Institute for Infocomm Research

I²R

END OF STUDIES INTERNSHIP DEFENSE

Visual tasks representation for remote perception and guidance using a wearable device

Gouneau Joceran, supervised by Ng Lai Xing

7th of September 2023

Content

1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Content

1. **Introduction**
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Introduction

Institute for Infocomm Research

I²R

"To create digital world innovations for a thriving and resilient Singapore harnessing AI, Connectivity and Cybersecurity."

Human-in-the-loop systems; Augmented reality

Task Assistant using a wearable device

CREATING GROWTH, ENHANCING LIVES

Content

1. Introduction
2. **Context**
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Content

1. Introduction
2. Context
   - **2.1. Problem**
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Problem

source: https://ego4d-data.org/

- Episodic Memory
- Hand-Object Interactions
- Audio-Visual Diarization
- Social Interactions
- Forecasting

CREATING GROWTH, ENHANCING LIVES

Problem

Given a video V and a timestamp t, the model should be able, given V up to t, to predict the next active objects :

$$\{(\hat{b}_i, \hat{n}_i, \hat{v}_i, \hat{\delta}_i, \hat{s}_i)\}_{i=1}^{N}$$

Where :

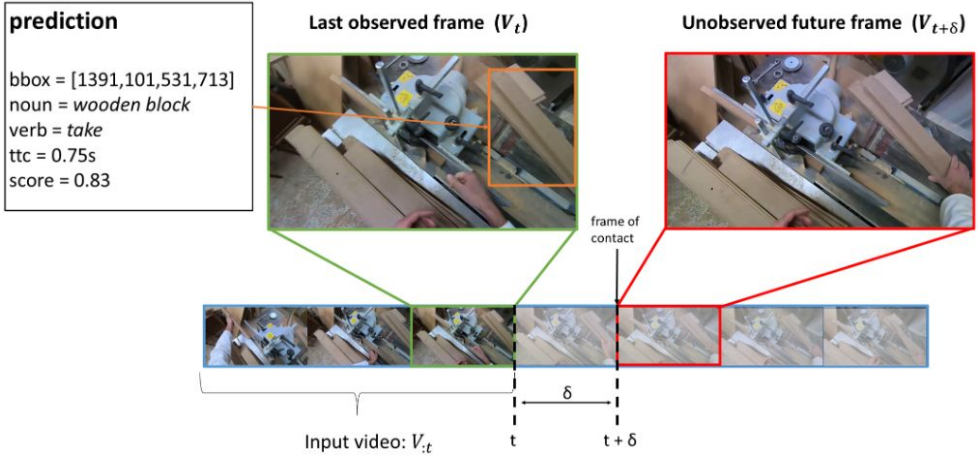
$\hat{b}_i$ is the bounding box at t.

$\hat{n}_i$ is a *name*.

$\hat{v}_i$ is a *verb*.

$\hat{\delta}_i$ is the *time to contact* from t.

$\hat{s}_i$ is a confidence score.



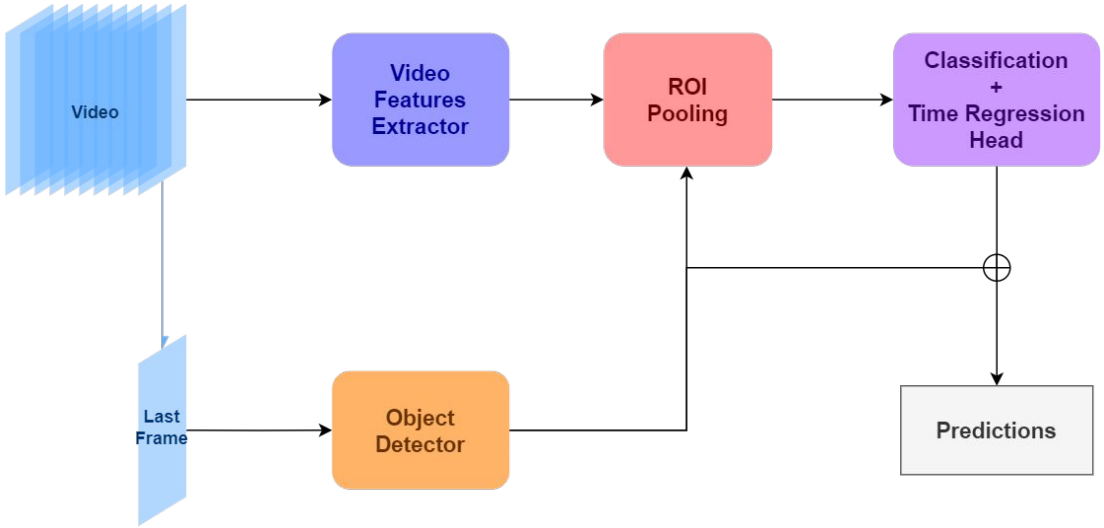
F. Ragusa et al. *Stillfast: An end-to-end approach for short-term object interaction anticipation*, 2023.

Content

1. Introduction
2. Context
   - 2.1. Problem
   - **2.2. Current Solutions**
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Current Solutions

Model Name	Object Detector	Video Backbone	Dataset Version	$\mathbf{mAP}_{\text{Overall}}$
Baseline V1	Faster RCNN	SlowFast	V1	2.45
InternVideo	DINO DETR	VideoMAE	V1	3.40
Baseline V2	Faster RCNN	SlowFast	V2	3.61

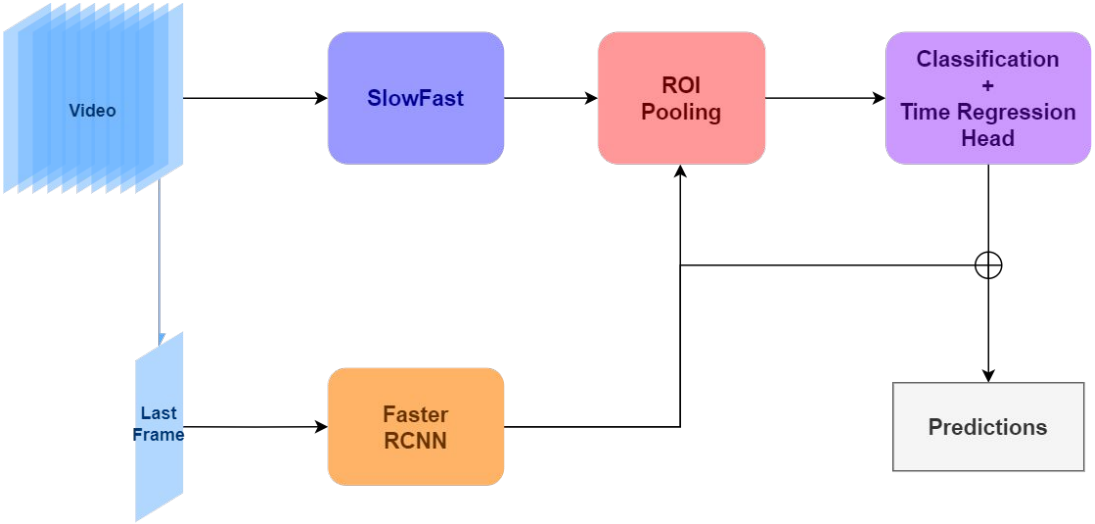


CREATING GROWTH, ENHANCING LIVES

Content

1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - **2.2.1. Baseline**
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Baseline

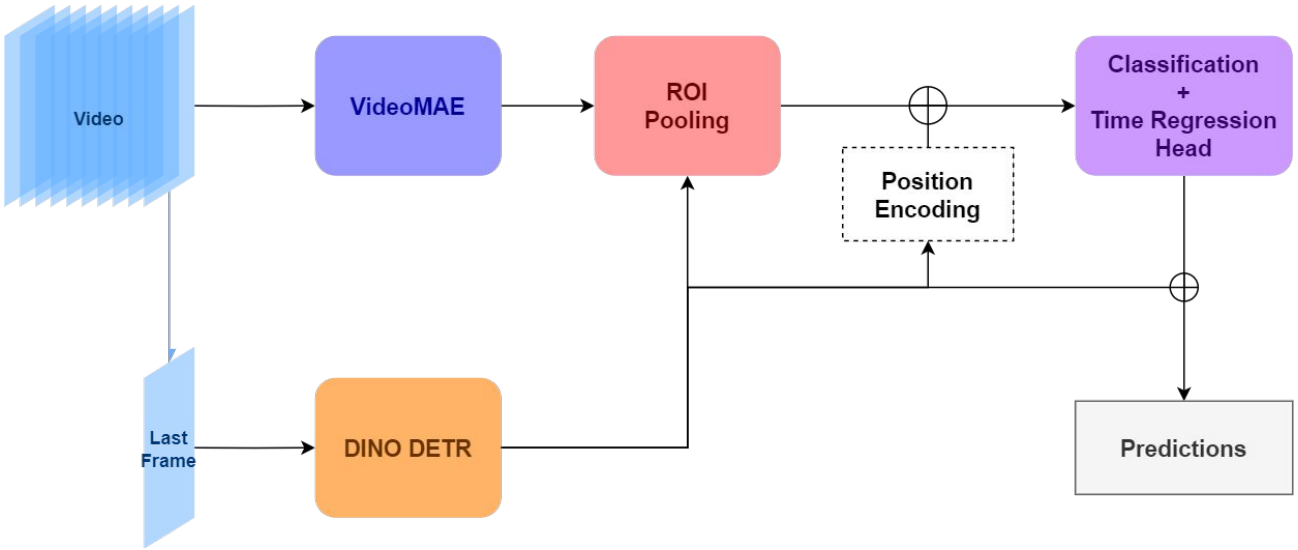


CREATING GROWTH, ENHANCING LIVES

Content

1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - **2.2.2. InternVideo**
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

