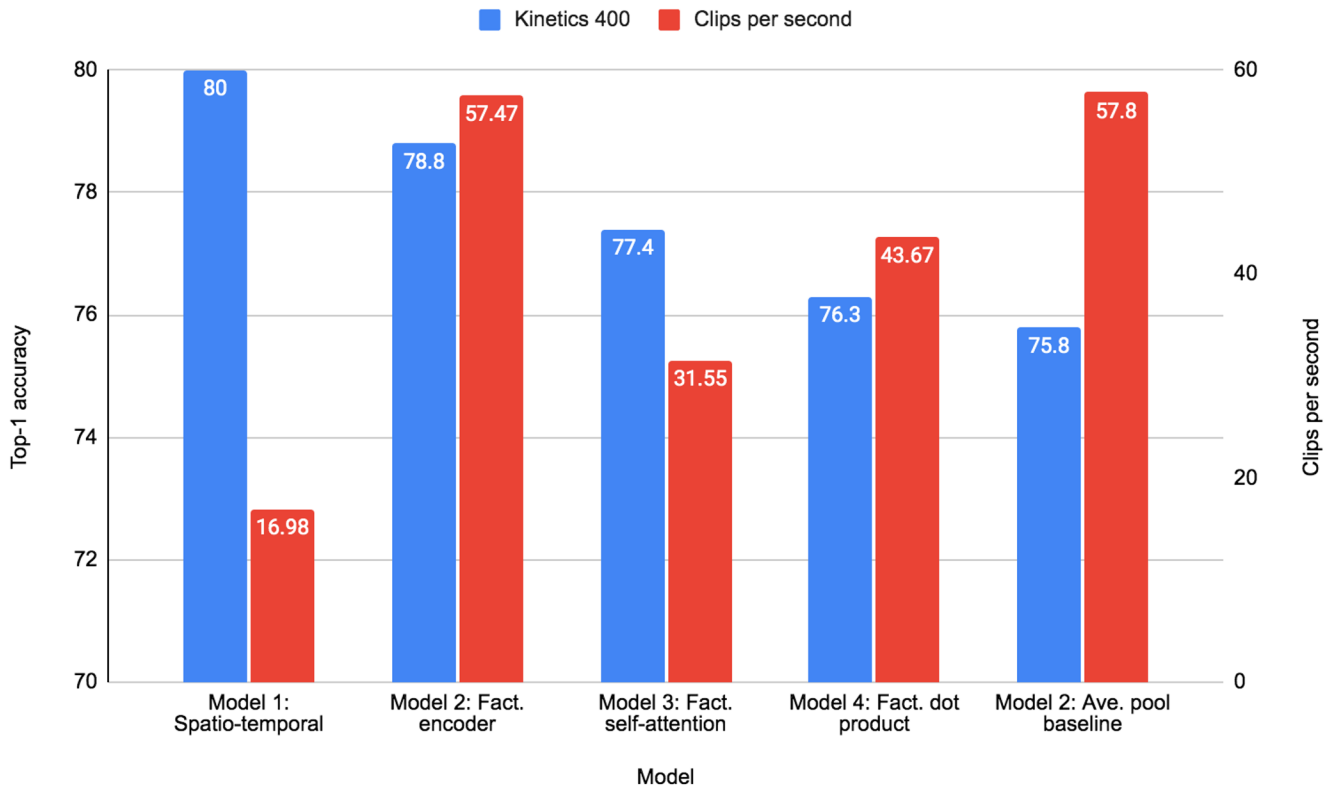
	Top-1 accuracy
Uniform frame sampling	78.5
Tubelet embedding	
Random initialisation [22]	73.2
Filter inflation [6]	77.6
Central frame	79.2

