Temporal Modular Networks for Retrieving Complex Compositional Activities in Videos

Bingbin Liu, Serena Yeung, Edward Chou, De-An Huang, Li Fei-Fei, Juan Carlos Niebles

STANFORD VISION & LEARNING

Google Cloud

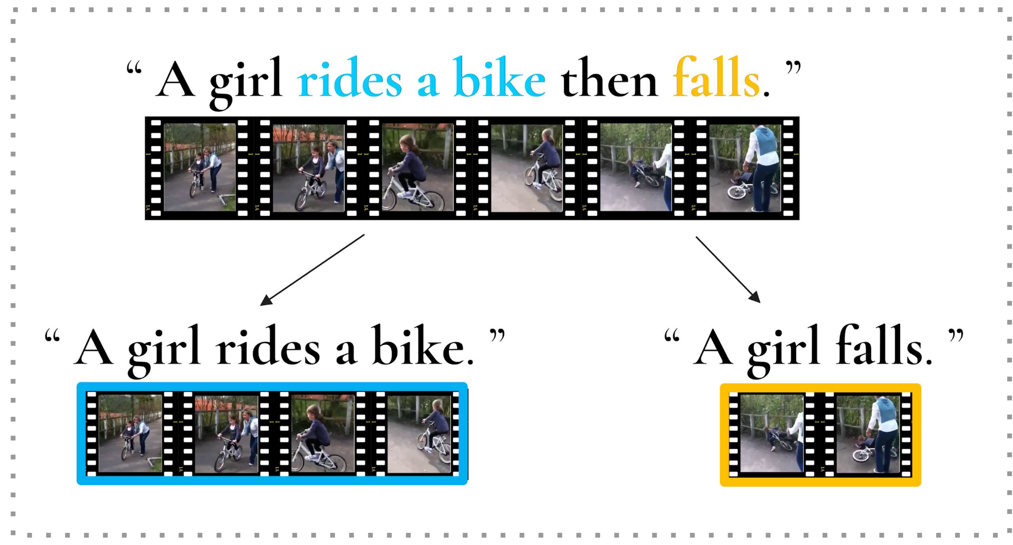
1. Motivation

One challenge in video understanding is to **scale to the long tail of complex activities** without requiring large amounts of data for new activities.

An insight of this project is that these activities are often **compositional**, where different complex activities may be composed of shared smaller units.

Key observations:

- A modular network of reusable modules with shared parameters can improve scalability.
- Leveraging structures in natural languages can enhance temporal video understanding.



