LONG BEACH
CALIFORNIA
June 16-20, 2019

Grounding Spoken Words in Unlabeled Video

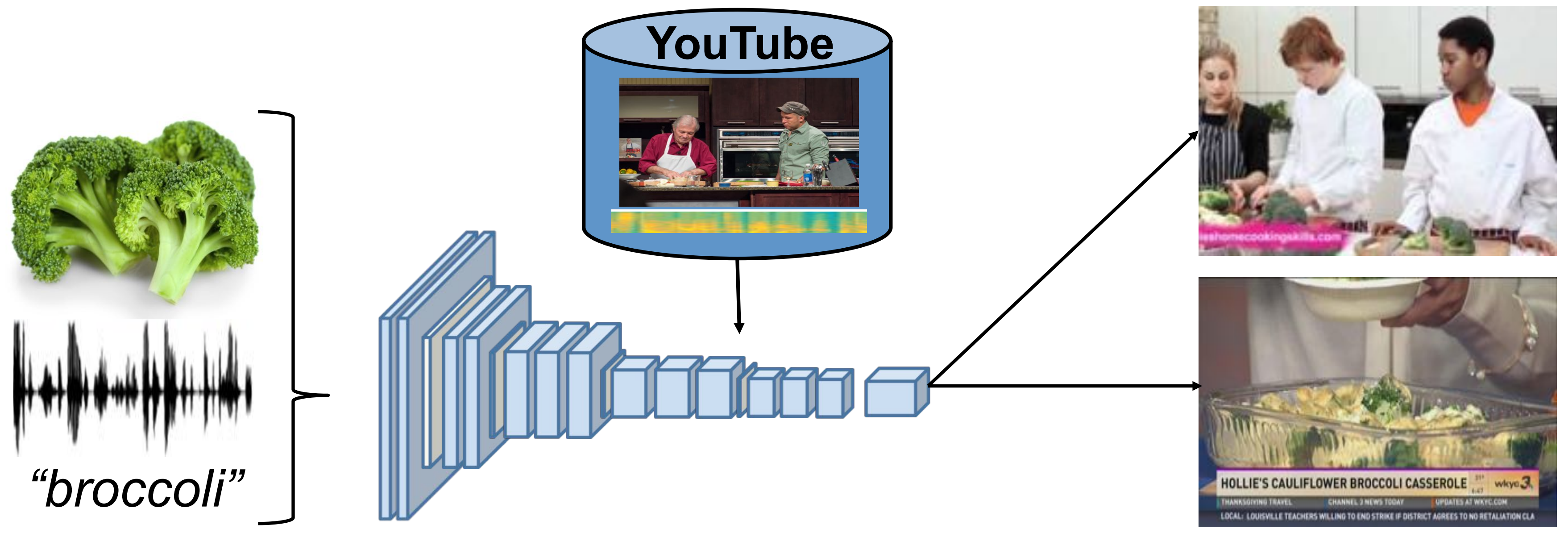
Angie Boggust, Kartik Audhkhasi, Dhiraj Joshi, David Harwath, Samuel Thomas, Rogerio Feris
Dan Gutfreund, Yang Zhang, Antonio Torralba, Michael Picheny, James Glass

PROJECT OVERVIEW

Goal: Leverage descriptive video data from cooking shows for self-supervised learning of objects & actions

Motivation: Learn spoken language and visual perception without labels/annotations just like human babies

Applications: Media indexing and search, visual object detection, scene understanding, etc.