[1] Carreira and Zisserman. CVPR 2017.
[2] Feichtenhofer et al. NeurIPS 2016

Google Research

Model Variants

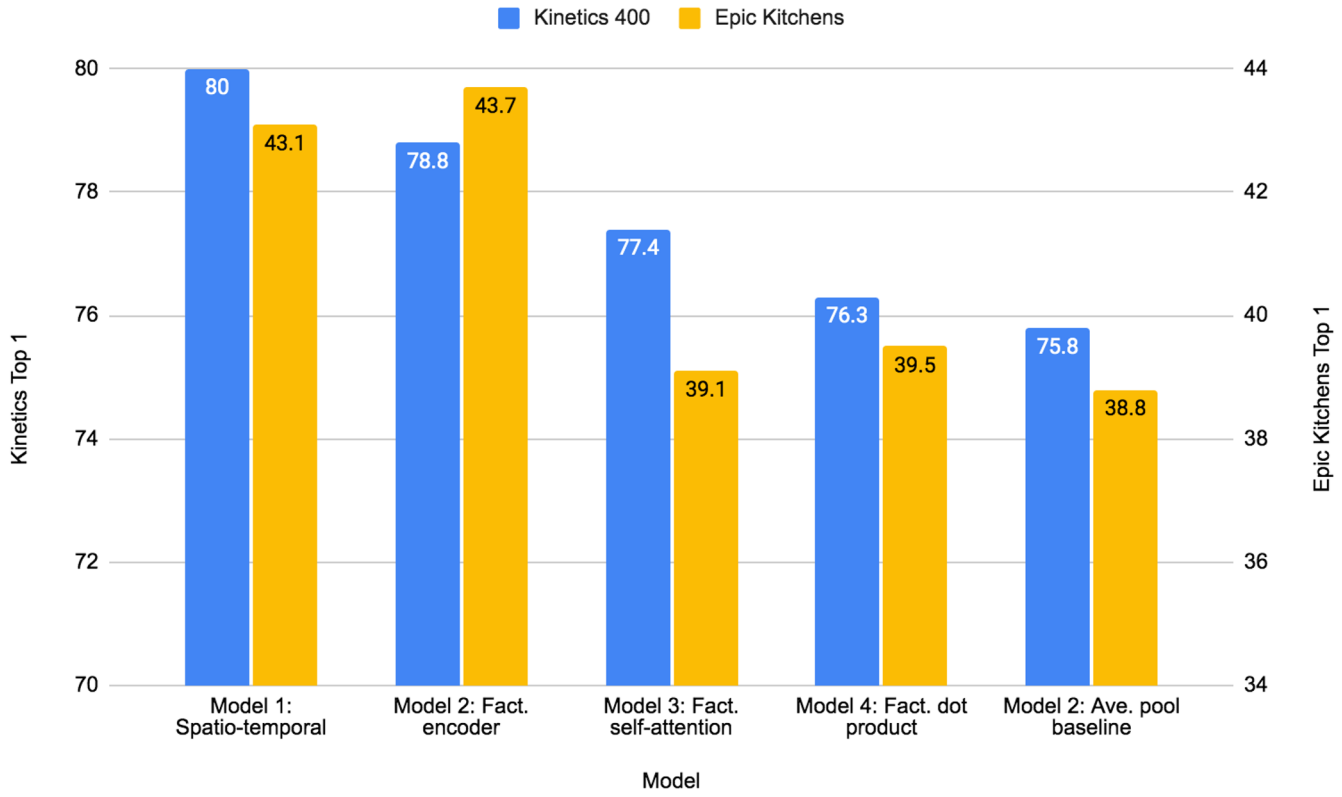
- Tokens fixed across models
- Unfactorised model works best on larger datasets (ie Kinetics), but slowest.



Google Research

Model Variants

- Factorised encoder works best on smaller datasets (ie Epic Kitchens) as it overfits less.



Google Research

Regularisation

- Video datasets are not as large as ImageNet / ImageNet21k / JFT
  - Original ViT paper didn't get good performance on ImageNet.
- Strategies
  - Use pretrained image models from ImageNet-21K or JFT
  - For smaller datasets, we use further regularisation methods, inspired by DeIT.

	Top-1 accuracy
Random crop, flip, colour jitter	38.4
+ Kinetics 400 initialisation	39.6
+ Stochastic depth [28]	40.2
+ Random augment [10]	41.1
+ Label smoothing [58]	43.1
+ Mixup [79]	43.7

