VLTinT: Visual-Linguistic Transformer-in-Transformer for Coherent Video Paragraph Captioning

K. Yamazaki, K. Vo, S. Truong, B. Raj, N. Le
{kyamazak, khoavoho, sangt, thile}@uark.edu, bhiksha@cs.cmu.edu

SCAN ME

Problem

Video paragraph captioning (VPC) aims to generate a descriptive multi-sentence caption of an untrimmed video given temporal boundaries. This is a challenging task due to three aspects.

1. **Representation**: video has spatial and temporal dimensions, which makes harder than image understanding.
2. **Coherence**: sentence descriptions of events should be logically connected to one another.
3. **Alignment**: visual stimuli should be linked to its text description.

Contributions

We formulate the VPC task with video encoder and caption decoder. Based on the inspection of the previous works, we proposed:

1. **Novel video representation** based on vision and language features and their interactions.
2. **Novel Transformer-in-Transformer** design to simultaneously model intra- and inter-event dependencies in an end-to-end fashion producing a coherent paragraph.
3. **VL contrastive loss** function to better align both visual and linguistic information.

Results

v_PAGuZzrzSO4

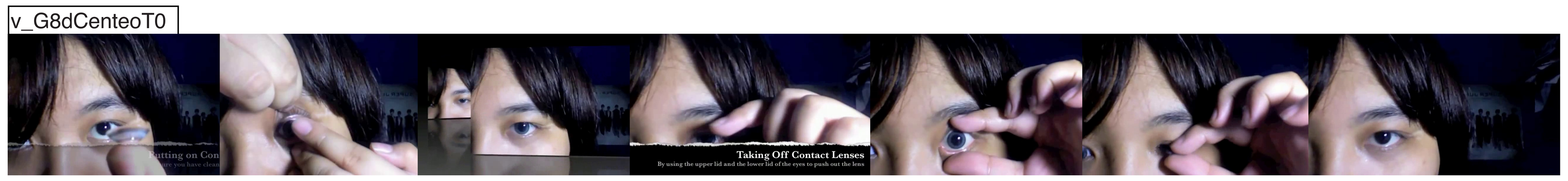
VTrans: A man is riding a horse down a river. The man then gets up and throws the calf down and grabs the horse and runs back to the horse. He gets back on his horse and gets back on his horse .

MART: A man is seen standing on a horse and throws a rope around. The man throws the calf down and the man chases after it. He ties the calf up and walks back to the horse .

VLTinT: A man is riding a horse in a rodeo ring. He lassos a calf. He ties the calf up and ties it up.

GT: A cowboy is riding a horse in a barn. He lassos a small calf. He dismounts, tying the calf and celebrating.

v_G8dCenteoT0



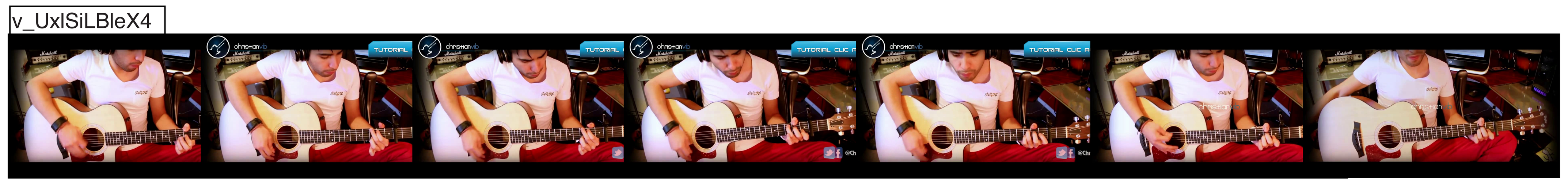
VTrans: The person then puts eye on the contact lens. The woman puts the contact lens in her eye. The person puts a contact lens in the eye.

MART: A woman is seen looking at the camera. She holds up a contact lens and puts it in her eye. She then puts the contact into the camera.

VLTinT: A close up of a eye is shown with a person's eye. A person is then seen putting a contact lens in her eye. The person then takes a contact lens out of her eye.

