EVOQUER: Enhancing Temporal Grounding with Video-Pivoted BackQuery Generation

PennState

Kwai

Yanjun Gao[1], Lulu Liu[1], Jason Wang[1], Xin Chen[2], Huayan Wang[3], Rui Zhang[1]

Pennsylvania State University[1], Kwai.Inc[2,3]

{yug125, lzl5409, jjw6188, r.zhang}@psu.edu[1], xinchen.hawaii@gmail.com[2], wanghuayan@kuaishou.com[3]

Background -- Temporal Grounding

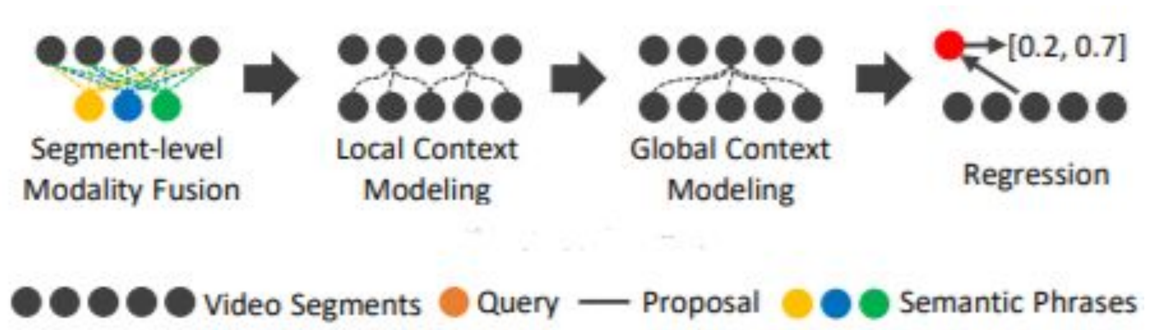
- Predicting a **time interval** corresponding to **natural language query input**
- Addresses the **temporal, semantic alignment** between vision and language as well as tasks like **visual storytelling**
- Recent work has emphasized modeling the semantic mapping of **verbs and nouns** to **actions and objects**

Contribution -- ***EVOQUER*** for Temporal Grounding

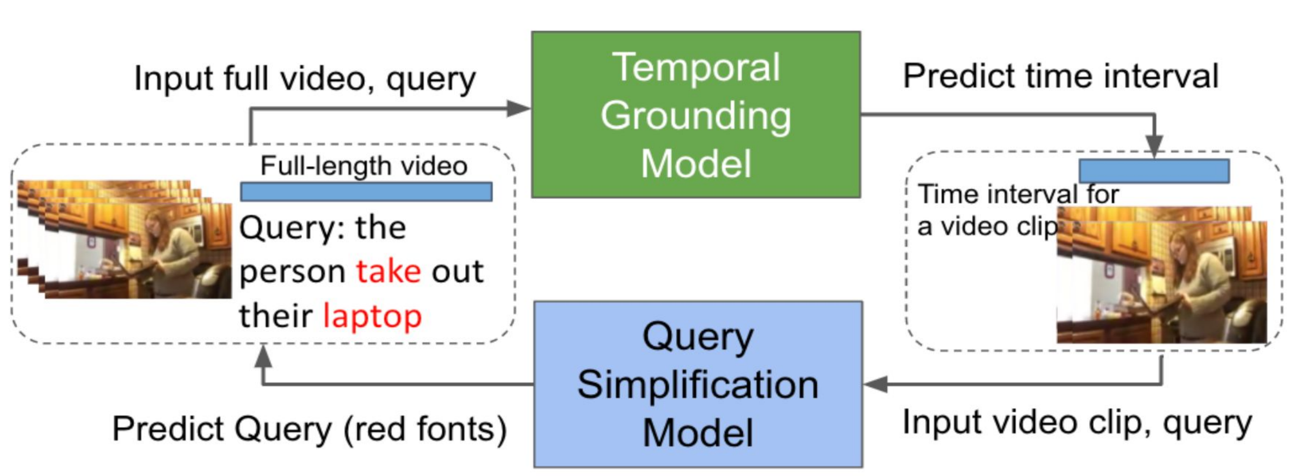
- Utilizes a closed-loop system, borrowing the idea of feedback-error-learning (FEL)
- Adapts a **video-pivoted query simplification task**
- Pairs a state-of-the-art temporal grounding model (LGI) with a video machine translation model

INTRODUCTION

- **Strongly Supervised Learning for Temporal Grounding (LGI -- Mun, Cho, Han, 2020)**



- **Visual Pivoted Translation**
  - Provide more fine-grained semantic discrepancy between video features and text
- **EVOQUER: Enhancing Temporal Grounding with VideO-Pivoted BackQUERy Generation**