5.3% gain on Epic Kitchens

Google Research

State-of-the-art Results on 5 Datasets

(a) Kinetics 400

Method	Top 1	Top 5	Views
blVNet [16]	73.5	91.2	–
STM [30]	73.7	91.6	–
TEA [39]	76.1	92.5	10 × 3
TSM-ResNeXt-101 [40]	76.3	–	–
I3D NL [72]	77.7	93.3	10 × 3
CorrNet-101 [67]	79.2	–	10 × 3
ip-CSN-152 [63]	79.2	93.8	10 × 3
LGD-3D R101 [48]	79.4	94.4	–
SlowFast R101-NL [18]	79.8	93.9	10 × 3
X3D-XXL [17]	80.4	94.6	10 × 3
TimeSformer-L [2]	80.7	94.7	1 × 3
ViViT-L/16x2	80.6	94.7	4 × 3
ViViT-L/16x2 320	**81.3**	**94.7**	4 × 3
Methods with large-scale pretraining			
ip-CSN-152 [63] (IG [41])	82.5	95.3	10 × 3
ViViT-L/16x2 (JFT)	82.8	95.5	4 × 3
ViViT-L/16x2 320 (JFT)	83.5	95.5	4 × 3
ViViT-H/16x2 (JFT)	**84.8**	**95.8**	4 × 3

(b) Kinetics 600

Method	Top 1	Top 5	Views
AttentionNAS [73]	79.8	94.4	–
LGD-3D R101 [48]	81.5	95.6	–
SlowFast R101-NL [18]	81.8	95.1	10 × 3
X3D-XL [17]	81.9	95.5	10 × 3
TimeSformer-HR [2]	82.4	**96.0**	–
ViViT-L/16x2	82.5	95.6	4 × 3
ViViT-L/16x2 320	**83.0**	95.7	4 × 3
ViViT-L/16x2 (JFT)	84.3	96.2	4 × 3
ViViT-H/16x2 (JFT)	**85.8**	**96.5**	4 × 3

(c) Moments in Time

	Top 1	Top 5
TSN [69]	25.3	50.1
TRN [83]	28.3	53.4
I3D [6]	29.5	56.1
blVNet [16]	31.4	59.3
AssembleNet-101 [51]	34.3	62.7
ViViT-L/16x2	**38.0**	**64.9**

(d) Epic Kitchens 100 Top 1 accuracy

Method	Action	Verb	Noun
TSN [69]	33.2	60.2	46.0
TRN [83]	35.3	65.9	45.4
TBN [33]	36.7	66.0	47.2
TSM [40]	38.3	**67.9**	49.0
SlowFast [18]	38.5	65.6	50.0
ViViT-L/16x2 Fact. encoder	**44.0**	66.4	**56.8**

(e) Something-Something v2

Method	Top 1	Top 5
TRN [83]	48.8	77.6
SlowFast [17, 77]	61.7	–
TimeSformer-HR [2]	62.5	–
TSM [40]	63.4	88.5
STM [30]	64.2	89.8
TEA [39]	65.1	–
blVNet [16]	65.2	**90.3**
ViViT-L/16x2 Fact. encoder	**65.4**	89.8

Google Research

Conclusion

- Family of pure-transformer architectures for video
- Showed how to regularise models appropriately to train on smaller datasets. Detailed ablations in paper
- State-of-the-art results on 5 video datasets

- A Arnab *et al*. ViViT: A Video Vision Transformer. Arxiv 2103.15691, 2021. [PDF]

Google Research

ECCV'20 ONLINE
23-28 AUGUST 2020

Uncertainty-Aware Weakly Supervised Action Detection from Untrimmed Videos

Anurag Arnab, Chen Sun,
Arsha Nagrani, Cordelia Schmid

Google Research

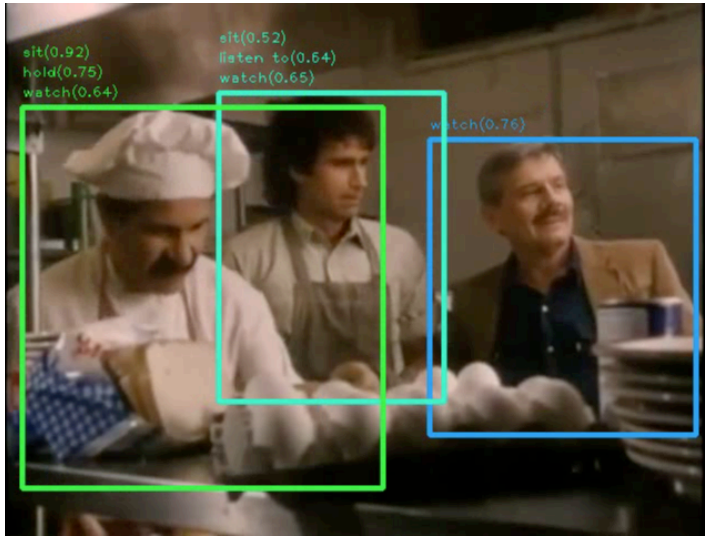
Introduction

- Spatio-temporal action detection
  - Bounding box in space and time around action of interest
- Most approaches extend detectors, such as Faster-RCNN and SSD, temporally.
- In this paper, we only use cheap, video-level labels