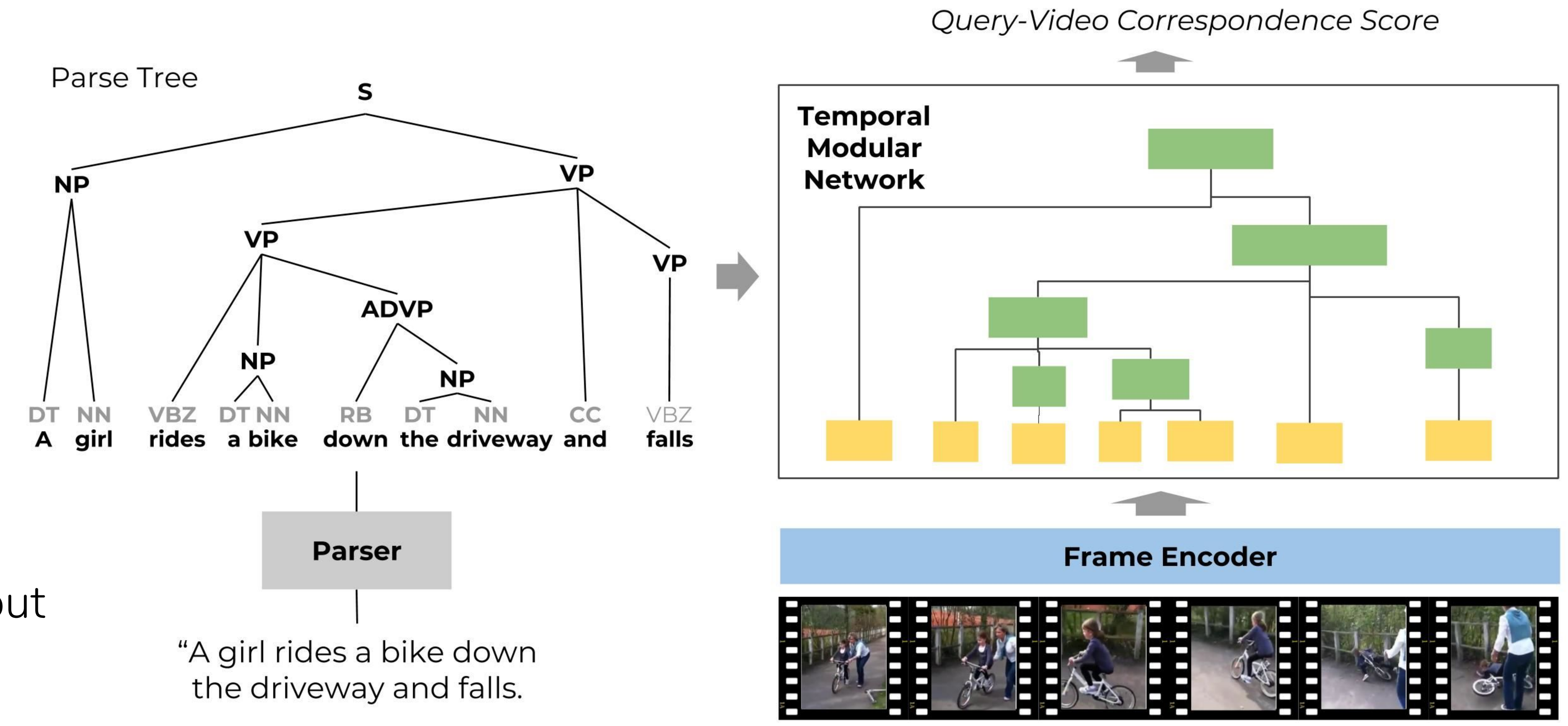
2. Overview

The work focuses on natural language intra-video retrieval, which aims at locating the query in the input video.

Given an input query-video pair, the proposed framework will:

1) **Dynamically assemble** a network based on the structure of the query's parse tree; and

2) **Temporally locate** the query in the input video by segment-level correspondence.



