Consistency based semi supervised learning for Video dataset

Sabaina Haroon
University of Central Florida
Orlando Florida
sabainaharoon@knights.ucf.edu

Abstract

Collecting annotations is a hard task and especially when it comes to a video, it needs to be annotated on the level of frames which adds much strain to the already arduous task. As it is hard to annotate data, we find unlabeled data easily. Semi-supervised learning aims to leverage unlabeled data to get a better understanding of labeled data. This paper aims at implementing the concept of semi supervised learning for videos using consistency regularization not only for classification but also for localization regression in detecting an action.

1. Introduction

Recently a lot of research focus is being given to the domain of semi supervised learning and self-supervised learning, where the idea is to take in use unlabeled data that is easy to collect and is in abundance and learn about the nature and distribution of seen data from unseen data as humans do. On the other hands, collecting labeled data with classification and localization annotations is a time consuming and arduous task to do. Hence labeled data is always limited in availability whereas deep learning networks learn well by seeing more, which cannot be achieved just by looking into labeled data. There are different works being done in the context of learning from unlabeled data. One very common method is Pseudo labeling, where a model is trained over labeled data, predicts over unlabeled data and then use high confident features from this pseudo labeled data to retrain the model. This process seems appealing, but it can be time consuming since one must iteratively filter the unlabeled data for pseudo labels and retrain the model with new added dataset.

Consistency regularization methods have also been used for semi supervised learning where a model learns to be robust towards perturbations applied to the input. Semi supervised learning with consistency regularization has generated state of the art results previously for object detection in images. For this research assignment, we are exploiting the work done in the domain of images for extending the idea to action detection in videos. The task of annotating data becomes even more difficult for videos. As a single clip has hundreds of frames in it, depending on frame rate, and annotating actions inside a video due to motion of objects, becomes even more hard. As can be seen by the availability of limited datasets, such as UCF101, kinetics or Charades, for videos action recognition.

If we succeed in implementing and transferring the knowledge learnt from the domain of semi supervised learning for object detection to videos, this can become a breakthrough in action related tasks. We will be able to focus and work on the unexplored aspects of video recognition that were previously constrained by the unavailability of enough amount of data for models to learn from.

2. Related Work

In following section, we talk about some of the research works done in the field of semi supervised learning for video processing.

2.1. Semi-supervised learning of feature Hierarchies

This work [1] learns feature hierarchies for object detection in a video. A basic detector learnt over labelled data is used to get candidate detection boxes from unlabeled videos. These candidate boxes are used for rescoring based on their confidence value. Output from the detector is divided in three categories based on their confidence as, confidence-positive, confidence-negative and hard samples. Features learned for confidence-positive and confidence-negative videos are fed into the detector for retraining, retrained detector is used to obtain confidences for hard samples, samples from this set are now rescored into positive, negative and hard sets again. Rescoring process is repeated until hard sets are no more classified as confidence positive and confidence negative sets. This might learn good features but has the same issue as pseudo semi supervised learning where iterative process of adding new dataset to model and retraining it takes time.

2.2. SS-AL

This method [2] combines active learning and

semi-supervised learning in such a way that as semi supervised process, pseudo labels are generated for a video for iterative model training, but these pseudo labels are generated for samples selected through active learning. Active learning marks samples important for investing resources for annotating and through pool-based uncertainty sampling, filters out other for obtaining their labels from an annotator trained using labeled samples selected through active learning. As an enhancement to the model extracted features from proposed neural network are passed through linear SVM classifier

2.3. SSL of object detectors from videos

In this work [3], tracking along the frames is done from confident annotations for sparsely labeled videos, Once tracked, detector is updated and retrained with tracked samples. This process alone is prone towards semantic drift; noisy tracked samples when added to the model for retraining might take it away from correct labeling. These are handled by filtering out temporally inconsistent detections, and high confidence false positives using decorrelation errors. This proposed technique can take in a sparsely labeled video where a few frames are annotated and successfully annotates all the frames using tracking. Detectors trained using these heavily annotated video samples along with labelled videos can annotate confidently fully unlabeled videos using SSL.

2.4. Consistency Based Semi-Supervised Object Detection

This method [4] extends from consistency regularization where perturbation is applied to an image and model is trained for robustness of noises or unknown data. CSD considers as input, labeled data and unlabeled data, where an image whether labeled or not is flipped horizontally and consistency loss learns regularity between feature representations of flipped and original image. Consistency based classification and localization both are performed using Jensen Divergence loss for confidence consistency and L2 loss for localization regression. This method can be applied both to single stage detectors where entire image is considered for bounding box considerations, and RPN multi-stage detector where a region-specific localization is learnt. Semi supervised learning through this method produces state of the art results for object detection in images compared to other SSL approaches.

3. Method

For this assignment, we have to extend the idea of state-of-the-art semi supervised learning through consistency regularization in the domain of images for action detection in videos. Consistency regularization based semi supervised learning can be implemented using both single stage and two stage detectors whereas my

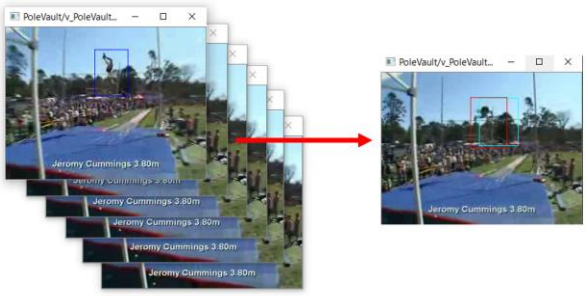

Figure 1: action detection through CSD-3D for central frame of video. Red bounding box is ground truth and aqua blue is predicted.

implementation is based upon single stage detector called single shot detector SSD with consistency regularization and consistency classification learning.

Following are the main steps summarized for this method.

3.1. Consistency Based Single Shot 3D Detector

CSD-3D uses same architecture as followed by original paper [5] but modifies it for 3d convolutions for temporal learning along with spatial learning, instead of 2d convolution that exploits spatial features only.

This model will be based on multi box learning where we fit bounding boxes of different aspect ratios at different feature scales by connecting layers from six different scales to multi box model which would predict specific numbers of bounding boxes. Different size bounding boxes are used at different scales to learn detections. As compared to other single stage detectors this multi box network is fast, instead of considering entire image for localization regression feature scalability helps us to focus only at specific regions in a frame.

3.1.1. Head Model: I3D

I3d learns the spatial temporal features of input video. For the purpose of multi box detection, I used 18 layers of I3D that are subdivided in 4 conv3d layers, 5 pooling layers with all pooling layers are max except last one which is average pooling layers.

3.1.2. Extra Layers

Features learnt from i3d are sent for to an extra layers model that consist of 10 layers broken down into 8 conv2d layers and 2 pooling layers.

3.1.3. Multi Box Model

This model defines the basis of single shot detector by fitting bounding boxes of different sizes at different scales

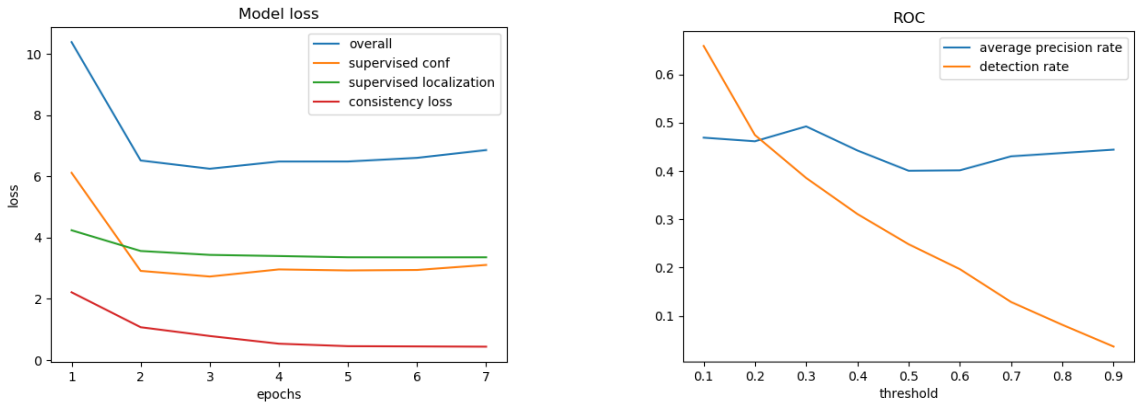

Figure 2: Model behavior for CSD-3D with first feature scale selected from 13th layer of I3D. 24 classes used.