Annotation

Labels: *sit, hold, watch, talk to, listen to*

Prediction



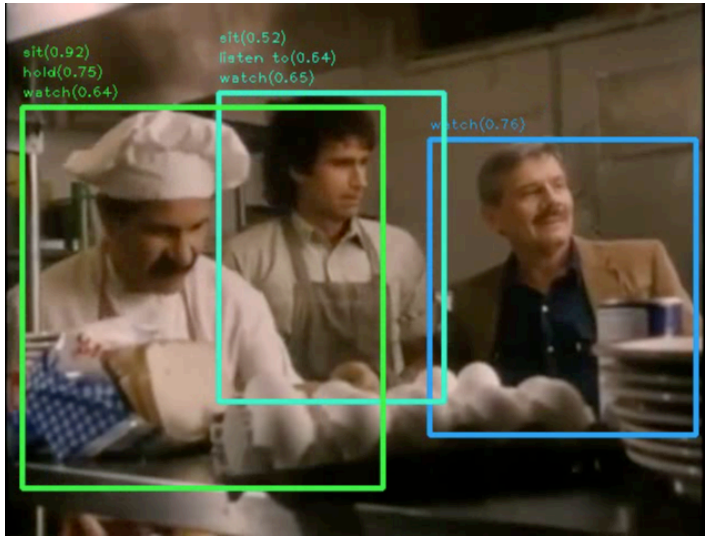
Weaker supervision

- Labelling bounding boxes per frame is too expensive
- Temporal boundaries of actions are ambiguous, annotators often do not agree with each other
- Only use cheap, video-level labels.

Annotation

Labels: *sit, hold, watch, talk to, listen to*

Prediction



Approach Overview

- Leverage off-the-shelf, per-frame person detectors to obtain person tubelets.
- Multiple Instance Learning
  - Each bag is formed from all tubelets in the video
- Due to noise, and violations of the MIL assumptions, predict the uncertainty for each bag as well.

Multiple Instance Learning

- Have a bag of examples, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$.
- Only know label for the whole bag, y.
- Key assumption is that one or more instances in the bag have label y.
- Want to train an instance-level classifier.
    - Classify each instance in the bag.

Multiple Instance Learning

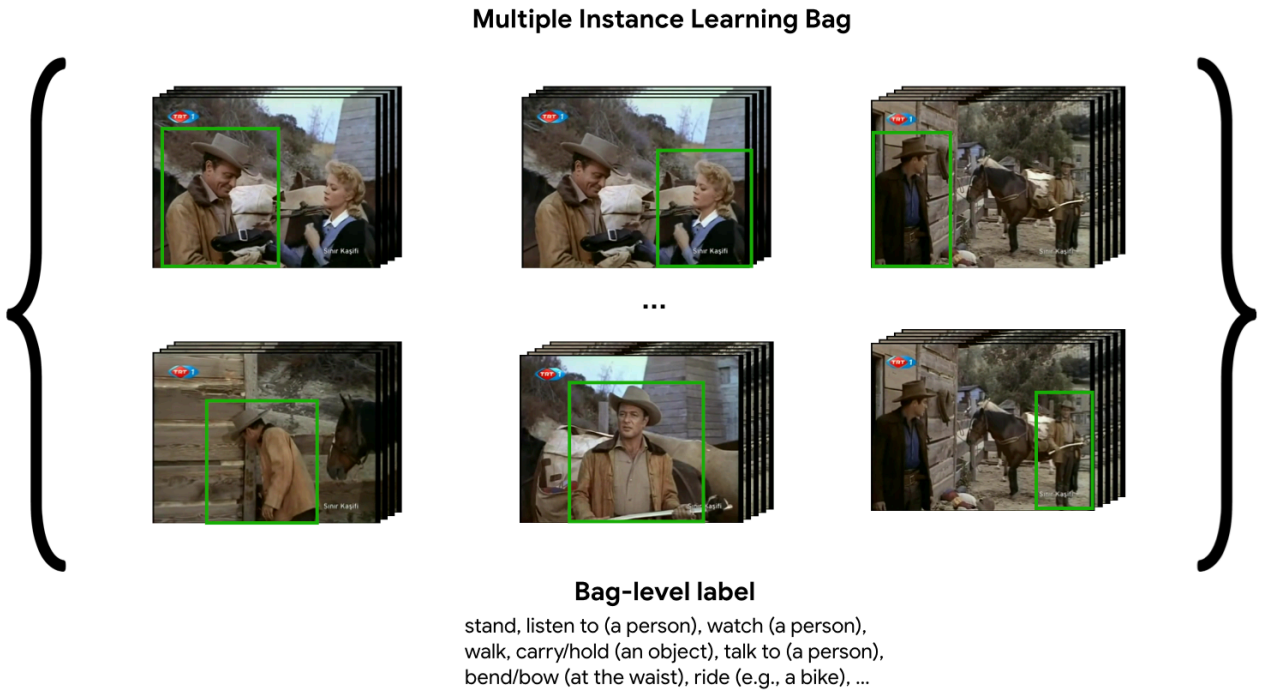
- Have a bag of examples, $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$.
- Only know label for the whole bag, y.
- Want to train an instance-level classifier.
- Aggregate instance-level predictions into a bag-level prediction.