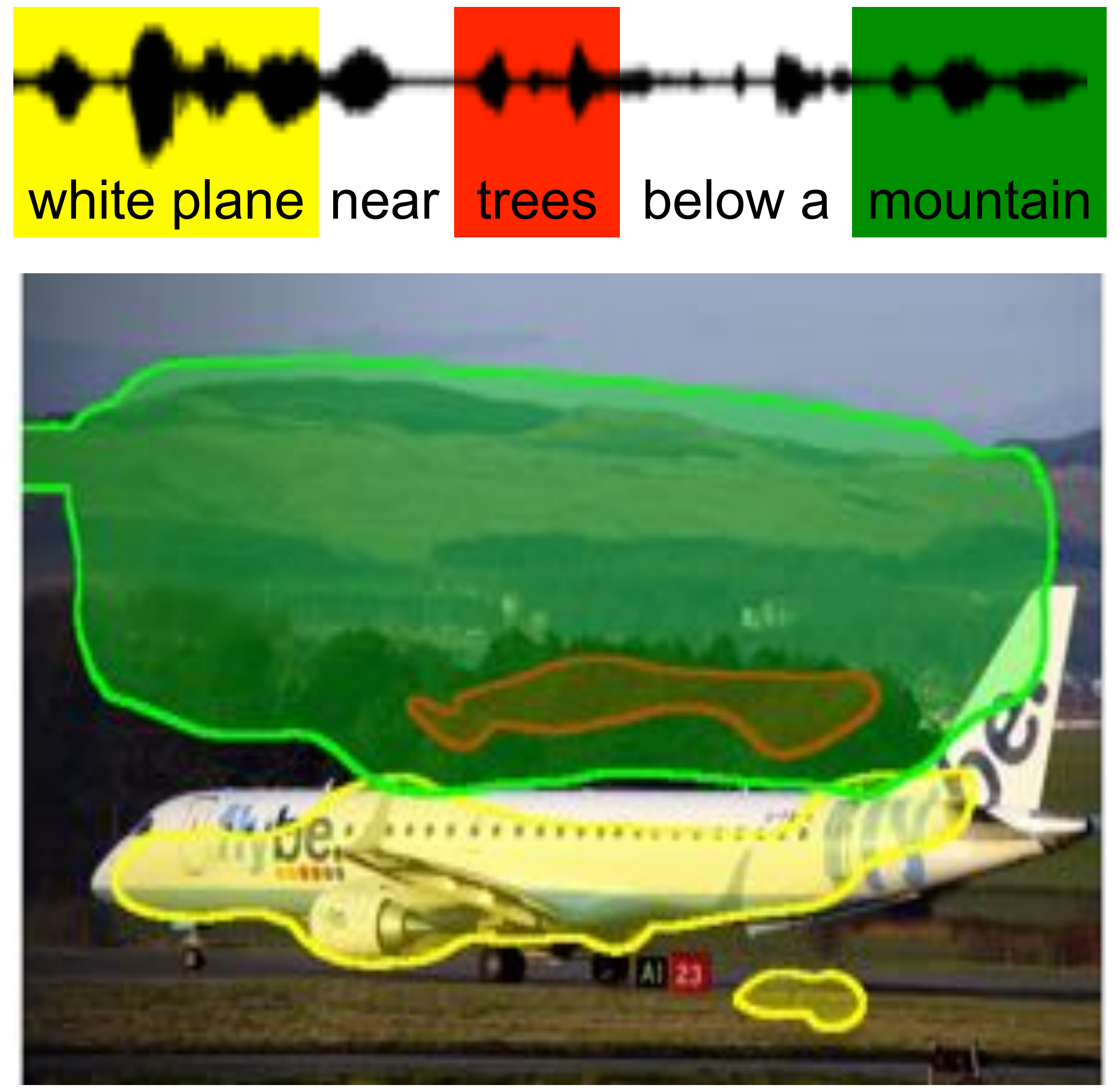


PRIOR WORK: DAVENET

Prior work introduces the DAVEnet architecture which learns to associate speech with images [1].

DAVEnet consists of two parallel convolutional branches which take in RGB images and log-Mel frequency spectrograms respectively and map them to a shared feature space.

We develop models that ground the visual and audio tracks of real world videos to one another.