InternVideo

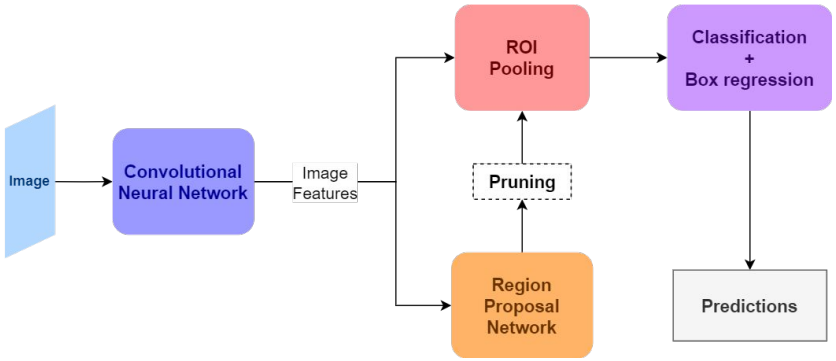


CREATING GROWTH, ENHANCING LIVES

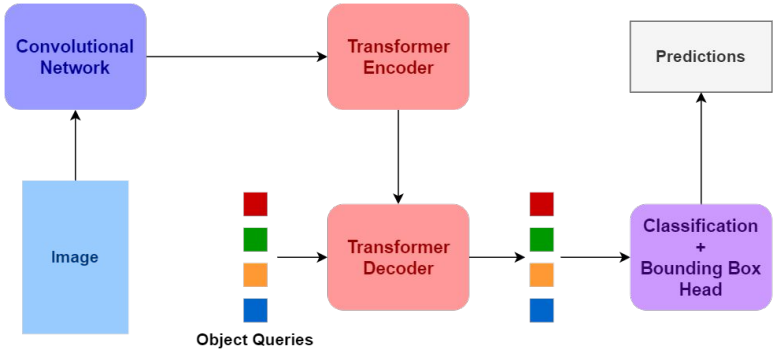
Content

1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - **2.3. Transformers for Detection**
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Transformers for Detection



Faster RCNN



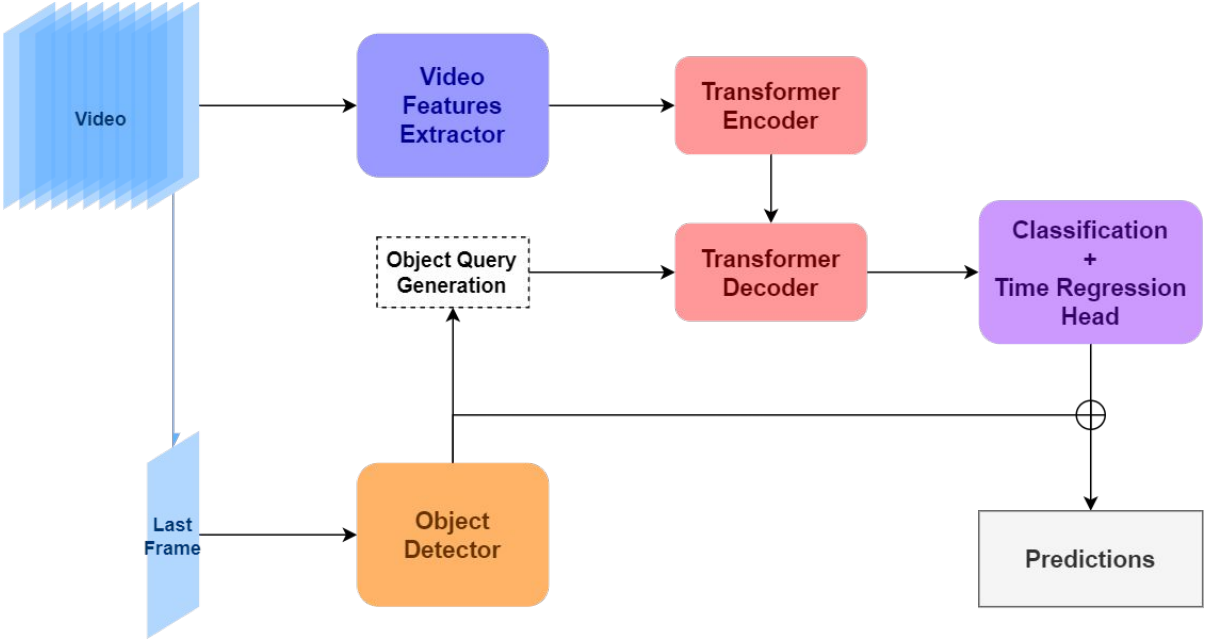
DETR

CREATING GROWTH, ENHANCING LIVES

Content

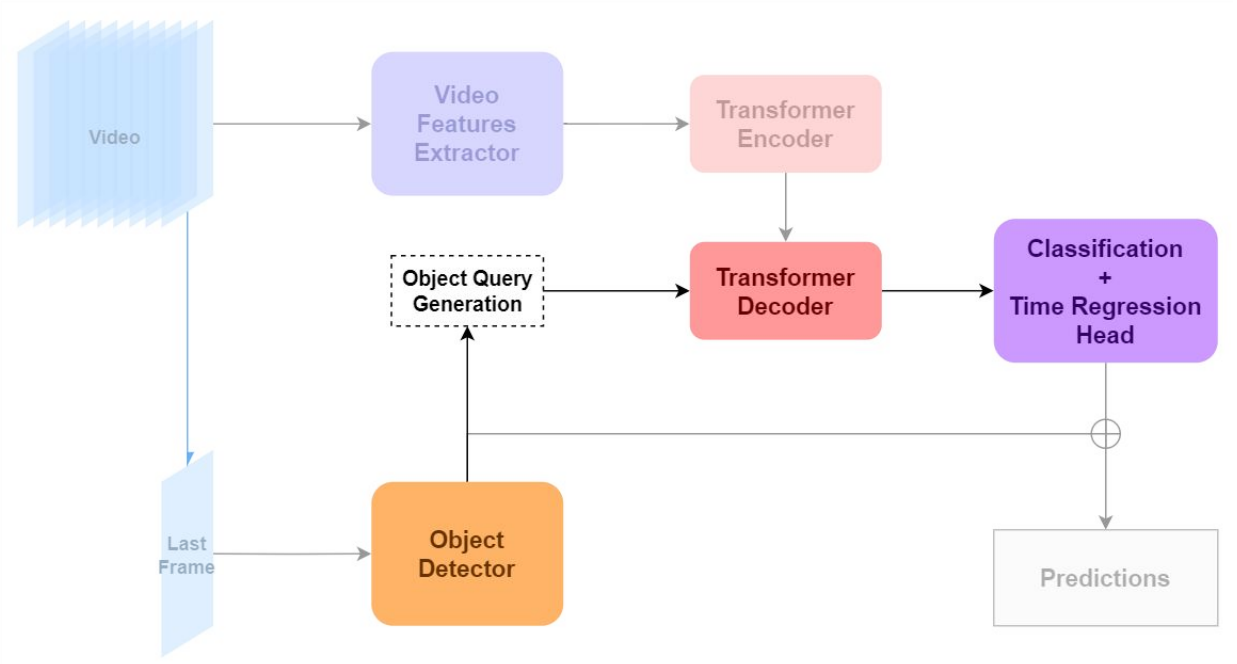
1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. **Method**
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Method



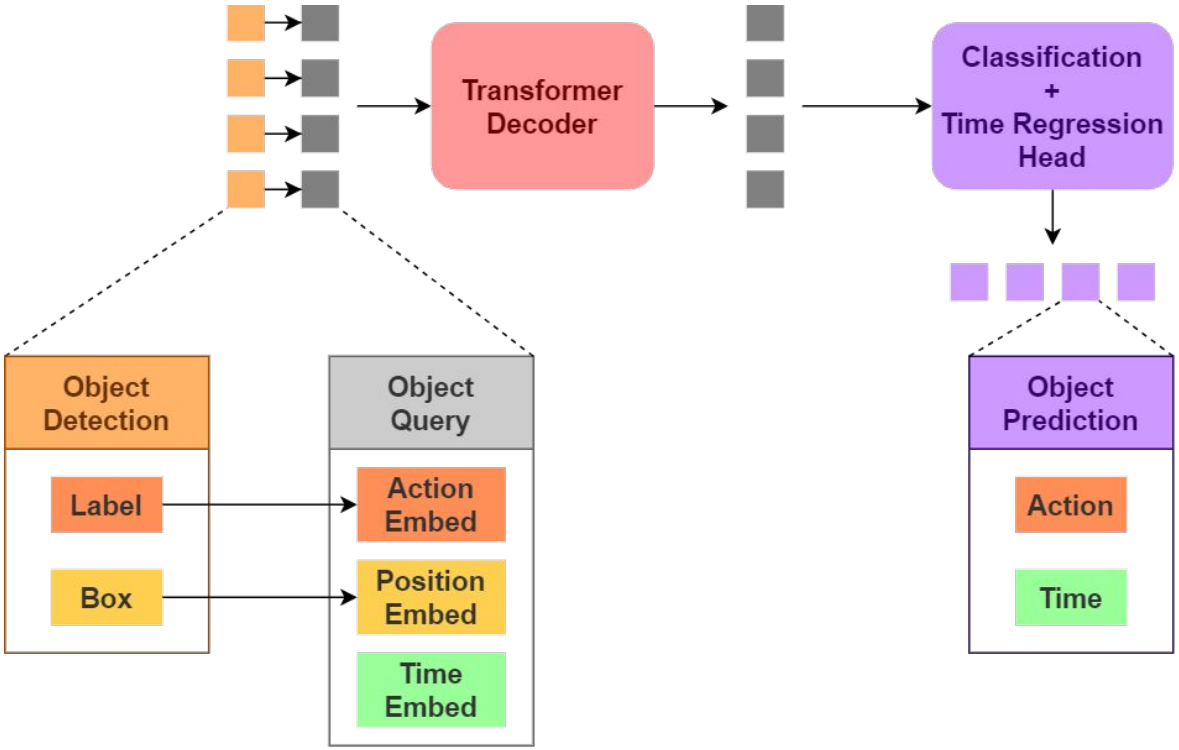
CREATING GROWTH, ENHANCING LIVES

Method



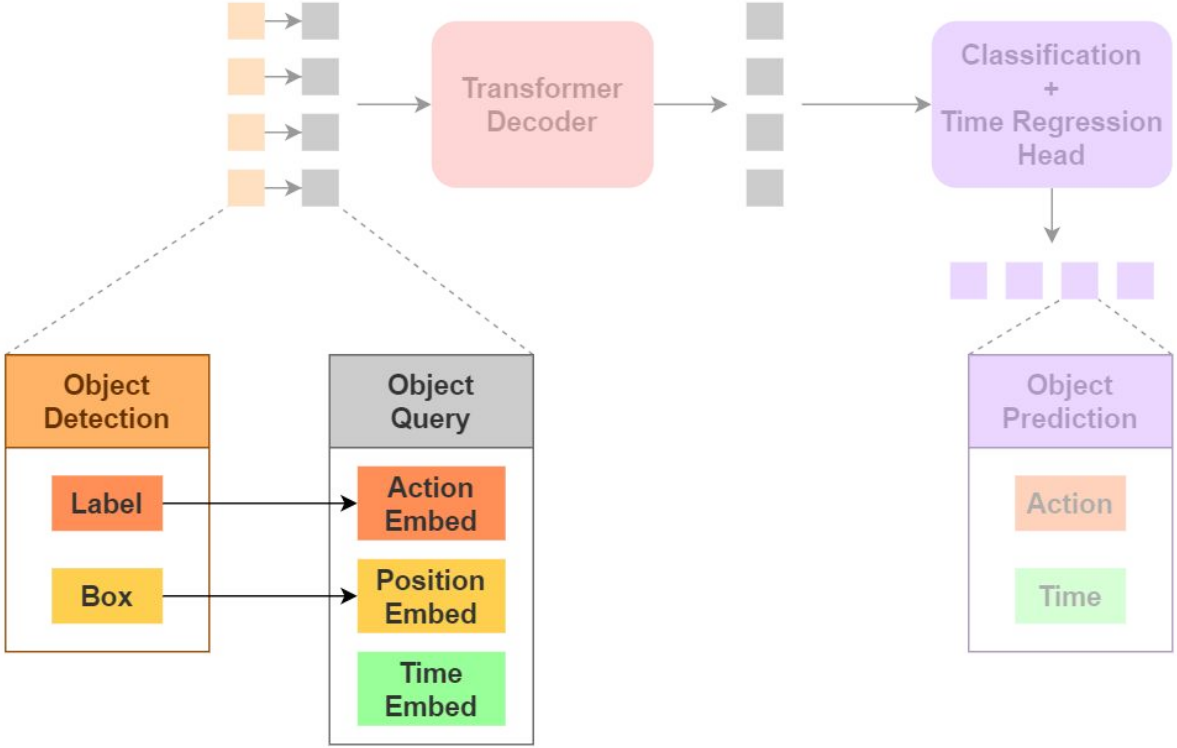
CREATING GROWTH, ENHANCING LIVES

Method



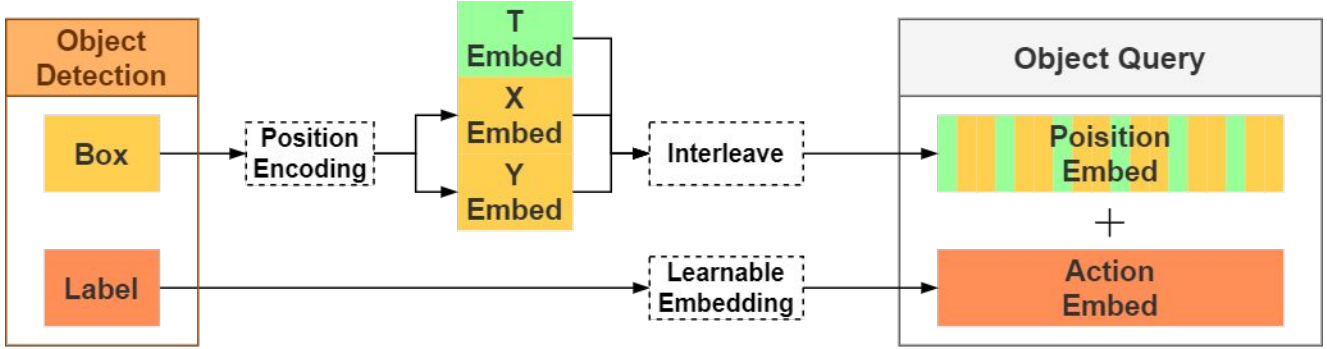
CREATING GROWTH, ENHANCING LIVES

Method

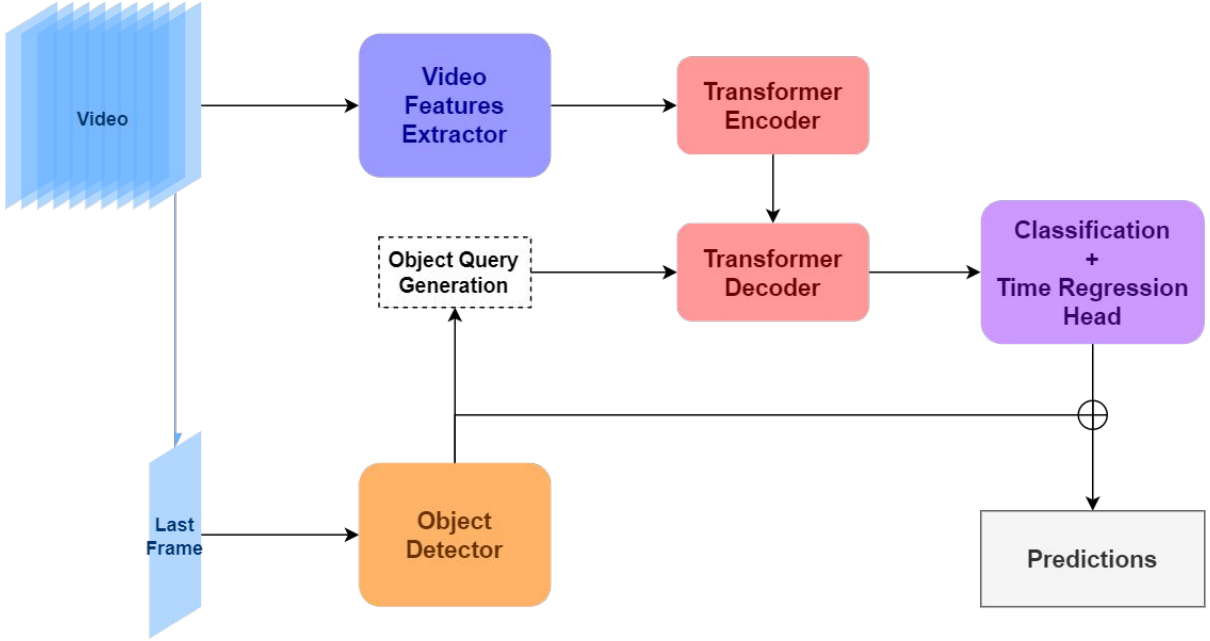


CREATING GROWTH, ENHANCING LIVES

Method

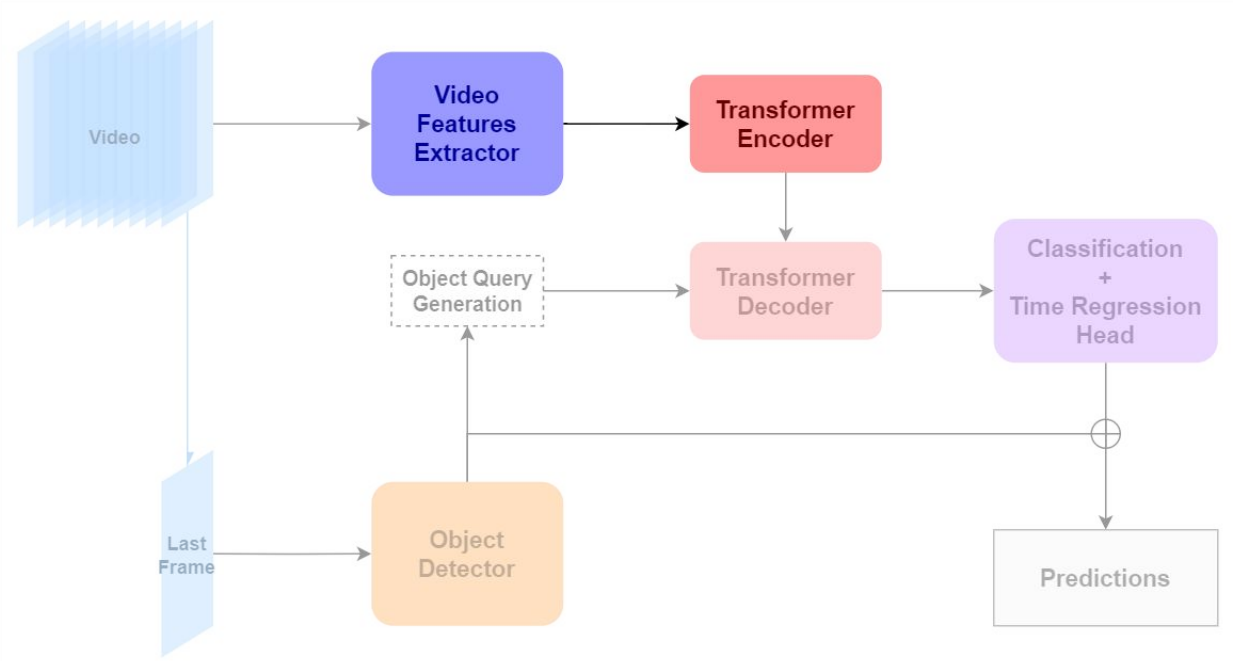

Object Detection
Box
Label
Position Encoding
T Embed
X Embed
Y Embed
Interleave
Learnable Embedding
Object Query
Poisition Embed
+
Action Embed

Method



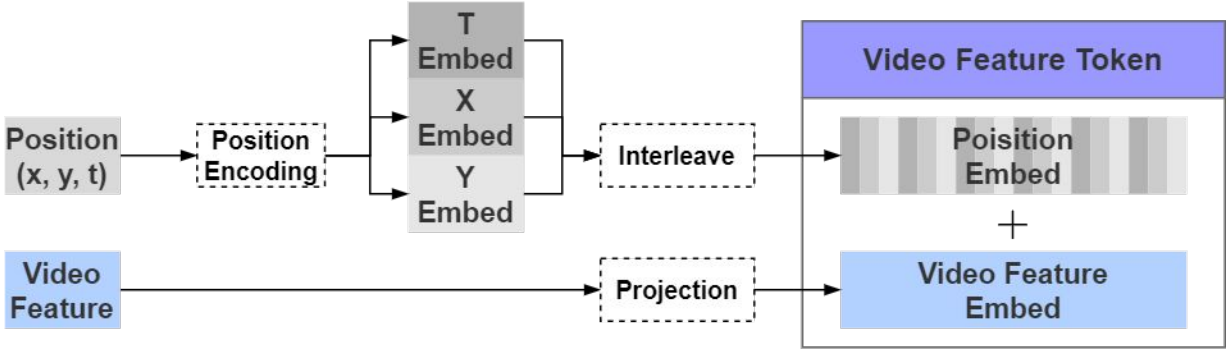
CREATING GROWTH, ENHANCING LIVES

Method



CREATING GROWTH, ENHANCING LIVES

Method



Content

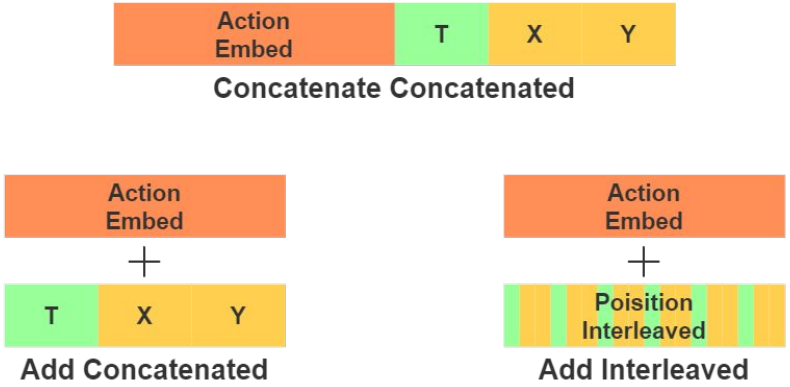
1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. **Experiments**
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Content

1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - **4.1. Jupyter Notebook Quickstart**
