FACULTAD DE CIENCIAS FISICAS Y MATEMATICAS
UNIVERSIDAD DE CHILE

RELELA
Representations for
Learning and Language

Thesis talk 1

BENCHMARKING VIDEO ACTION FEATURES FOR THE VIDEO TEMPORARY SENTENCE GROUNDING TASK

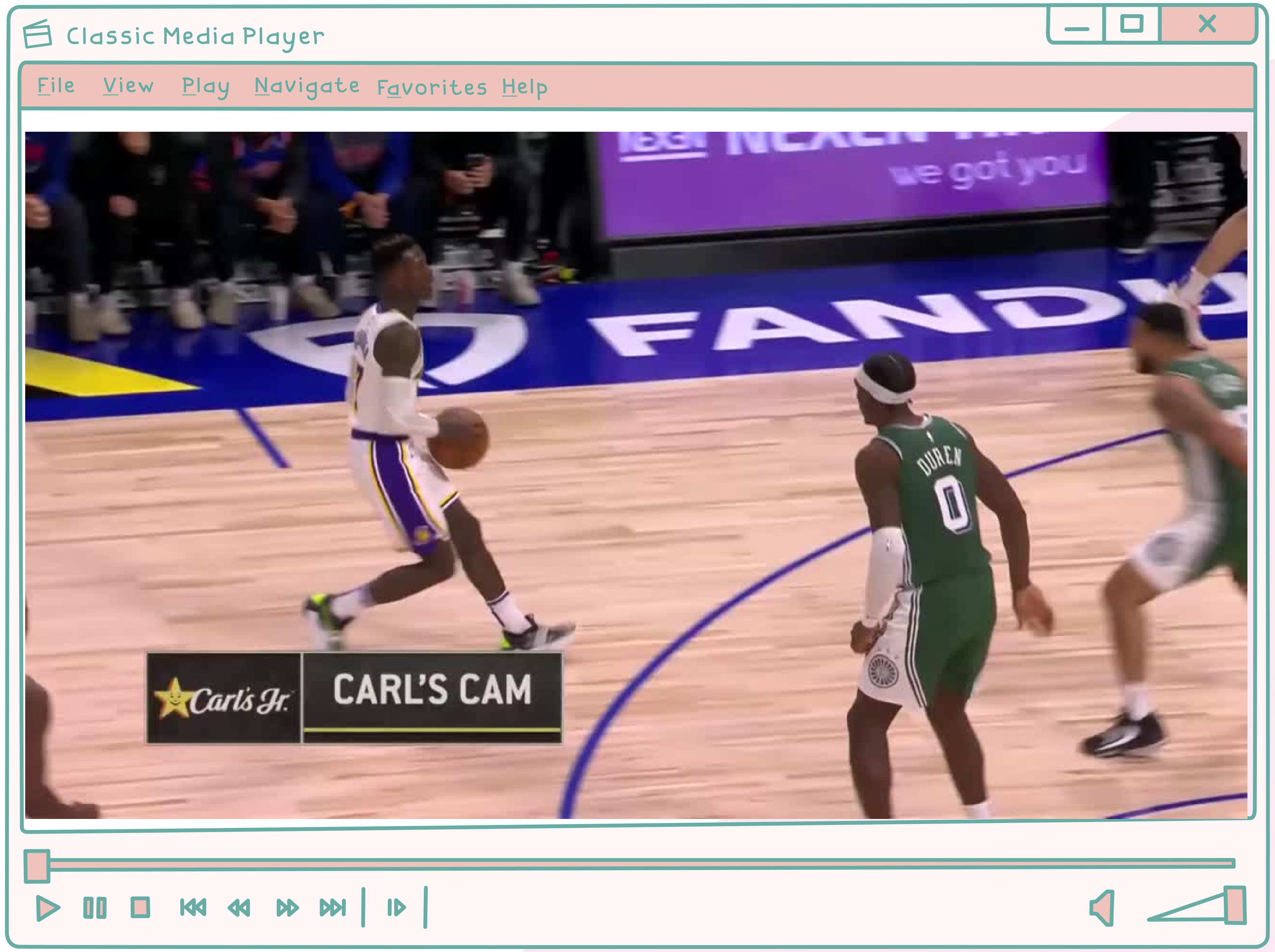
Presented by: Ignacio Meza De la Jara

Index

What we'll see?

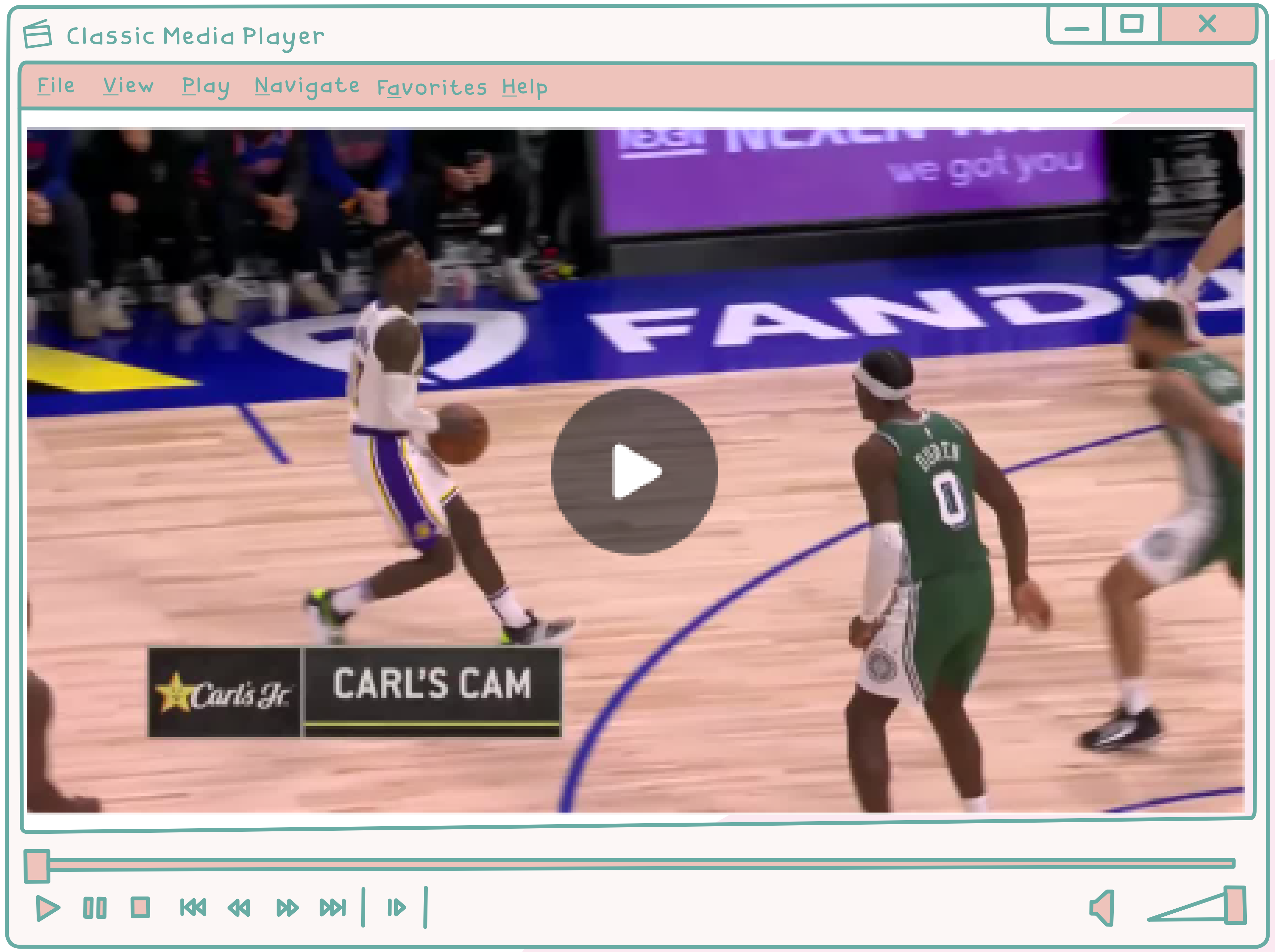
- What is TSGV?
- What kind of action classificators exist?
- Problem statement and Goals
- Methodology
- Contributions

Let's see a video...



