Compositional Event Detection Using Weak Supervision

Mark Cramer, Aasavari Kakne and Sundararajan Renganathan
{mdcramer, adkakne, rsundar} @stanford.edu

cs229
Machine Learning
December 12th, 2019

Abstract

Detecting domain-specific events in videos, such as commercials or interviews, is of great interest to video processing applications. The task is challenging as building models generally requires prohibitive amounts of hand-labelled data and computation. While human-generated labels are typically high quality, they're expensive to produce.

Rekall[1] provides a programming interface and data model for detecting domain-specific events in videos. Rekall queries combine the outputs of pre-trained models (e.g. object detection, facial recognition, etc.) in order to identify events, obviating the need to train new models.

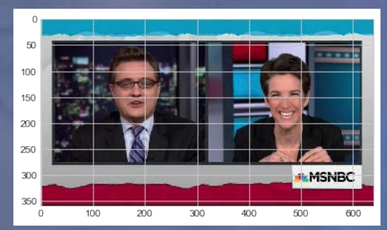
We use the outputs of Rekall queries as a form of weak supervision to train new models, reducing the need to build end-to-end models from scratch using laboriously hand-labeled data.

Dataset

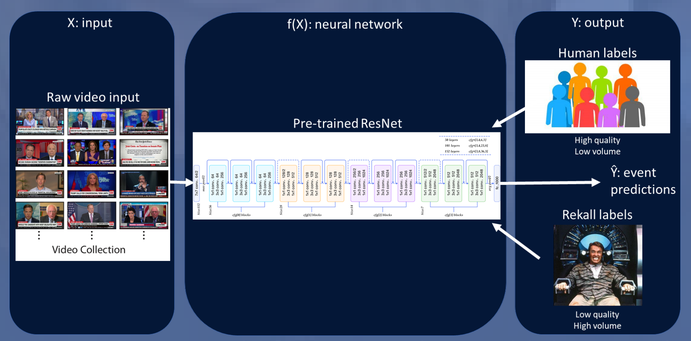
64,000 cable news video frames with ground truth and Rekall labels and an additional 280,000 frames with only Rekall labels

Task: Detecting Commercials

28.6% of video frames were true labeled as commercials. No sound, closed captions or other annotations were available.



Deep Learning approach: ResNet-50 Model

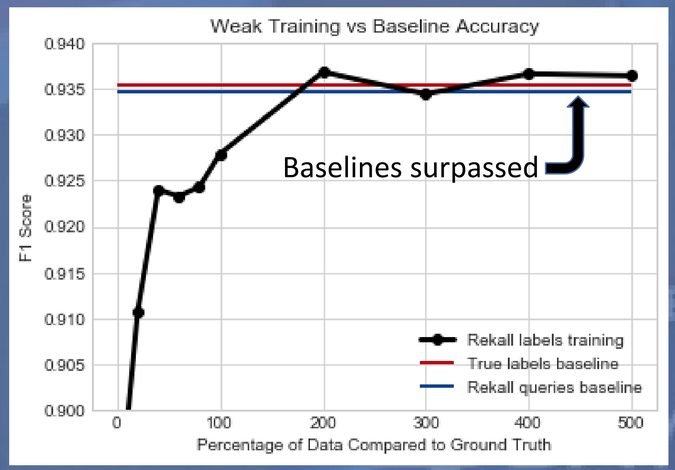


ResNet architecture with image inputs and training labels generated by humans versus Rekall

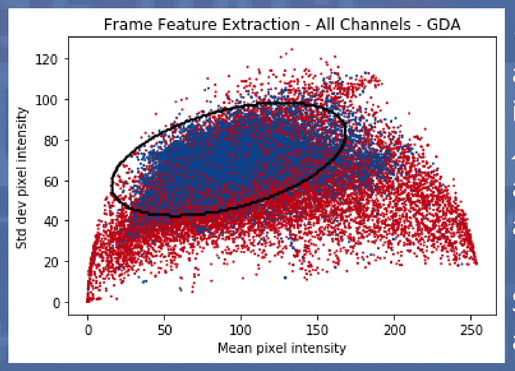
Setup:

- Began with ResNet-50 model pre-trained on ImageNet
- Fine-tuned all 25 million parameters in PyTorch
- Gradually increased the quantity of Rekall labels used for training
- Computed F1 scores as the figure of merit for the models
- Compared models trained using Rekall labels against the Rekall queries themselves and the model trained on ground truth labels

Finding: We determine that weak supervision data in the form of Rekall labels outperforms the models trained on ground truth data and the Rekall queries themselves. The graph is slightly noisy as we ran every experiment only once.

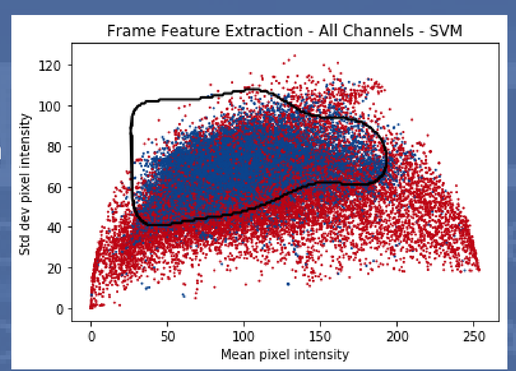


Classic ML: GDA versus SVM (RBF kernel)



← GDA 82% accuracy and 58% F1 with independent Σs and $X \in \mathbb{R}^8$ including mean and std dev for R, G and B channels

SVM 85% accuracy → and 70% F1 with same



Discussion and Future Work

Generalization to other tasks: Since the Rekall-labelled data outperformed the baselines for the commercial detection task, we may now expand to conversation detection and shot scale detection.

Fine-tuning final layer of ResNet: Adjusting weights throughout the entire ResNet caused training to be very slow, so we will explore results after freezing most of the layers. We may also compare performance using a smaller ResNet-18.

Reference

[1] D. Y. Fu, W. Crichton, J. Hong, X. Yao, H. Zhang, A. Truong, A. Narayan, M. Agrawala, C. Ré, and K. Fatahalian. Rekall: Specifying video events using compositions of spatiotemporal labels, 2019