Comparison of Deep Learning Models for Semantic Segmentation on Domain Specific Data in Food Processing

Nicolas Loerbroks   Piyawat Suwanvithaya   Isabel Schwende*   Marko Simic
Elie Gatete Magambo

LeapMind Inc.
Tokyo, Japan
www.leapmind.io

1. Introduction

In recent years, deep convolutional neural networks (CNN) have set the state-of-the-art for semantic segmentation. The idea of first downsampling (encoding) the input image with convolutional layers, before upsampling (decoding) it to arrive back at the original image resolution was first proposed in [9] and later in [1]. In the former approach called FCN, the upsampling layers are learned, while in the latter called SegNet, upsampling makes use of the pooling indices obtained during encoding. A different approach was applied in PixelNet [2] where instead of upsampling, features are processed for every pixel individually.

However, the reported results are commonly based on large public datasets covering images in the indoor/outdoor or medical domains [7] [11] [5] [3][13] [12] [8] [10] while the performance of these methods on limited domain specific datasets remains an open question to both the research community and practitioners.

In this paper, we present experimental results obtained using deep semantic segmentation for a domain specific task. The aim of the task is an accurate localization of certain bones of the leg part of raw pork meat to automate an essential aspect of the food processing pipeline. For this particular dataset it is crucial for the model to incorporate both very fine-grained and overall positional features and we argue that for that reason PixelNet [2] is preferable over the more typical Encoder-Decoder approach. We show PixelNet's superior performance to the SegNet architecture [1]. We further evaluate the influence of dataset size, data augmentation techniques, and pretraining [6].

2. Dataset

The dataset consists of 1915 photographic images in resolution 1024x768 of the leg part of pork meat. The dataset is split into 1,852 images for training and 63 images for testing. All images contain 2 labeled objects of interest: a small bone in the center part (hip bone, blue) and a lengthy larger bone in the lower part (tail bone, red) (see Figure 1). An accurate identification of the object boundaries is very challenging even to the human eye. Manual annotation was performed by experts. Furthermore, this dataset is challenging from a deep learning perspective, since (1) the relative position between the 2 objects is fixed (hip bone always above tail bone), and (2) there are other bones present in the image that have to be classified as background. Therefore, we need to build a model that is both able to learn very high level semantic features of the images, and able to create a very fine-grained segmentation mask around the objects.

3. SegNet and PixelNet

SegNet is a typical representative of the Encoder-Decoder approach to semantic segmentation which is also applied in the current state-of-the-art [4]. PixelNet stands out compared to those approaches as it does not require upsampling. Instead, during inference, for every pixel, the feature value corresponding to it after every convolutional layer are concatenated to hypercolumns that are then fed to a series of FC-layers. In this way the final prediction is more directly based on the early layers of the CNN (including the input), rather than on the learned statistics of the images when using the decoding approach. This allows for a more fine-grained prediction around the edges which is highly desired for our given dataset.

4. Experiments

We train both the SegNet and PixelNet architectures with the VGG16 architecture and investigate the effects of pretraining, data augmentation and amount of training data. To investigate the latter we employ a subset of the training set

[*] Currently affiliated with Mobius Labs GmbH, Berlin, Germany

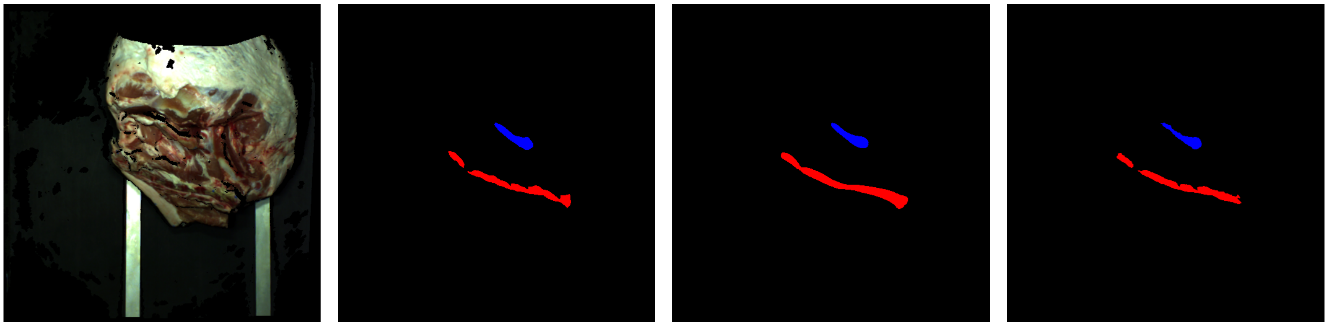
(a) Input image (b) Ground truth labels (c) SegNet output (d) PixelNet output

Figure 1: Exemplary qualitative experimental results for SegNet and PixelNet (best viewed in color and high resolution).