Let's see a video...

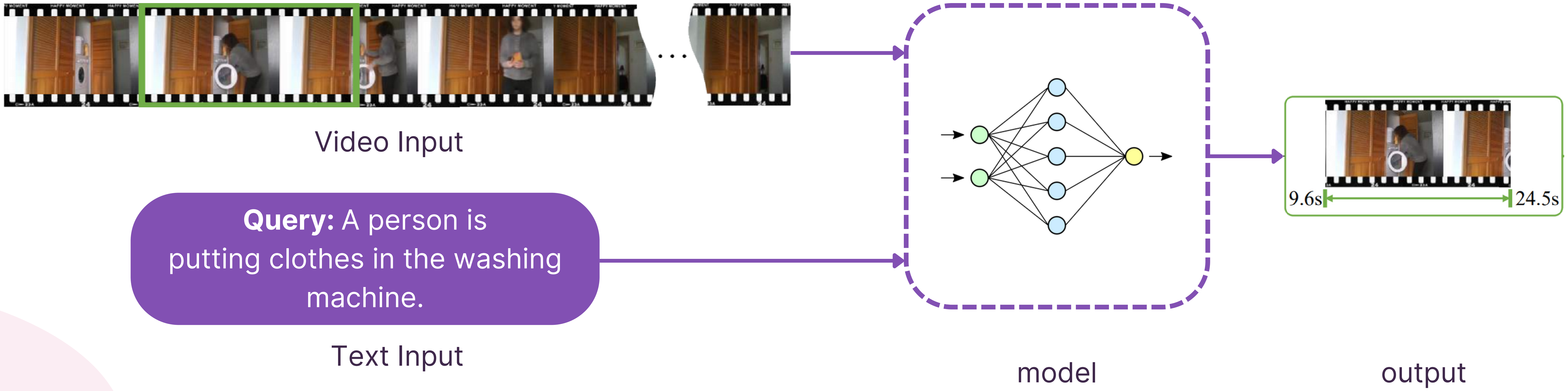
- It is interesting to automatically find a relevant moment in a video.
- This can be ambiguous... When is an action initiated?
- How could we relate a query to a frame?



- Reduce time spent searching for interesting events.
- Automation can be useful for many tasks

What is Temporary Sentence Grounding in Videos?

Given an untrimmed video, temporal sentence grounding in videos is to retrieve a video segment, also known as a temporal moment, that semantically corresponds to a query in natural language.



Video Input

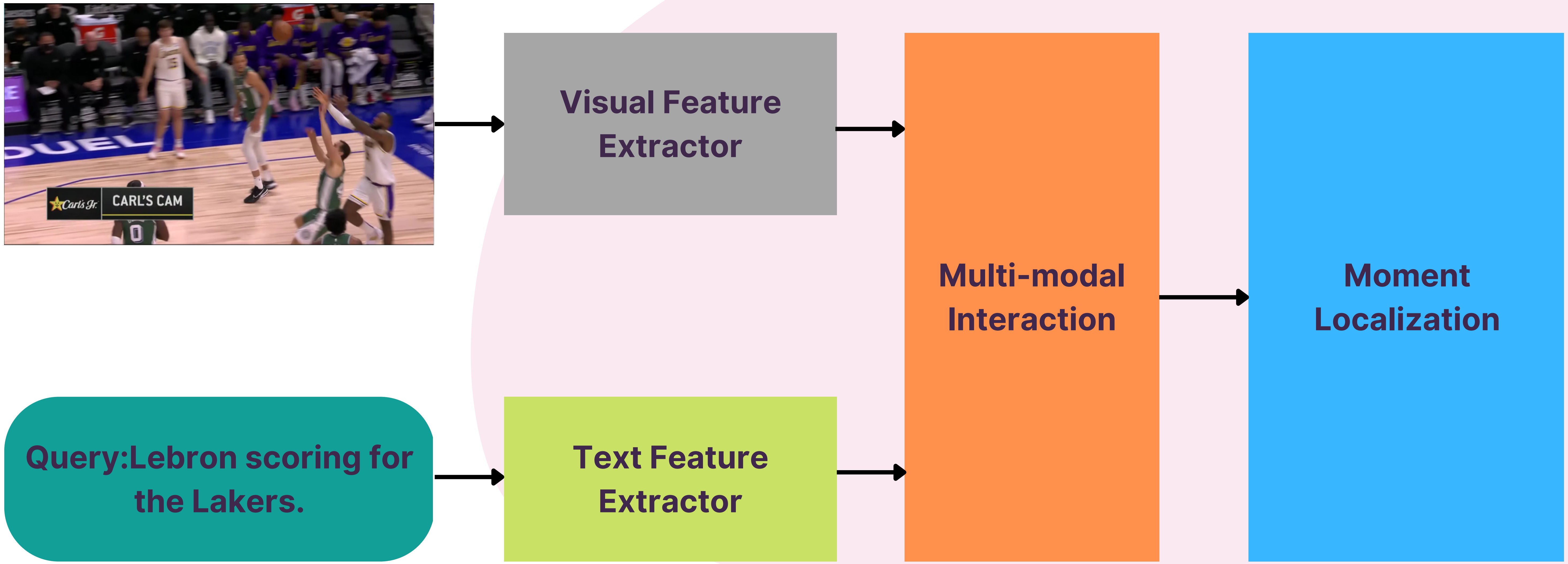
Query: A person is putting clothes in the washing machine.

Text Input

model

output

What is Temporary Sentence Grounding in Videos?



Visual Feature Extractor

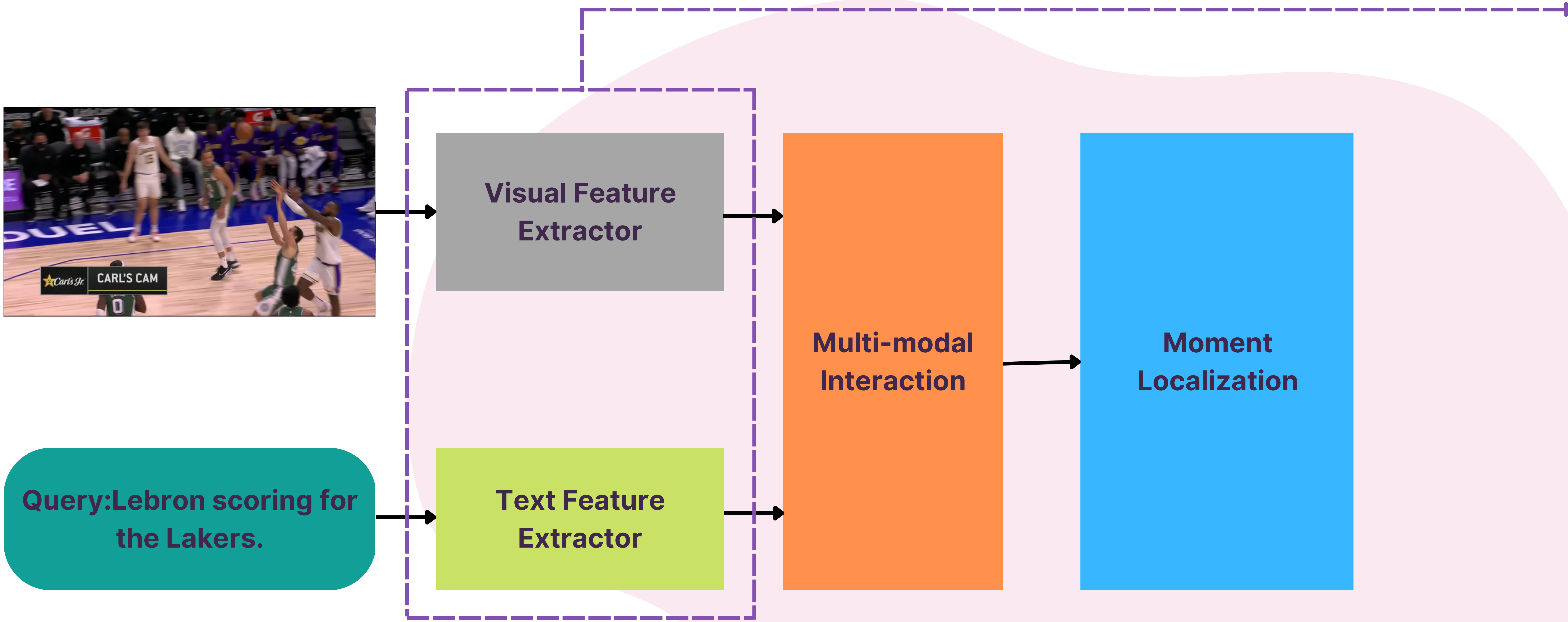
Query:Lebron scoring for the Lakers.