GT: A woman holds a contact lens on her finger. She puts the contact lens into her eye. She opens her eye with her fingers and takes the contact lens out.

v_UxlSiLBleX4



VTrans: A man is playing a guitar. He is playing the guitar. He stops playing the guitar .

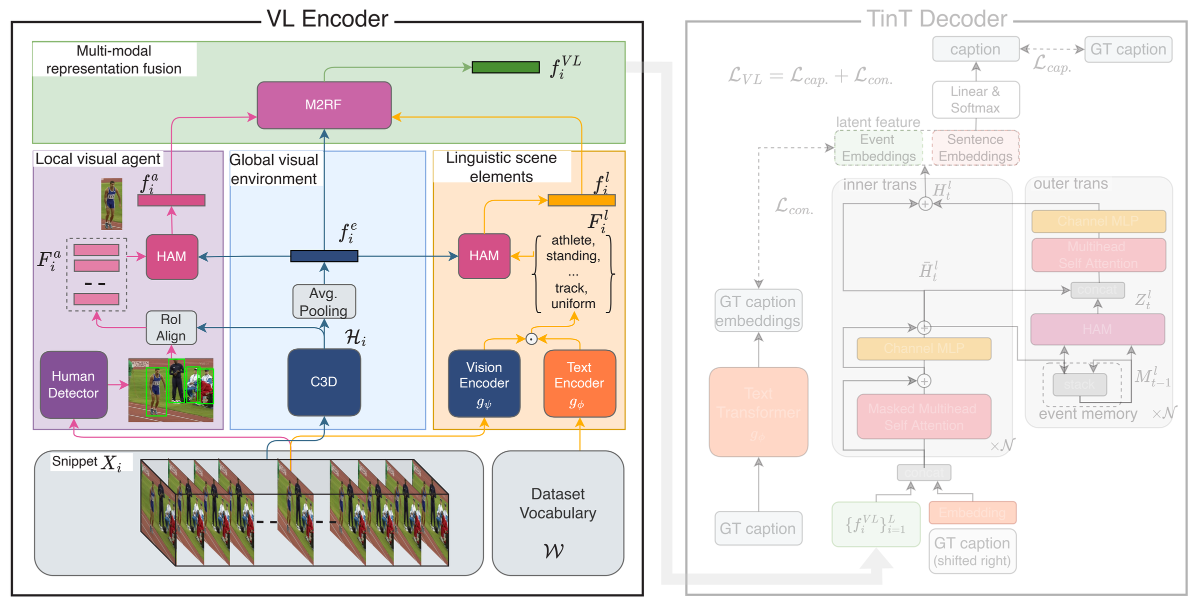
MART: A man is seen sitting on a stool holding a guitar and playing a guitar. The man continues playing the guitar while the camera captures his movements. The man finishes the song and smiles .

VLTinT: A man is sitting down playing an acoustic guitar. He is playing the guitar. He finishes playing the guitar and smiles .

GT: A man is sitting down in a chair. He begins to play an acoustic guitar. He finishes playing the guitar and standing up.

Qualitative comparison on ActivityNet Captions between our VLTinT and baselines. **Red text** indicates the captioning mistakes, **purple text** indicates repetitive patterns, and **blue text** indicates some distinct expressions. Overall, VLTinT can generate more descriptive captions with fine-grained details. Compared to baselines, which prone to use high-frequency words for their caption, VLTinT can use expressive but less frequently appearing words, e.g., "guitar" vs. "acostic guitart" in the example.

