Convolutional Neural Networks to Image Segmentation

Felipe Augusto Lima Reis
PUC Minas - Pontifícia Universidade Católica de Minas Gerais
R. Walter Ianni 255 - Bloco L - Belo Horizonte, MG, Brasil
falreis@sga.pucminas.br

Abstract

Image segmentation refers to the partition of an image into a set of regions to cover it, to represent a meaningful area. Before the use of deep neural networks, the best-performing methods mostly was made using hand engineered features [5]. This paper will evaluate two Deep Neural Networks for semantic segmentation: Segnet and U-Net. These DNNs are both "fully convolutional" networks that output a result with the same size as the input. The models are trained and tested using KITTI Road Dataset. The results were also evaluated using KITTI Dataset Toolkit. The tests showed that Segnet had a better performance in image segmentation than U-Net. U-Net, however, was easier to train, smaller and provide faster results than Segnet.

1. Introduction

Image segmentation refers to the partition of an image into a set of regions to cover it, to represent meaningful areas [8]. The goal is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze [3].

Segmentation has two main objectives: the first one is to decompose the image into parts for further analysis and the second one is to perform a change of representation [8]. Also, segmentation must follow some characteristics to identify regions, as it follows:

- Regions of an image segmentation should be uniform and homogeneous with respect to some characteristic, such as gray level, color, or texture [8];
- Region interiors should be simple and without many small holes [8];
- Adjacent regions of a segmentation should have significantly different values with respect to the characteristic on which they are uniform [8];
- Boundaries of each segment should be smooth, not ragged, and should be spatially accurate [8].

Semantic pixel-wise segmentation is an active topic of research [5]. Before the use of deep neural networks, the best-performing methods mostly was made using hand engineered features [5]. The success of deep convolutional neural networks for object classification led researchers to use these techniques to learn new capabilities, such as segmentation [5].

This paper will evaluate two different Deep Neural Networks for semantic segmentation and compare the results over Kitti Road Dataset [11]. The chosen neural networks are Segnet [5] and U-NET [18].

The organization of this paper is as follows. In the next Section we show related works to this paper. Section 3 contains information about the dataset used in this work as other datasets used previously to start the project. Section 4 explains the methods used to develop this project, Section 5 describe the experiments and shows the results and Section 6 concludes this paper and gives some final considerations and ideas for future works.

2. Related Work

2.1. Superpixels

A segmentao de imagens consiste em dividir uma imagem em um conjunto de regies logicamente agrupadas, de modo a reunir reas que contm informao relevante dentro dos grupos [8]. Nessa tarefa, tomamos os *pixels* como unidades bsicas de processamento [23]. O agrupamento de pixels em unidades maiores permite um tipo de segmentao chamado de *oversegmentation* [23]. O uso de superpixels possibilita o aumento da velocidade de processamento posterior, uma vez que a quantidade de pixels diminui consideravelmente em relao a imagem original.

A utilizao de superpixels possibilita a reduo de itens a serem processados, entretanto pode causar perda de informao importante. No entanto, para alguns casos, a perda de qualidade pode se justificar em relao ao ganho de velocidade obtido utilizando esse tipo de operao. Essa relao consiste ento em um *trade-off* entre ambas as caractersticas, sendo viveis em alguns cenrios de processamento em tempo

real ou para dispositivos com baixo desempenho.

Alguns mtodos de gerao de superpixels so utilizados para segmentao de imagens e deteco de bordas, como os mtodos EGB [10] e SLIC [2]

2.2. SEGNET

SEGNET is a deep encoder-decoder architecture for multi-class pixelwise segmentation [5]. The SEGNET architecture consists of a sequence of non-linear processing layers (encoders) and a corresponding set of decoders followed by a pixel-wise classifier [5] [22]. Typically, each encoder consists of one or more convolutional layers with batch normalization and a ReLU non-linearity, followed by non-overlapping max-pooling and sub-sampling [5] [22]. The sparse encoding due to the pooling process is upsampled in the decoder using the max-pooling indices in the encoding sequence [5] [22]. Figure 1 presents the architecture of SEGNET.

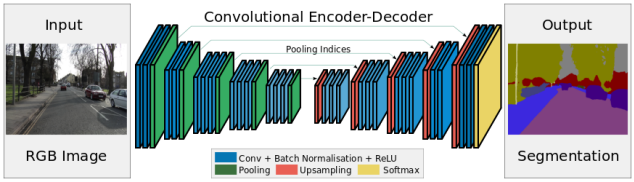

Figure 1. SEGNET architecture. *Image adapted from SEGNET project website* [22] [5]

2.3. U-NET

U-NET is a Convolutional Networks for Biomedical Image Segmentation [18] [17]. Although U-NET was developed for biomedical image segmentation, its architecture can be trained to segment other types of image. In this project, we will use U-NET to classify images from BSDS500.

U-NET architecture consists of the repeated application of two 3×3 convolutions, each followed by a rectified linear unit (ReLU) and a 2×2 max pooling operation with stride 2 for downsampling [18]. Every step in the expansive path consists of an upsampling of the feature map followed by a 2×2 convolution, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3×3 convolutions, each followed by a ReLU [18]. At the final layer a 1×1 convolution is used. In total the network has 23 convolutional layers [18]. Figure 2 presents the architecture of U-NET.

3. Data

KITTI Vision Benchmarking Suite is an project of Karlsruhe Institute of Technology and Toyota Technological Institute at Chicago to provide an real-world computer vision benchmark for autonomous driving platform Annieway [12]

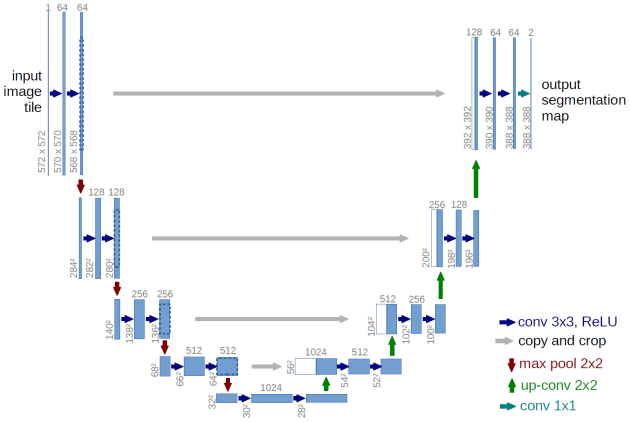

Figure 2. U-NET architecture. *Image adapted from U-NET project website* [17] [18]

[14]. KITTI contains benchmarks and datasets for the following area of interests: stereo, optical flow, visual odometry, 3D object detection and 3D tracking [14].

One of the benchmark in KITTI suite is the Road/Lane Dataset Evaluation [11]. The road and lane estimation benchmark consists of 289 training and 290 test images, in four different categories of road scenes [11]:

- uu - urban unmarked (98 training images and 100 test images) [11];
- um - urban marked (95 training images and 96 test images) [11];
- umm - urban multiple marked lanes (96 training images and 94 test images) [11];
- urban - combination of the three above [11].

Ground truth has been generated by manual annotation of the images and is available for two different road terrain types: road - the road area (the composition of all lanes), and lane (the ego-lane, the lane the vehicle is currently driving on) [11] [14]. Ground truth is provided for training images only [11].

As the dataset do not provide test groundtruth, the results must be evaluated using a benchmarking tool provided with the dataset [14]. This tool performs road and lane estimation in the bird's-eye-view space [11] [14]. The metrics used are Maximum F1-measure, Average precision as used in PASCAL VOC challenges, Precision, Recall, False Positive Rate, False Negative Rate, F1 score and Hit Rate [11] [14].

3.1. Other Datasets

3.1.1 CamVid Dataset

3.1.2 BSDS500 Dataset

Berkeley Segmentation Data Set contains 500 natural images and its respective ground-truths, annotated by humans [4]. The images are explicitly separated into disjoint train, validation and test subsets [4].

To evaluate the quality of the segmentation methods, the results will be evaluated with BSDS500 benchmarking tool, provided with the Dataset [4]. BSDS500 dataset uses the Precision and Recall Method to evaluate the results [4].

4. Methods

4.1. Transfer Learning

Transfer learning is a technique in machine learning that stores knowledge gained while solving one problem, adapt and apply it to a different but related problem. As the growing of neural networks usage, it becomes reasonable to seek out methods that avoid "reinventing the wheel", and instead are able to build on previously trained networks' results [16] [24].

In this work is expected to use transfer learning to speed up the training process. For that, it will be used a pre-trained VGG-16 (Very Deep Convolutional Networks for Large-Scale Image Recognition) [20]. The pre-trained VGG-16 will be provided by Keras, a Python Deep Learning Library [6]. Keras is a high-level neural networks API of running on top of TensorFlow [1], CNTK [19], or Theano [21] [6].

VGG-16 provided by Keras contains weights pre-trained on ImageNet Dataset [7]. ImageNet contains a fixed-size 224 224 RGB image. VGG-16 passes the image through a stack of convolutional (conv.) layers, where it's used filters with a very small receptive field: 3×3 [20]. The convolution stride is fixed to 1 pixel and the spatial padding of convolutional layer input is such that the spatial resolution is preserved after convolution [20]. Spatial pooling is carried out by five max-pooling layers, which follow some of the convolutional layers [20]. Max-pooling, them, is performed over a 2×2 pixel window, with stride 2 [20]. Figure 5 presents the architecture of VGG-16.