Text Feature Extractor

Multi-modal Interaction

Moment Localization

What is Temporary Sentence Grounding in Videos?



Visual Feature Extractor

Query:Lebron scoring for the Lakers.

Text Feature Extractor

Multi-modal Interaction

Moment Localization

Why should we use encoders?

- For the text and video we must obtain a computer-understandable representation of the inputs.

- Our general purpose is to find a relevant moment. A relevant moment is considered a moment where an action occurs, that is why we use models that manage to abstract information.

Visual Feature Extractor

- Typically, action classifiers are used to generate the action feature of the video.
- Many different types of classifiers are currently available.
- There is no work on the impact on TGSV using the different classifiers.

Textual Feature Extractor

- Different models of embeddings are used.
- Pre-trained models such as GloVe or BERT are generally used to obtain the characteristics.
- There are studies that prove the impact of different techniques.

Index

What we'll see?

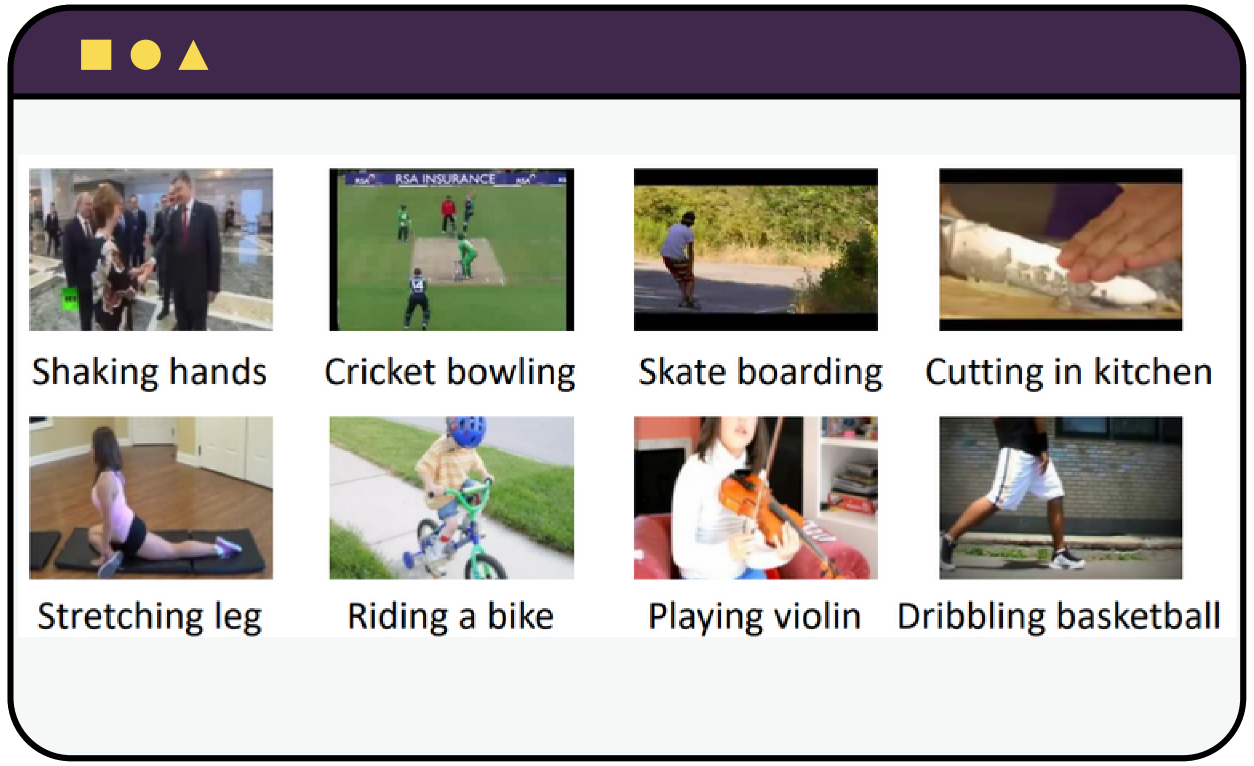
- What is TSGV?
- What kind of action classificators exist?
- Problem statement and Goals
- Methodology
- Contributions

What is a video action classifier?

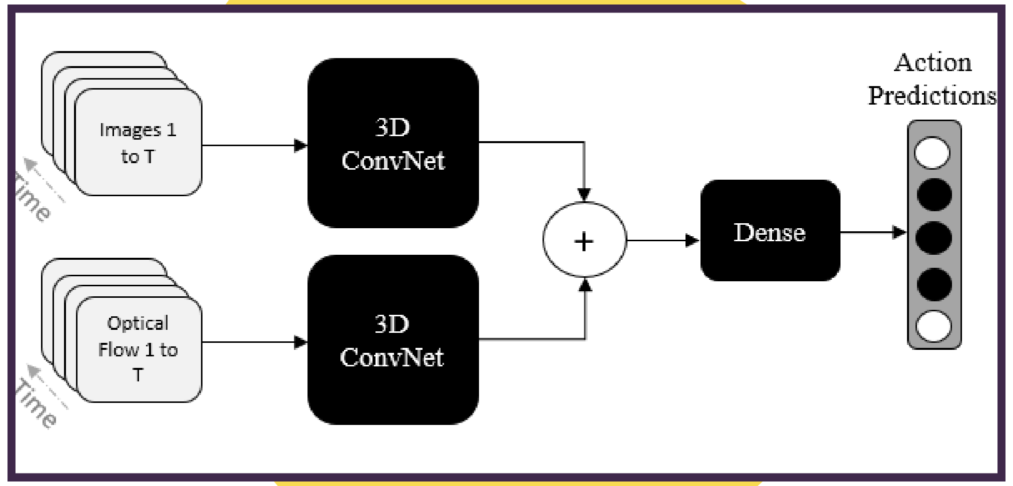
- Action classification is a task that attempts to classify human actions using trimmed videos.

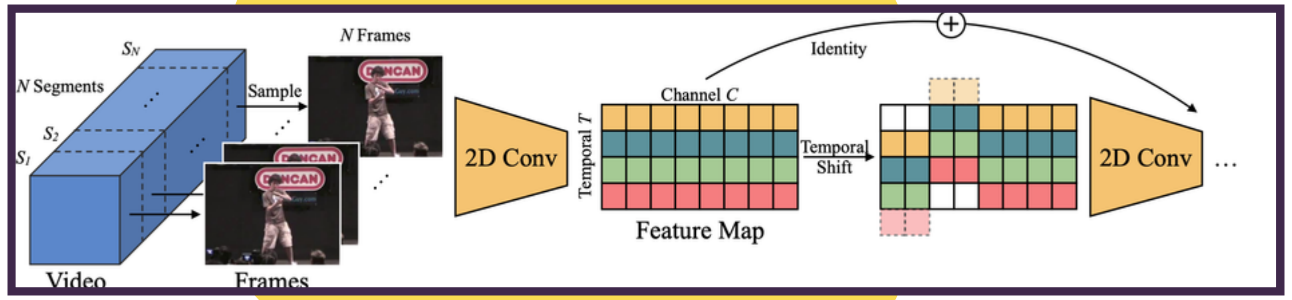
- The problem is difficult because human actions are often composite concepts and the hierarchy of these concepts is not well defined.

