Video Action Transformer Network

Based on: Girdhar el al.[2019]
Advisor: Maria Bravo

Christian Leininger

Block-Seminar on Deep Learning
for Bio-Medical Data Analysis

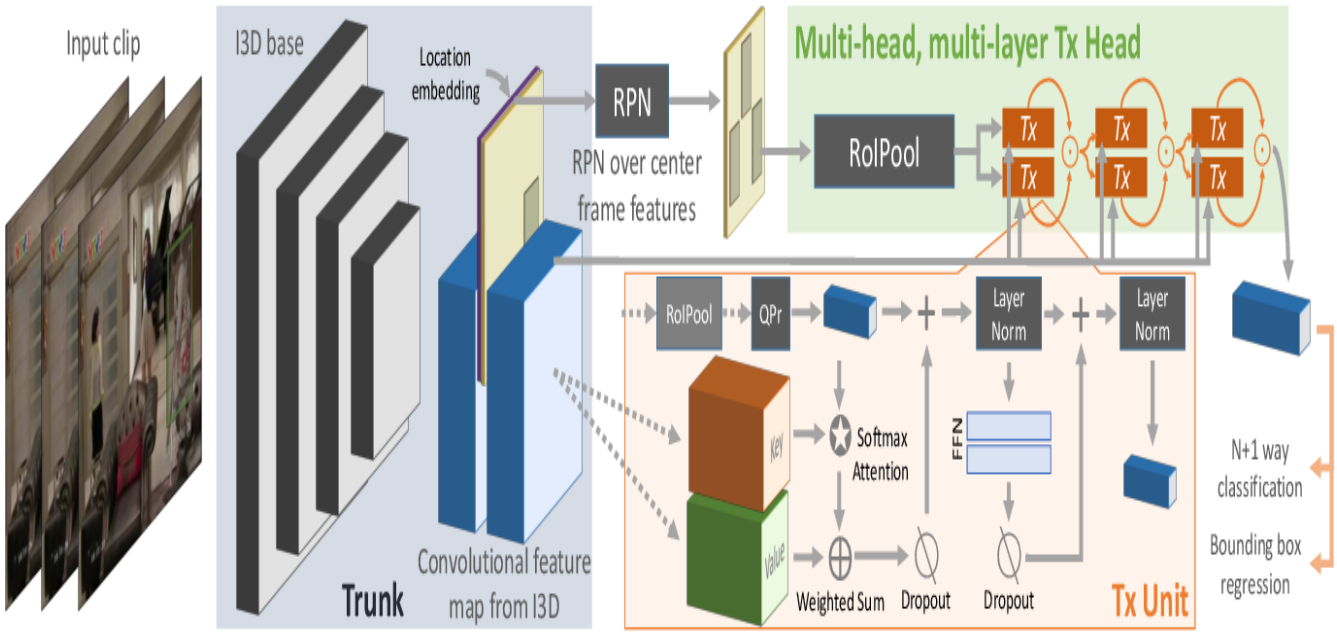
April 4, 2020

Objective of this work

- Localize all Humans and recognize their actions in a Video Clip
- Overcome the limitations of RNNs with self-attention
- Extract contextual Information without supervision

Video Action Transformer Network

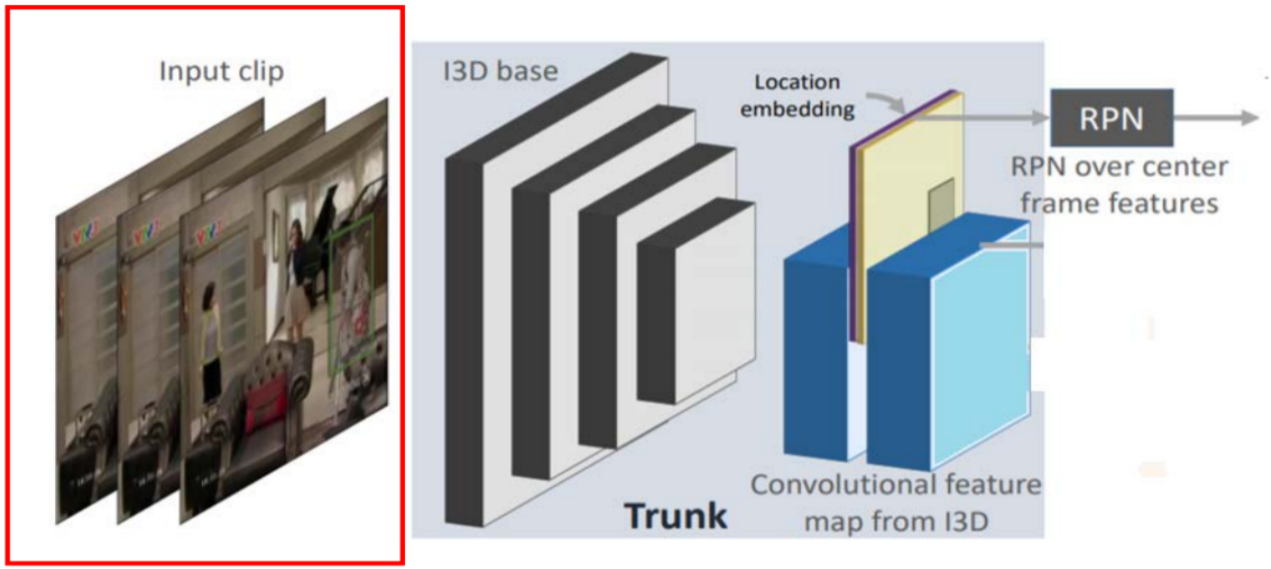
Video Action Transformer



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Overview of the proposed model

Video Action Transformer Network

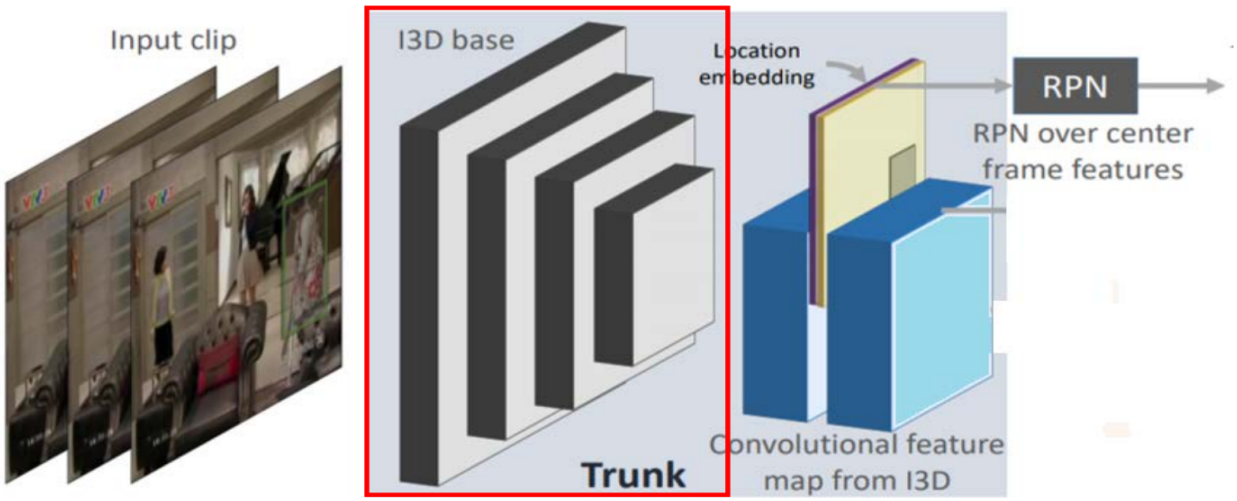
Trunk Region



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: T-frame Input of the model