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Jupyter Notebook Quickstart

- pre-extracted video features by Omnivore:
  - frozen
  - time only dependent



Model Name	$\mathbf{mAP}_{\text{Box, Noun}}$	$\mathbf{mAP}_{\text{Box, Noun, Verb}}$	$\mathbf{mAP}_{\text{Box, Noun, TTC}}$	$\mathbf{mAP}_{\text{Overall}}$
Notebook Baseline	29.11	10.08	**7.16**	2.61
Concat Concatenate		10.00	6.57	2.38
Add Concatenate		9.81	5.92	2.21
Add Interleaved		**10.38**	6.68	**2.69**

CREATING GROWTH, ENHANCING LIVES

Content

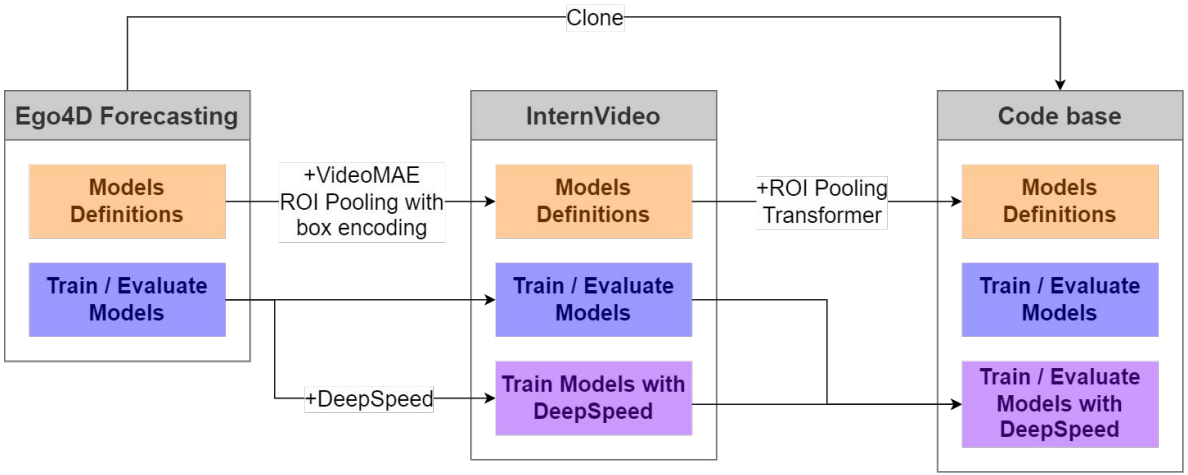
1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - **4.2. Scaling Up to the Full Dataset**
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Content

1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - **4.2.1. Building the Codebase**
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Building the Codebase

