LongVLM: Efficient Long Video Understanding via Large Language Models(ECCV2024)

- https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/04936.pdf

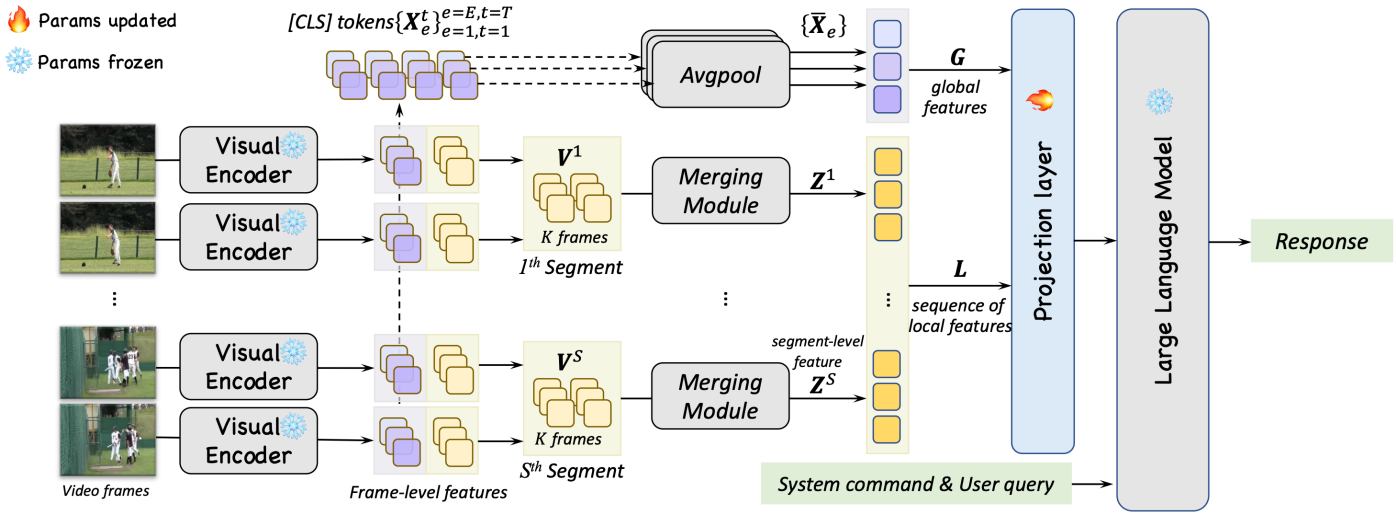
概要

- VideoLLMの問題点としてlocal featureを見落としていることがある
- そこでlocal feature, global featureをそれぞれ計算して最後にmergeするようなアーキテクチャを考えた

related work

- Video LLM
  - VideoChatGPT
  - Valley
  - VideoChat
  - Video-LLaMA
  - Video-ChatCaptioner
  - MovieChat
- long term video processing
  - A.: Temporal alignment networks for long-term video(CVPR2022)
  - Revisiting the" video" in video-language understanding(CVPR2022)

LongVLM



(元論文より)

- 全体図はVisual Encoder, projection layer, LLMの3層から構成
- 学習するのはProjecton Layerのみ
- Visual EncoderはCLIP-ViT-LのImage Encoderのこと
- 全体の流れは以下
  - 動画をフレームごとに分割してImage Encoderに入れて特徴量を2つ得る(詳しくは後述)
  - 得られた特徴量の片方(図の紫)を全フレーム間で足してglobal featureを得る
  - 得られた特徴量の他方(図の黄色)を数フレームから成るセグメント間で足してセグメントごとに特徴量を計算して、セグメントごとの特徴量をくっつけてlocal featureを得る
  - global feature, local featureをprojection layerに入れて最終的なfeatureを得る
  - LLMに入れる

visual encoderの詳細

- 入力動画 : $\mathcal{V} \in \mathbb{R}^{T \times H \times W \times 3}$
- これをencoderに入れると$\{X^t, P^t\}_{t=1}^{T}$を得る
- ここで、$P^t \in \mathbb{R}^{N \times d}$は画像のパッチごとの特徴量で$N$がパッチトークンの個数
- $X^t \in \mathbb{R}^{E \times d}$は[CLS]tokenで$E$は選択されたencoder layerの個数

local feature

- S個のsegmentに動画を分割して、各segmentはKフレームとする(つまり$SK = T$)
- このときs番目のセグメントの特徴量はpatch featureを集めて$V^s = \{P^t\}_{t=(s-1)K}^{sK}$
- これからセグメントの特徴量$Z^s = g(V^s)$を得る

- Z^sをすべてconcatenateしてlocal feature Lを得る

global feature

- X^t, 図の紫部分
- これをすべて足して平均を取ることでglobal feature$G$を得る

英語

- revolutionize : 革命を起こす
- corpora : corpusの複数形