Video Action Transformer Network

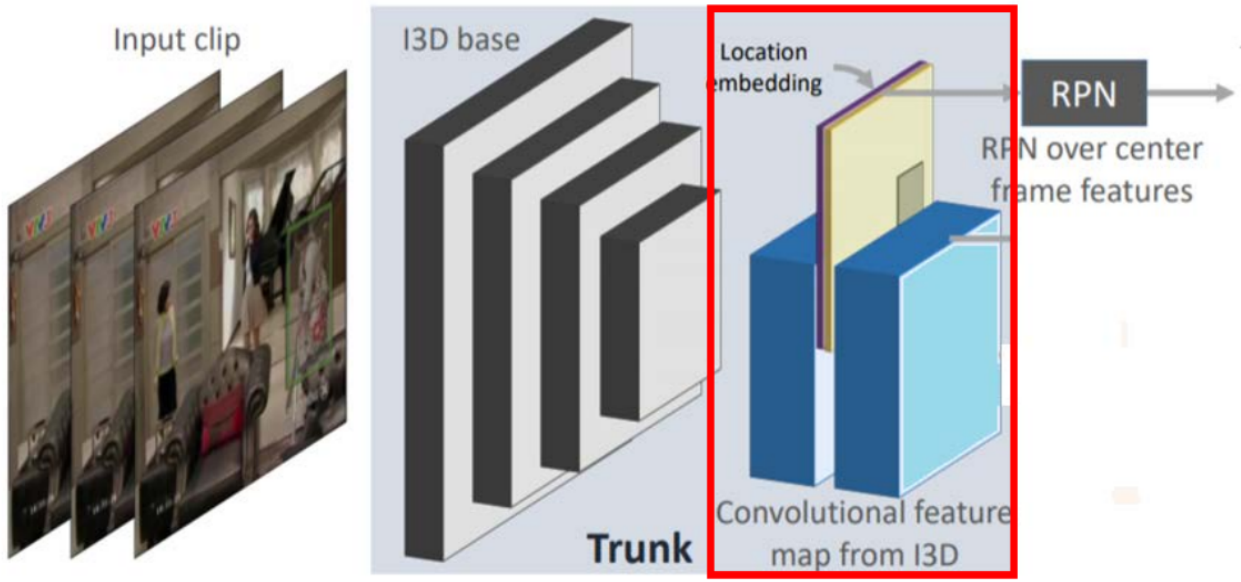
Trunk Region



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: I3D Base of the Trunk Region

Video Action Transformer Network

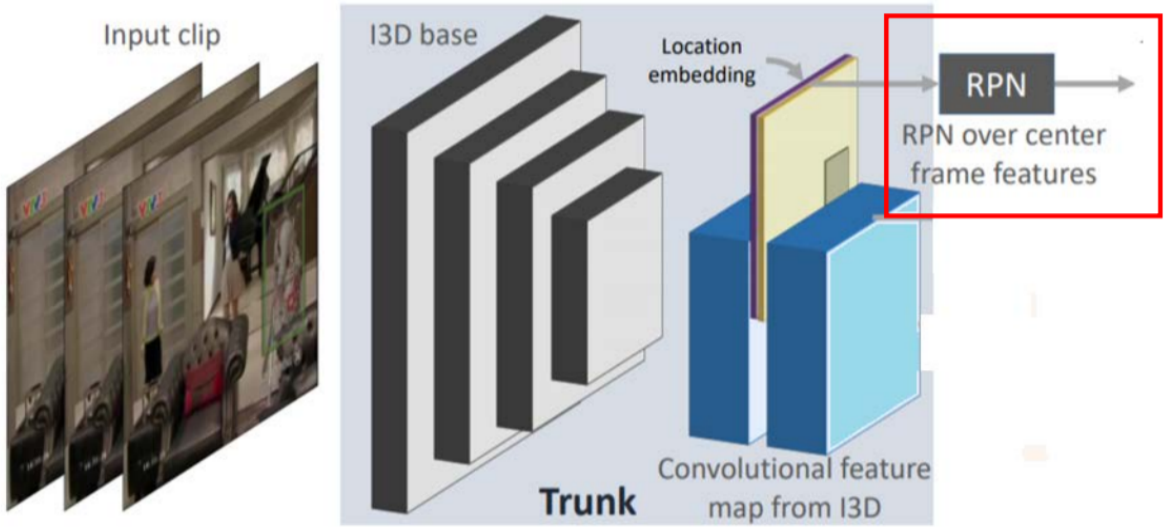
Trunk Region



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: I3D Base of the Trunk Region