of the base and extra layer models, it learns localizations for bounding boxes using the same concept of scalability and sizes as used in SIFT detector but with deep learning. In this step of the network we take connections from different scales of previous models, output from 7th layer having feature map of 38x38 and 18th layer having feature map of 19x19 from I3D network. Output from every second layer of extra layers is also selected for bounding box fitting having feature scales of 10x10, 5x5, 3x3 and 1x1 maps.

Outputs from these layers are fed into 2d conv layer having output channels equal to the number of bounding boxes we are fitting at a scale. For instance, we learn 4 bounding boxes for 38x38 feature map, 4 to 6 bounding boxes are learnt at each scale using ReLU activation with conv2d layer. In total model learns 8732 boxes against every video and filters out the most relevant predictions during multi-box loss calculations.

All the above steps are done for both confidence layers and localization layers. We get 8732x4 bounding box features learnt through localization layers and 8732 x #classes confidence scores learnt for each bounding box. Each of the 8732 bounding boxes will have softmax confidence scores for all the classes.

When input is passed through a network it returns output tensor consisting of confidence and localization activation layers with a prior tensor, confidence tensor having softmax applied to its activation from the output tensor, localization tensor with softmax applied, flipped confidence tensor with softmax applied and flipped localization tensor with softmax.

3.2. Loss calculations

The formulation of correct type of loss functions and their calculation methodologies will define the strength of our semi supervised learning model. We are using four different losses to calculate total loss. Two from supervised samples and two from unsupervised and supervised samples.

3.2.1 Supervised Multi Box Loss

For this loss prior bounding box localizations are calculated using prior box calculations from this paper [5] where variances of bounding boxes are defined along two dimensions. I am assuming that human labels have higher variance mostly along one direction and smaller in the other having a rectangular bounding box so, I selected variances for prior boxes as {0.1, 0.2} whereas this is a hyperparameter which has to be adjusted with one having the full knowledge of the datasets. Right now, I am empirically trying these variances but for future these variances can be calculated from ground truth labelled examples of a video set. Prior bounding boxes also use aspect ratios depending upon the size of feature scale. This again is dataset dependent and must be tuned having fine understanding of the semantics of the dataset.

We have 8732x4 box predictions from our network, 8732x4 box predictions calculated using priors. And ground truth bounding boxes. Jaccard overlap is calculated between ground truth bounding boxes and prior boxes and those prior boxes with highest overlap confidence are used to find loss with predicted bounding boxes from the network.

Classification loss for bounding boxes having overlapping scores above a set threshold are calculated using hard negative mining where we filter out negative boxes' features having a very low match with priors and ground truth boxes.

3.2.2 Consistency Loss

A video is fed into our CSD model to calculate localization and classification features. Also, this video is flipped horizontally and passed through the same network. Now we have two new sets of features learnt from the flipped video. Our consistency loss will force our model to learn similar features for such kind of perturbed inputs and become robust against them.