VL Encoder

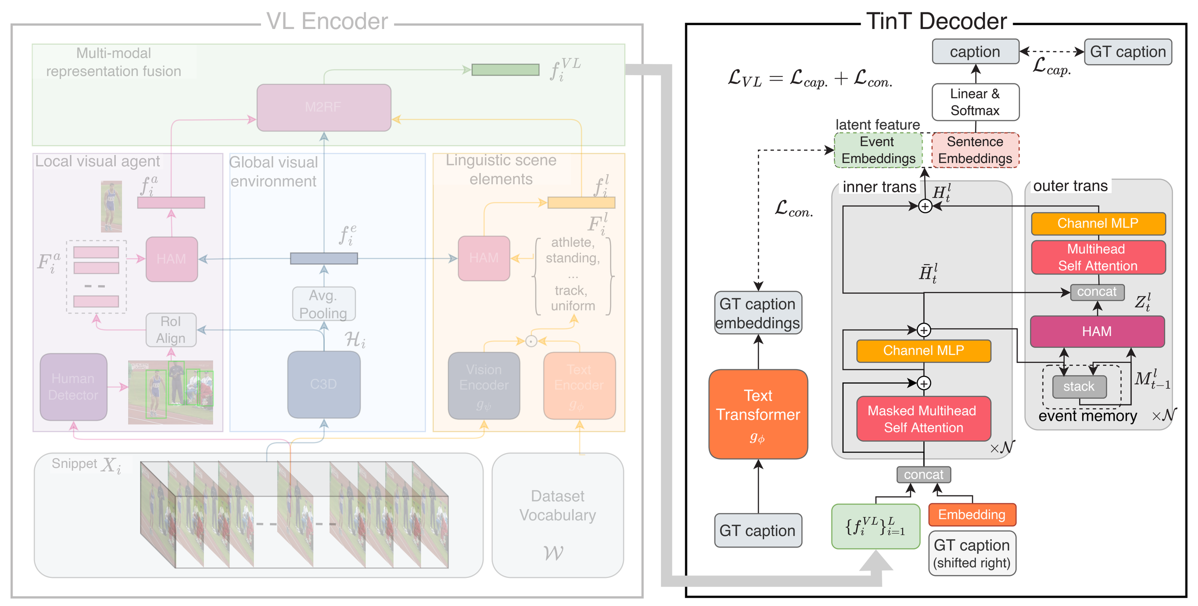


VL Encoder: We modeled the scene with three modalities and their interactions:

- **global visual environment**: provides the visual semantic information from the entire spatial scene.
- **local visual main agents**: provides the visual features of the main human agents, who actually contribute to the formation of the event.
- **relevant linguistic scene elements**: provides additional contextual details of the scene as text-based feature.

Our Multi-modal Representation Fusion (M2RF) module is used to model the interaction of the multiple modalities and generate a representation for an event in the video.

TinT Decoder

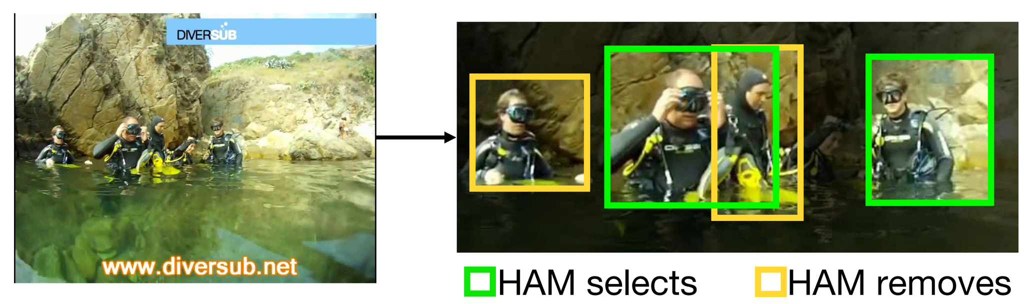


TinT Decoder: We enhance the unified encoder-decoder transformer with the autoregressive outer transformer to better model inter-sentence relationships and produce coherent paragraph caption.

- **inner transformer**: taking video features and textual tokens, it produces the sentence description of the event.
- **outer transformer**: stores the internal video and textual embeddings of the inner-transformer and selectively utilizes them according to the current input.

VL loss: The loss function for our model consists of a captioning loss and a contrastive loss. The contrastive loss helps ensure the alignment of the event embedding and the ground truth caption.

Relative Feature Selection



We use Hybrid Attention Module (HAM) to select salient features from list of features. Above shows the example that we can eliminate trivial agents while keeping the key agents who actually commit the action in the scene. We applied HAM for local visual agent features, linguistic scene elements, and internal embedding of TinT.

A Future Direction

Future investigations might include further examining linguistic feature in video understanding and exploring the VL Encoder in other video analysis problems. Further application of TinT Decoder in sequential modeling is also an important direction for the future research.

Source Code

The source code of this work is available on our github:
`https://github.com/UARK-AICV/VLTinT`