Video Action Transformer Network

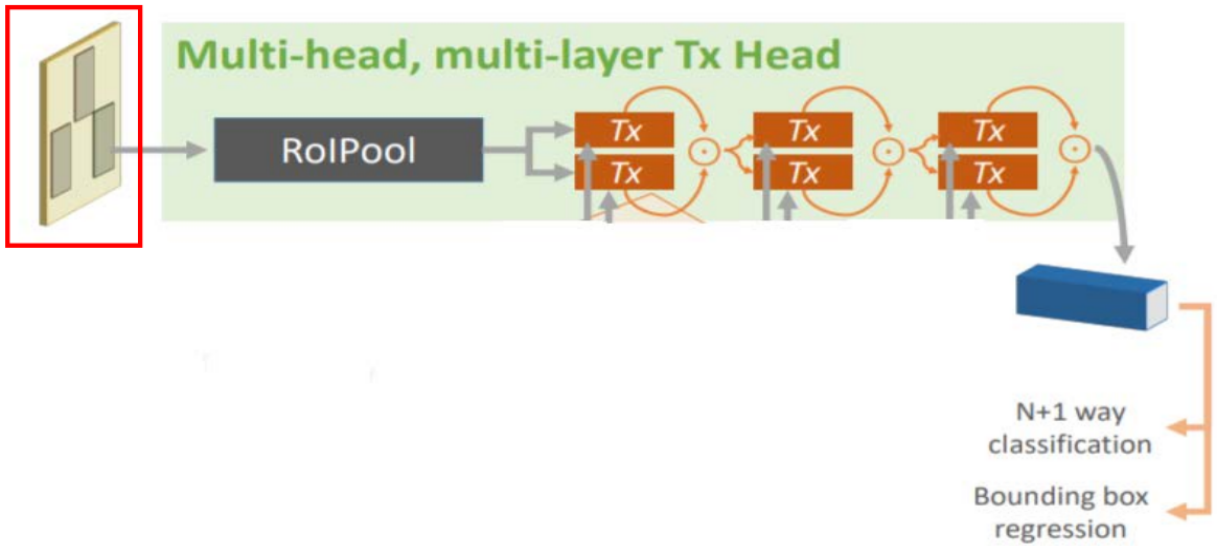
Trunk Region



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Region Proposal Network of the Trunk Region

Video Action Transformer Network

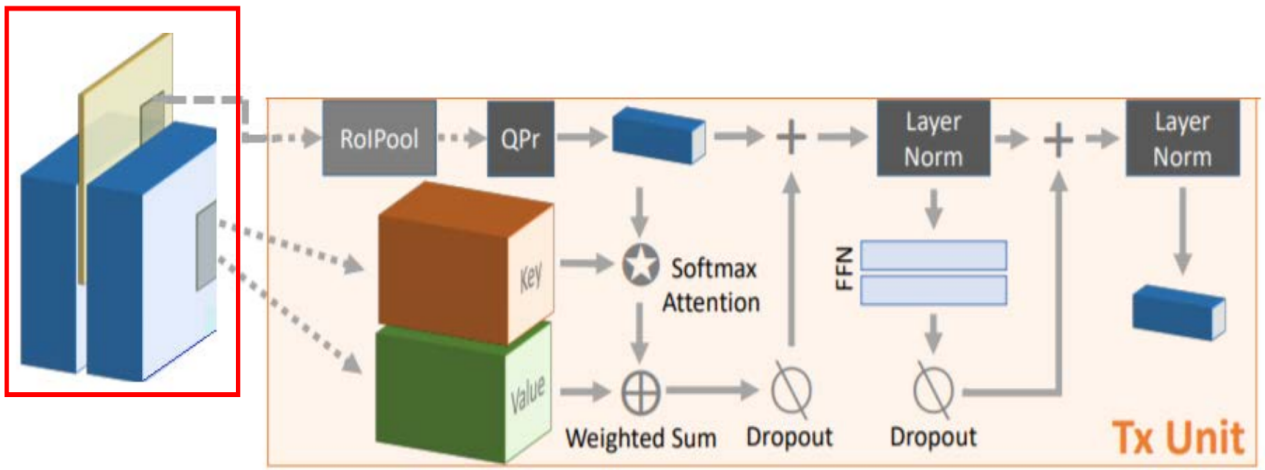
Head Region



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Action Transformer of the Head Region

Video Action Transformer Network

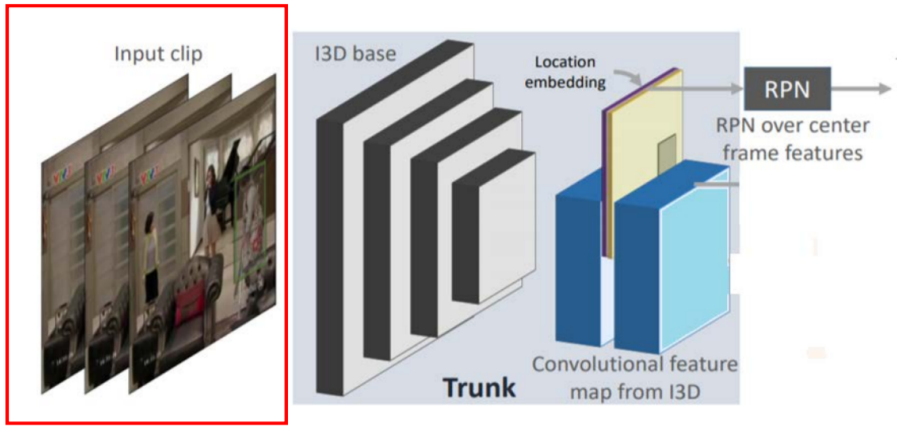
Action Transformer Tx-Unit



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Tx-Unit

Video Action Transformer Network

Trunk Region

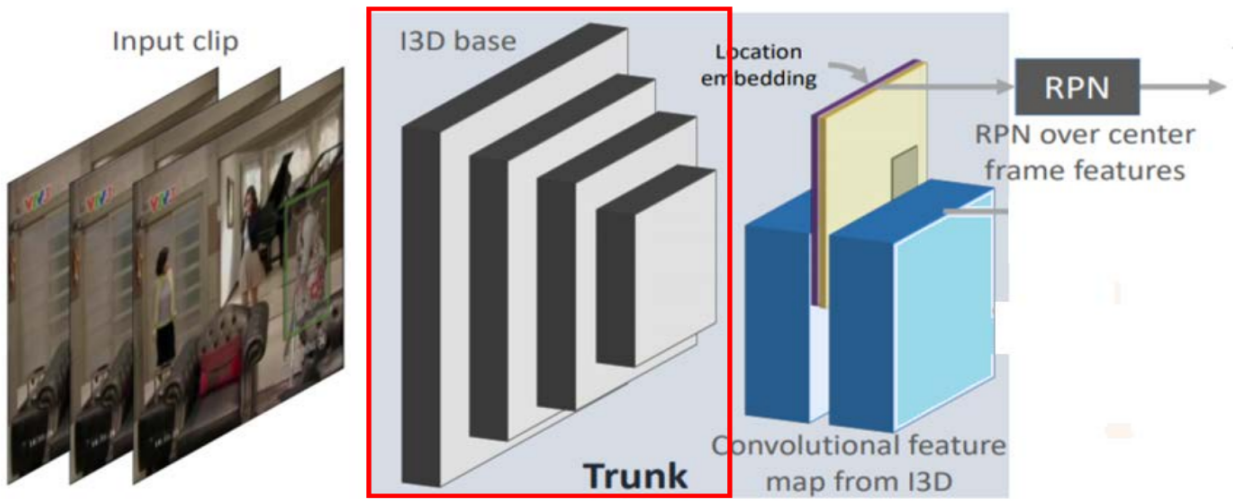


Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: T-frame Input of the model