Experimental setup				Train	Test
Model	Pre-trained	Aug-mented	# train images	mIoU in %	mIoU in %
SegNet	-	-	756	82.2	55.5
SegNet	Y	-	756	78.6	69.6
SegNet	Y	Y	756	76.6	71.2
SegNet	Y	Y	1,852	72.5	73.4
PixelNet	-	-	756	82.8	49.8
PixelNet	Y	-	756	84.6	70.3
PixelNet	Y	-	1,852	76.3	73.1
PixelNet	Y	Y	756	78.4	71.6
PixelNet	Y	Y	1,852	80.2	80.2

Table 1: mIoU results obtained with different experimental configurations.

containing 756 images. When indicated we apply on-the-fly data augmentation employing color, brightness, contrast and sharpness changes as well as rotation up to 15 degrees (flipping the image by 90 degrees was found to decrease performance which is expected as it compromises the relative positions between the two objects). All images are normalized and cropped to a resolution of 512x512 due to memory constraints. For pretrained models the convolutional layers are initialized with weights obtained by training on PASCAL VOC [7] for which the convolutional layers were in turn pretrained on ImageNet as it is common practice. We evaluate the models using the mean Intersection over Union (mIoU) of the two target classes. All models are trained for up to 50 epochs until the IoU on the validation set (random subset of test set) is saturated.

4.1. Results and Discussion

Comparing Figure 1(c) with 1(d), it can be seen that PixelNet provides a more fine-grained segmentation mask. Furthermore, Table 1 shows consistently larger mIoU values for the Pixelnet models over the SegNet ones. Pretraining on PASCAL VOC has a significant effect on the performance of the models resulting in a jump of 14% and 20% in mIoU for SegNet and PixelNet respectively, which is somewhat surprising considering the immense differences in domain between PASCAL VOC and the given dataset. Doubling the amount of training data only slightly increases the performance for SegNet, but has a very severe effect on the performance of PixelNet. Additionally, in the case of no pretraining and less data, PixelNet overfits greatly while train and test accuracy are aligned in the case of pretraining and data augmentation on the large dataset. This again showcases the expressive power of the PixelNet model.

5. Conclusion and Future work

In this work we have successfully applied deep learning for semantic segmentation to a challenging industry application in the domain of food processing. We compared the SegNet and PixelNet architectures and showcased the benefits of PixelNet for a dataset that requires both very fine-grained and high level features to arrive at good predictions. We further showed the importance of pretraining for an extremely domain specific dataset.

In future work we are planning to additionally compare the performance on this dataset with more recent segmentation methods, particularly DeepLab-v3 [4]. Another aspect will be the deployment of the models to edge devices such as FPGA or mobile GPU to allow for real-time, low cost processing in which case the computational constraints also need to taken into account when choosing an appropriate architecture.

6. Acknowledgement

We would like to thank our customer MAYEKAWA MFG. for entrusting us with this project, providing us with the data and allowing us to publish our findings.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[2] A. Bansal, X. Chen, B. Russell, A. Gupta, and D. Ramanan. Pixelnet: Representation of the pixels, by the pixels, and for the pixels. *arXiv:1702.06506*, 2017.

[3] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, 30(2):88–97, 2009.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[5] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset. In *CVPR Workshop on the Future of Datasets in Vision*, volume 1, page 3, 2015.

[6] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660, 2010.

[7] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal voc challenge 2007, 2007.

[8] M. Hatt, B. Laurent, A. Ouahabi, H. Fayad, S. Tan, L. Li, W. Lu, V. Jaouen, C. Tauber, J. Czakon, et al. The first miccai challenge on pet tumor segmentation. *Medical image analysis*, 44:177–195, 2018.

[9] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional models for semantic segmentation. In *CVPR*, 2015.

[10] O. Maier, B. H. Menze, J. von der Gablentz, L. Häni, M. P. Heinrich, M. Liebrand, S. Winzeck, A. Basit, P. Bentley, L. Chen, et al. Isles 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral mri. *Medical image analysis*, 35:250–269, 2017.

[11] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014.

[12] K. Sirinukunwattana, J. P. Pluim, H. Chen, X. Qi, P.-A. Heng, Y. B. Guo, L. Y. Wang, B. J. Matuszewski, E. Bruni, U. Sanchez, et al. Gland segmentation in colon histology images: The glas challenge contest. *Medical image analysis*, 35:489–502, 2017.

[13] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010.