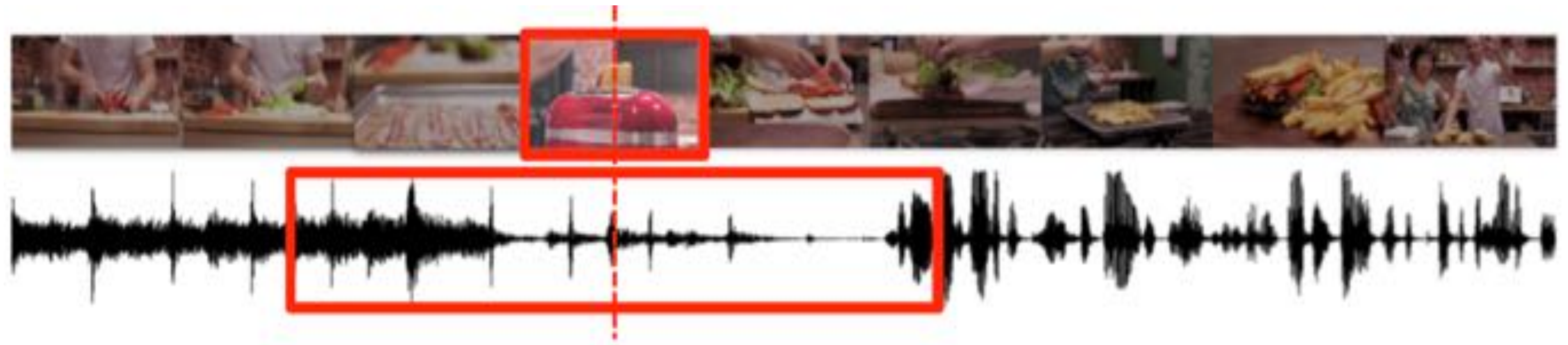
VIDEOS

Datasets: We use cooking show videos because they provide a natural example of aligned audio and visual content. We use videos from the YouCook2 and YouTube-8M datasets.

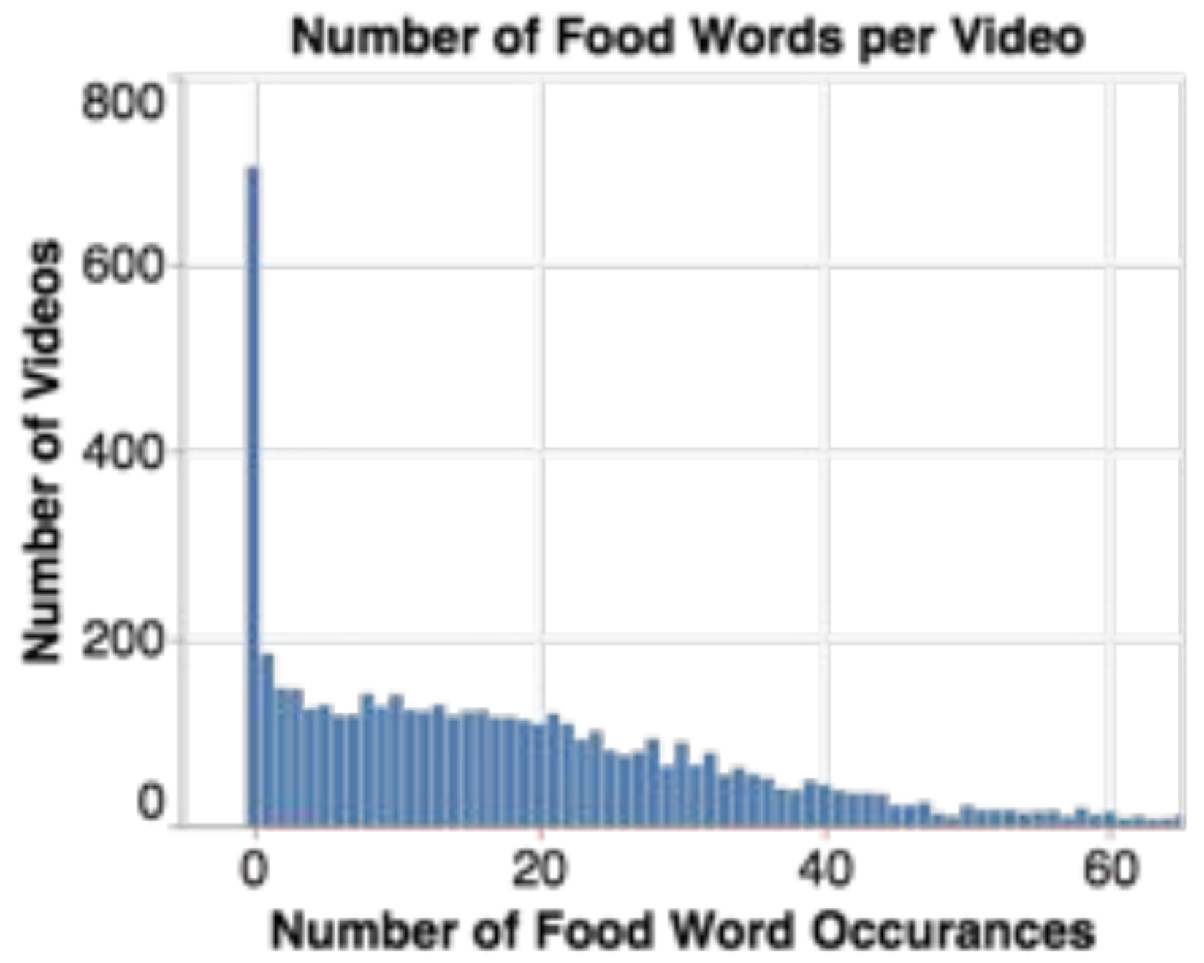
YouCook2: 2000 YouTube cooking show videos from across 89 recipe types [2].