Input Shape 64x400x400

Video Action Transformer Network

Trunk Region

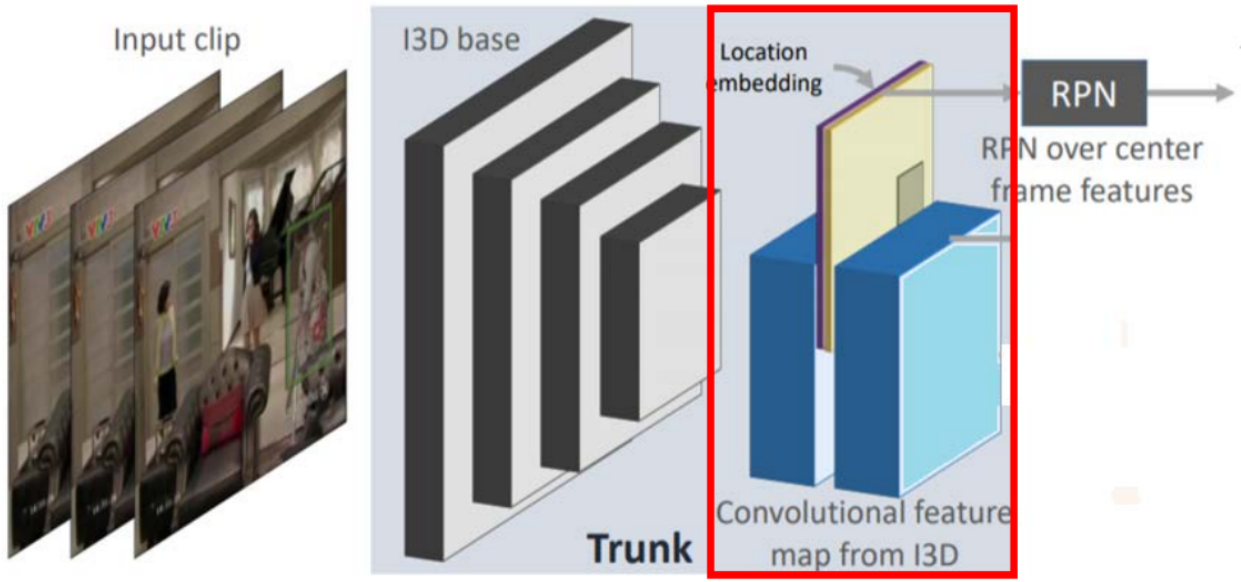


Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Layer wights of I3D base are pre-trained on Kinematics-400

downsampled feature map of Shape 16x25x25

Video Action Transformer Network

Trunk Region

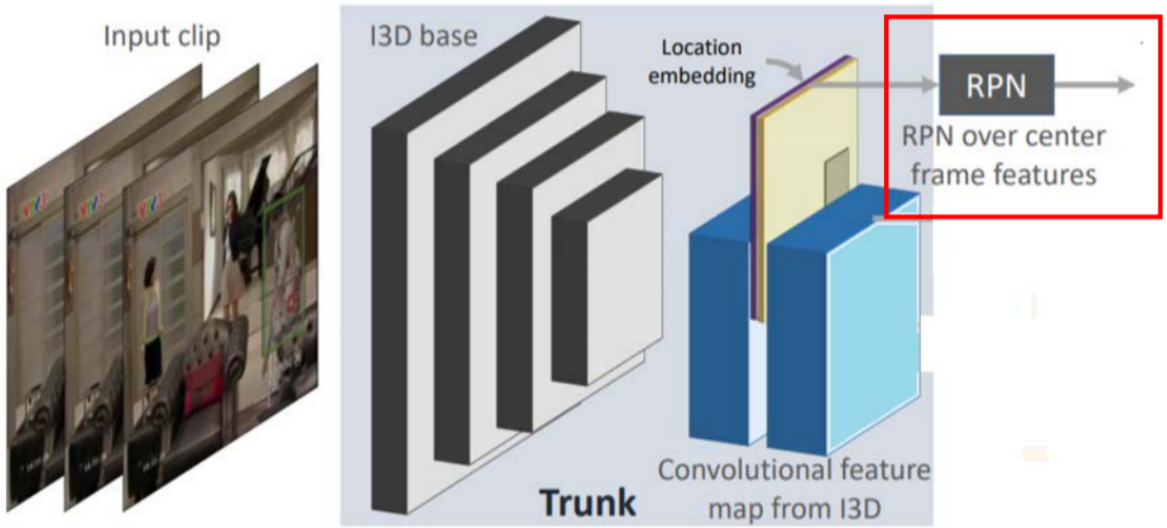


Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf

Figure: feature map of the Mixed_4 *fLayer* $\frac{T}{4} x \frac{W}{16} x \frac{H}{16}$

Video Action Transformer Network

Trunk Region

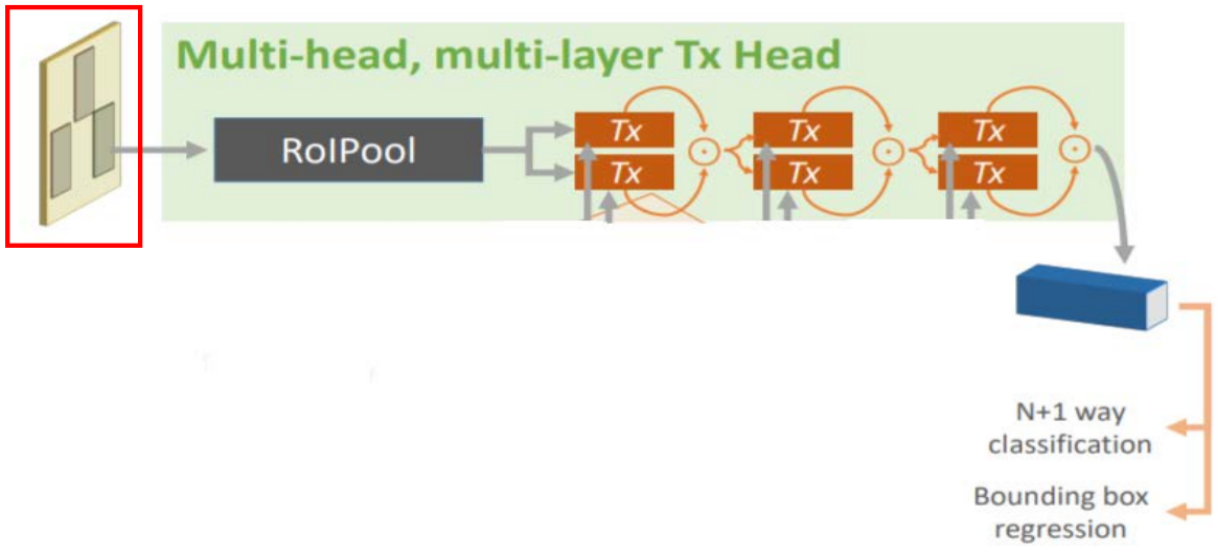


Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Region Proposal Network of the Trunk Region