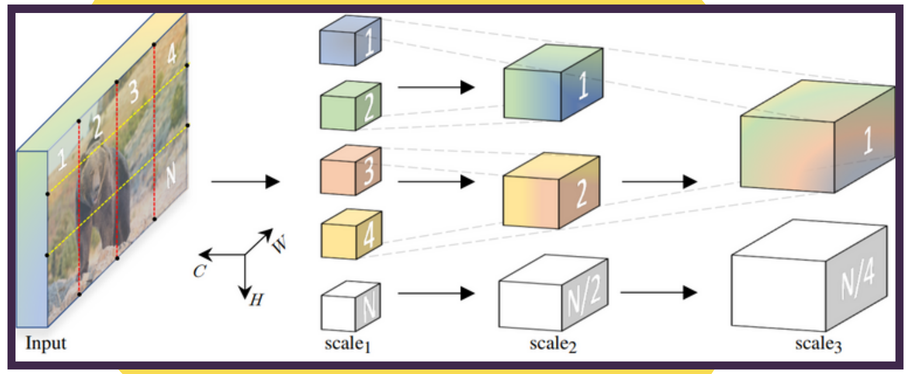
- There are different natures of solutions to address the problem



What kind of action classificators exist?







Holistic CNN

Temporal Reasoning

Holistic Transformer

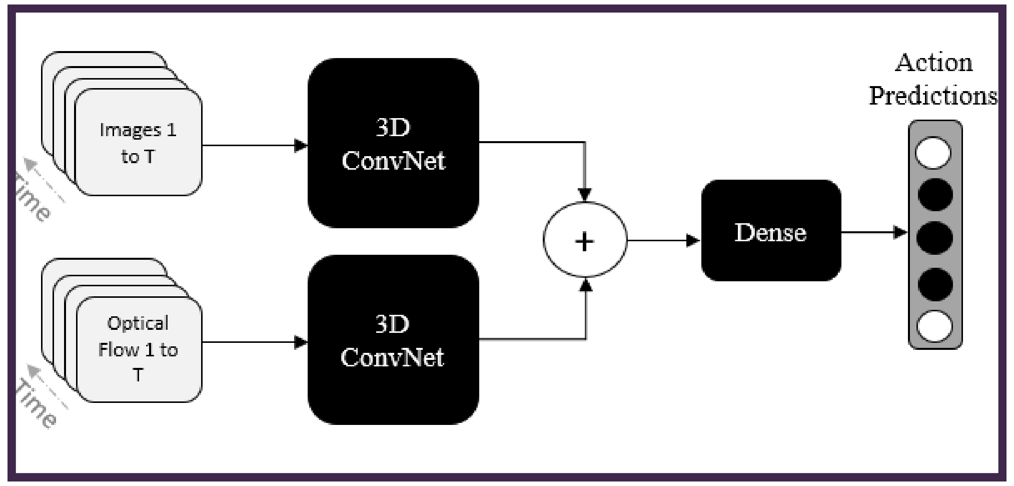
Holistic CNN

These are models that mainly use only convolutional networks to classify actions.

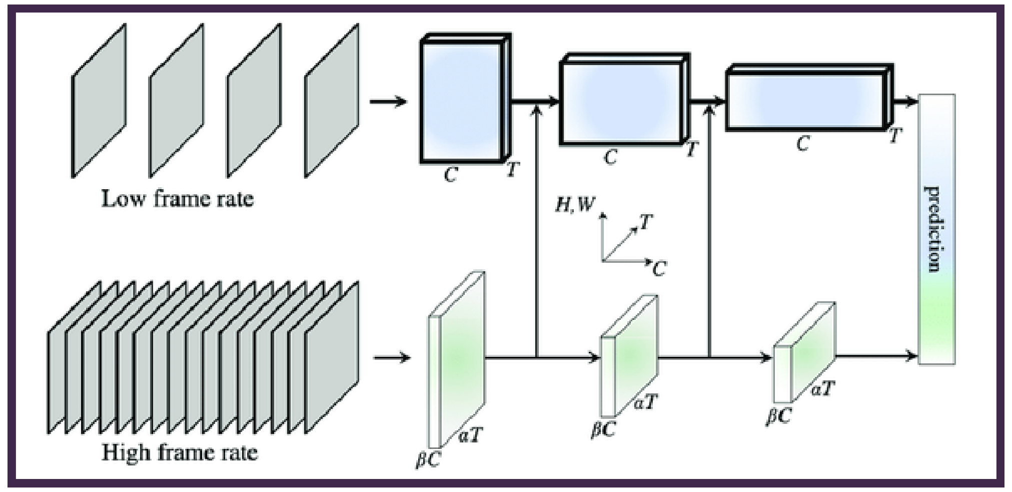
Within their architecture they have convolutional networks that manage to capture temporal information from the videos.

The convolutional network obtains the temporal information directly through a 3D CNN.

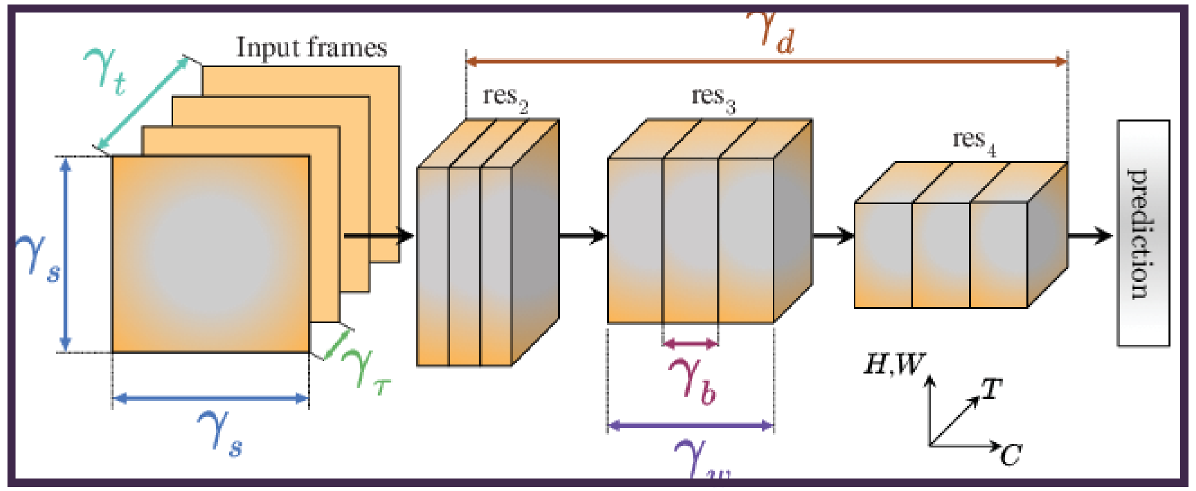
Some Relevants Works:



I3D



SlowFast



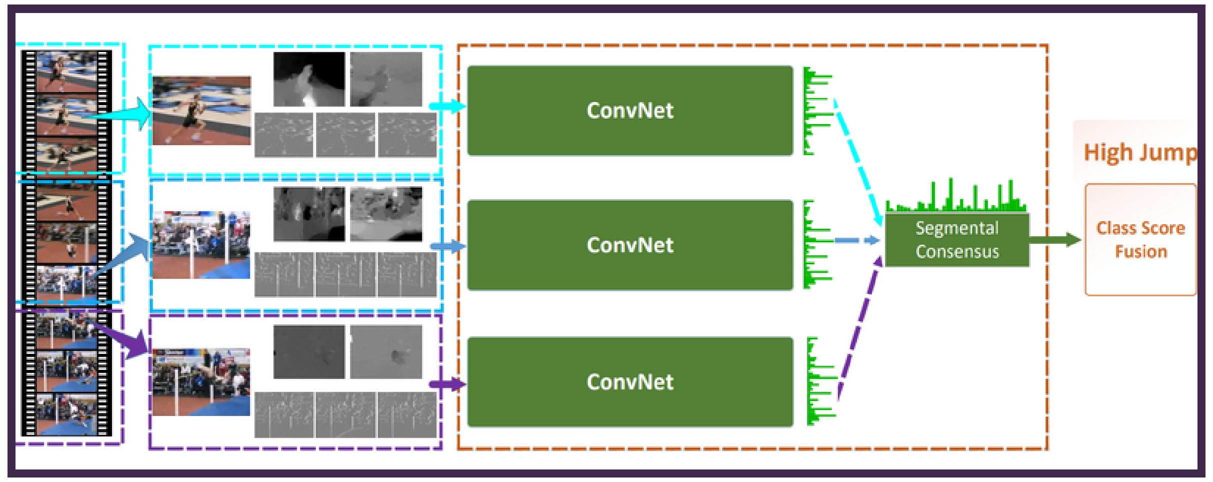
X3D

Temporal Reasoning

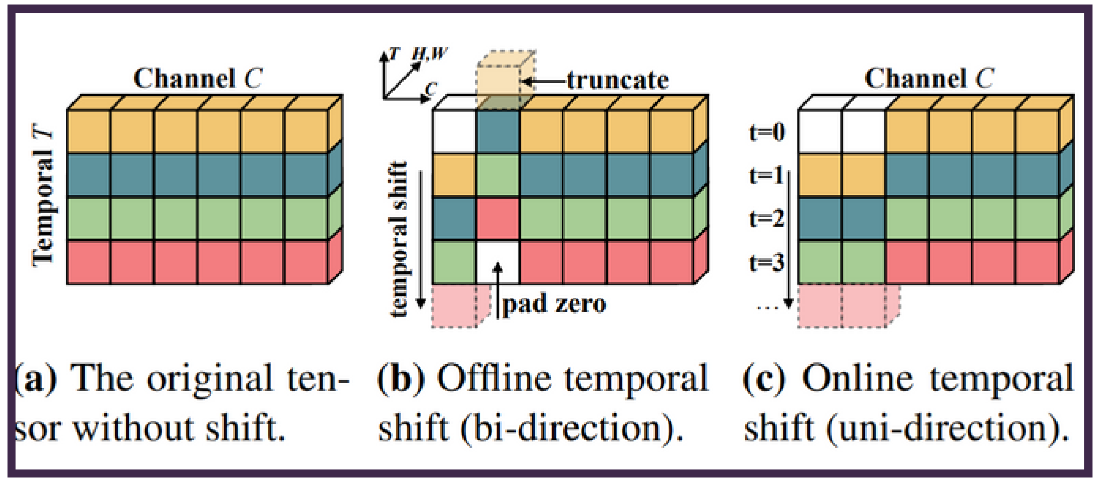
Their main objective is to reduce the operational cost of action classification.

Temporal Reasoning networks use methodologies that focus on efficient frame sampling and combining temporality with frame channel information.

Some Relevants Works:



TSN

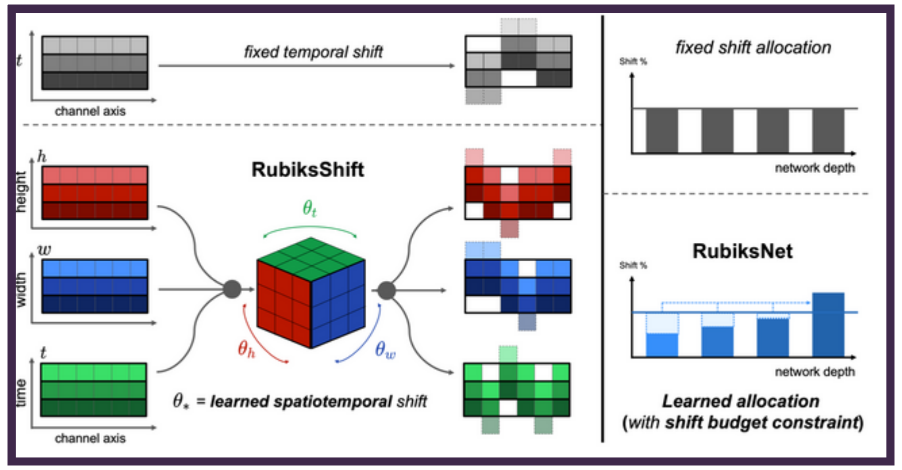


(a) The original tensor without shift.

(b) Offline temporal shift (bi-direction).

(c) Online temporal shift (uni-direction).

TSM



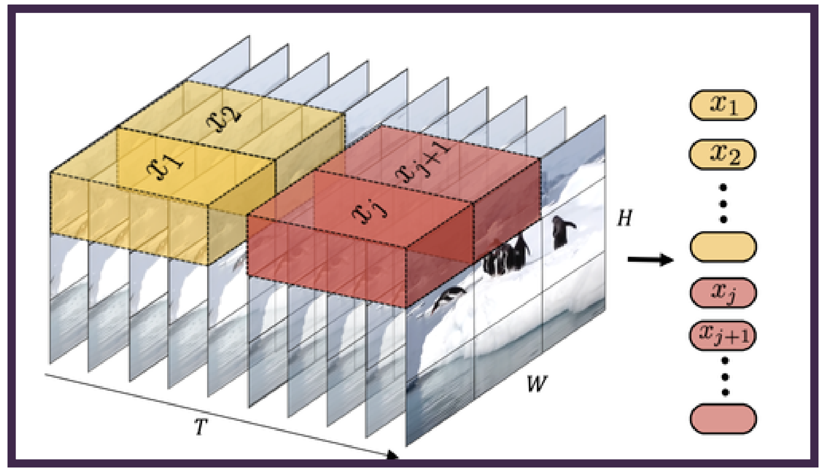
RubiksNet

Holistic Transformer

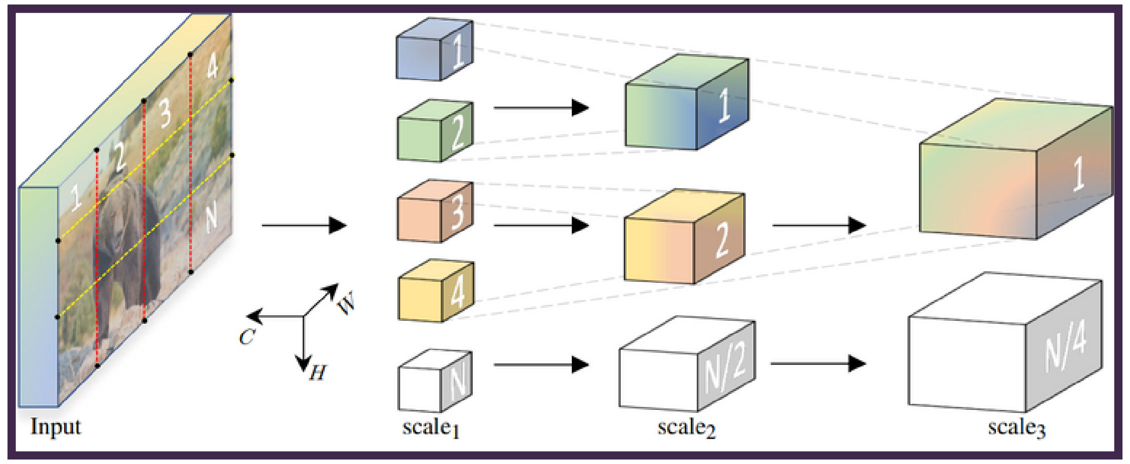
Combine the processes already known from CNNs with Transformers.