$$p(y_l = 1 | x_1, x_2, \ldots, x_n) = g(p_i, p_2, \ldots, p_n)$$

- Use standard loss function
- Common aggregation functions: max, log-sum-exp, average, attention

Multiple Instance Learning (MIL)

- All the person tubelets within a video form a “bag”
  - Person tubelets are detections linked over at most K frames.
- The standard MIL assumption is that at least one tubelet in the bag has the video-level label.



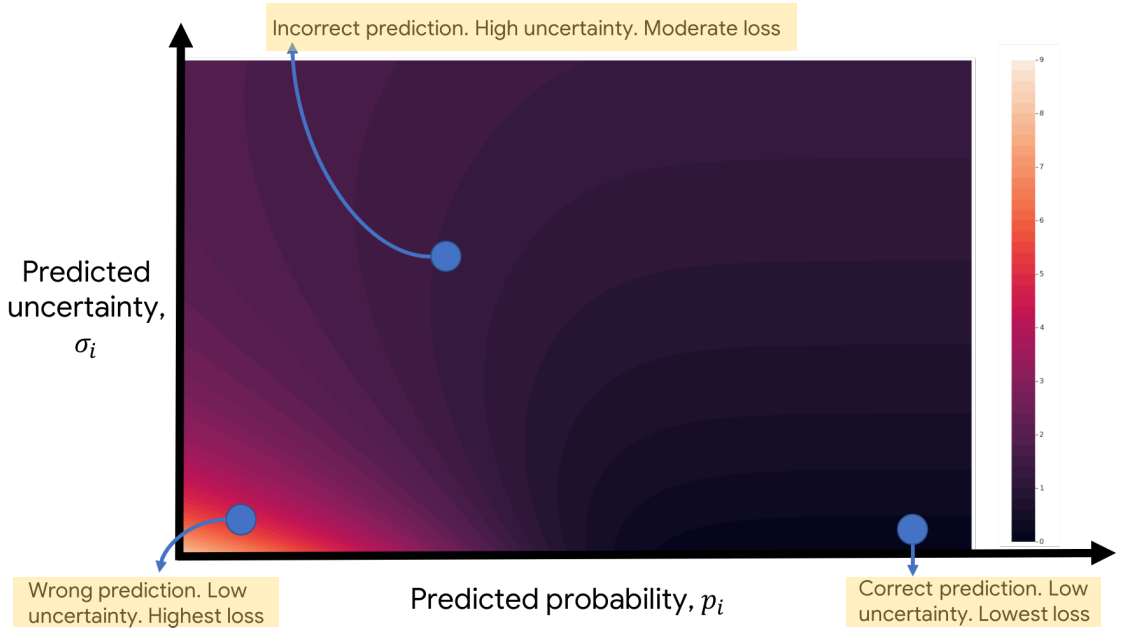
Label Noise and Violations of MIL Assumption

- MIL assumption is often violated
- Sampling bags
  - Cannot fit a whole bag in memory
  - Particularly as videos get longer
  - Uniformly sample tubes
- Person detector errors
  - Due to domain gap
  - False positives as some datasets don’t label actors exhaustively

All person detections besides the pole-vaulter are considered false-positives in this dataset.