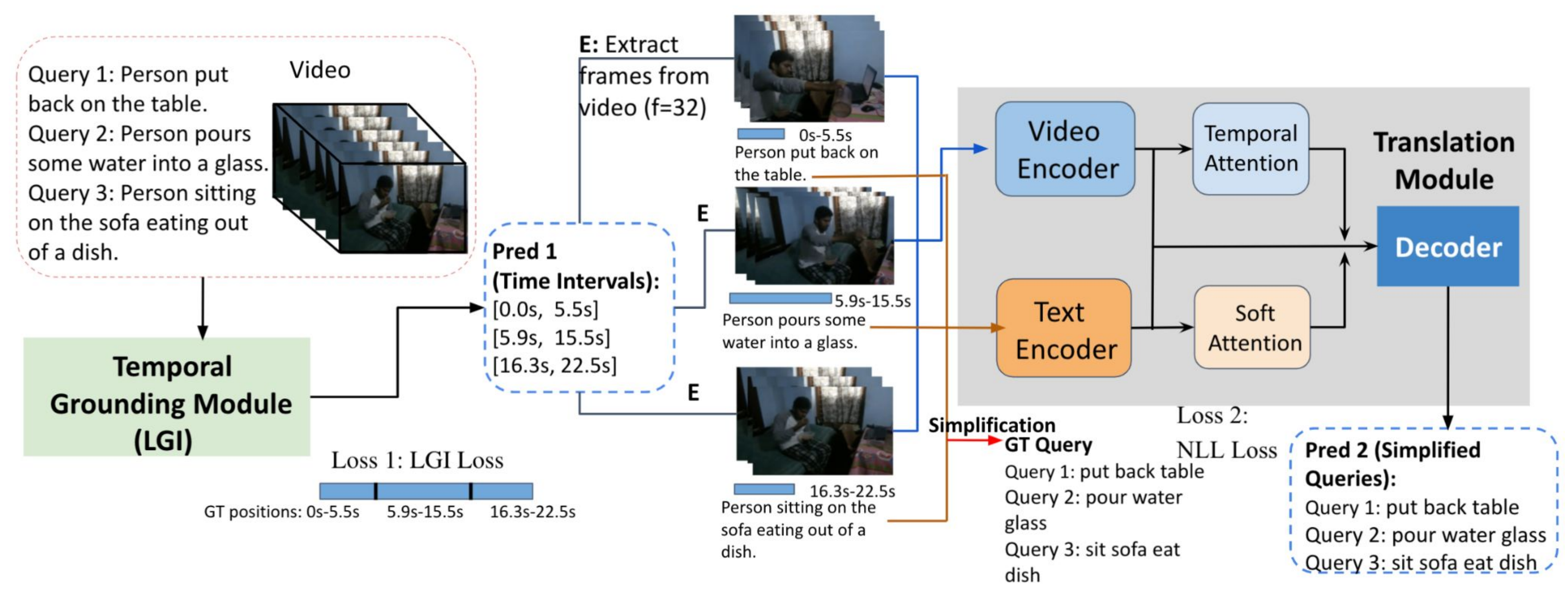


EVOQUER -- METRICS

- **R@tIoU:** The recall at different thresholds between prediction and ground truth. Thresholds were set as 0.3, 0.5, and 0.7.
- **mIoU**: The average of temporal interval recall from all threshold levels.
- **Jaccard similarity**: The intersection over union between prediction and ground truth.
- **BLEU:** A standard evaluation metric for machine translation that measures n-gram word overlap.

EVOQUER FRAMEWORK

- **(LGI) Input:** A set of queries and their full-length video clips
- **(LGI) Output:** Prediction of time intervals for the queries
- **(Translation) Input:** Queries and 32 frames video features trimmed by the predicted time intervals
- **(Translation) Output:** Prediction of simplified queries with only verbs and nouns



- **LGI Loss:** Loss for predicting time intervals using LGI model
- **NLL Loss:** Loss for simplifying queries using VMT translation framework
- Extract verbs and nouns as simplified version of queries
- Combine NLL loss with LGI loss to update the networks
- **Optional Setting: VSE Loss (Faghri et.al 2017)**

Experiment with an alternative setting of the translation module: generate simplified queries from video input and apply VSE loss to enforce the mapping between video features and text features

DATASET -- CHARADES-STA

- Charades-STA is a widely used benchmark dataset for temporal grounding, consisting of 9,848 ~ 30 second videos.
- 27,848 text queries -- Maximum of 10 words per query
- Train/Valid/Test Split (%) -- 50/25/25 respectively

TEMPORAL GROUNDING RESULTS

- **Performance on Charades-STA test set (LGI model and EVOQUER MODEL)**

Model	R@0.3	R@0.5	R@0.7	mIoU
LGI	71.54	**58.08**	34.68	50.28
EVOQUER	**71.57**	57.81	**35.73**	**50.48**
EVOQUER +VSE	70.46	57.81	35.51	50.16

Table 1: Results on Charades-STA test set from the LGI model and two EVOQUER variants.

