What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Datasets

De-An Huang\textsuperscript{1}, Vignesh Ramanathan\textsuperscript{2}, Dhruv Mahajan\textsuperscript{2},
Lorenzo Torresani\textsuperscript{2,3}, Manohar Paluri\textsuperscript{2}, Li Fei-Fei\textsuperscript{1}, and Juan Carlos Niebles\textsuperscript{1}
\textsuperscript{1}Stanford University, \textsuperscript{2}Facebook, \textsuperscript{3}Dartmouth College
Problem Background

- The emphasis on temporal modeling is the main difference between videos and images.
- The scene and objects in a frame are almost sufficient for the tasks. (i.e. Action Recognition).

(a) knocking ball  
(b) Pushups

How important is the temporal information for the video tasks?
Problem Background

• If an existing model (i.e. C3D) trained on videos utilizes temporal information while classifying a new video?

• Naïve Approach: Repeat a single frame $n$ times to form a new clip.

  Result in almost 25% performance drop

  ◆ significantly alter the temporal distribution.
  ◆ potentially remove critical frames in the video that are important for recognizing the action.
Related Work: *C3D*

![Diagram](image)

Figure 1. **2D and 3D convolution operations.** a) Applying 2D convolution on an image results in an image. b) Applying 2D convolution on a video volume (multiple frames as multiple channels) also results in an image. c) Applying 3D convolution on a video volume results in another volume, preserving temporal information of the input signal.

Input: $16 \times 3 \times 112 \times 112$
Problem Formulation: *class-agnostic temporal generator*

- To hallucinate the motion from subsampled frames to compensate the temporal distribution.
Problem Formulation: *class-agnostic temporal generator*

- Train a temporal generator that utilize the spatial relations among sub-sampled frames to recover the information.

- Offsets the difference in temporal distribution between video and sub-sampled frames.
Problem Formulation: *motion-invariant frame selector*

\( \{X_i\} : \) A set of candidate frames

**Max Response:** frame is most confident about its prediction.

\[ i^* = \arg \max_i \phi(X_i), \quad \phi(X_i) = \max_c f_c(X_i) \]

i.e. \( f(X_i) = [0.1, 0.2, 0.3, 0.1, 0.1, 0.1, 0.1] \), \( \phi(X_i) = 0.3 \)

**Oracle:** remove “cheat” by looking ground truth
Experiment: Datasets and Setup

- **Datasets**
  - UCF101: 101 action categories, 13320 videos.
  - Kinetics: 400 action classes, 306245 videos.

- **Setup**
  - Train C3D model, temporal generator, frame selector on training set.
  - For test videos, randomly sample a 16-frame clip.
Experiment: Results

- Uniform: naively sub-sampling
- + Temporal Gen.Only: using uniform sampled frames as generator input
- Original Videos: Original accuracy

- Kinetics needs more temporal information. (5%-47%, 6%-79%)
- We do not need entire clip.
- Frame selection is important.
- Importance of temporal generator.
Experiment: Results

- Some classes use temporal information
Experiment: Results

- Perceptual loss
Experiment: Results

Temporal generator successfully offsets the temporal distribution difference on 77% of UCF101 classes and 75% of the Kinetics classes.
Experiment: Results
Conclusion

- Provide in-depth quantitative and qualitative analysis of the video model and dataset.
- The analysis framework is critical to design better models and collect better datasets.