Listen to Look: Action Recognition by Previewing Audio
Ruohan Gao$^{1,2}$, Tae-Hyun Oh$^{2}$, Kristen Grauman$^{1,2}$, Lorenzo Torresani$^2$
$^1$The University of Texas at Austin $^2$Facebook AI Research

Action Recognition in Untrimmed Video

- **Goal:** Efficient and accurate **clip-level** and **video-level** action recognition in untrimmed video
- **Our Idea:** Previewing Audio

We propose a framework for efficient action recognition in untrimmed video that uses audio as an **efficient preview** of the accompanying visual content at the **clip-level** and **video-level**.

**Motivation:**
- High temporal redundancy across clips
- Efficiency and accuracy

**Clip-Level Preview**
- Clip-level preview replaces the costly analysis of video clips with a more efficient processing of **image-audio pairs** through distillation.
- **Teacher model:** Knowledge distillation
- **Student model:** By processing only a single frame and the clip’s audio, we get an estimate of the expensive video descriptor for the full clip.

**Video-Level Preview**
- We iteratively predict where to “listen at” and “listen to” next to select the key moments for efficient video-level recognition.
- **Uniform across clips:** High temporal redundancy within a clip

Evaluation Results

**Datasets:**
- Kinetics-Sound (Arandjelovic & Zisserman 2017)
- UCF-101 (Soomro et al. 2012)
- ActivityNet (Heilbron et al. 2015)
- Mini-Sports1M (Karpathy et al. 2014, a subset of Sports1M)

**Clip-level preview results:**

**Video-level preview results:**

- Our approach strikes a favorable balance between accuracy and speed.
- We outperform all sota frame selection methods given the same computational budget.

**Qualitative results:** 5 uniformly selected moments and the first 5 visually useful moments selected by our method for two videos of actions throwing discus and rafting in ActivityNet.

**Project page:**
http://vision.cs.utexas.edu/projects/listen_to_look/
Code/Model are available!