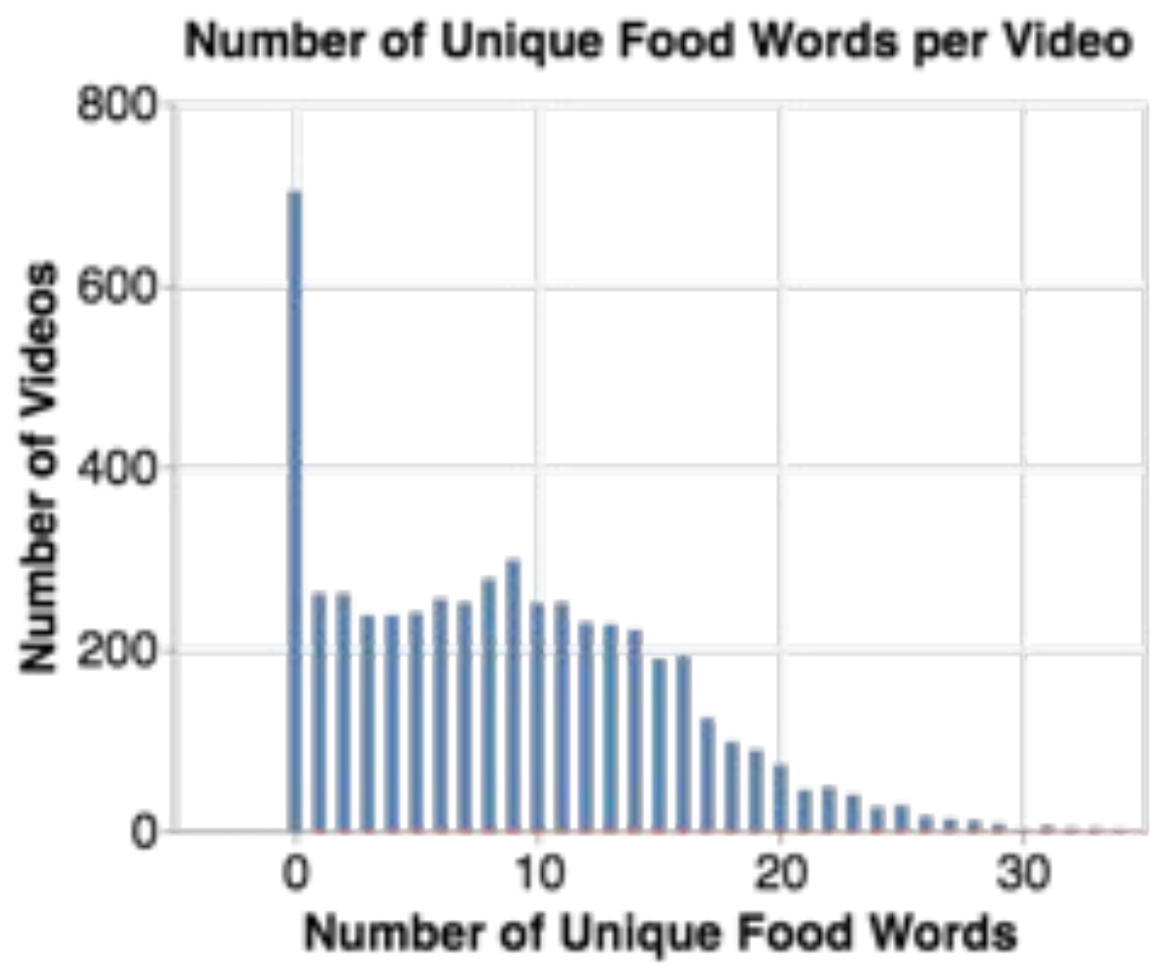
YouTube-8M: 3500 English-tagged YouTube videos from the *baking, cooking, cooking show, cuisine, dish,* and *food* categories [3].

