
Zuxuan Wu¹, Caiming Xiong², Yu-Gang Jiang³, Larry S. Davis¹
¹University of Maryland, ²Salesforce Research, ³Fudan University

Overview

Video classification framework: averages scores from 25 uniformly sampled frames.

Scores Mean-pooling

Do we really need the same computation budget for all videos?

How to average scores in an online setting?

Key Observation: (1) Different video clips have different computational requirements. (2) 74% FLOPs can be saved when testing with 112*112 rather than 224*224.

Our Idea: Use economical features by default, and learn when to compute powerful features

Method

Goal: Learn video-specific fine-feature usage policies

We use a coarse-to-fine framework, with two LSTMs interacting with each other. In particular, it consists of:

- a coarse LSTM, operating on features computed at a coarse scale with lightweight CNNs.
- a conditional gating module, to decide whether to examine the incoming video frame more carefully to obtain finer details
- a fine LSTM, aggregating powerful features from a computationally expensive CNN model

Trained in a fully-differential manner using with Gumbel-Softmax to learn the gating function.

Learned model is suitable for dynamic inference in both offline and online settings!

Results

Offline settings: 51.8% and 51.3% computational savings on average

Online settings: best trade-off: computational cost and accuracy