Query Tensor $\in 25x25$

Video Action Transformer Network

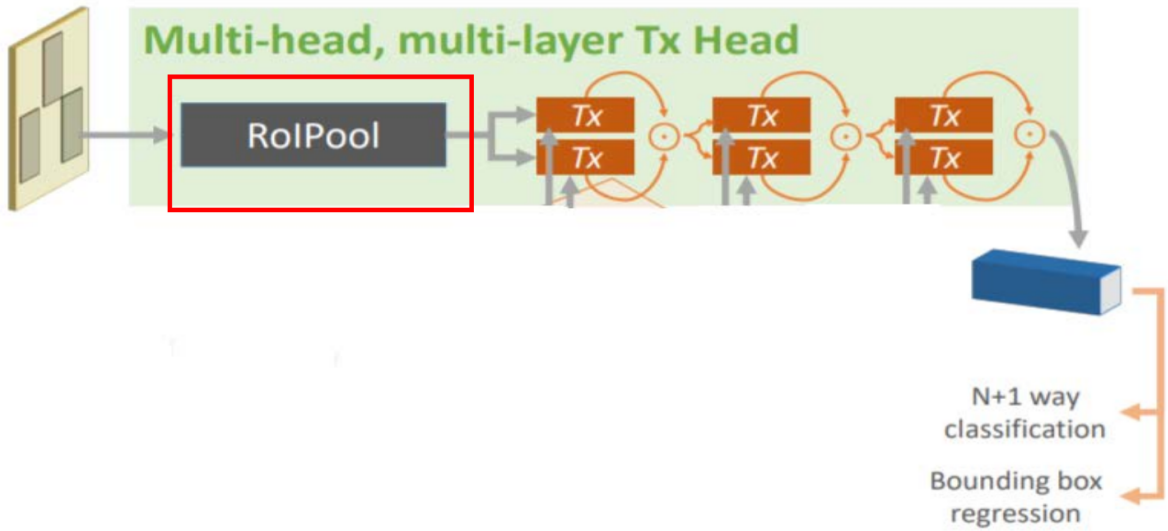
Head Region



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Region Proposals as Query Input of the Head Region

Video Action Transformer Network

Head Region

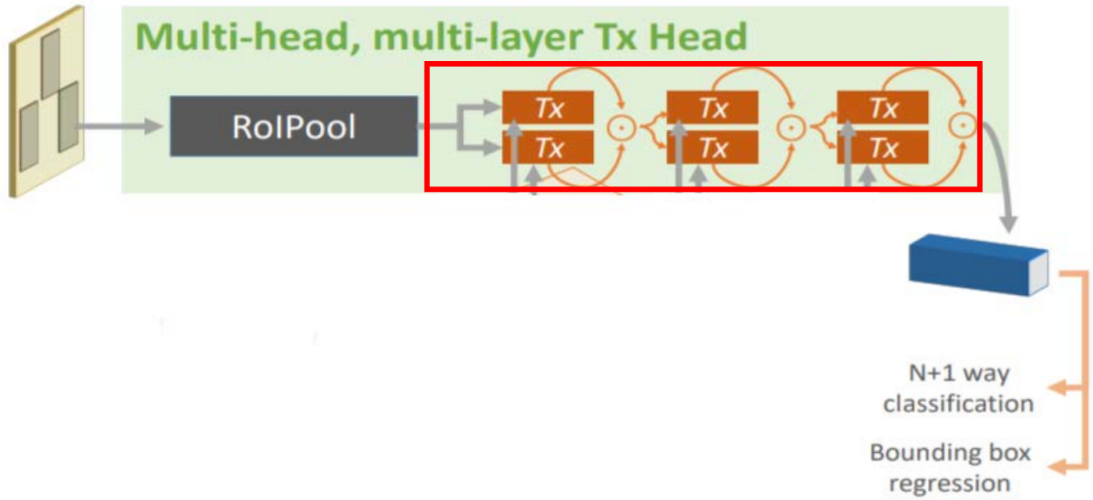


Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Region of Interest Pooling Operation