To calculate classification consistency loss I am using

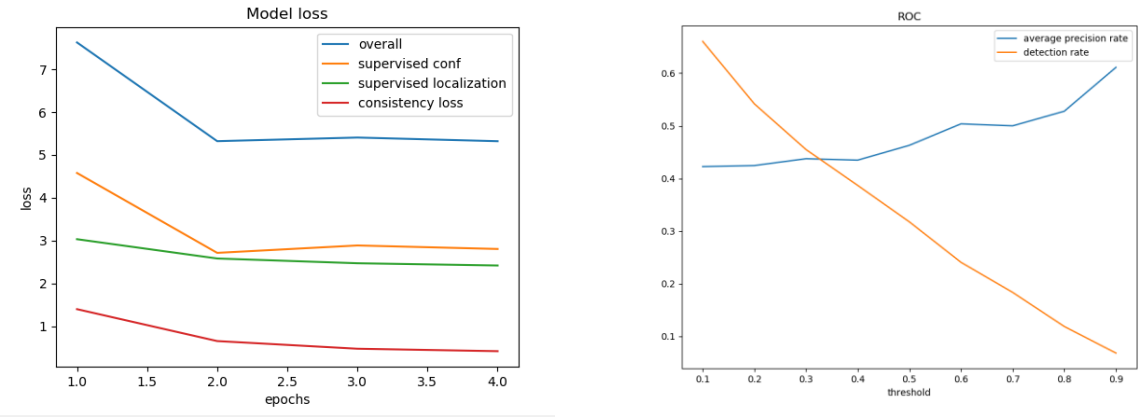

Figure 3: Model behavior for CSD-3D with first feature scale selected from 7th layer of I3D. 24 classes used

Jenson Shannon, JSD, loss. This JSD classification loss is further divided into two losses, where for one we consider original video as ground truth and calculate flipped video feature loss from it. Whereas for second type we consider flipped version as ground truth and calculating loss with original features as predictions.

For consistency localization loss I am using the same loss formula as from the baseline research paper for this assignment [4]. Where we calculate L2 loss between original 8732x4 bounding box features learnt and flipped features learnt. Horizontal dimension of loss in L2 is added instead of subtraction because it is flipped across rows only.

3.2.3 Ramp Weight

Ramp down and ramp up weighting strategies are used for consistency loss. Weight for consistency loss is very small for starting iterations until the model is being given ample time to learn from supervised losses. Over the iterations supervised loss decrease and weightage for consistency loss increase with a ramp until both reach comparable values. Ramp down weightage is now used for consistency loss.

3.3. Background Elimination

Along with the total classes in our dataset one extra class is added as background class through network training. Scores learnt for this class help us to eliminate all other classes for a sample whose highest score is less than background score. A mask is applied over localization and confidence layers for both original and flipped videos, this masks samples out foreground classes which then are used for consistency loss calculation.

3.4. Data loader

For a single batch both labelled and unlabeled videos are returned from data loader. Data loader checks if a video is annotated or not, if annotated it returns video sample for 30 frames, annotations of center frame and a flag telling if video loaded is supervised or not. For unlabeled videos it returns video sampled at 30 frames and bool flag returning true for semi supervised sample.

In nutshell, my method uses a batch size of 30, a batch is loaded through data loader, this batch is passed onto three-tire model to calculate features and confidence scores. Flipped batch set is also passed through the network in the same way. Output from the network is used for multi box supervised loss calculation and confidence, localization, flipped confidence and flipped localization scores are used for background elimination and then consistency loss calculations. Model loss term contains supervised losses and consistency losses weighted with ramps weights.

4. Experimentation

To evaluate the performance of our network we used UCF101 video dataset. This dataset consists of more than 13000 videos, 101 classes of human doing different actions such as applying makeup, playing basketball, pole vault, skiing and others. 24 classes out of these are annotated with human labels. Videos from this dataset are collected from YouTube.

Dataset is divided into approximately 9500 training samples and remaining set is used for testing. I performed following different kinds of experiments to study the behavior of my method

- Results calculated by extracting first bounding box scale features from 13th layer of I3D network
- Results calculated by getting first bounding box scale features from 7th extracting of I3D network
- Model tested with 24 given classes with one background class
- Model tested with 101 classes with one background class
- Model trained for only one annotation box per video but tested on all ground truth annotation boxes to evaluate whether model has learnt multiple annotations for a frame by itself through unsupervised learning or not.