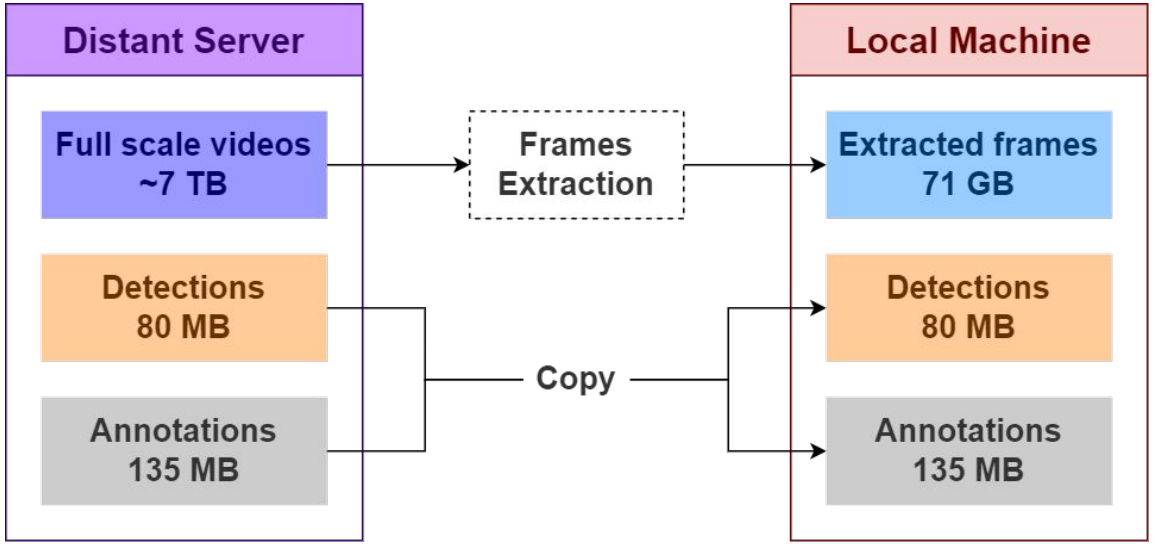
- Ego4D Forecasting:
  - Official Training / Evaluation scripts;
  - SlowFast definition (pre-trained on Kinetics-400);
  - Baseline definition;
- InternVideo:
  - Modified Training Pipeline using DeepSpeed;
  - VideoMAE definition (pre-trained on Ego4D V1);



Content

1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - **4.2.2. Accessing and Preprocessing the Data**
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Accessing and Preprocessing the Data

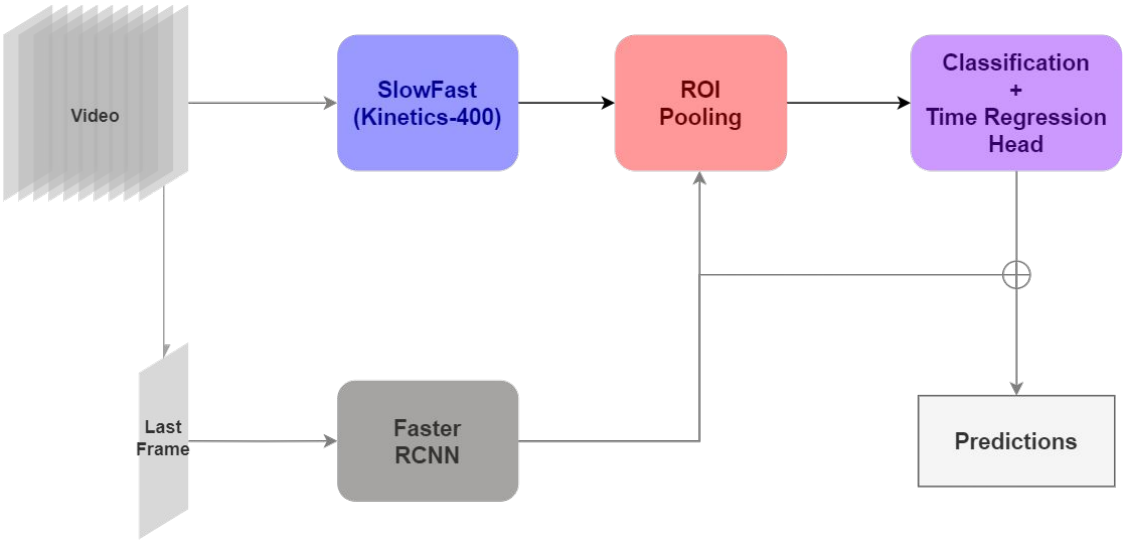


CREATING GROWTH, ENHANCING LIVES

Content

1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - **4.2.3. Establishing Our Baseline**
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Establishing Our Baseline

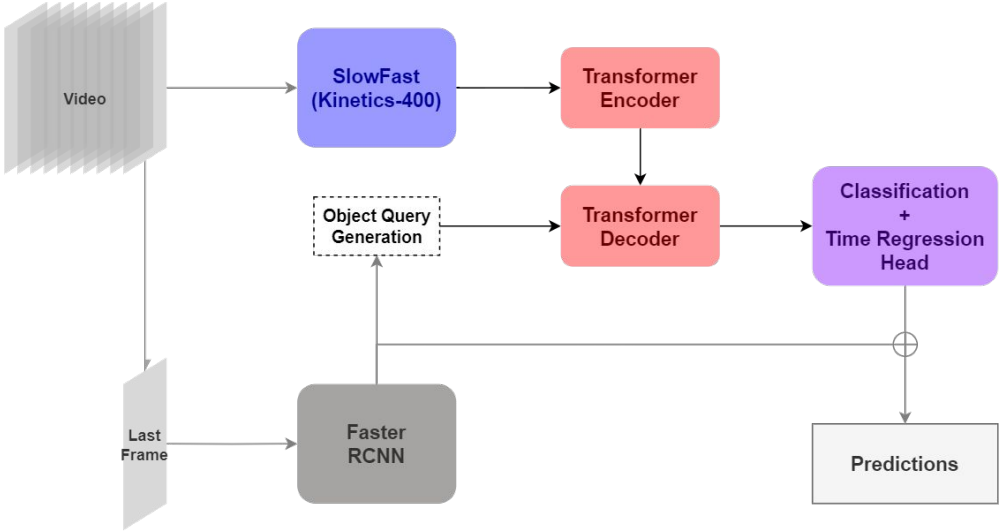


Subset	$\mathbf{mAP}_{\text{Box, Noun}}$	$\mathbf{mAP}_{\text{Box, Noun, Verb}}$	$\mathbf{mAP}_{\text{Box, Noun, TTC}}$	$\mathbf{mAP}_{\text{Overall}}$
validation	24.79	8.86	7.58	2.66
test	26.15	9.48	8.11	3.36

Content

1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - **4.3. Using SlowFast as a Backbone**
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Using SlowFast as a Backbone



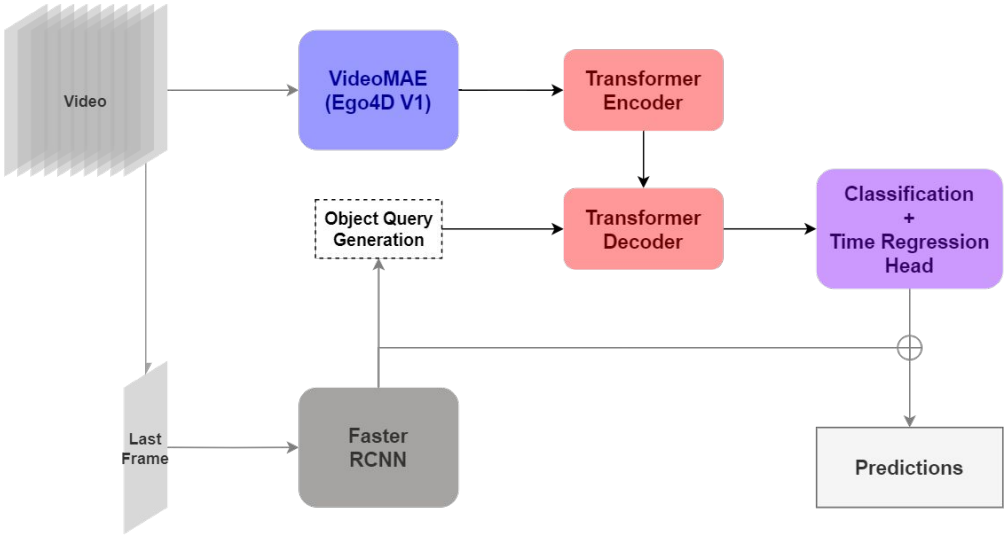
Subset	$\mathbf{mAP}_{\text{Box, Noun}}$	$\mathbf{mAP}_{\text{Box, Noun, Verb}}$	$\mathbf{mAP}_{\text{Box, Noun, TTC}}$	$\mathbf{mAP}_{\text{Overall}}$
validation	24.79	8.83	7.21	3.24
test	26.15	9.90	7.62	3.28

CREATING GROWTH, ENHANCING LIVES

Content

1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - **4.4. Using VideoMAE as a Backbone**
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Using VideoMAE as a Backbone



Subset	$\mathbf{mAP}_{\text{Box, Noun}}$	$\mathbf{mAP}_{\text{Box, Noun, Verb}}$	$\mathbf{mAP}_{\text{Box, Noun, TTC}}$	$\mathbf{mAP}_{\text{Overall}}$
validation	24.79	10.48	8.70	3.92
test	26.15	11.25	9.22	4.75

