Fully Convolutional Multi-Class Multiple Instance Learning

Deepak Pathak, Evan Shelhamer, Jonathan Long & Trevor Darrell
UC Berkeley
{pathak,shelhamer,jonlong,trevor}@cs.berkeley.edu

Abstract

Multiple instance learning (MIL) can reduce the need for costly annotation in tasks such as semantic segmentation by weakening the required degree of supervision. We propose a novel MIL formulation of multi-class semantic segmentation learning by a fully convolutional network. In this setting, we seek to learn a semantic segmentation model from just weak image-level labels. The model is trained end-to-end to jointly optimize the representation while disambiguating the pixel-image label assignment. Fully convolutional training accepts inputs of any size, does not need object proposal pre-processing, and offers a pixelwise loss map for selecting latent instances. Our multi-class MIL loss exploits the further supervision given by images with multiple labels. We evaluate this approach through preliminary experiments on the PASCAL VOC segmentation challenge.

1 Introduction

Convolutional networks (convnets) are achieving state-of-the-art performance on many computer vision tasks but require costly supervision. Following the ILSVRC12-winning image classifier of Krizhevsky et al. (2012), progress on detection (Girshick et al., 2014) and segmentation (Long et al., 2014) demonstrates that convnets can likewise address local tasks with structured output.

Most deep learning methods for these tasks rely on strongly annotated data that is highly time-consuming to collect. Learning from weak supervision, though hard, would sidestep the annotation cost to scale up learning to available image-level labels.

In this work, we propose a novel framework for multiple instance learning (MIL) with a fully convolutional network (FCN). The task is to learn pixel-level semantic segmentation from weak image-level labels that only signal the presence or absence of an object. Images that are not centered on the labeled object or contain multiple objects make the problem more difficult. The insight of this work is to drive the joint learning of the convnet representation and pixel classifier by multiple instance learning. Fully convolutional training learns the model end-to-end at each pixel. To learn the segmentation model from image labels, we cast each image as a bag of pixel-level-instances and define a pixelwise, multi-class adaptation of MIL for the loss.

MIL can reduce the need for bounding box annotations (Cinbis et al., 2014; Song et al., 2014), but it is rarely attempted for segmentation. Oquab et al. (2014) improve image classification by inferring latent object location, but do not evaluate the localization. Hoffman et al. (2014) train by MIL fine-tuning but rely on bounding box supervision and proposals for representation learning. Most MIL problems are framed as max-margin learning (Andrews et al., 2002; Felzenszwalb et al., 2010), while other approaches use boosting (Ali & Saenko, 2014) or Noisy-OR models (Heckerman, 2013). These approaches are limited by (1) fixed representations and (2) sensitivity to initial hypotheses of the latent instance-level labels. We aim to counter both shortcomings by simultaneously learning the representation to maximize the most confident inferred instances. We incorporate multi-class annotations by making multi-class inferences for each image. When an image / bag contains multiple classes the competition of pixelwise models help to better infer the latent instance-level classes.

We investigate the following ideas and carry out preliminary experiments to these ends:

- We perform MIL jointly with end-to-end representation learning in a fully convolutional network. This eliminates the need to instantiate instance-label hypotheses. FCN learning and inference can process images of different sizes without warping or object proposal pre-processing. This makes training simple and fast.
- We propose a multi-class pixel-level loss inspired by the binary MIL scenario. This tries to maximize the classification score based on each pixel-instance, while simultaneously taking advantage of inter-class competition in narrowing down the instance hypotheses.
- We target the under-studied problem of weakly supervised image segmentation. Our belief is that pixel-level consistency cues are helpful in disambiguating object presence. In this way weak segmentation can incorporate more image structure than bounding boxes.

2   FULLY CONVOLUTIONAL MIL

A fully convolutional network (FCN) is a model designed for spatial prediction problems. Every layer in an FCN computes a local operation on relative spatial coordinates. In this way, an FCN can take an input of any size and produce an output of corresponding dimensions.