Processing: We process each video into frame-audio pairs which can be used in the DAVEnet architecture. We extract frames at a rate of 1 frame-per-second and pair each frame with the 2 seconds of audio centered around it.



Analysis: 300 food nouns were manually selected from STT transcripts of the videos. Each video in our dataset contains on average 10.5 unique food words and 22 total food words. 16% of the time the food in the scene is referred to in the surrounding 20 second window.

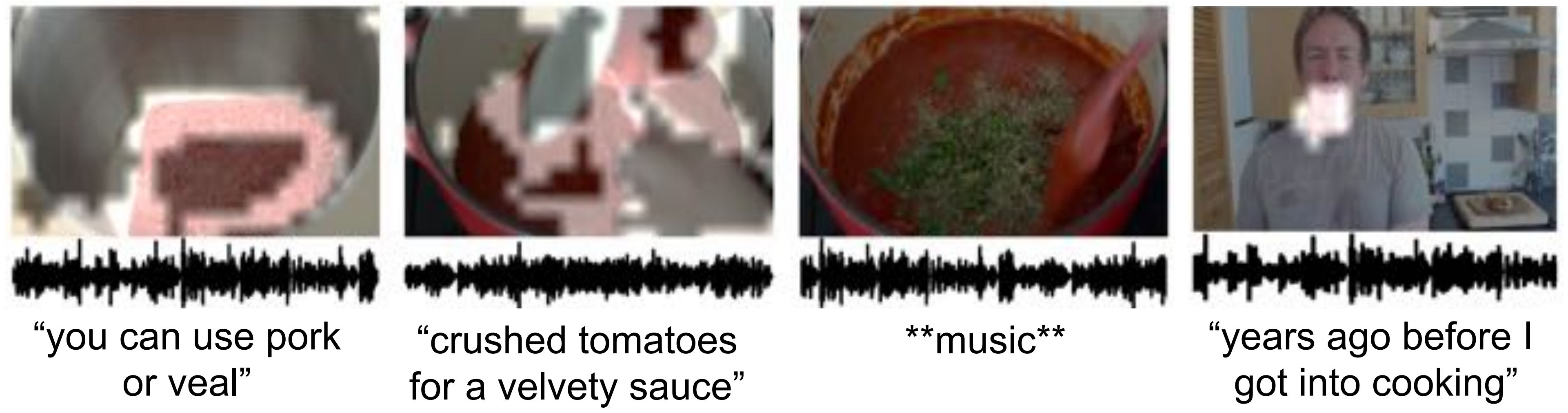




RESULTS

Unsupervised Learning: We train DAVEnet on 1M frame-audio pairs from YouCook2 and YouTube-8M. We evaluate on 1000 YouCook2 validation pairs that encapsulate a food word.

Audio Recall@10: 20.3% **Video Recall@10: 19.4%.**


“you can use pork or veal”
“crushed tomatoes for a velvety sauce”
music
“years ago before I got into cooking”

Semi-supervised Learning: Using a small labeled dataset containing food objects in the visual and audio channels as specified by both the IBM food concept detector and IBM Watson Speech-To-Text system, we fine tune the unsupervised model and increase performance.

Audio Recall@10: 27.2% **Video Recall@10: 23.3%.**

This suggests the ability to continue to improve our model’s cross-modal learning capabilities and provides a soft upper bound on performance.

SUMMARY

Key Take-Away: We set a benchmark for unsupervised cross-modal learning of audio-visual concepts from unannotated instructional video.

Future Work: guided frame-audio pair extraction, audio/visual alignment modeling, and utilization of a larger corpora of descriptive video

Relevant Papers:

1. D. Harwath, A. Recasens, D. Suris, G. Chuang, A. Torralba, and J. Glass, “Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input,” ECCV 2018
2. L. Zhou, C. Xu, and J. Corso, “Towards automatic learning of procedures from web instructional videos,” arXiv: 1703.09788, 2017
3. S. Abu-Al-Haija et al, “YouTube-8M: A Large-Scale Video Classification Benchmark,” arXiv:1609.08675, 2016