Query processed by RoIPool in to $\in [7x7]$

Video Action Transformer Network

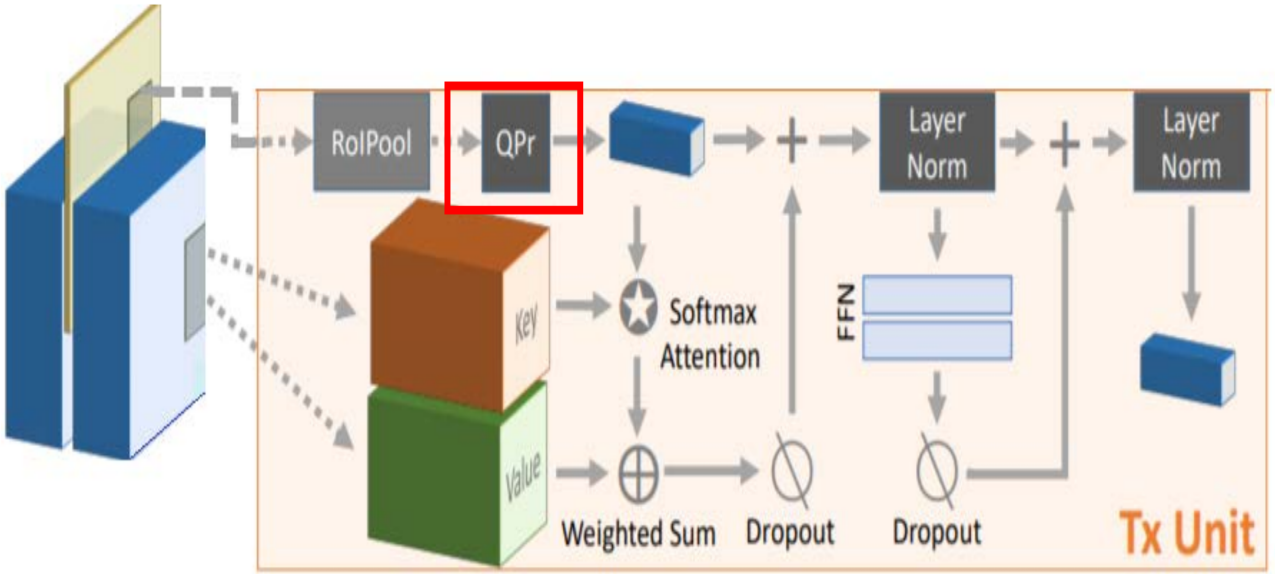
Head Region



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Tx-Unit stackt

Video Action Transformer Network

Head Region

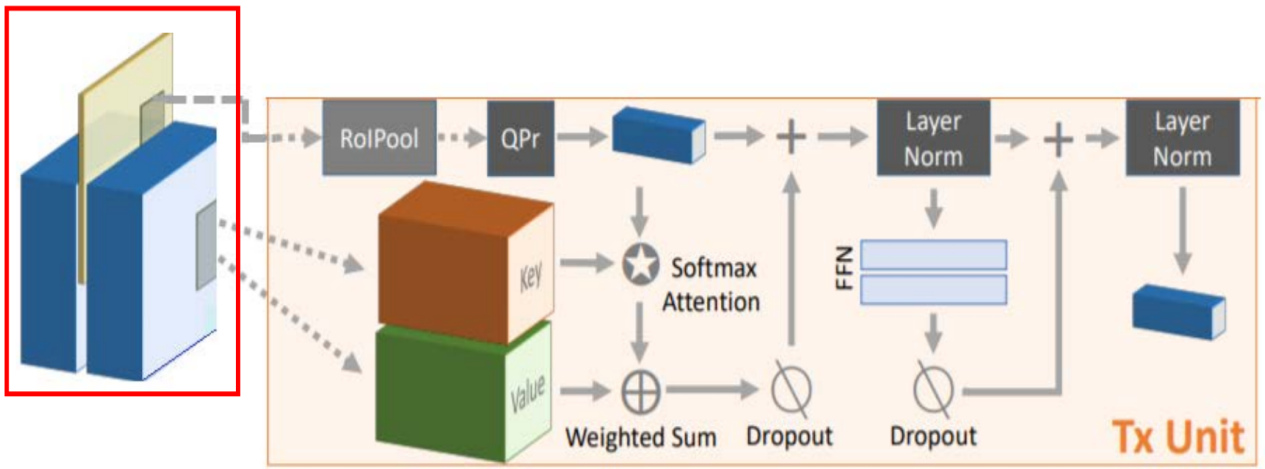


Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Query Preprocessor

The HighRes/LowRes query preProcesser transforms the Query in to $Q^r \in [1x1x128]$

Video Action Transformer Network

Head Region

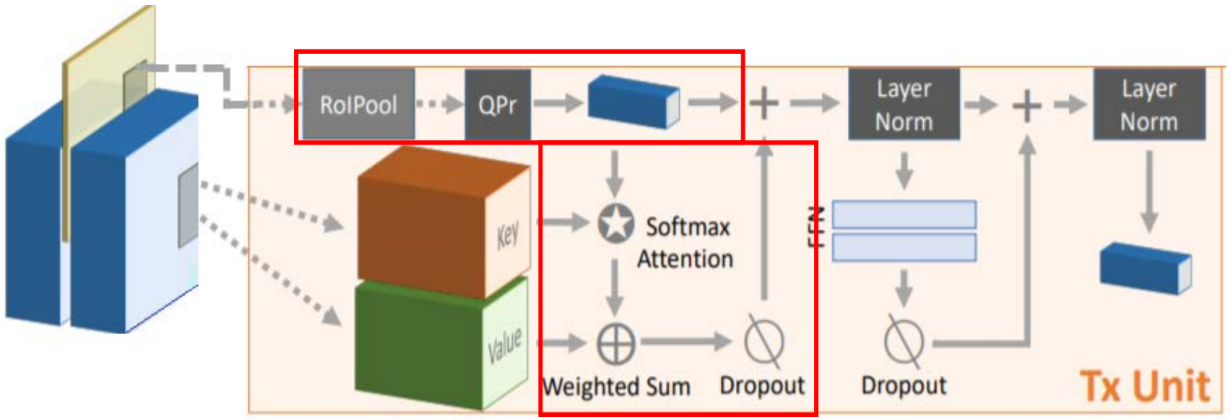


Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Linear projection of Key and Value

Key Tensor $\in [16x25x25x128]$

Video Action Transformer Network

Head Region



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Softmax Attention and Weighted Sum

Video Action Transformer Network

Attention Equation

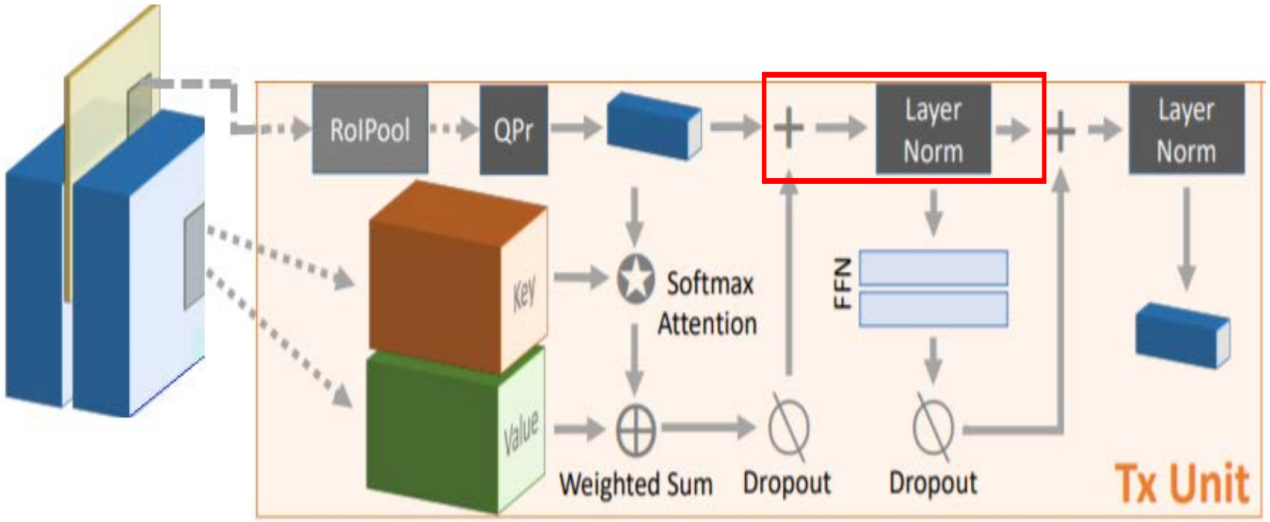
$$a_{xyt}^{(r)} = \frac{Q^{(r)} K_{xyt}^{T}}{\sqrt{D}}; A^{(r)} = \sum_{x,y,t} \left[\text{Softmax}\left(a^{(r)} \right) \right]_{xyt} V_{xyt}$$

Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Attention Equation

$KQ^r \in [16x25x25x1] A^r \in [1x1x128]$

Video Action Transformer Network

Head Region



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Layer Norm

Video Action Transformer Network

Head Region

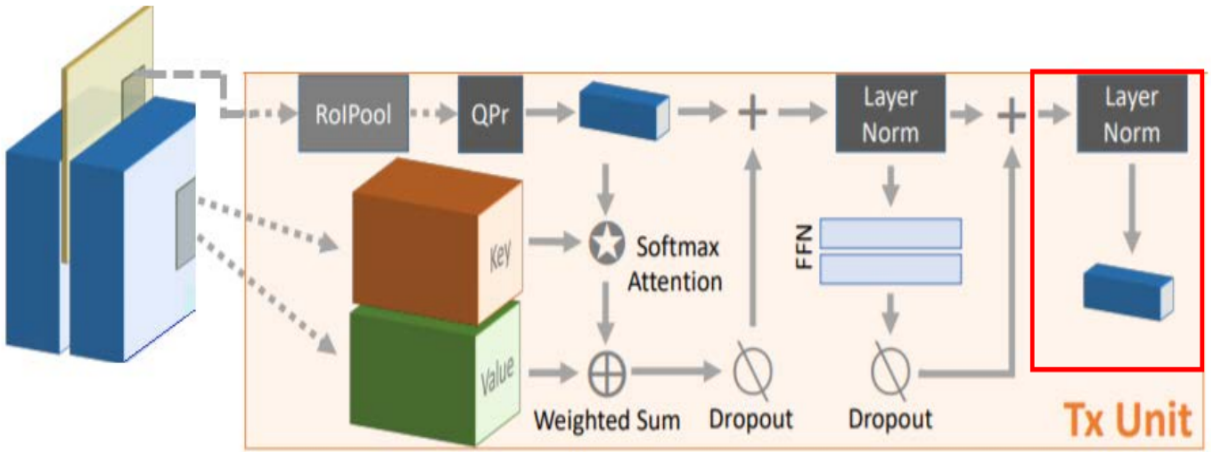
$$Q^{(r)'} = \text{LayerNorm}\left(Q^{(r)} + \text{Dropout}\left(A^{(r)}\right)\right)$$

$$Q^{(r)''} = \text{LayerNorm}\left(Q^{(r)'} + \text{Dropout}\left(\text{FFN}\left(Q^{(r)'}\right)\right)\right)$$

Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Layer Norm

Video Action Transformer Network

Head Region



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Action predicted and bounding box for the human

Experiments

- Atomic Visual Actions (AVA) Dataset
  - Training set 211 K
  - Validation 57 K
  - Test 117 K
- 1 FPS from 430 15 min movie clips
- Key-Frame is labels for action and ground-truth boxes
- Evaluation Metric frame levelmean average precision at IOU threshold of 0.5

Experiments

- Action classification (ground-truth boxes)
- Localization performance (action agnostic)
- Overall performance (compare to State of the art)

Video Action Transformer Network

Case 1: Action classification with and without ground-truth boxes

Trunk	Head	QPr	GT Boxes	Params (M)	Val mAP
I3D	I3D	-		16.2	21.3
I3D	I3D	-	✓	16.2	23.4
I3D	Tx	LowRes		13.9	17.8
I3D	Tx	HighRes		19.3	18.9
I3D	Tx	LowRes	✓	13.9	28.5
I3D	Tx	HighRes	✓	19.3	27.6

Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Evalute the performance of the Action Transformer Head

Video Action Transformer Network

Case 2: Localization performance (action agnostic)

RoI source	QPr	Head	Val mAP	
			IOU@0.5	IOU@0.75
RPN	-	I3D	92.9	77.5
RPN	LowRes	Tx	77.5	43.5
RPN	HighRes	Tx	87.7	63.3

Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Evaluate the localization performance of the Transformer Head

Video Action Transformer Network

Overall Performance Compare to other State of the Art models

Method	Modalities	Architecture	Val mAP	Test mAP
Single frame [16]	RGB, Flow	R-50, FRCNN	14.7	-
AVA baseline [16]	RGB, Flow	I3D, FRCNN, R-50	15.6	-
ARCN [42]	RGB, Flow	S3D-G, RN	17.4	-
Fudan University	-	-	-	17.16
YH Technologies [52]	RGB, Flow	P3D, FRCNN	-	19.60
Tsinghua/Megvii [23]	RGB, Flow	I3D, FRCNN, NL, TSN, C2D, P3D, C3D, FPN	-	21.08
Ours (Tx-only head)	RGB	I3D, Tx	24.4	24.30
Ours (Tx+I3D head)	RGB	I3D, Tx	24.9	24.60
Ours (Tx+I3D+96f)	RGB	I3D, Tx	**25.0**	**24.93**

Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf

Figure: The Action Transformer Network model outperforms other models by a significant margin

Video Action Transformer Network

Performance depending on numbers of heads and layers

#layers↓	#heads→	2	3	6
	2	27.4	28.7	27.6
	3	28.5	28.8	27.7
	6	29.1	28.3	26.5

Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Different Head combination affect the performance

Video Action Transformer Network

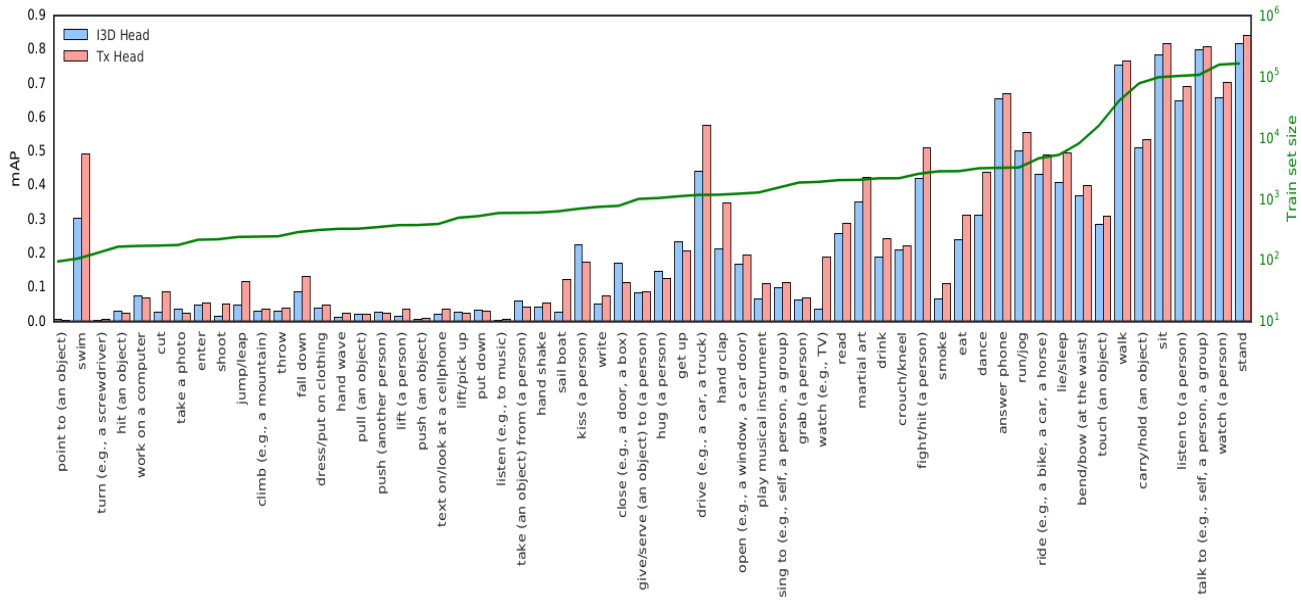
Data Augmentation and pre-trained layers