CREATING GROWTH, ENHANCING LIVES

Content

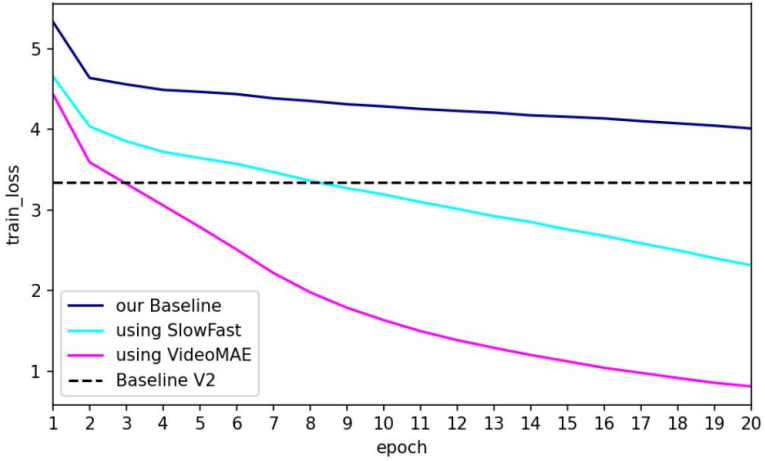
1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. **Discussion**
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. Conclusion

Content

1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - **5.1. Comparison**
   - 5.2. Areas of Improvement
6. Conclusion

Comparison

Model	$\mathbf{mAP}_{\text{B, N}}$	$\mathbf{mAP}_{\text{B, N, V}}$	$\mathbf{mAP}_{\text{B, N, TTC}}$	$\mathbf{mAP}_{\text{Overall}}$
Baseline V1	20.45	6.78	6.17	2.45
InternVideo	24.60	9.19	7.64	3.40
Using SlowFast	**26.15**	9.90	7.62	3.28
Our Baseline		9.48	8.11	3.36
Baseline V2		9.45	8.69	3.61
Using VideoMAE		11.25	**9.22**	4.75
StillFast	25.06	13.29	9.14	5.12
GANO	25.67	**13.60**	9.02	**5.16**

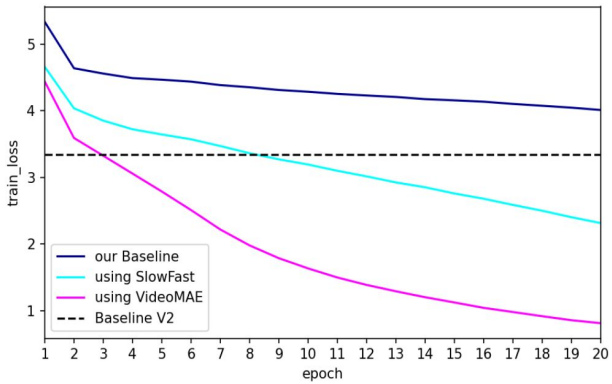


Content

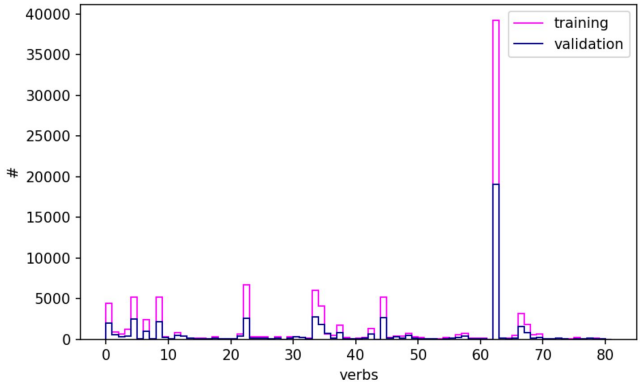
1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - **5.2. Areas of Improvement**
6. Conclusion

CREATING GROWTH, ENHANCING LIVES

Areas of Improvement



Epoch	Subset	$\mathbf{Loss}_{\text{Total}}$	$\mathbf{Loss}_{\text{Verb}}$	$\mathbf{Loss}_{\text{TTC}}$	$\mathbf{mAP}_{\text{Overall}}$
20	train	0.82	0.27	0.05	
	validation	5.61	3.61	0.20	3.92
2	train	3.59	1.82	0.17	
	validation	4.02	2.13	0.19	2.70



Subset	Training		Validation	
Detections	GT	Pred	GT	Pred
$\mathbf{mAP}_{\text{Box, Noun}}$	100	46.01	100	24.79
$\mathbf{mAP}_{\text{Box, Noun, Verb}}/\mathbf{mAP}_{\text{Box, Noun}}$	84.15	77.14	42.41	42.28
$\mathbf{mAP}_{\text{Box, Noun, TTC}}/\mathbf{mAP}_{\text{Box, Noun}}$	45.31	41.80	38.30	35.09
$\mathbf{mAP}_{\text{Overall}}/\mathbf{mAP}_{\text{Box, Noun}}$	39.02	33.71	18.52	15.81

CREATING GROWTH, ENHANCING LIVES

Content

1. Introduction
2. Context
   - 2.1. Problem
   - 2.2. Current Solutions
     - 2.2.1. Baseline
     - 2.2.2. InternVideo
   - 2.3. Transformers for Detection
3. Method
4. Experiments
   - 4.1. Jupyter Notebook Quickstart
   - 4.2. Scaling Up to the Full Dataset
     - 4.2.1. Building the Codebase
     - 4.2.2. Accessing and Preprocessing the Data
     - 4.2.3. Establishing Our Baseline
   - 4.3. Using SlowFast as a Backbone
   - 4.4. Using VideoMAE as a Backbone
5. Discussion
   - 5.1. Comparison
   - 5.2. Areas of Improvement
6. **Conclusion**

Conclusion

Thank you !

Institute for Infocomm Research

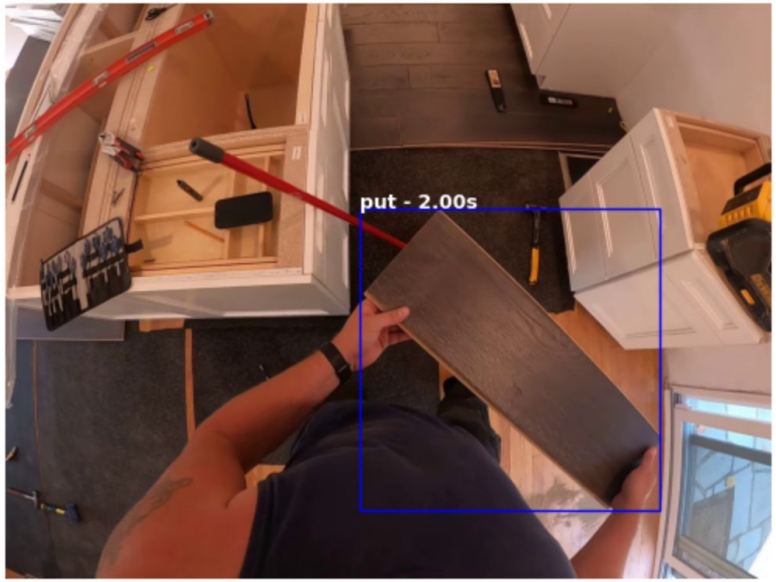
I²R

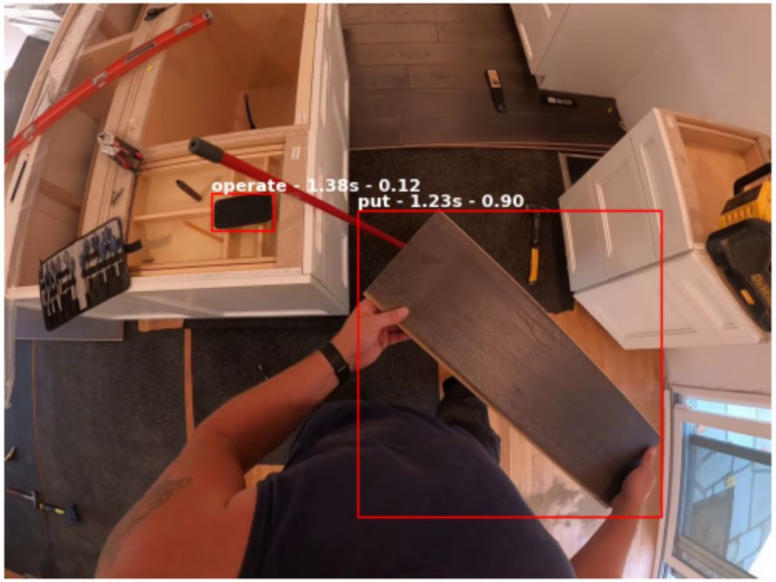
END OF STUDIES INTERNSHIP DEFENSE

Visual tasks representation for remote perception and guidance using a wearable device