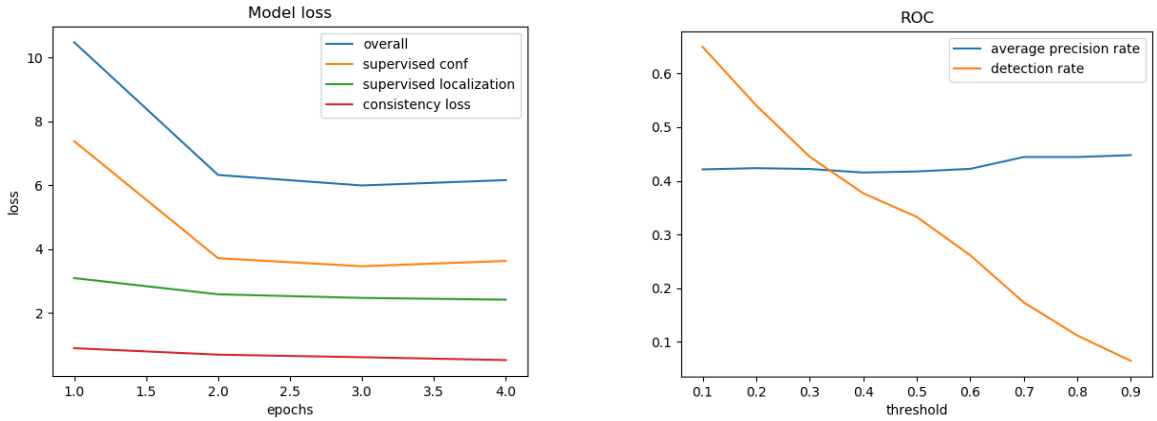

Figure 4: Model behavior for CSD-3D trained with 101 classes and first feature scale selected from 7th layer of I3D

4.1. Detector for testing

Input video forward passes through the model and is sent to detector with prior boxes, confidence and localization activations. Detector matches the predicted bounding box locations with the prior boxes and selects highest scored predictions. Non maximum suppression is done over these predictions to stop having unnecessary bounding boxes around the same location. Detector selects top-5 confidence predictions along with predicted bounding boxes for each class.

Once detector return top-5 predictions against each class, test class filters out further top predictions based on a specific class confidence threshold. All the predictions above a threshold are used for calculating overlap with ground truth bounding boxes for a video. Note that model was trained using only one annotation box per video since we only have human annotation boxes, I assume that using some of the human annotation boxes, model should be able to learn and predict annotations for more than one human present in a video frame. Therefore, predictions are compared with all the ground truth bounding boxes for true positives, false positives and detection rate.

4.2. Learning Rate Decay

Learning rate for all the experiments done are decayed by a factor of 10 after specified number of iterations. For these experiments I found best learning rates if decayed for the first time after 100 iterations and then 2 decays are performed at the end of iterations

4.3. Results

Results are shown for this method using loss curves for supervised and unsupervised losses and ROC curves showing detection accuracy and average precision curve for confidence thresholds for 0.1 to 0.9.

4.3.1 Losses

Figure 2. shows loss curves for the model that was trained using 24 classes and prior bounding box values given as greater variance and expect ratio along x-axis and smaller variance along y-axis. First feature scale selected for training this model was selected from the 13th layer of I3D base model. Model was trained using SGD optimizer at first which was optimizing loss to the value of 10, whereas using Adam optimizer with decaying learning rates, model reached uniform loss in the first two epochs compared to SGD which achieved uniformity after 5th epoch. Loss gets stagnant at the value of approximately 7 and does not decrease further.

Figure 3. shows results for our model where I3D's 7th layer is used as first feature scale for bounding box predictions. Here loss is decreased a few points, and this is the minimum loss that I got through all experimentations. Compared with the baseline SSD results[1] for VOC loss drops down to the value of approximately 3 but after 120000 iterations whereas for our video dataset using Adam optimizer loss closer to this range is achieved during first 12000 iterations. Model is trained using 24 labelled classes with one additional background class.

Figure 4. shows results for model trained using full 101 classes and one background class. First feature scale for this model is 7th layer from I3D. Loss for 101 classes is slightly higher, it might be happening because of background masking where for consistency learning, feature maps for certain classes are filtered out before loss calculation, since model may think of unlabeled classes as background classes they get lower score and are masked out before learning

4.3.2 ROC curves

Images at the right side of figure 2,3 and 4 show ROC

[1] https://github.com/Hakuyume/chainer-ssd/blob/master/images/loss_curve.png

curves for detection accuracy along with precision shown for detection at each threshold. For every 10^{th} factor of threshold from 0.1 to 0.9, detection rate and Precision rate is shown for the samples detected at each threshold level.