	I3D head	Cls-specific bbox-reg	No Data Aug	From Scratch
Val mAP	21.3	19.2	16.6	19.1

Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Data Augmentation and pre-trained

Video Action Transformer Network

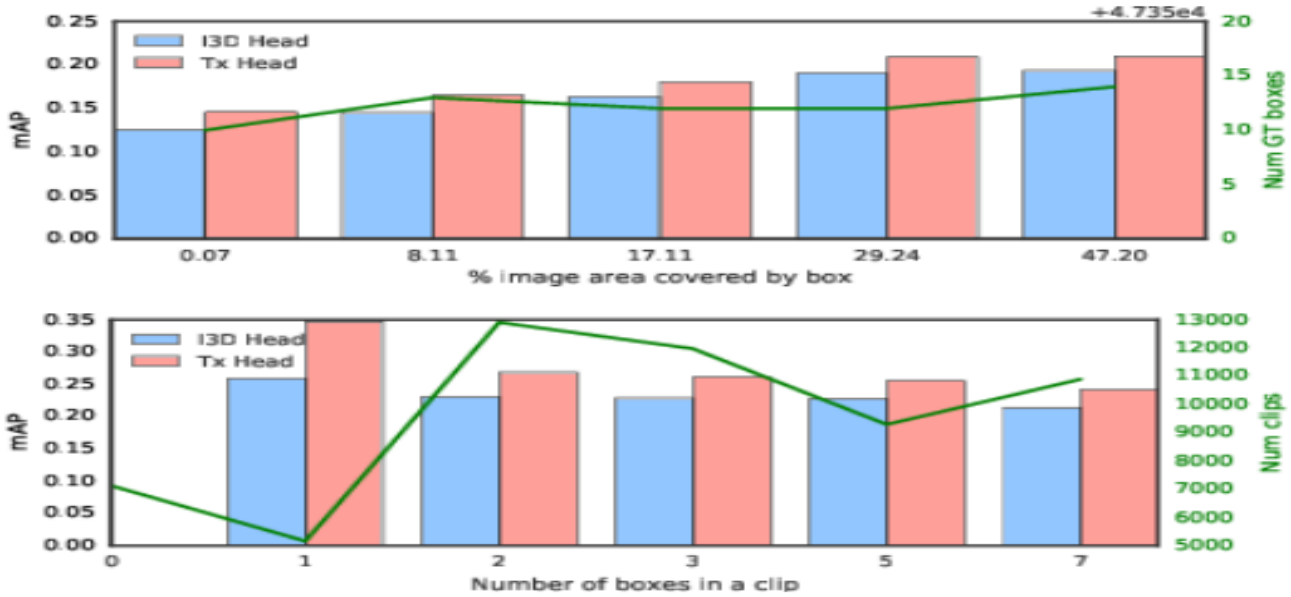
Correlation of performance with train-set size



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: trend suggests a positive correlation of performance with train-set size

Video Action Transformer Network

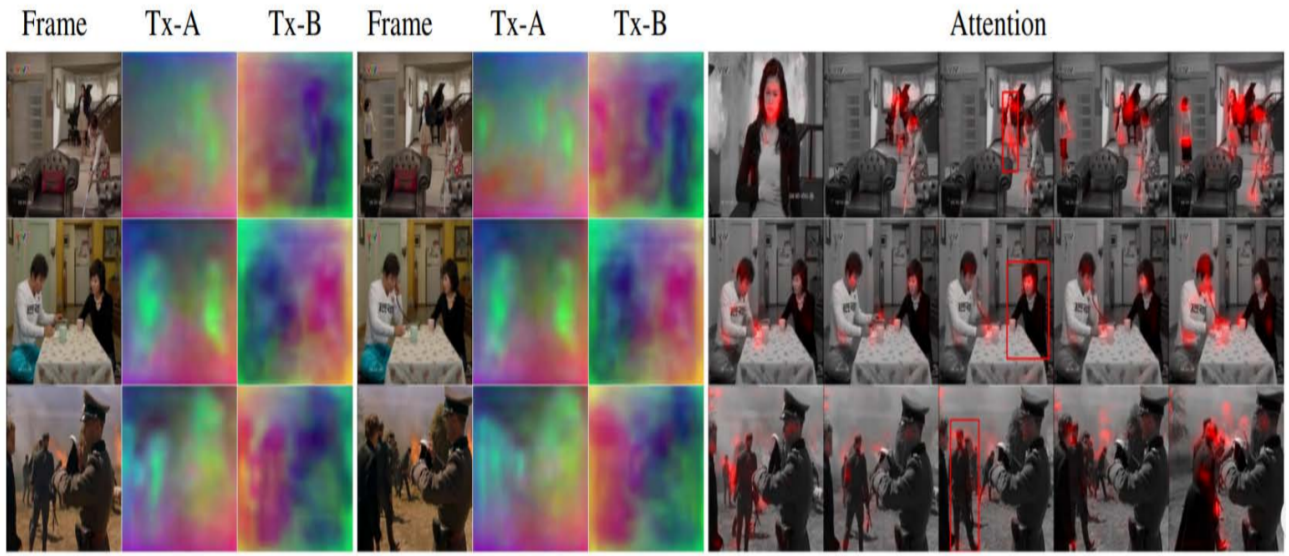
Correlation of performance with Number of Boxes



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: The trend suggests a positive correlation of performance with train-set size

Video Action Transformer Network

Embedding and Attention



Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf

Figure: 'Key' embeddings as color-coded 3D PCA projection for two of the six heads in our 2-head 3-layer Tx head

Video Action Transformer Network

Examples of Correct Predictions

Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf
Figure: Actions with contect like 'watching a person', 'holding an object'

Video Action Transformer Network

Examples of Incorrect Predictions

Source: https://www.crcv.ucf.edu/wp-content/uploads/2019/03/adcv/ppt.pdf

Figure: The 'smoking' action class obtains low performance even with large amount of training data

Conclusion

- Embeddings and attention maps learned without supervision have a semantic meaning
- Model is able to learn spatio-temporal context from othe r humans and objects
- Outperforms other state of the art models
- Dataset or Objective far from solved

Thank you for your Attention