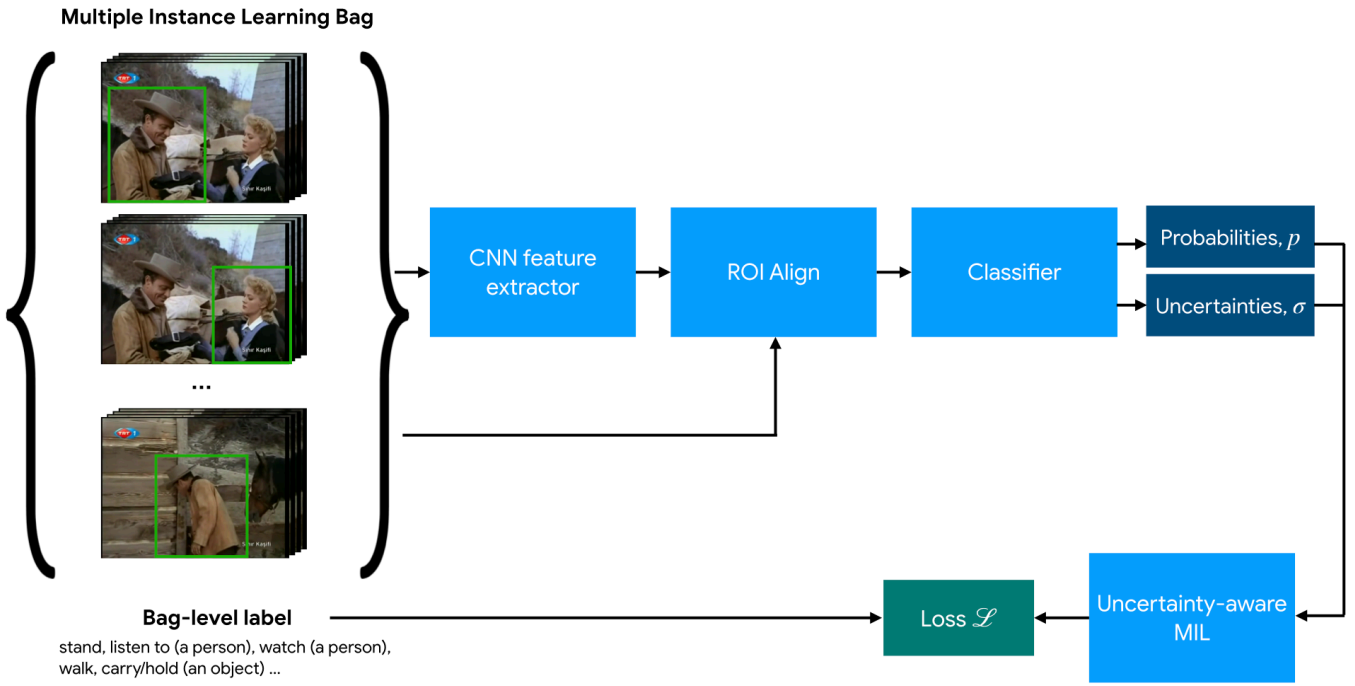
Uncertainty Estimation

- Predict uncertainty for each instance in the bag
- Intuition:
  - When possible, predict correct label with low uncertainty
  - Otherwise, predict incorrect label with high uncertainty.
- $L(x, y, \sigma) = \frac{1}{\sigma^2}\mathcal{L}_{ce}(x, y) + \lambda \log(\sigma^2)$