It can be seen from the ROC graphs of all models, that detection accuracy decreases with the increase in threshold: as the confidence increases less videos are detected. Whereas precision rate remains same or sometimes increase with the decrease in detection as less new samples are detected with more confidence.

5. Discussion

Results produced for this assignment are not very strong. There are couples of reasons behind that which can be fixed in future work. Firstly, due to time constraint, only number of models could have been experimented, whereby training model with I3D takes lot of time, around 15 to 18 hours depending on the number of frames. All the models were trained only using center frame which is a factor for delimiting performance for instance most of the times pole vault was classified also as basketball class with high confidence score. The reason as looking at the videos seems that in middle frame activity of both persons is same , both are jumping towards a specific point, in basketball a person is jumping with basketball towards a higher goalpost, in pole vault a person is jumping with a pole towards a vault.

Lastly I have taken features from 7^{th} layer and 18^{th} layer of I3D. 7^{th} layer had scale 38x38, extracting features from such an early scale was to let model learn the localized features also, whereas this comes with a tradeoff as at such an early layer model has not learnt temporal features of a video fully. Another factor in the hindrance of not achieving a good accuracy was prior boxes calculations. My code was based on the SSD code from VOC dataset. They have experimented their code for general orientations of bounding boxes present in their data and have set variances and aspect ratios for prior boxes calculations accordingly. Whereas, since prior knowledge is considered as a form of the ground truth to compare results with, it should be selected after due considerations and experimentations. Selecting right type of aspect ratios and variances along the dimensions of the box can increase the model's performance.

6. Conclusion

In this assignment, our objective was to perform action detection for UCF101 video dataset, for central frame of the video, by extending concepts from Consistency based semi-supervised learning for Object detection. I have implemented Single-Shot Detector with consistency based semi-supervised learning using I3D model instead of VGG model used by referenced paper. However, there is a huge room for improvement in results due to different factors mentioned in section 5. From related work section, it can be seen that most of the semi supervised work done for action classification and detection is using iterative labeling and model training processes based on prediction confidences. Iterative processes can be quite slow as we don't know how many times, we have to train our network, also training models for videos take considerable amount of time compared to images which makes iterative process a bad candidate for semi supervised learning in videos. If right amount of time and effort is dedicated to this model and method in future work, it can become a powerful and efficient tool for semi-supervised learning.

As, I have discussed in section 5, by extracting features from earlier layers, localization information can be learnt better as overall shapes of objects are preserved in those layers, but the model at this stage has not learn temporal information, (this fact is backed by the loss curves where classification loss is mostly greater than localization loss) in future we can modify this method as such that classification can be learnt offline using any state-of-the-art object classifier for videos, and multi box method with priors is only used for localization and consistency learning. Also, right now we are feeding features to extra layers and multi box model from i3d for only center frame, but for multiple frames we might be able to extend this idea in future to be able to learn 3d bounding boxes for a video.

References

[1] Yang, Yang et al. "Semi-supervised Learning of Feature Hierarchies for Object Detection in a Video." *2013 IEEE Conference on Computer Vision and Pattern Recognition* (2013): 1650-1657.

[2] Sabata, Tomás et al. "Semi-supervised and Active Learning in Video Scene Classification from Statistical Features." *IAL@PKDD/ECML* (2018).

[3] Ishan Misra, Abhinav Shrivastava and Martial Hebert. Watch and Learn: Semi-Supervised Learning of Object Detectors from Videos, *Proceedings of IEEE Computer Vision and Pattern Recognition (CVPR)*, 2015

[4] Jeong, Jisoo et al. "Consistency-based Semi-supervised Learning for Object detection." *NeurIPS* (2019).

[5] Liu, Wei et al. "SSD: Single Shot MultiBox Detector." *ECCV* (2016).