To use transfer learning, it will be needed to make some data transformations. First, it will need to adapt SEGNET architecture to uses VGG-16 weights. Also, the BSDS500 images and ground-truths will be reduced to the shape of 224×224. As BSDS500 images are not square, a process step will transform the shape, adding black borders before reducing the image. This step will avoid deformations in the shape of the image and the ground-truth.

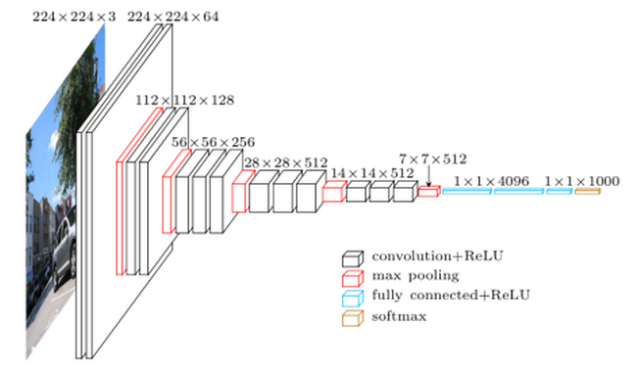

Figure 3. VGG-16 architecture. *Image adapted from TechKingdom website[13]*

4.2. Data Augmentation

Data augmentation consists of a range of transformations that can be applied to the dataset to increase the number of data with the target of improving the accuracy and robustness of classifiers [9]. The problem with small datasets is that models trained with them do not generalize well [15].

Data augmentation also can act as a regularizer in preventing overfitting in neural networks and improve performance in imbalanced class problems [25]. According to Wong et al. [25], data augmentation is better to perform in data-space instead of feature-space, as long as label preserving transforms are known [25].

Once BSDS500 contains only 200 images for training and 100 images for validation, the Neural Network may not generalize well and learn enough information from the dataset. Then, it's necessary to provide a range of transformation to add some generated images for training and validation.

To provide data augmentation, the images and the ground-truth will be rotated 12 times, 30 degrees each. Also, the images will be flipped and rotated 12 times each. Then, each image will transform into 24 possible images. Then, 200 images for the training set will become 4800 training images and the validation set will contains 2400 images. The number of images is not too big but can help the DNN predict with more accuracy.

4.3. Segnet Model

4.4. U-NET Model

5. Experiments

6. Conclusion

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia,

R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, Nov 2012.

[3] S. A. Ahmed, S. Dey, and K. K. Sarma. Image texture classification using artificial neural network (ann). In *2011 2nd National Conference on Emerging Trends and Applications in Computer Science*, pages 1–4, March 2011.

[4] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(5):898–916, 5 2011.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, abs/1511.00561, 2015.

[6] F. Chollet et al. Keras. `https://keras.io`, 2015.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[8] D. Domnguez and R. R. Morales. *Image Segmentation: Advances*, volume 1. 2016.

[9] A. Fawzi, H. Samulowitz, D. Turaga, and P. Frossard. Adaptive data augmentation for image classification. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3688–3692, Sept 2016.

[10] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, Sep 2004.

[11] J. Fritsch, T. Kuehnl, and A. Geiger. A new performance measure and evaluation benchmark for road detection algorithms. In *International Conference on Intelligent Transportation Systems (ITSC)*, 2013.

[12] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[13] P. Kochakarn. Machine learning with python: Image classifier using vgg16 model - part 1: Theory. `https://www.techkingdom.org/single-post/2017/11/07/Machine-Learning-with-Python-Image-Classifier-using-VGG16-Model---Coming-Soon`, 2017.

[14] K. I. of Technology and T. T. I. at Chicago. The kitti vision benchmark suite. `http://www.cvlibs.net/datasets/kitti/`, 2018.

[15] L. Perez and J. Wang. The effectiveness of data augmentation in image classification using deep learning. *CoRR*, abs/1712.04621, 2017.

[16] L. Y. Pratt. Discriminability-based transfer between neural networks. In *Proceedings of the 5th International Conference on Neural Information Processing Systems*, NIPS'92, pages 204–211, San Francisco, CA, USA, 1992. Morgan Kaufmann Publishers Inc.

[17] V. P. Recognition and F. o. E. Image Processing, Dept. of Computer Science. U-net: Convolutional networks for biomedical image segmentation. `https://lmb.informatik.uni-freiburg.de/people/ronneber/u-net/`, 2018.

[18] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pages 234–241. Springer, 2015. (available on arXiv:1505.04597 [cs.CV]).

[19] F. Seide and A. Agarwal. Cntk: Microsoft's open-source deep-learning toolkit. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 2135–2135, New York, NY, USA, 2016. ACM.

