Accurate and Fast Compressed Video Captioning

Yaojie Shen*, Xin Gu*, Kai Xu, Heng Fan, Longyin Wen, Libo Zhang†

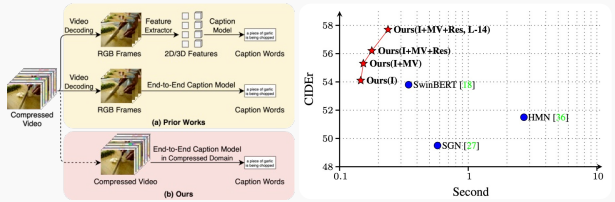
Overview

Main idea

- Minimizing redundant information processing, which is common in traditional methods
- Utilizing I-frames, motion vectors, and residuals in compressed videos to improve the speed of video captioning

Our Approach

- We propose a new, innovative method for video captioning operating directly in the compressed domain
- Using distinct features of compressed video, our method learns from the entire video
- Our efficient end-to-end transformer processes less redundant information, improving inference and learning directly from compressed video

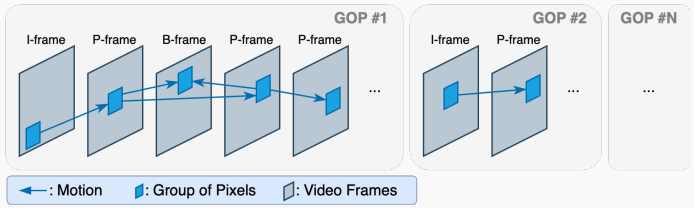


Results

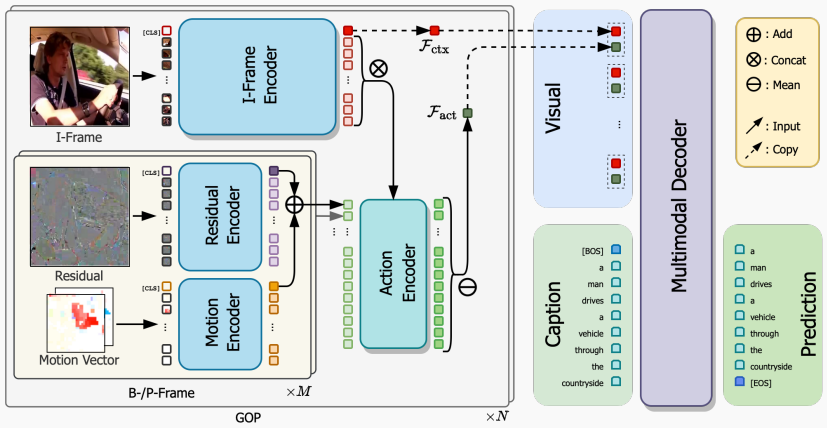
- Our approach significantly enhances the efficiency and performance of caption generation, leading to more accurate and relevant captions
- Our method achieves competitive performance on three datasets, and nearly **2×** faster than the fastest existing SOTA

Architecture

- The GOP structure of encoded video: each GOP starts with an I-frame, followed by several B-frames or P-frames



- We follow the structure of GOP in compressed domain to design our model. Our model takes the I-frames, and motion vectors and residuals of B-/P-frames as inputs



- The architecture of our proposed Compressed Video Captioner:
  - Left: Compressed Video Transformer extracts video representations for each GOP using a large visual backbone for I-frames and two small Vision Transformers for motion vectors and residuals. Features are fused with an action encoder
  - Right: Multimodal Decoder with causal masks is utilized for learning captions

Main Results

Video Captioning on MSVD, MSRVTT and VATEX

Method	MSVD (BLUE@4/METEOR/ROUGE/CIDEr)	MSRVTT (BLUE@4/METEOR/ROUGE/CIDEr)
SGN	52.8 35.5 72.9 94.3	40.8 28.3 60.8 49.5
HMN	59.2 37.7 75.1 104.0	43.5 29.0 62.7 51.5
SwinBERT	58.2 41.3 77.5 120.6	41.9 29.9 62.1 53.8
Ours	55.9 39.9 76.8 113.0	43.1 29.8 62.7 56.2
Ours (ViT/L-14)	**60.1 41.4 78.2 121.5**	**44.4 30.3 63.4 57.2**

Method	VATEX (BLUE@4/METEOR/ROUGE/CIDEr)
ORG-TRL	32.1 22.2 48.9 49.7
SwinBERT	**38.7 26.2 53.2 73.0**
Ours	31.4 23.2 49.4 52.7
Ours (ViT/L-14)	35.8 25.3 52.0 64.8

Inference Speed

Method	Data	Inference Time ↓	CIDEr ↑
SGN	RGB Video Frames	2818 ms	51.5
HMN	RGB Video Frames	339 ms	49.5
SwinBERT	RGB Video Frames	339 ms	53.8
Ours	I-frame+MV+Res	**178 ms**	**56.2**

Visualization