Self-Adaptive Sampling for Efficient Video Question-Answering on Image--Text Models

Wei Han, Hui Chen, Min-Yen Kan, Soujanya Poria
Singapore University of Technology & Design, Singapore
National University of Singapore, Singapore

Introduction

• Video Question Answering (ViQA): Given a short video, answer the question based on the video

• Image—Text Models (ITMs): a subclass of visual language models (VLMs) that accept image sequences and text as input and generate text outputs, such as CLIP. To process video input, a series of frames must be first sampled from that video.

Related Works

• Current Sampling Strategy
  • Learning-free sampling is cost-effective but hard to reach optimal
  • Learning-based sampling can adapt to different question input, but requires additional computational cost (huge) and difficult to converge

Research Question

• Can we move the sampling stage offline (decouple it from the main network)?
• Can we find a simple yet effective formulation for the offline sampling?
• Is question-aware sampling always required (can we design a question-agnostic one)?

Method

• Most Implied Frames (MIF)
  • A captioner and a scorer to calculate scores for each frame
  • Choose the frames of the highest scores as sampled ones

• Most Dominant Frames (MDF)
  • Based on previous analysis, we can move one step forward by even discarding the question-aware component
  • Sampling scores are calculated on visual feature similarity

Results

• Both MIF and MDF achieve good performance
• MDF is competitive to MDF, showing that question-aware sampling is not necessary