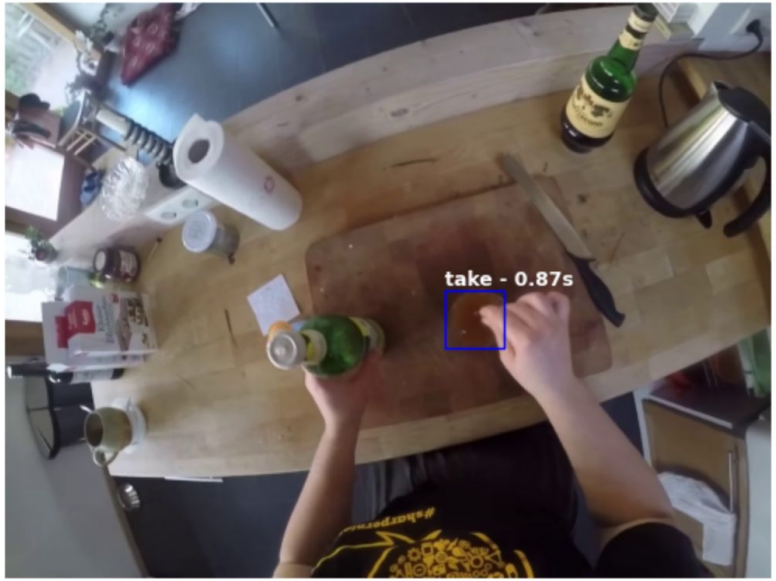
Gouneau Joceran, supervised by Ng Lai Xing