Network Architecture

- Fast-RCNN style detector
- Use person tubelets as proposals

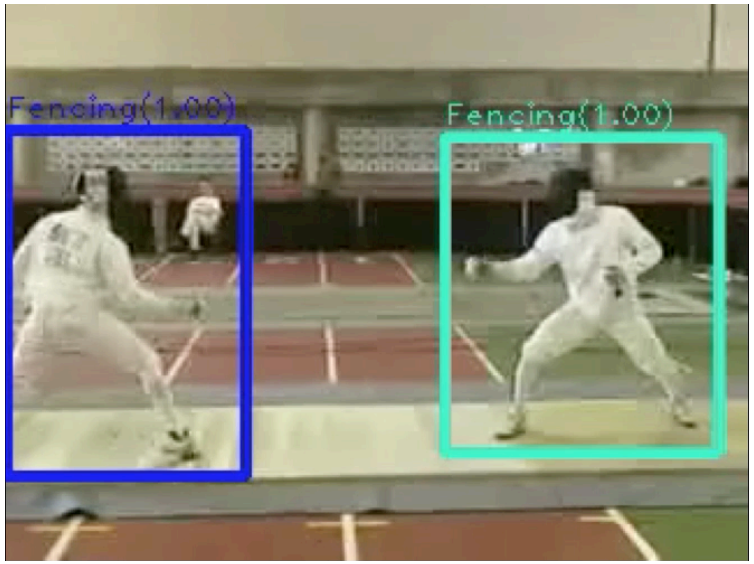


Evaluation Datasets

- UCF101-24
  - Most common dataset
  - Sports videos from YouTube, 24 classes
  - Many “background people” not doing the labelled action
  - Evaluate Video AP
- AVA
  - 60 atomic actions, from 15 minute movie clips
  - Keyframes at 1Hz, are labelled. Predict actions at keyframe given temporal context
  - Evaluate Frame AP

UCF101-24 Ablation

	Video AP	
	0.2	0.5
Weakly supervised baseline	54.3	29.7
MIL - LSE pooling	60.1	33.1
MIL - mean pooling	60.3	33.0
MIL - max pooling	60.7	33.5
MIL - max pooling, uncertainty	61.7	35.0
Fully supervised	69.3	43.6



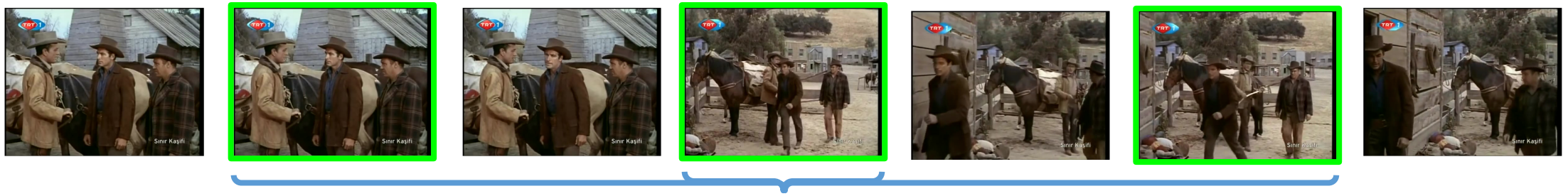
- Big domain gap between COCO and UCF
- Detector, trained only on COCO, has 47% recall and 21% precision on UCF training set.
- Sampling tubelets is necessary: Average of 33.1 tubelets per video, V100 GPU can hold 16.

UCF101-24 Comparison

	Video AP at 0.2	Video AP at 0.5
Fully supervised		
Peng *et al.* [35]	42.3	35.9
Hou *et al.* [17]	47.1	–
Weinzaepfel *et al.* [50]	58.9	–
Saha *et al.* [38]	63.1	33.1
Singh *et al.* [41]	73.5	46.3
Zhao *et al.* [52]	78.5	50.3
Singh *et al.* [40]	79.0	50.9
Kalogeiton *et al.* [19]	77.2	51.4
Ours	69.3	43.6
Weakly supervised		
Escorcia *et al.* [8]	45.5	–
Chéron *et al.* [6]	43.9	17.7
Ours	61.7	35.0

AVA

- AVA labels keyframes at 1Hz, videos are 15 minutes long.
- Vary the subclip of the video from which we take clip-level annotation
- Problem gets harder as the subclip duration is increased.



AVA Results

	Sub-clip duration (seconds)						
	FS	1	5	10	30	60	900
Frame AP	24.9	22.4	18.0	15.8	11.4	9.1	4.2

Conclusion

- Weakly-supervised spatio-temporal action detection with Multiple Instance Learning
- Predict uncertainty to better handle noise and violations of standard MIL assumption.

- A Arnab *et al*. Uncertainty-Aware Weakly Supervised Action Detection from Untrimmed Videos. ECCV 2020 [PDF]

Conclusion

- A Arnab *et al*. ViViT: A Video Vision Transformer. Arxiv 2103.15691, 2021. [PDF]
- A Arnab *et al*. Uncertainty-Aware Weakly Supervised Action Detection from Untrimmed Videos. ECCV 2020 [PDF]
- Contact: anurag.arnab@gmail.com