Long-Term Feature Banks for Detailed Video Understanding

Abstract

We propose a long-term feature bank–supportive information extracted over the entire span of a video–to augment state-of-the-art video models that otherwise would only view short clips of 2-5 seconds.
Augmenting 3D convolutional networks with a long-term feature back yields state-of-the-art results.

1. Introduction

Without the ability to use the past to understand the present, we, as human observers, would not understand what we are watching.

In this paper, we propose the idea of a *long-term feature bank* that stores a rich, time-indexed representation of the entire movie.

The long-term feature bank is inspired by works that leverage long-range temporal information by using precomputed visual features.

2. Related Work

- **Deep networks** are the dominant approach for video understanding. *Two-stream networks & 3D convolutional networks*
- **Temporal and relationship models** include RNNs that model the evolution of video frames and multilayer perceptrons that model ordered frame features.
- **Long-term video understanding**
  - One strategy to overcome these constraints is to use precomputed features without end-to-end training.
  - Another strategy is to use aggressive subsampling or large striding.
  - To our knowledge, our approach is the first that enjoys the best of three worlds: end-to-end learning for strong short-term features with dense sampling and decoupled, flexible long-term modeling.
- **Spatio-temporal action localization** is an active research area.
- **Information 'bank'** representations have been used as image-level representations, for video indexing and retrieval, and for modeling information in text corpora.

Long-Term Feature Bank Models

> The ability to relate what is happening in the present to events that are distant in time.

3.1 Method Overview

> The authors' method can be used for the task of *spatio-temporal action localization*, where the goal is to detect all actors in a video and classify their actions.
> Most state-of-the-art methods combine a 'backbone' 3D CNN with a *region-based person detector*. To process a video, it is split into *short* clips of 2-5 seconds, which are *independently* forwarded through the 3D CNN to compute a feature map, which is then used with region proposals and region of interest

(RoI) pooling to compute RoI features for each candidate actor. This approach captures only short-term information.

The central idea in this method is to extend this approach with two new concepts:

1. A long-term feature bank that intuitively acts as a ‘memory’ of what happened during the entire video – we compute this as RoI features from detections at regularly sampled time steps.
2. A feature bank operator (FBO) that computes interactions between the short-term RoI features (describing what actors are doing now) and the long-term features.

3.2 Long-Term Feature Bank

The goal of the long-term feature bank, *L*, is to provide relevant contextual information to aid recognition at the current time step.
For the task of spatio-temporal action localization, we run a person detector over the entire video to generate a set of detections for each frame.
In parallel, we run a standard clip-based 3D CNN over the video at regularly spaced intervals, such as every one second.
We then use RoI pooling to extract features for all person detections at each time-step processed by 3D CNN.
Intuitively, *L* provides information about when and what all actors are doing *in the whole video* and it can be efficiently computed in a single pass over the video by the detector and 3D CNN.

3.3 Feature Bank Operator

Our model references information from the long-term features *L* via a *feature bank operator* (FBO).

??? What is L centered at the current clip?

> Intuitively, the feature bank operator (FBO) computes an updated version of the pooled short-term features S_t by relating them to the long-term features.
>
> **Batch vs. Casual**

3.4 Implementation Details

> **Backbone**, a standard 3D CNN architecture from recent video classification work. **ResNet50**, **I3D technique**
> **RoI Pooling**.
> **Feature Bank Operator Instantiations**. The feature bank operator can be implemented in a variety of ways.
>
> - LFB NL: out default bank operator is an attention operator. Intuitively, we use S_t to attend to features in L_t, and add the attended information back to S_t via a shortcut connection.
> - LFP Avg/Max:
>
> **Training**, we treat 3D CNN and detector that are used to compute *L* as *fixed* components that are trained offline, but still on the target dataset, and not updated subsequently.
> **A Baseline Short-Term Operator** To validate the benefit of incorporating long-term information, we also study a 'degraded' version of our model that does not use a long-term feature bank.

4. Experiments on AVA

4.1 Implementation Details

Person Detector We use Faster R-CNN with a ResNeXt-101-FPN backbone for person detection.
Temporal Sampling
Training. We train our models using synchronous SGD with a minibatch size of 16 clips on 8 GPUs (i.e. 2 clips per GPU), with batch normalization layers frozen.
Inference

4.2 Ablation Experiments

Temporal Support
Feature Decoupling
FBO Function Design
FBO Input Design
Non-Local Block Design
Model Complexity
Example Predictions
Backbone and Testing
Comparison to Prior Work

5. Experiments on EPIC-Kitchens

5.1 Impementation details

Long-Term Feature Banks
Adaption to Segment-Level Tasks

5.2 Quantitative Evaluation

6. Experiments on Charades

6.1 Impementation details

> **Training and Inference**

6.2 Quantitative Evaluation

7. Discussion

> In conclusion, we propose a **Long-Term Feature Bank** that provides long-term supportive information to video models. We show that enabling video models with access to long-term information, through an LVB, leads to a large performance gain and yields state-of-the-art results on challenging datasets like AVA, EPIC-Kitchens, and Charades.