Allows for a more robust capture of temporality.

Some Relevants Works:



ViViT



MviT

Index

What we'll see?

- What is TSGV?
- What kind of action classificators exist?
- Problem statement and Goals
- Methodology
- Contributions

Problem statement 1

Video encoders form a crucial part of the temporal localization task, however, there is no evidence of the impact of each type of feature generated.

Research Questions

Q1: Are the characteristics of the encoder used relevant for the temporal localization of actions?.

Q2: The nature of the datasets on which the action classifiers were trained may affect temporal localization performance?.

Q3: Is the state of the art video encoders the best choice for temporal localization of actions?.

Problem statement 2

While multiple models exist for the localization task, the problem of temporal localization still present an opportunity for improvement using recent techniques.

Research Questions

Q1: Does a good end-to-end architecture mitigate the embedding extraction process?
Q2: What aspects should be considered to generate good multi-modal representations?
Q3: How can labeling uncertainty be modeled?

Goals

1. Studying the impact of different video classifiers on the task of temporal localization of a relevant moment.

2. Creation of an end-to-end model for the localization of relevant moments based on Deep learning.

Index

What we'll see?

- What is TSGV?
- What kind of action classificators exist?
- Problem statement and Goals
- Methodology
- Contributions

Methodology

Benchmark

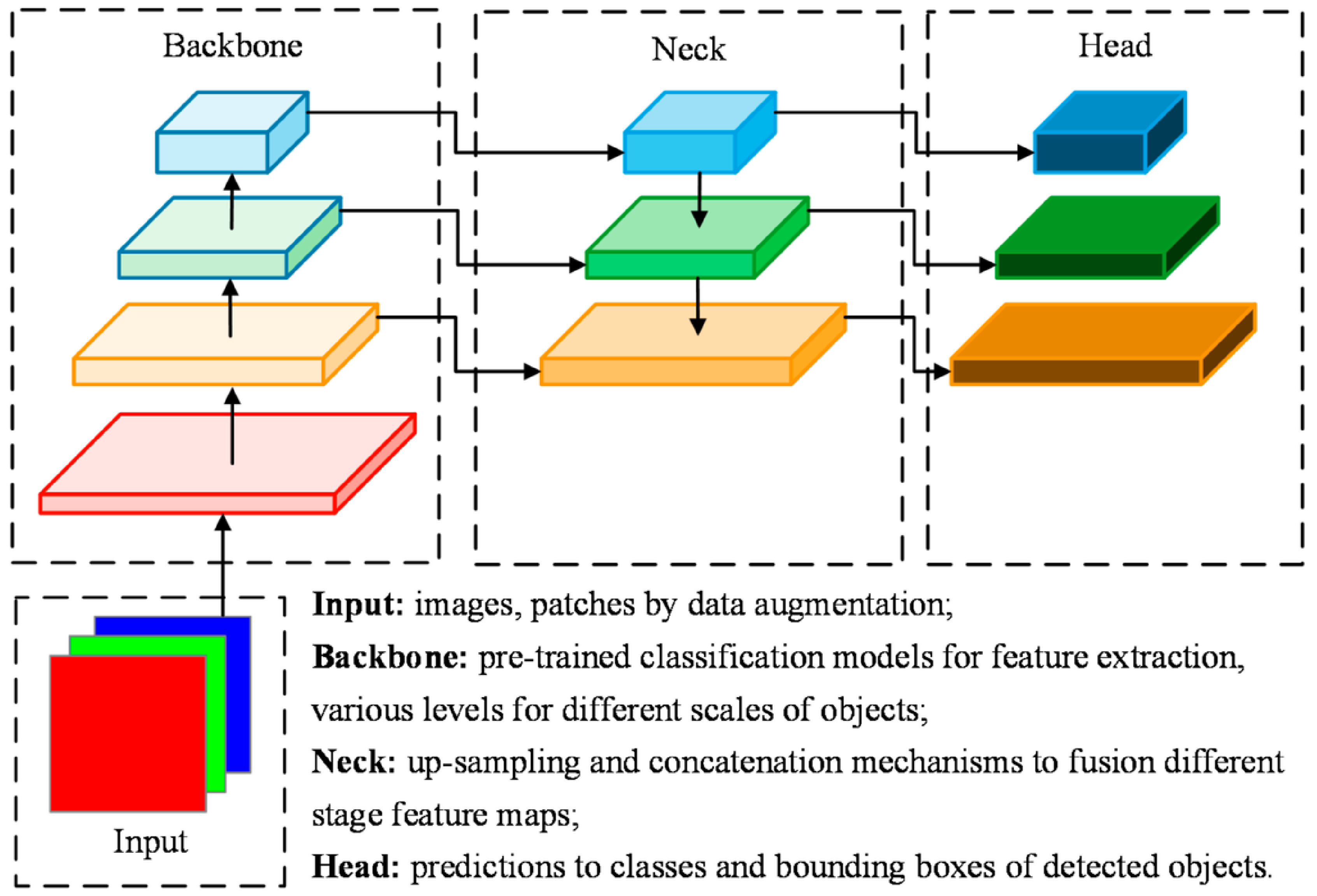
Focused on obtaining features from videos for comparison.

Model Creation

Focused on the creation of a model capable of finding a relevant time through a natural language query.

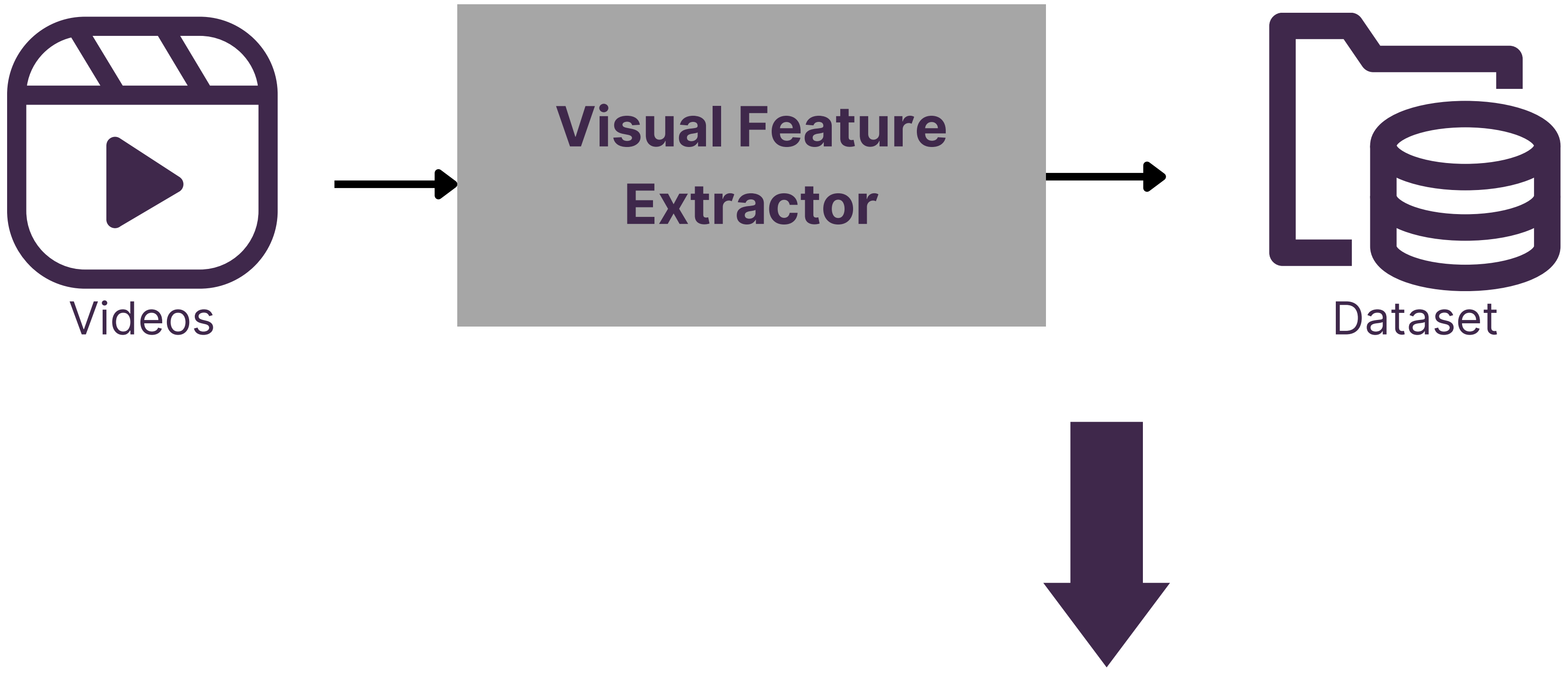
Methodology: Benchmark

The features are obtained by modifying the outputs of the stock classification models omitting the neck and head of the networks.