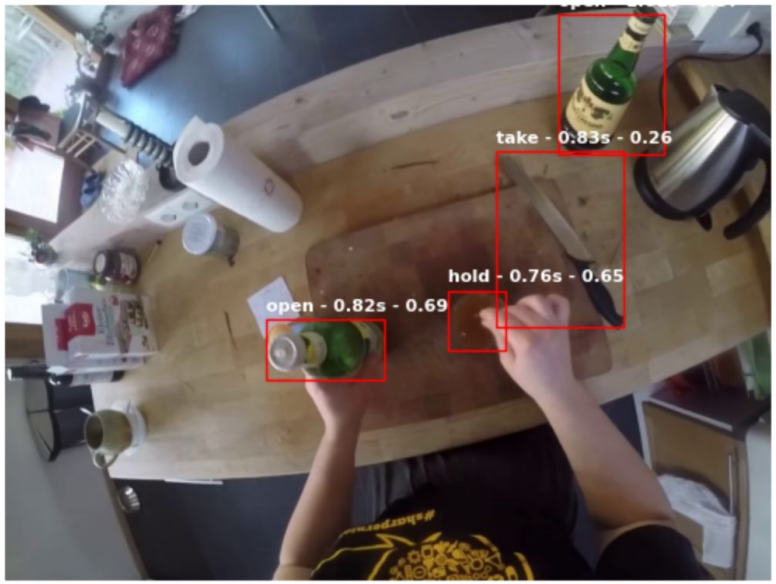
7th of September 2023



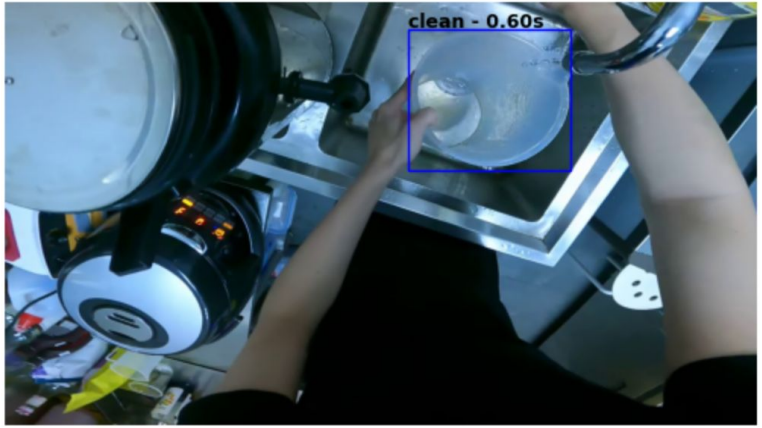


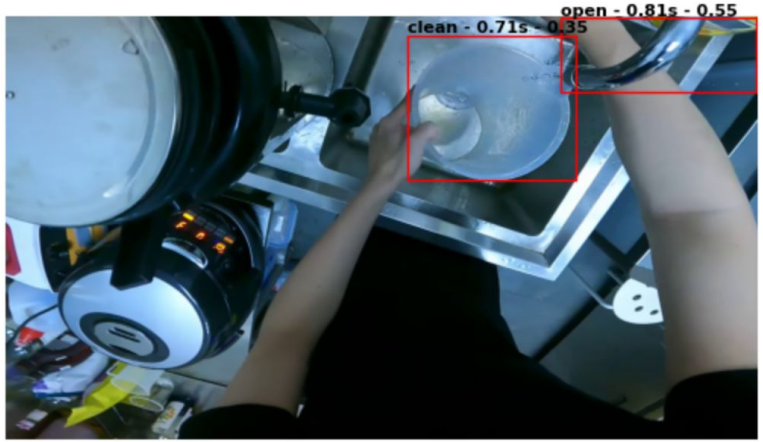
Training - Using VideoMae



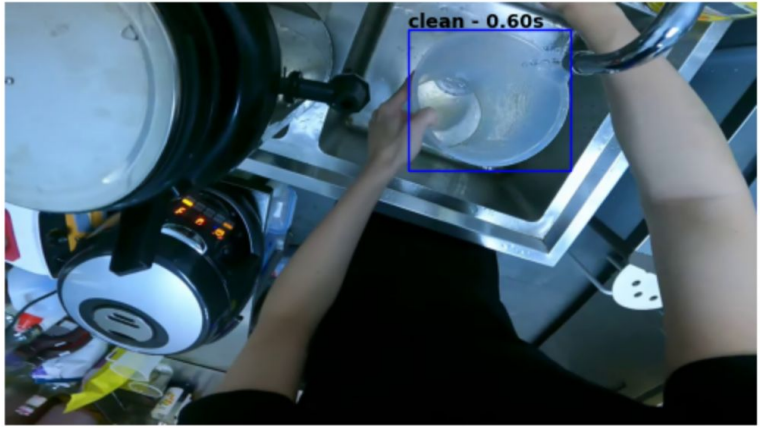


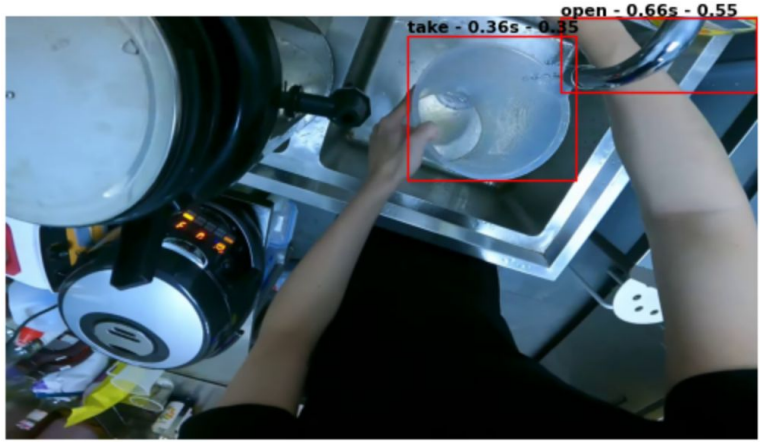
Training - Using VideoMae





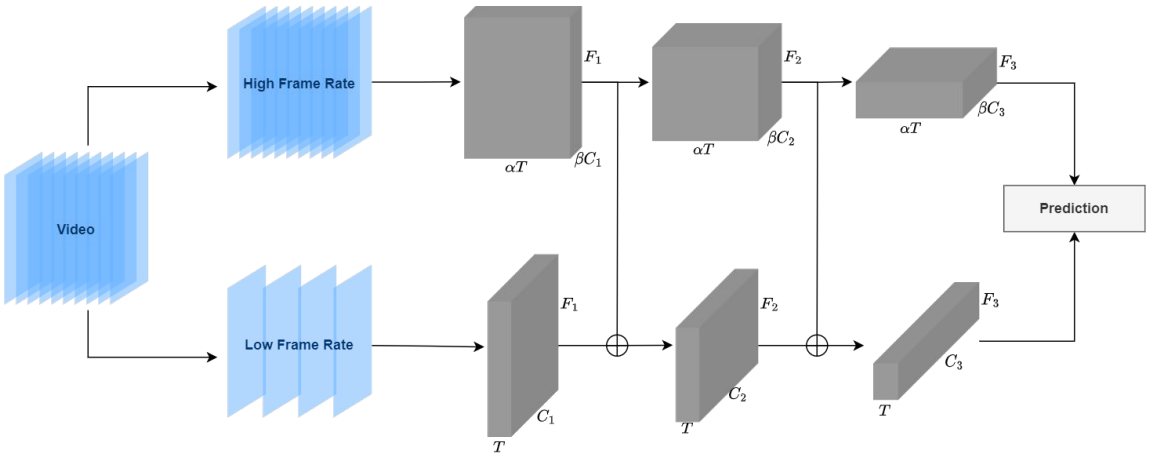
Validation - Using VideoMAE





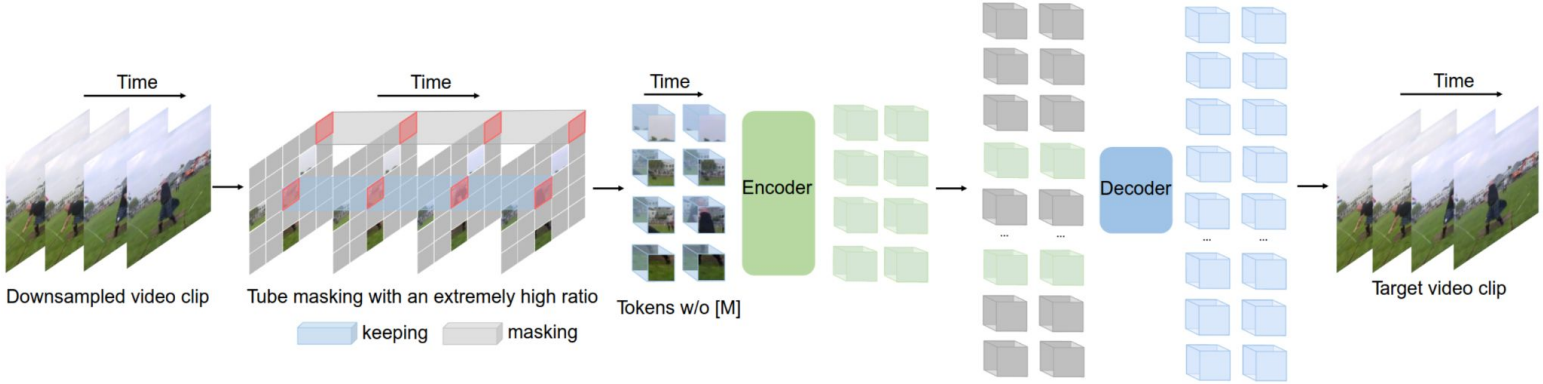
Validation - Using SlowFast

SlowFast



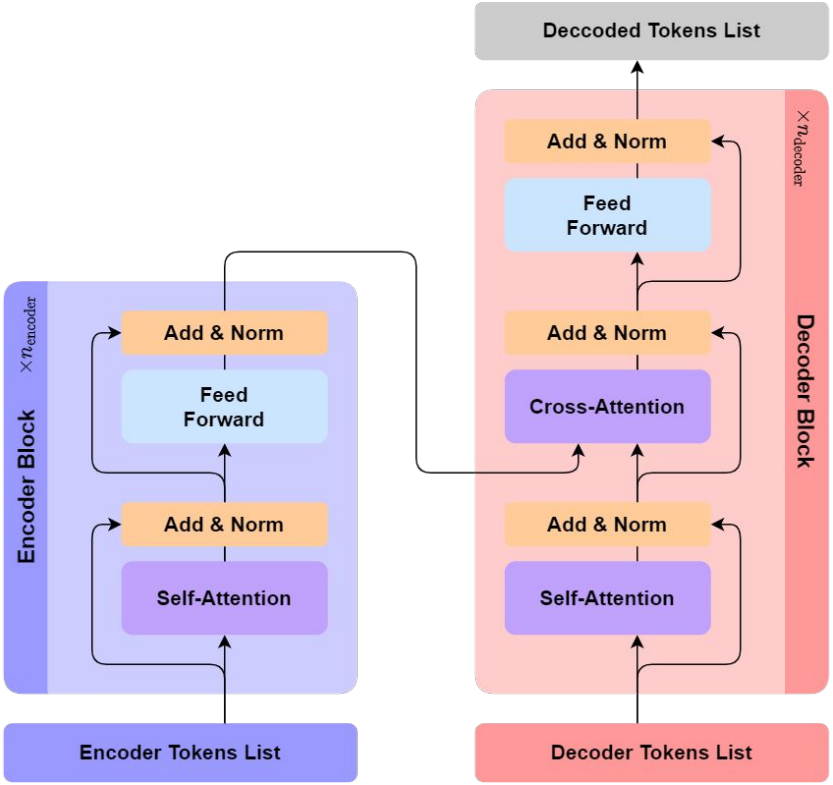
CREATING GROWTH, ENHANCING LIVES

VideoMAE

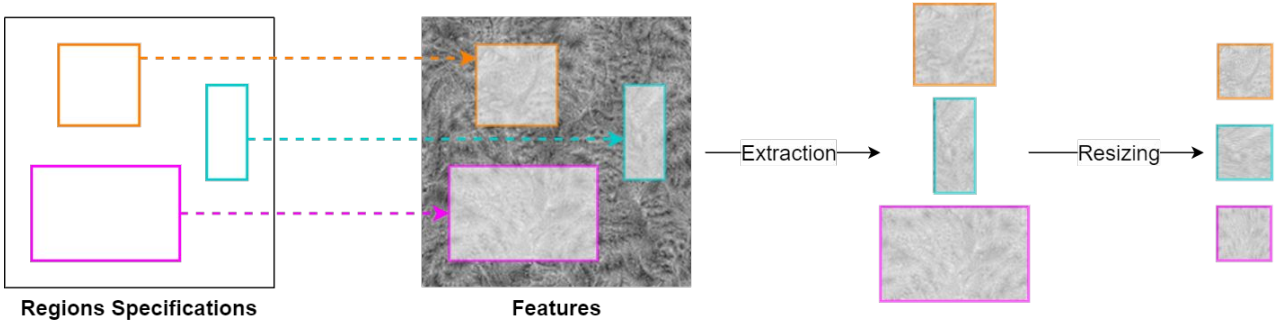


Z. Tong et al. *Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training*, 2022

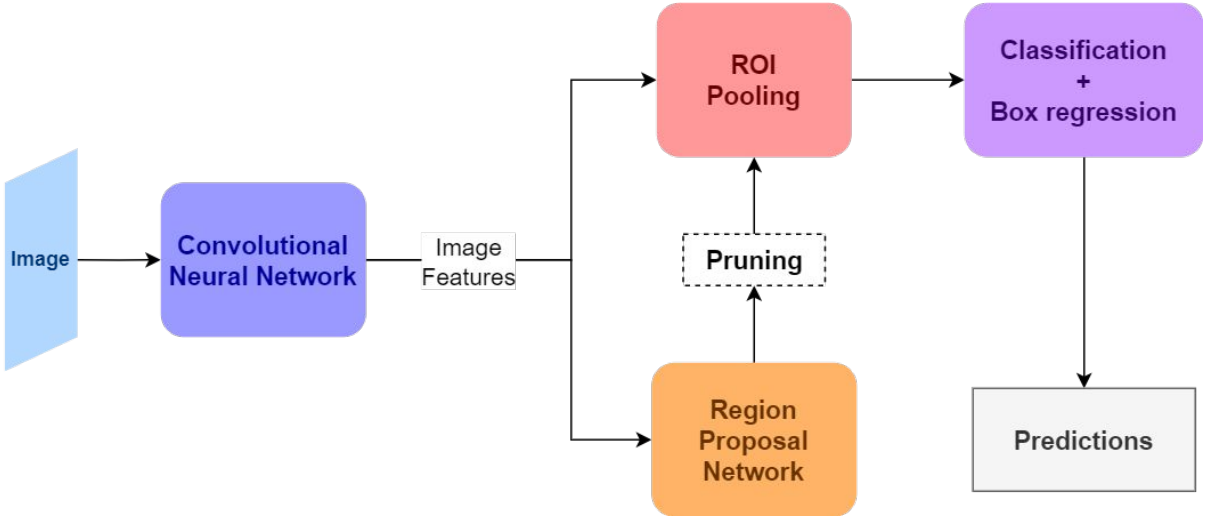
Transformer



ROI Pooling



Faster RCNN



Ground Truth Attribution

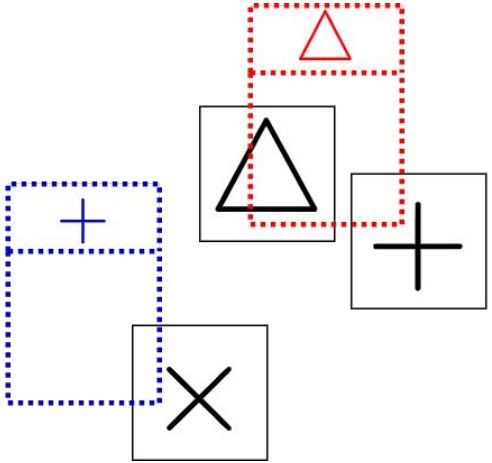


Table of Correspondence		
Detections		
Ground Truth		

Packing and Un-packing Data