3. Temporal Modular Network (TMN)

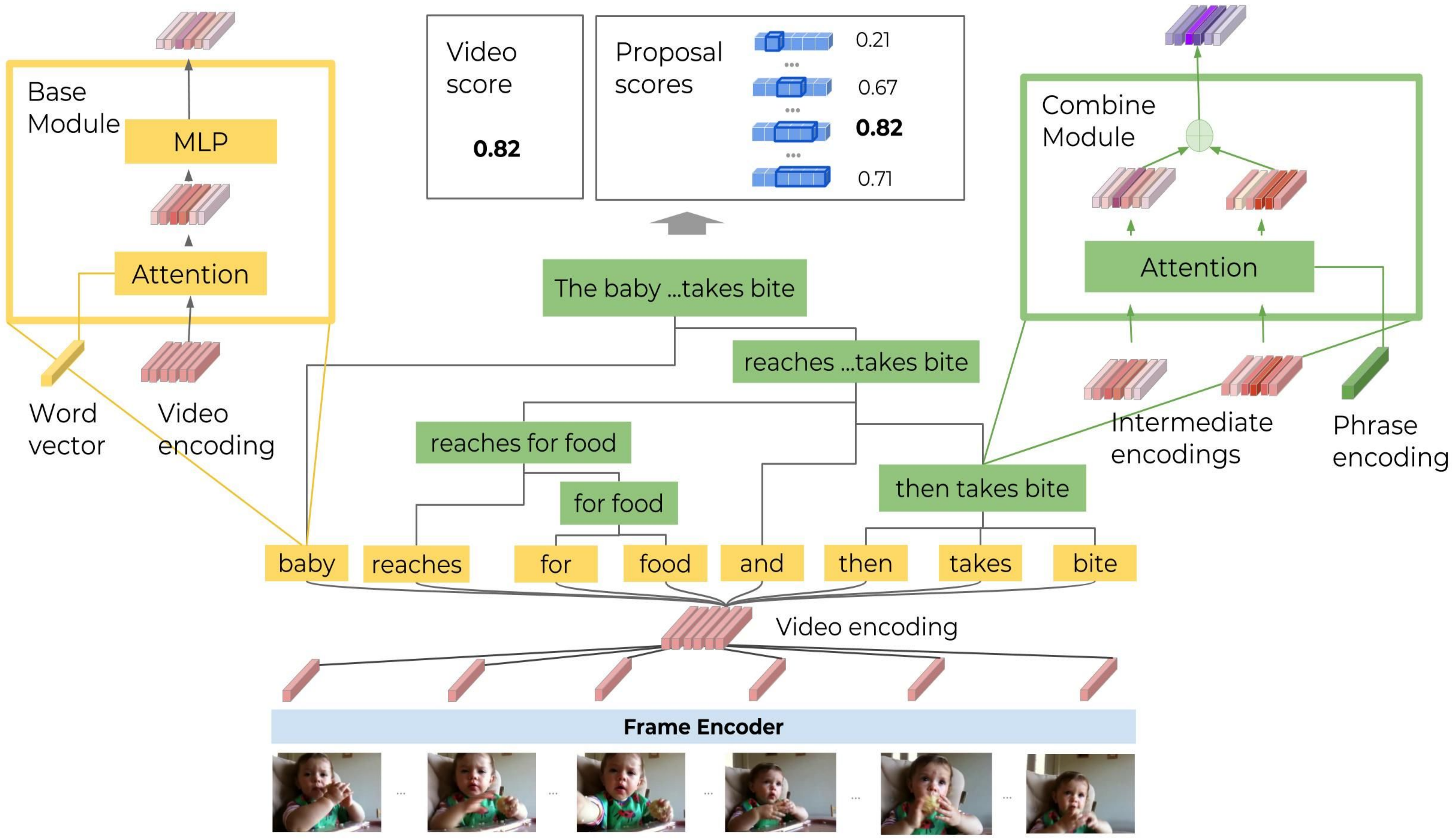
The proposed **temporal modular network** takes as input a query-video pair and performs intra-video retrieval in three stages:

1. Leveraging **structures in natural languages** by adapting the network structure using query-dependent parse trees.

2. Building instance-specific network from reusable modules:

- **Base modules** (yellow) which take in word vectors and video encodings.
- **Combine modules** (green) which pass information in lower-level feature maps up in the compositional structure.

3. **Temporal localization** from segment level correspondence scores.



i) Handheld game system is switched on.

ii) A dog sniffs another dog and then jumps away.

iii) Woman turns away.

iv) A cheerleader runs away.

4. Experiments

We conduct experiments on the **DiDeMo** dataset.

Training: network modules and scoring layers are jointly trained given query-video pairs, using both *intra-video* negatives for temporal accuracy and *inter-video* negatives for scene semantics.

Rank loss is used to better fit the *intra-video* retrieval setting, by penalizing less on segments with less accurate temporal bounds but containing correct semantics.

Results: TMN outperforms the baseline on different modalities:

Feature	Model	Rank@1	Rank@5	mean IoU
RGB	MCN	13.10	44.82	25.13
	TMN	**18.71**	**72.97**	**30.14**
Flow	MCN	18.35	56.25	31.46
	TMN	**19.90**	**75.14**	**31.95**
Fuse	MCN	19.88	62.39	33.51
	TMN	**22.92**	**76.08**	**35.17**

Ablation study:

- Temporal attention
- Rank loss
- Use of tree structures
- Type of structure

Model	Rank@1	Rank@5	mean IoU
MCN [16] (i.e. no tree structure)	19.88	62.39	33.51
const + max pool + rank loss	21.89	75.69	34.24
dep + combine attention + BCE loss	20.41	75.38	32.86
dep + combine attention + rank loss	21.67	75.98	33.94
const + combine attention + BCE loss	21.60	75.81	34.40
const + combine attention + rank loss	**22.92**	**76.08**	**35.17**