Automated Visual Weak Supervision for Object Recognition in Videos

Swetha Revanur and Vishnu Sarukkai (srevanur, sarukkai) @ stanford.edu

Motivation

- As cameras become an increasingly ubiquitous part of modern life, videos offer a valuable source of data for a variety of applications. However, it is intractable for humans to manually observe most of this incoming content.
- It is crucial to be able to semantically parse the video, starting with entity classification.

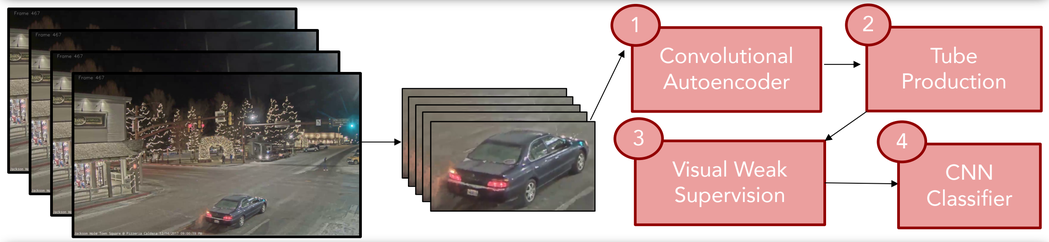
Weak Supervision and Related Works

- Use a variety of heuristics and a small number of labeled data points to predict labels for a large number of unlabeled points. These weakly-labeled points can then be added to a training set for a future classifier.
- Trading certainty in the labels for a larger training set.
- Though convolutional architectures have been used for video classification, they historically assume large training sets.
- Any weak supervision approaches rely on captions to generate latent probability distributions for objects in the image or create surrogate classes. However, all of these methods require manual heuristic generation.
- There is an existing automated weak supervision pipeline called Reef, however it was built for textual data.

Objectives

- Creating an automated pipeline for automating the generation of heuristics for visual weak supervision.
- Exploring the trade-off between the number of weakly- labeled data points used in training a subsequent classification model for videos.

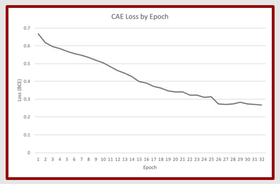
Pipeline



Phase I: Image Encoding

A. Convolutional Autoencoder

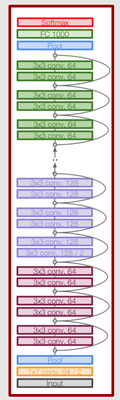
- The unsupervised CAE aims to extract a 1D high-level feature representation from a 3D frame.



- CAEs aim to maximize the information contained in a 1D "encoded" form of the image by training an encoder and decoder and optimizing a binary cross-entropy loss.

B. Transfer Learning Approach

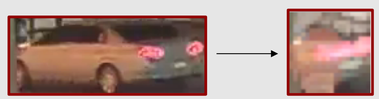
- Extracting the outputs of a forward pass just before the final fully connected layer from a pre-trained ResNet-18 and ResNet-50.
- Compared efficacy of the CAE and ResNet approaches by measuring Reef results.



Phase II: Automated Visual Weak Supervision with Reef

A. Action Tube Production

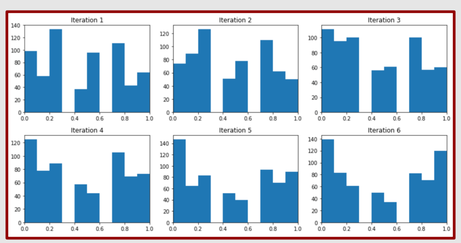
- When a vehicle drives through the intersection in the video, it appears in multiple frames. We can link 30 frames to produce action tubes that describe the motion of the vehicle over time.



- Task of vehicle classification in the large-scale video can now be feasibly reduced to the task of classifying action tubes.
- We randomly selected 30 tubes to be in the labeled validation set, and allocated 300 tubes for the unlabeled train set.

B. Automated Heuristic Generation

- Using the labeled dataset, Reef generates a variety of heuristics using *k*-nearest neighbors, decision trees, and logistic regression models.
- Heuristics are synthesized from these models, low-confidence heuristics are pruned, and data points with low-confidence labels are iteratively passed into Reef to generate more heuristics.



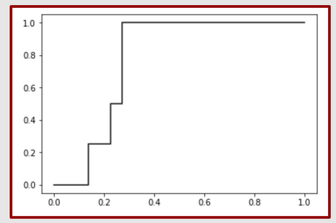
Phase III: Binary Vehicle Classifier

Classification Task Definition

We now have reduced our task to using a weakly labeled dataset to classify videos as containing cars vs. trucks. We use the outputs from the convolutional autoencoder and labels from Reef as the input to this step.

A. CNN-Based Classification

- Following weak supervision with Reef, all tubes in the previously unlabeled training set have been tagged with noisy labels.
- The binary classifier is a convolutional neural network (CNN) that accepts the two-dimensional input of the encoded frames in an action tube.
- Achieved an AUC of 0.773. AUROC shown to the right.



Experimental Results and Discussion

- We attempted to modulate the volume of weak labels being fed into the binary classifier. Validation results are shown below; note the relative gains in accuracy once weakly-supervised labels were added.

Weak Volume	lr = 1e-5	lr = 1e-4	lr = 1e-3	lr = 1e-2
0	0.500	0.667	0.667	0.333
100	0.769	0.808	0.769	0.769
200	0.739	0.739	0.739	0.739
300	0.712	0.712	0.712	0.712

- **Problem:** Though the datasets we used were extremely small (drawing on a training set of only 24 videos when not augmented by weakly-labeled points), the performance of the more complex models lagged behind benchmarks, at times performing not much better than a random model.
- **Solution:** Add data and train for more epochs.

Acknowledgements

Special thanks to Paroma Varma and Daniel Kang of the Stanford InfoLab.