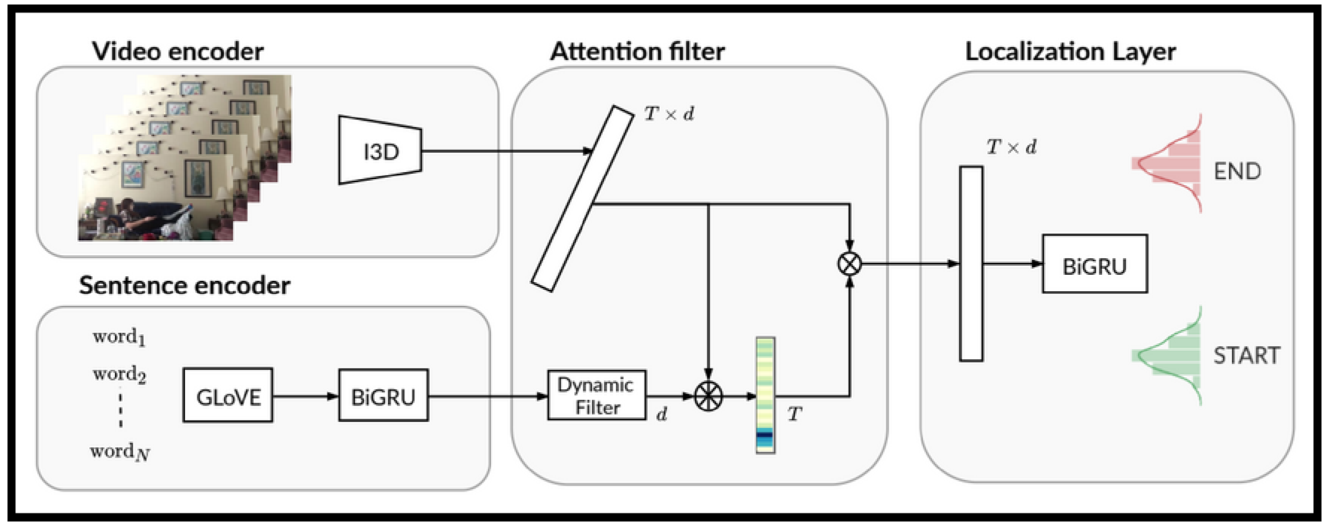


Input: images, patches by data augmentation;
Backbone: pre-trained classification models for feature extraction, various levels for different scales of objects;
Neck: up-sampling and concatenation mechanisms to fusion different stage feature maps;
Head: predictions to classes and bounding boxes of detected objects.

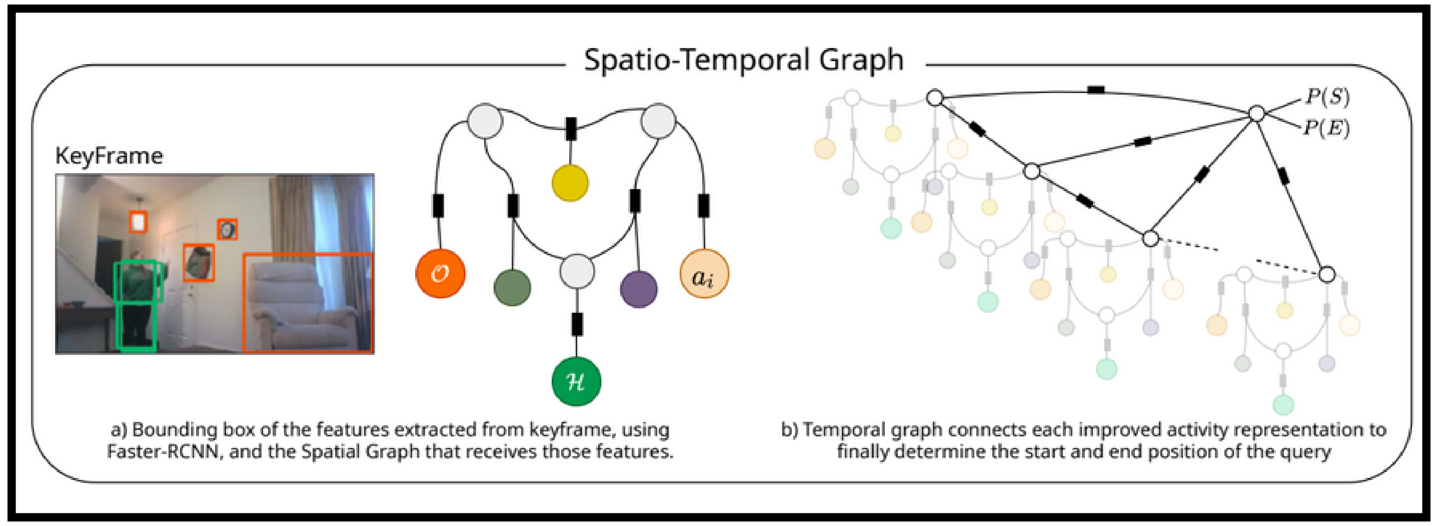
Methodology: Benchmark



Encoders are kept fixed during feature extraction and are only used for feature extraction.

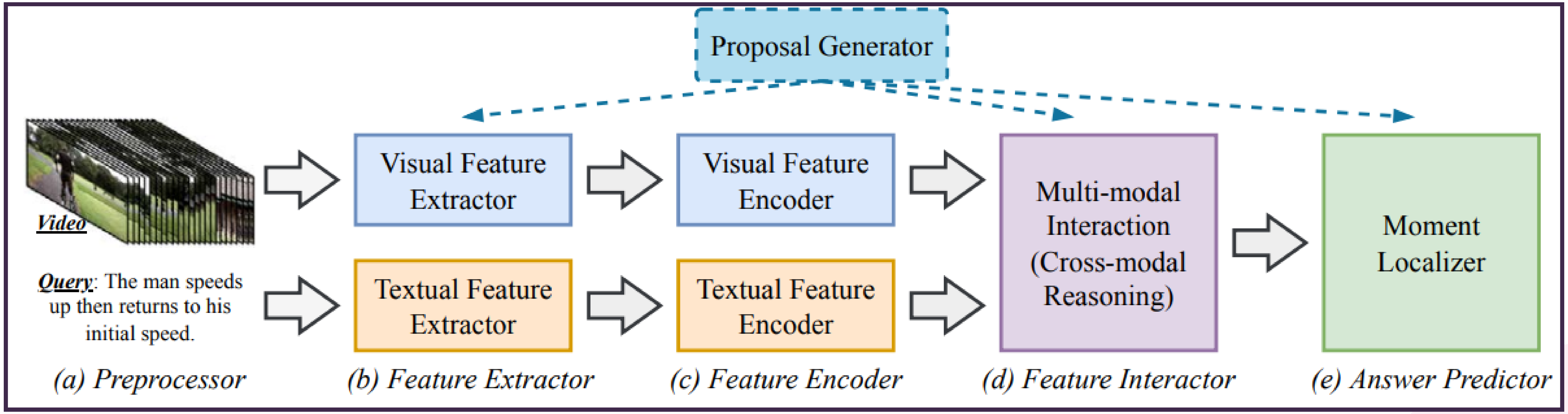


TMLGA


Spatio-Temporal Graph
KeyFrame
$\mathcal{O}$
a_i
$\mathcal{H}$
$P(S)$
$P(E)$
a) Bounding box of the features extracted from keyframe, using Faster-RCNN, and the Spatial Graph that receives those features.
b) Temporal graph connects each improved activity representation to finally determine the start and end position of the query