[20] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[21] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, 5 2016.

[22] C. Vision and U. Robotics Group at the University of Cambridge. Segnet. `http://mi.eng.cam.ac.uk/projects/segnet/`, 2018.

[23] M. Wang, X. Liu, Y. Gao, X. Ma, and N. Q. Soomro. Superpixel segmentation: A benchmark. *Signal Processing: Image Communication*, 56:28 – 39, 2017.

[24] K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, May 2016.

[25] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell. Understanding data augmentation for classification: when to warp? *CoRR*, abs/1609.08764, 2016.

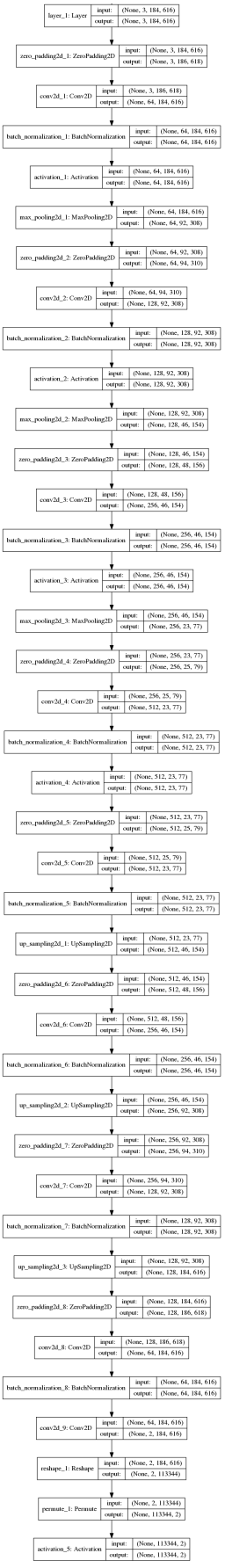

Figure 4. Segnet architecture for KITTI Road/Lane Dataset.

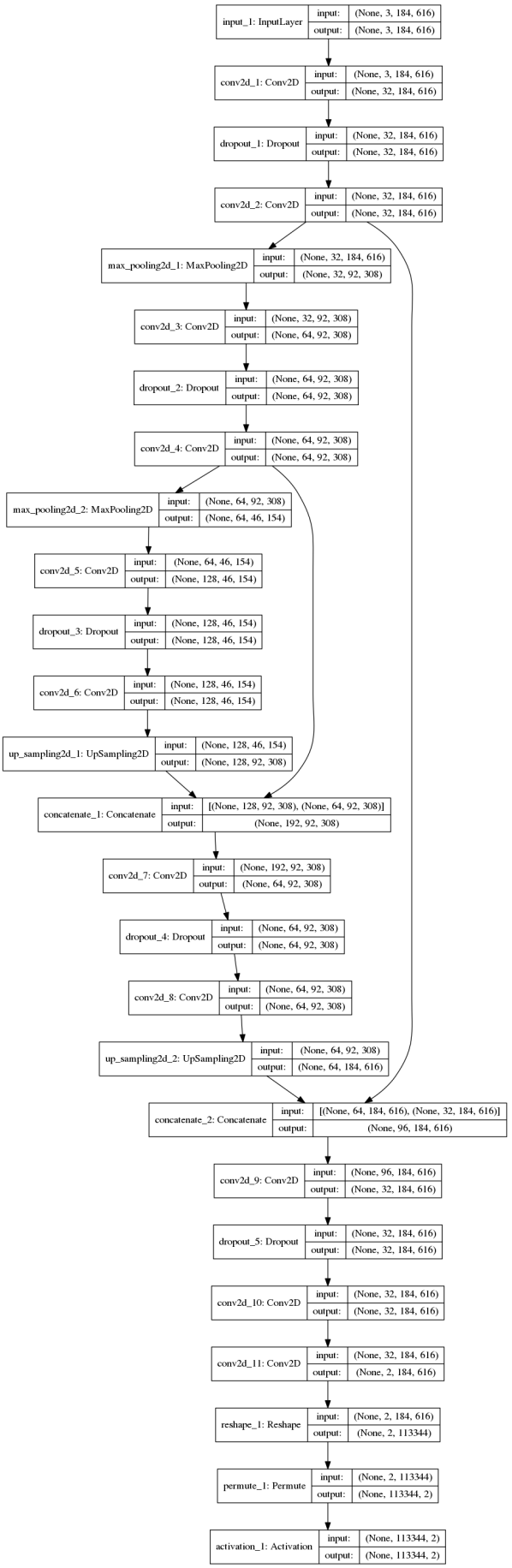

Figure 5. U-NET architecture for KITTI Road/Lane Dataset.