A Review on Deep Learning Techniques Applied to Semantic Segmentation

Yuan An

In the last article, some common networks of semantic segmentation is introduced in brief. In the rest of some sections, I know about a common training technique–transfer learning, and data pre-processing and augmentation approaches.

Transfer Learning

There are two reasons for difficult to training a deep neural network from te very beginning are that a dataset of sufficient size is needed (not usually available) and reaching convergence needs to long. Even if meet the above conditions, it is often helpful to start with pre-trained weights instead of random initialized ones [1, 3].

Yosinski *et al.* [5] also proved that transferring features even from distant tasks can be better than using random initialization.

By fine-tuning the weights of a pre-trained network, transfer learning can make training a deep neural network faster and more efficient.

But transfer technique is not completely straightforward. On the one hand, there are architectural constraints that must be met to use a pre-trained network, so it is common to reuse already existing network architectures. On the other hand, the difference of process between transfer learning and training networks from the very beginning is sightly.

Data augmentation is a common technique that has been proven to benefit the training of machine learning models in general and deep architectures in particular; either speeding up convergence or acting as a regularizer, thus avoiding overfitting and increasing generalization capabilities [4].

Datasets

In the section 3 of [2], the authors introduce some common used dataset, including 2D datasets, 2.5D datasets and 3D datasets. Some instance of mentioned datasets will be presented in the following.

2D Datasets

PASCAL Visual Object Class (VOC) consists of a ground-truth annotated dataset of images and five different competitions: classification, detection, segmentation, action classification, and person layout(see Figure 1).

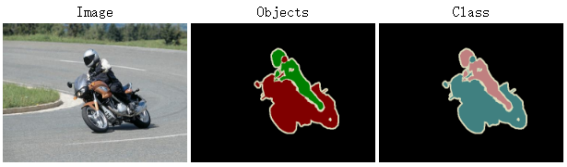

Figure 1. Instances from PASCAL VOC datasets.

Semantic Boundaries Dataset (SDB) is an extended version of PASCAL VOC which provides semantic segmentation ground truth for those images that were not labelled in VOC (see Figure 2).

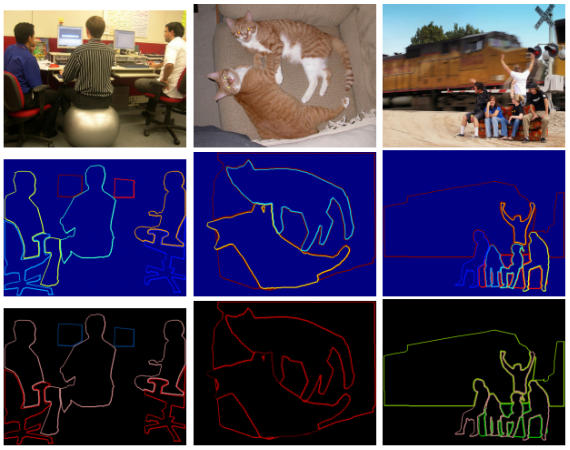

Figure 2. Instances from SDB.

SYNTHetic Collection of Imagery and Annotations (SYNTHIA) is a large-scale collection of photo-realistic renderings of a virtual city, semantically segmented, whose purpose is scene understanding in the context of driving or urban scenarios (see Figure 3).

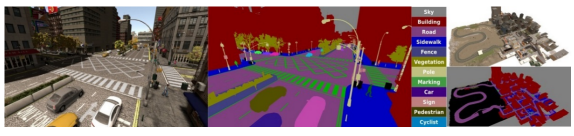

Figure 3. Instances from SYNTHIA.

KITTI is one of the most popular datasets for use in mobile robotics and autonomous driving. It consists of hours of traffic scenarios recorded with a variety of sensor

modalities, including high-resolution RGB, grayscale stereo cameras, and a 3D laser scanner (see Figure 4).

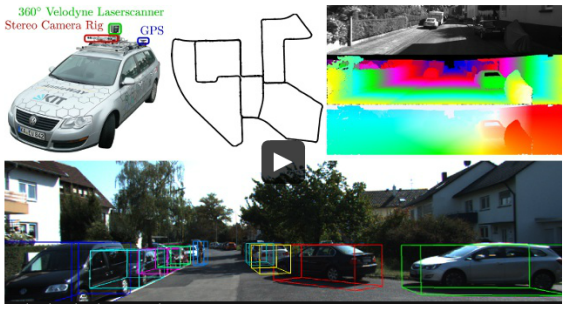

Figure 4. Instances from KITTI.

2.5D Datasets

With the advent of low-cost range scanners, datasets including not only RGB information but also depth maps are gaining popularity and usage. NYUDv2 consists of 1449 indoor RGB-D images captured with a Microsoft Kinect device.

SUN3D is similar to the NYUDv2, containing a large-scale RGB-D video database, with 8 annotated sequences.

The Object Segmentation Database (OSD) has been designed for segmenting unknown objects from generic scenes even under partial occlusions (see Figure 5).

Figure 5. Instances from OSD.

RGB-D Object Dataset is composed by video sequences of 300 common household objects organized in 51 categories arranged (see Figure 6).

3D Datasets

Pure 3D databases are scarce, this kind of datasets usually provide CAD meshes or other volumetric representations, such as point clouds. Generating large-scale 3D datasets for segmentation is costly and difficult, and not many deep learning methods are able to process that kind of data as it is.

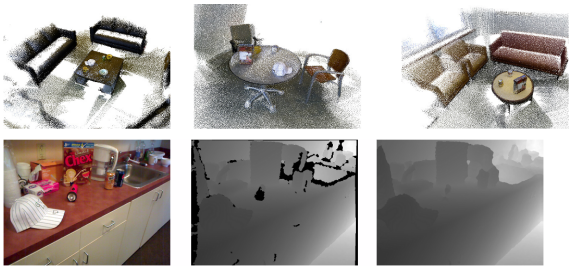

Figure 6. Instances from RGB-D Object Dataset.

The common used 3D datasets are including ShapeNet Part, Stanford 2D-3D-S, Benchmark for 3D Mesh Segmentation and Sydney Urban Objects Dataset.

By knowing those knowledge about training method, data preprocessing and augmentation technique, and dataset common used, it can make it easily to know semantic segmentation.

References

[1] A. Ahmed, K. Yu, W. Xu, Y. Gong, and E. Xing. Training hierarchical feed-forward visual recognition models using transfer learning from pseudo-tasks. In *ECCV*, pages 69–82, 2008. 1

[2] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. G. Rodríguez. A review on deep learning techniques applied to semantic segmentation. *arXiv preprint arXiv:1704.06857*, 2017. 1

[3] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, pages 1717–1724, 2014. 1

[4] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell. Understanding data augmentation for classification: When to warp? *arXiv preprint arXiv:1609.08764*, 2016. 1

[5] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *NIPS*, pages 3320–3328, 2014. 1