DoRi

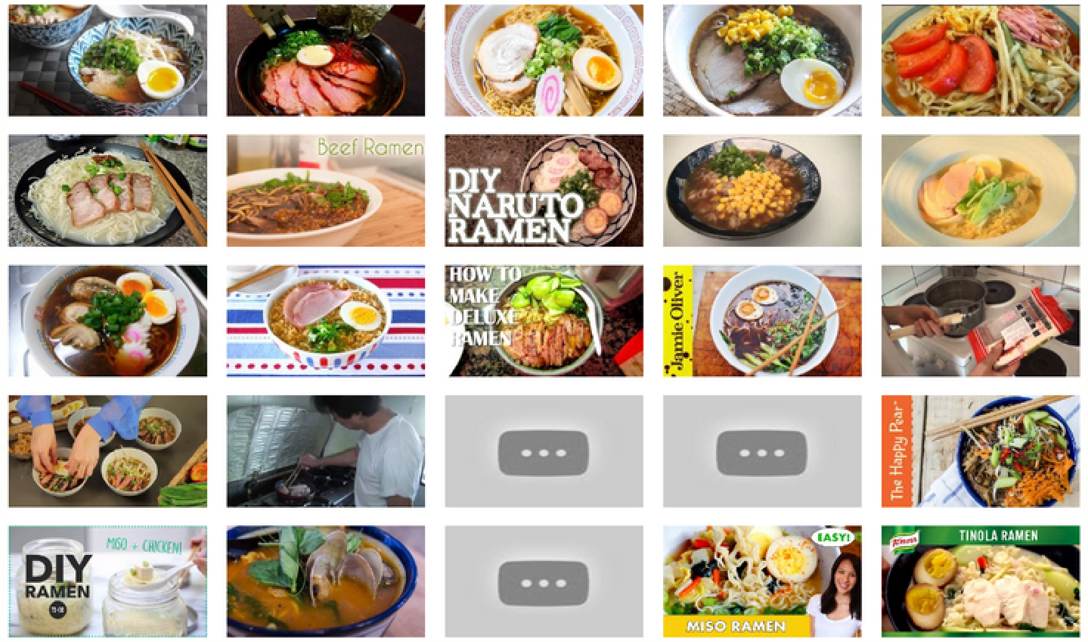
Methodology: Model Creation



Methodology: Datasets

Charades-STA

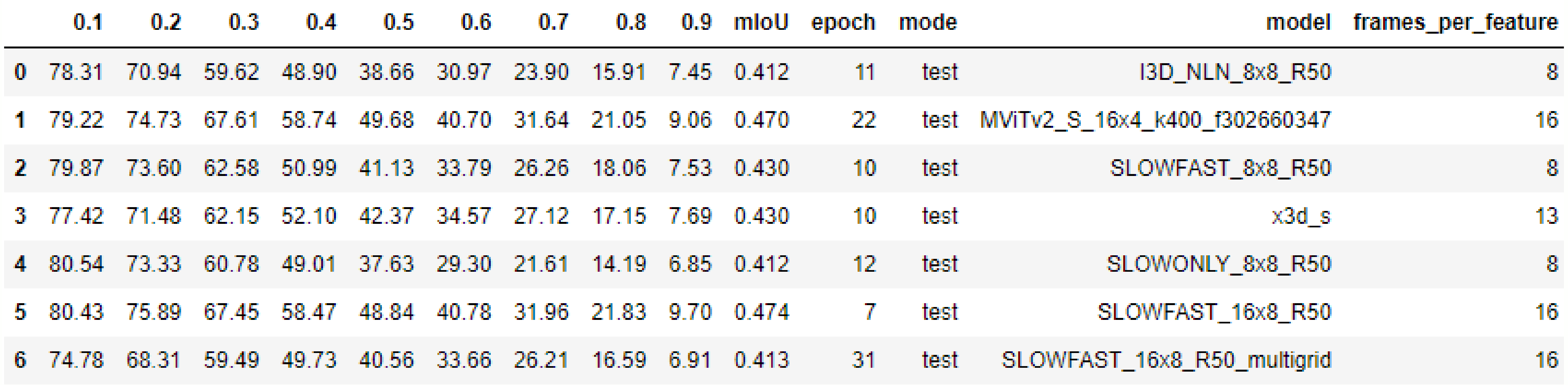
ActivityNet



YourCook II

Methodology: Preliminary Progress

- Implemented the extraction of most of the features to be analyzed in code.
- Tested part of Charades features with TMLGA model
- Features obtained for the Charades-STA and ActivityNet datasets



	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	mIoU	epoch	mode	model	frames_per_feature
0	78.31	70.94	59.62	48.90	38.66	30.97	23.90	15.91	7.45	0.412	11	test	I3D_NLN_8x8_R50	8
1	79.22	74.73	67.61	58.74	49.68	40.70	31.64	21.05	9.06	0.470	22	test	MViTv2_S_16x4_k400_f302660347	16
2	79.87	73.60	62.58	50.99	41.13	33.79	26.26	18.06	7.53	0.430	10	test	SLOWFAST_8x8_R50	8
3	77.42	71.48	62.15	52.10	42.37	34.57	27.12	17.15	7.69	0.430	10	test	x3d_s	13
4	80.54	73.33	60.78	49.01	37.63	29.30	21.61	14.19	6.85	0.412	12	test	SLOWONLY_8x8_R50	8
5	80.43	75.89	67.45	58.47	48.84	40.78	31.96	21.83	9.70	0.474	7	test	SLOWFAST_16x8_R50	16
6	74.78	68.31	59.49	49.73	40.56	33.66	26.21	16.59	6.91	0.413	31	test	SLOWFAST_16x8_R50_multigrid	16

Index

What we'll see?

- What is TSGV?
- What kind of action classificators exist?
- Problem statement and Goals
- Methodology
- Contributions

Contribution

Dataset of video features based on two of the main datasets used for the localization of relevant moments.

Comparison and analysis of the impact of different video classifier features on the temporal localization task.

A model for the temporal localization task with competitive results in the current state of the art.

References

- Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik & Kaiming He (2018): SlowFast Networks for Video Recognition. https://arxiv.org/abs/1812.03982, doi:10.48550/ARXIV.1812.03982
- Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Basura Fernando, Hongdong Li & Stephen Gould (2020): DORi: Discovering Object Relationship for Moment Localization of a Natural-Language Query in Video, doi:10.48550/ARXIV.2010.06260. Available at https://arxiv.org/abs/2010.06260.
- Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik & Christoph Feichtenhofer (2021): Multiscale Vision Transformers. https://arxiv.org/abs/2104.11227, doi:10.48550/ARXIV.2104.11227.
- Linxi Fan*, Shyamal Buch*, Guanzhi Wang, Ryan Cao, Yuke Zhu, Juan Carlos Niebles & Li Fei-Fei (2020): RubiksNet: Learnable 3D-Shift for Efficient Video Action Recognition.

- Cristian Rodriguez-Opazo, Edison Marrese-Taylor, Basura Fernando, Hongdong Li & Stephen Gould (2020): DORi: Discovering Object Relationship for Moment Localization of a Natural-Language Query in Video, doi:10.48550/ARXIV.2010.06260. Available at https://arxiv.org/abs/2010.06260.

Thesis talk 1

BENCHMARKING VIDEO ACTION FEATURES FOR THE VIDEO TEMPORARY SENTENCE GROUNDING TASK

Presented by: Ignacio Meza De la Jara