Distantly Supervised Road Segmentation

Wenjie Niu

May 29, 2018

Abstract

We present an approach for road segmentation that only requires image-level annotations at training time. We leverage distant supervision, which allows us to train our model using images that are different from the target domain. Using large publicly available image databases as distant supervisors, we develop a simple method to automatically generate weak pixel-wise road masks. These are used to iteratively train a fully convolutional neural network, which produces our final segmentation model. We evaluate our method on the Cityscapes dataset, where we compare it with a fully supervised approach. Further, we discuss the tradeoff between annotation cost and performance. Overall, our distantly supervised approach achieves 93.8% of the performance of the fully supervised approach, while using orders of magnitude less annotation work. [1]

1 Saliency Map by Global Average Pooling

This section describes an experiment we conducted in order to see the performance of the saliency map obtained by global average pooling (GAP), as well as how we actually collected training images.We require an image collection of road and non-road images for training a classification CNN with GAP. Instead of annotating from scratch, we take advantage of two publicly available large image databases: ImageNet [2] and Places [3]. For road images, we use ImageNet for collecting road images, as its labels are organized in an object centric way. We searched for labels with the keyword road or highway, which yielded the classes n02744323: arterial road and n02744323: divided highway, dual carriageway.Random samples are shown in Figure 1.

Figure 1: Examples of road images from ImageNet used for training our saliency detector

We do not use the Cityscapes training set as road images, as the saliency FCN would highlight the objects that consistently appear in the Cityscapes images. We prove this empirically as a part of our experiments. For non-road images, we need to collect outdoor scene images without road. We use Places, since unlike ImageNet, it organizes images according to scenes. We first filter out scene labels whose meta class corresponds to indoor scenes, which resulted in a remaining 205 outdoor labels. We manually examined these labels in order to exclude irrelevant classes (e.g., baseball field) and road classes (e.g., field road), which resulted in 120 classes. Random samples are shown in Figure 2.

Table 1: Experimental results on Cityscapes when using saliency only.

Method	mIOU
Low resolution, Generic road image	0.405
High resolution, Generic road image	0.092
Low resolution, Car centric road image	0.206
High resolution, Car centric road image	0.093

Figure 2: Examples of non-road images from Places used for training our saliency detector.

This is the only practically required manual labeling process in our work. In addition, we do not aim to obtain very accurate images corresponding to a label, as we do not have time to check all the images manually, so some noisy images are acceptable. For example, the road class contains images that do not look like a typical car centric road, as they are shot from a helicopter (see the left side of Figure 1).

Car centric images are from Cityscapes. General road images are from ImageNet.The results are shown in Table 1. Perhaps surprisingly, the lower resolution input yields a much better mIOU of 0.405 compared to 0.092 for the higher resolution case. Our manual inspection of some random samples indicated that the higher resolution saliency map tends to highlight objects that often appear on the road, such as cars or traffic signs. As shown in Table 1, the mIOU is lower compared to just using road images.

References

[1] S. Tsutsui, T. Kerola, and S. Saito, "Distantly supervised road segmentation," in *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017. 1

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2014. 1

[3] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition." *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017. 1