For weakly supervised MIL learning, the FCN allows for the efficient selection of training instances. The FCN predicts an output map for all pixels, and has a corresponding loss map for all pixels. This loss map can be masked, re-weighted, or otherwise manipulated to choose and select instances for computing the loss and back-propagation for learning.

We use the VGG 16-layer net (Simonyan & Zisserman, 2014) and cast it into fully convolutional form as suggested in Long et al. (2014) for semantic segmentation by replacing fully connected layers with corresponding convolutions. The network is fine-tuned from the pre-trained ILSVRC classifier weights i.e. pre-trained to predict image-level labels. We then experiment with and without initializing the last layer weights i.e. the classifier layer. These initializations, without MIL fine-tuning, act as the baselines (row 1 and 2 in Table). If there is no image-level pretraining, the model quickly converges to all background. Semantic segmentation requires a background class but the classification task has none; we simply zero initialize the background classifier weights.

3   MULTI-CLASS MIL LOSS

We define a multi-class MIL loss as the multi-class logistic loss computed at maximum predictions. This selection is enabled by the output map produced by FCN i.e. for an image of any size, the FCN outputs a heat-map for each class (including background) of corresponding size. We identify the max scoring pixel in the coarse heat-maps of classes present in image and background. The loss is then only computed on these coarse points, and is back propagated through the network. The alternating optimization in the binary MIL problem inspires this ignoring of the loss at non-maximally scoring points. The background class is analogous to the negative instances by competing against the positive object classes. Let the input image be I, its label set be $\mathcal{L}_I$ (including background label) and $\hat{p}_l(x, y)$ be the output heat-map for the lth label at location (x, y). The loss is defined as:

$$(x_l, y_l) = \arg \max_{\forall (x,y)} \hat{p}_l(x, y) \qquad \forall l \in \mathcal{L}_I$$

$$\implies \text{MIL LOSS} = \frac{-1}{|\mathcal{L}_I|} \sum_{l \in \mathcal{L}_I} \log \hat{p}_l(x_l, y_l)$$

Ignoring the loss at all non-maximally scoring points is key to avoid biasing the learning of the FCN to the background. Simultaneous training exploits multi-label images through inter-class confusion to help refine the intra-class pixel accuracy. At inference time, the MIL-FCN takes the top class prediction at every point in the coarse prediction and bilinearly interpolates to image resolution to obtain a pixelwise segmentation.

4   EXPERIMENTS

All results are on the PASCAL VOC segmentation challenge. We train and validate on the VOC 2011 train augmented by Hariharan et al. (2011) and val sets then evaluate on the completely held-

out VOC 2012 test set. The evaluation metric is intersection over union (IU), and is defined per class as the percentage of pixels in the intersection of ground truth segmentation mask, and the predicted mask out of the number of pixels in their union. The MIL-FCN model is initialized from the 16-layer VGG ILSVRC14 classifier (Simonyan & Zisserman, 2014) then fine-tuned by the MIL loss. Long et al. (2014) fine-tune from all but the output layer, as they have access to complete supervision. In our setting however, transferring the output layer parameters for the classes common to both PASCAL and ILSVRC improves results. Including these classifier parameters helps prevent degenerate solutions of predicting all background. We train our model with a learning rate 0.0001, momentum 0.9 and weight decay 0.0005. The training is quick and the network converges in less than 10,000 iterations.

Table 1: *Results on PASCAL VOC 2011 segmentation validation and 2012 test data*. Fine-tuning with the MIL loss achieves 96% relative improvement over the baseline.

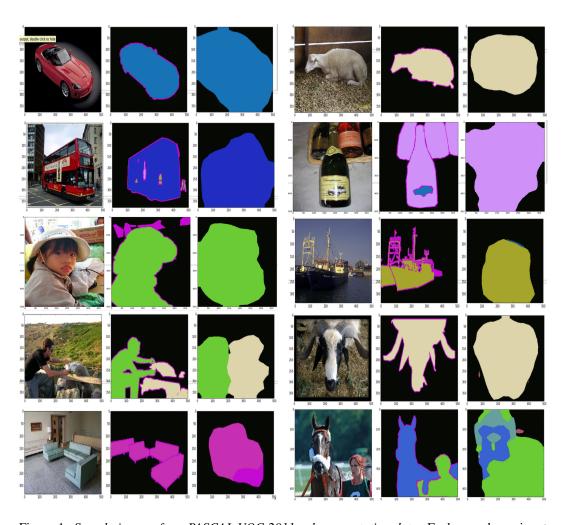| Approach	mean IU (VOC2011 val)	mean IU (VOC2012 test)
Baseline (no classifier)	3.52%	-
Baseline (with classifier)	13.11%	13.09%
MIL-FCN	**25.05%**	**25.66%**
Oracle (supervised)	59.43%	63.80%

Figure 1: *Sample images from PASCAL VOC 2011 val-segmentation data*. Each row shows input (left), ground truth (center) and MIL-FCN output (right).

Table 1 shows quantitative intersection-over-union (IU) scores while example outputs from MIL-FCN are shown in Figure 1. MIL-FCN achieves **96% relative improvement** over the baseline results when the classifier is fine-tuned from the common classes. These are preliminary but encouraging results.

5 DISCUSSION

We propose a novel model of joint multiple instance and representation learning with a multi-class pixelwise loss inspired by binary MIL. This model is learned end-to-end as a fully convolutional network for the task of weakly supervised semantic segmentation. It precludes the need for any kind of proposal or instance hypothesis mechanisms. Inference is fast ($\approx$ **1/5 sec**).

These results are encouraging, and can be improved further. Currently, the coarse output is merely interpolated; conditional random field regularization or super-pixel (Achanta et al., 2012) projection could refine the predictions. These grouping methods could likewise drive learning by selecting whole segments instead of single points for MIL training. Moreover, controlling convnet learning by manipulating the loss map could have further uses such as encouraging consistency across images for co-segmentation or hard negative mining.

REFERENCES

Achanta, Radhakrishna, Shaji, Appu, Smith, Kevin, Lucchi, Aurelien, Fua, Pascal, and Susstrunk, Sabine. Slic superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2012.

Ali, K. and Saenko, K. Confidence-rated multiple instance boosting for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

Andrews, Stuart, Tsochantaridis, Ioannis, and Hofmann, Thomas. Support vector machines for multiple-instance learning. In *Proc. NIPS*, pp. 561–568, 2002.

Cinbis, Ramazan Gokberk, Verbeek, Jakob, and Schmid, Cordelia. Multi-fold mil training for weakly supervised object localization. In *CVPR*, 2014.

Felzenszwalb, Pedro F, Girshick, Ross B, McAllester, David, and Ramanan, Deva. Object detection with discriminatively trained part-based models. *IEEE Tran. PAMI*, 32(9):1627–1645, 2010.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *In Proc. CVPR*, 2014.

Hariharan, Bharath, Arbeláez, Pablo, Bourdev, Lubomir, Maji, Subhransu, and Malik, Jitendra. Semantic contours from inverse detectors. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pp. 991–998. IEEE, 2011.

Heckerman, David. A tractable inference algorithm for diagnosing multiple diseases. *arXiv preprint arXiv:1304.1511*, 2013.

Hoffman, Judy, Pathak, Deepak, Darrell, Trevor, and Saenko, Kate. Detector discovery in the wild: Joint multiple instance and representation learning. *arXiv preprint arXiv:1412.1135*, 2014.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

Long, Jonathan, Shelhamer, Evan, and Darrell, Trevor. Fully convolutional networks for semantic segmentation. *arXiv preprint arXiv:1411.4038*, 2014.

Oquab, Maxime, Bottou, Léon, Laptev, Ivan, Sivic, Josef, et al. Weakly supervised object recognition with convolutional neural networks. In *Proc. NIPS*, 2014.

Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Song, Hyun Oh, Lee, Yong Jae, Jegelka, Stefanie, and Darrell, Trevor. Weakly-supervised discovery of visual pattern configurations. In *Proc. NIPS*, 2014.