- **Translation quality measuring by Jaccard similarity, BLEU Unigram (BLEU1) and Bigram (BLEU2)**

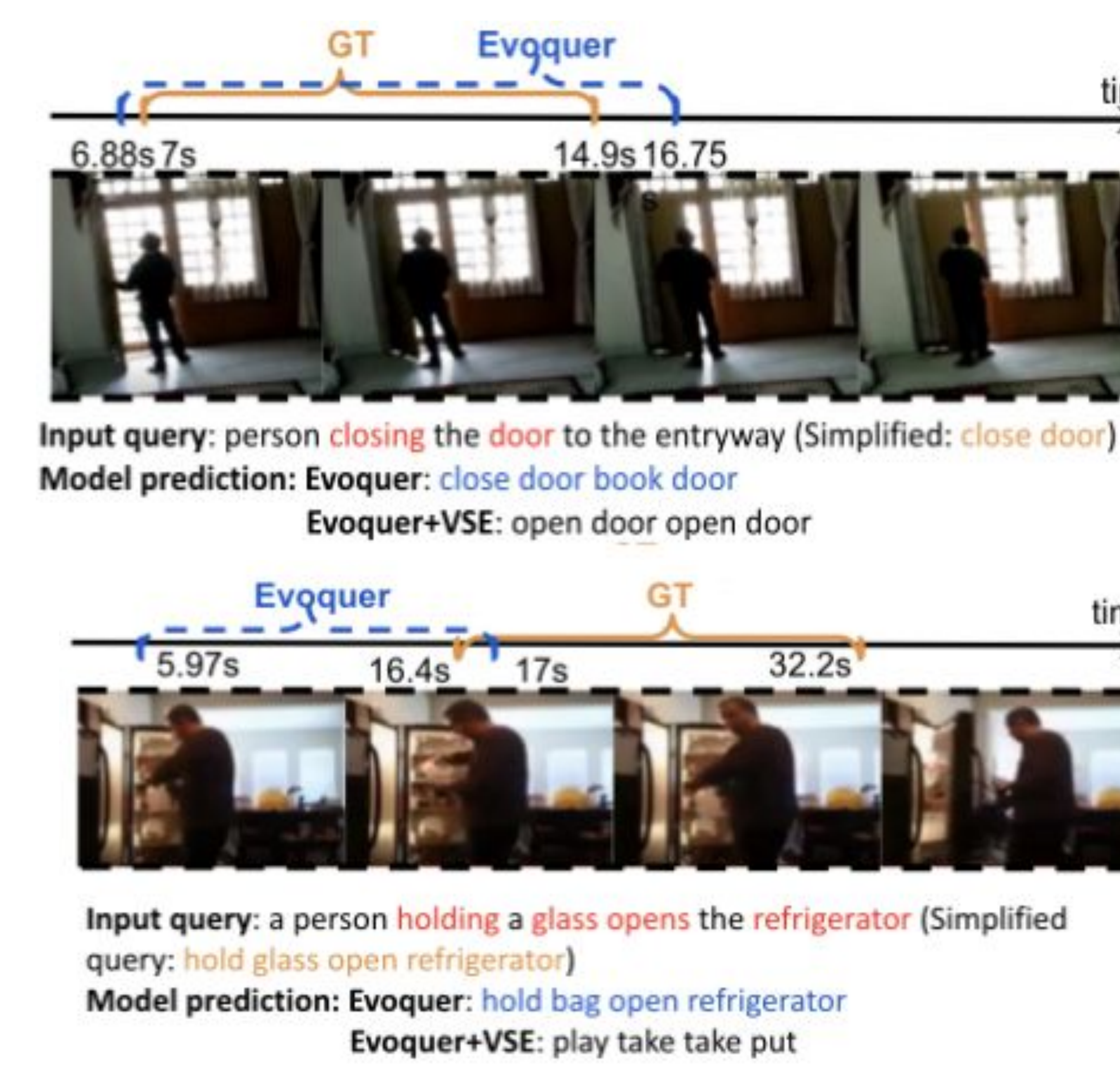
Model	JaccSim	BLEU1	BLEU2
EVOQUER	**51.98**	**53.04**	**42.47**
EVOQUER +VSE	6.37	7.96	1.20

RESULTS ANALYSIS

- **Statistics of samples showing improvements and drops for EVOQUER model compared with LGI model**

	Both >= R@0.3			Both
	EVOQUER ↑	EVOQUER ↓	Same	<R@0.3
Cnt.	441	362	1347	777

Table 2: Counts of samples that are scored by R@tIoU with four categories from comparison between EVOQUER and LGI model. Three of the categories are from samples where both models achieve recall equal and above threshold 0.3: samples that are improved (EVOQUER ↑), samples with performance drops (EVOQUER ↓), and equal performance with at least R@0.3 (Same). The fourth category is when both perform below R@0.3 (Both <R@0.3).



CURRENT AND FUTURE WORK

Current:

- On the left, EVOQUER successfully predicts time interval and simplified queries as ground truth.
- On the right, EVOQUER predicts **hold bag** rather than **hold glass** due to their similar shape and size.

Future:

- Extend EVOQUER on a multitude of other temporal grounding datasets such as ActivityNet, MSRVTT, DiDeMo.
- Explore other parameter settings to improve performance such varied learning rate, number of epochs, batch size, etc.

REFERENCES

- Mun, Jonghwan, Minsu Cho, and Bohyung Han. "Local-global video-text interactions for temporal grounding." *CVPR*. 2020.
- Faghri, Fartash, et al. "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives."