Semantic Segmentation of Mixed Crops using Deep Convolutional Neural Network

Anders Krogh Mortensen [a,*], Mads Dyrmann [b], Henrik Karstoft [c], Rasmus Nyholm Jørgensen [c], René Gislum [d]

[a] Department of Agroecology, Aarhus University, Aarhus, Denmark

[b] The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark

[c] Department of Engineering, Aarhus University, Aarhus, Denmark

[d] Department of Agroecology, Aarhus University, Slagelse, Denmark

[*] Corresponding author. Email: anmo@agro.au.dk

Abstract

Estimation of in-field biomass and crop composition is important for both farmers and researchers. Using close-up high resolution images of the crops, crop species can be distinguished using image processing.

In the current study, deep convolutional neural networks for semantic segmentation (or pixel-wise classification) of cluttered classes in RGB images was explored in case of catch crops and volunteer barley cereal. The dataset consisted of RGB images from a plot trial using oil radish as catch crops in barley. The images were captured using a high-end consumer camera mounted on a tractor. The images were manually annotated in 7 classes: oil radish, barley, weed, stump, soil, equipment and unknown. Data argumentation was used to artificially increase the dataset by transposing and flipping the images. A modified version of VGG-16 deep neural network was used. First, the last fully-connected layers were converted to convolutional layer and the depth was modified to cope with our number of classes. Secondly, a deconvolutional layer with a 32 stride was added between the last fully-connected layer and the softmax classification layer to ensure that the output layer has the same size as the input.

Preliminary results using this network show a pixel accuracy of 79% and a frequency weighted intersection over union of 66%. These preliminary results indicate great potential in deep convolutional networks for segmentation of plant species in cluttered RGB images.

Keywords: Pixel-wise classification, deep learning, computer vision

1. Introduction

Estimating total crop biomass or individual crop components in a mixed cropping system is important for agricultural research and farming. For example the assessment of either grass and/or clover biomass in a forage based feed system can be used to optimize animal feeding plans. Moreover the quantification of autumn catch crop biomass coupled with nitrogen (N) concentration will make it possible to estimate the N uptake. Nitrogen concentration can be rapidly measured through the relationship with chlorophyll content. Information of N uptake in autumn catch crops within and between fields can be used to differentiate spring N application(s) in the following crop; which should facilitate a higher N utilization and possibly reduce N leaching to ground and surface water.

Mounting a camera on either an unmanned aerial vehicle or a tractor can provide the farmer or researcher high resolution close-up images of the field and its crops. Processed, these images can give an estimation of the biomass and N-uptake in the field (Mortensen et al., 2015). In a field with mixed crops, an important first step in the process of estimating either total crop or individual crop component biomass is the ability to distinguish the different crop components. Current methods are traditionally based on hand crafted features and/or morphology and are as a result either very slow (Mortensen et al., 2015) or very low capacity (Himstedt, 2009).

Recent development in deep learning and in particular deep convolutional neural networks have shown impressive results in tasks such as image classification, speech recognition and lately in semantic segmentation (or pixel-wise classification).

In this paper, we explore convolutional neural networks for semantic segmentation in the context of mixed crops. In contrast to regular scenes, images of mixed crops are often much more complex and cluttered, but with fewer classes. We used a state-of-the-art convolutional network architecture for semantic segmentation (Long et al., 2015) to explore the potential and challenges in this case. The performance was evaluated on images of an oil radish plot trial with barley, grass, weed, stump and soil.

2. Materials and Methods

2.1. Image acquicition

The dataset consists of images, which were acquired from a plot experiment at Foulum Research Center, Denmark. The full plot experiment consisted of 36 plots (9 treatments with 4 repetitions), but only the repetitions of one of the treatments was photographed. The photographed plots (3 m x 15 m) contained oil radish as a catch crop and some amounts of voluntarily seeded barley, grass, weed and stump. The plots were photographed approximately every week over a period of 8 weeks. A Sony a7 with a 35 mm lens was used to photograph the plots. The camera was mounted on a

boom on the front of a tractor at approximately 3 m height and covering approximately 3 m x 2 m on the ground (figure 1). Prior to photographing the sample areas, yellow labeling sticks were used to mark the corners of the areas. The plant samples were 0.5 x 0.5 m^2 and in the sides of the plot (figure 2). Therefore, the camera was mounted to a side to center the camera above the cutting areas as well as possible (figure 2). The camera was further mounted in a gimbal to stabilize it. Directly above the camera a RTK GPS antenna was mounted. The camera was triggered from a PC with a fixed time interval (2.6 s). The camera trigger time and the GPS position (1 Hz) were recorded by the PC for offline geotagging of the images.

Figure 1. Photograph of experimental setup mounted on a tractor. The camera was mounted to one side, since the plant samples were taken from the side of the plot experiment. Each plot was visited twice in order to photograph both sides of the plot.

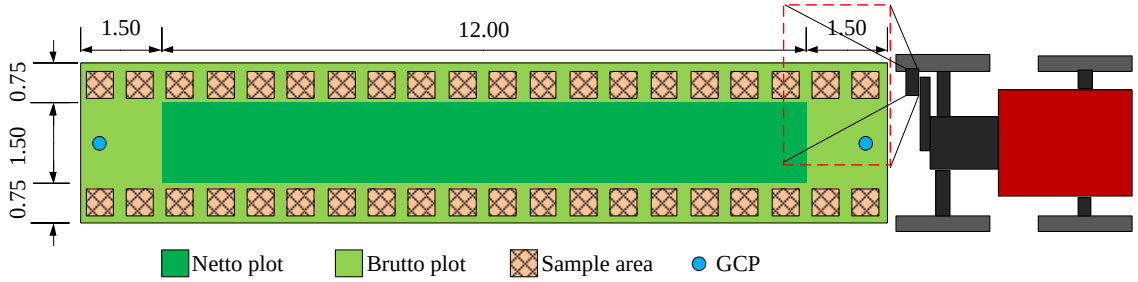

Figure 2. Overview of a photographed plot. The dashed red rectangle indicated the field of view of the camera. Driving from right to left, the top row of sample areas was photographed. Driving from left to right, the bottom row of sample areas was photographed. All numbers are in meters.

2.2. Plant samples

During the first 7 weeks, 4-5 plant samples were taken from each plot after it was photographed. In the 8th week, the plots were only photographed. The extra week of photographing the plots was made to have an after photo of all the sample areas, which could be used for easy identification of the areas. After the plant samples were collected, they were stored in a cooling room at 4°C. The samples were stored until they could be fractioned into four fractions: oil radish, barley/grass, weed and stump. After being fractioned, they were weighed and placed in a forced air drying oven at 60 °C to evaporate all water from the samples. After 48 hours, the samples were removed from the oven and weighed again. From the weight before and after drying the wet matter, dry matter and %-dry matter were calculated. For each sample the %N and %C were determined using combustion method (DUMAS). The measurements were performed using Vario EL III instrument from Elementar (www.elementar.de). Between the dry matter determination and the %N and %C measurements, the samples were placed in dry storage. Figure 3 shows boxplots of the above-ground dry matter and total N for fraction each week.

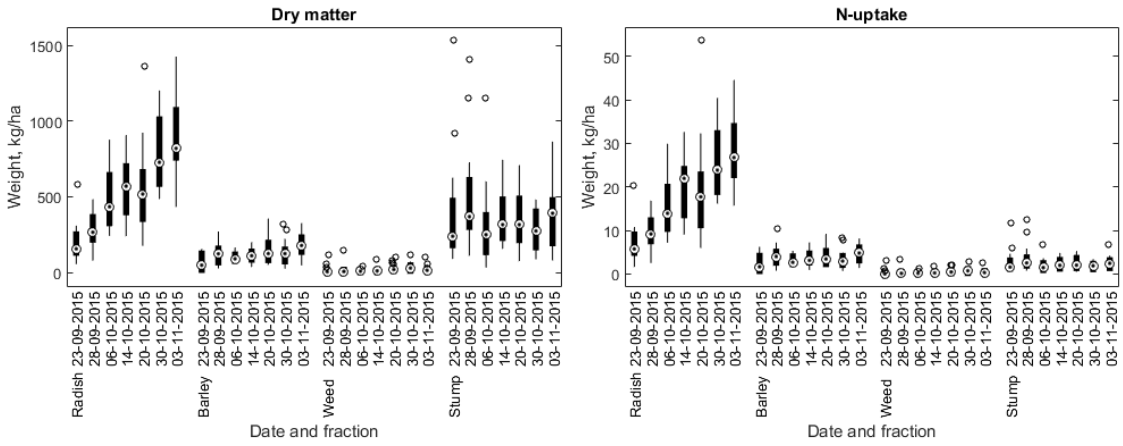

Figure 3. Boxplot of dry matter (left) and N-uptake (right) distributions as a function of sampling data and fraction.

2.3. Network architecture

Following the same procedure as Long et al. (2015), VGG-16D (Simonyan & Zisserman, 2015), which was developed for image classification, was used as a base model for the network architecture (see figure 4). It consists of 15 convolutional layers, 5 max pooling layers and a softmax classification layer. All convolutional layers use 3x3 kernels and a stride of 1. Convolutional layer 1 and 2, 3 and 4, 5-7, 8-13 and 14-15 have 64, 128, 256, 512 and 4096 kernels, respectively. In the base network model, the final three convolutional layers were fully connected layers; however, these were converted to convolutional layers for the purpose of handling arbitrary input image size. In the last convolutional layer, conv 16, the number of kernels has been changed from 1000 to 7 to reflect our number of classes. The max pooling layers all use a 2x2 kernel with a stride of 2. Due to the 5 max pooling layers with a stride of 2, the size of the output of the converted base network was down sampled by a factor of 32 (2^5= 32). To perform pixel-wise classification, a deconvolutional layer was inserted in the network after the three converted convolutional layers and before the softmax classification layer (figure 4).

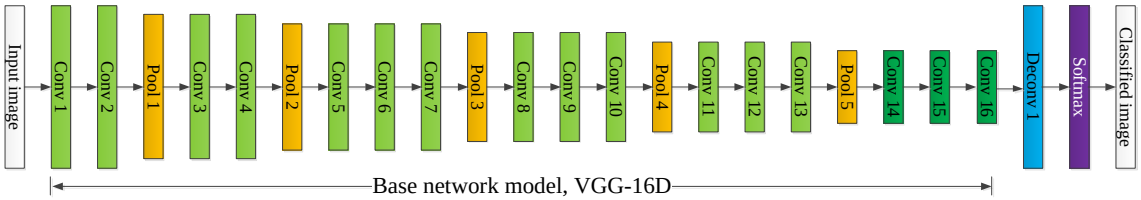

Figure 4. Network architecture. The green blocks are convolutional layers. The dark green blocks are fully connected layers in the base network model, but converted to convolutional layers in our network. Orange blocks are max pooling layers. The blue block is the deconvolutional layer. The purple block is the softmax classification layer. The first and last layer is the input image and the pixel-wise classified image, respectively.

2.4. Training

From the captured images, 1600 x 1600 pixels centered on the sample areas are extracted (figure 5). These extracts form the dataset. At the time of writing, only 3 of the images have been manually annotated resulting in a rather small dataset. To artificially increase the data set, data argumentation was used in the training. Each image was rotated 0, 90, 180 and 270 degrees as well as being flipped diagonally before performing the same set of rotations. This resulted in a factor 8 increase of the number of training examples. Due to the large memory requirements of the network, each 1600 x 1600 extract was divided into 16 patches of 400 x 400 pixels each. 75% of the 352 patches were used for training and the remaining 25% was used for soft-validation.

Instead of training the network from scratch, we transfer the learned weights from Long et at. (2015) and instead only fine-tune our network. Fine-tuning on the pre-trained weights allows for faster convergence and using a smaller dataset. A batch size of 1 training image patch, a fixed learning rate of 10^{-9} and a weight decay of $5{\cdot}10^{-4}$ were used to update the weights after each iteration. The network was validated after each epoch and the model was saved after each 10 epochs.

Figure 5. Example of images from a plot taken before (left) and after (right) taking a plant sample. The before image is capture in week 5, and the after image is captured in week 6. The solid red square indicates the sample area before and after taking the sample. The dashed square indicates an previously harvested sample area.

3. Results and Discussion

The network was set to be trained for 40 epochs, but after about 10 epochs it started to overfit and it was therefore manually terminated after 23 epochs (figure 6). The network model saved after epoch 10 was used for generating the results. Evaluating the fine-tuned network on the validation set yielded 79% pixel accuracy and 66% frequency weighted intersection over union. Comparing the ground truth annotated image to the predicted image, the coarse features of input image such as radish leafs and soil were predicted quite well. However, the finer features of the input image such as barley, grass or stump were often only captured crudely (figure 7). This is most likely due to the large effective stride of the network (32 px). A smaller stride will most likely improve the performance of the network and its ability to capture the finer features. State-of-the-art methods propose two ways: 1) extracting and DE convolve features from the earlier layers of the network and combine them with the deconvolution of the last layer (Long et al., 2015). 2) Reducing the stride of one or more of the pooling layers (Chen et al., 2014). The latter also proposes adding a conditional random field to the end of the network to improve the performance of their stride 8 network. The downside of this method is, the extra time that it takes to solve the conditional random field optimization.

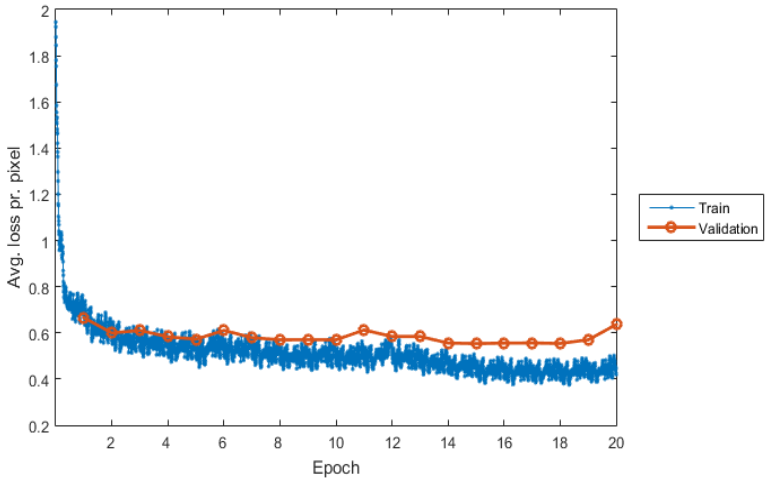

Figure 6. Training and validation loss as a function of epoch.

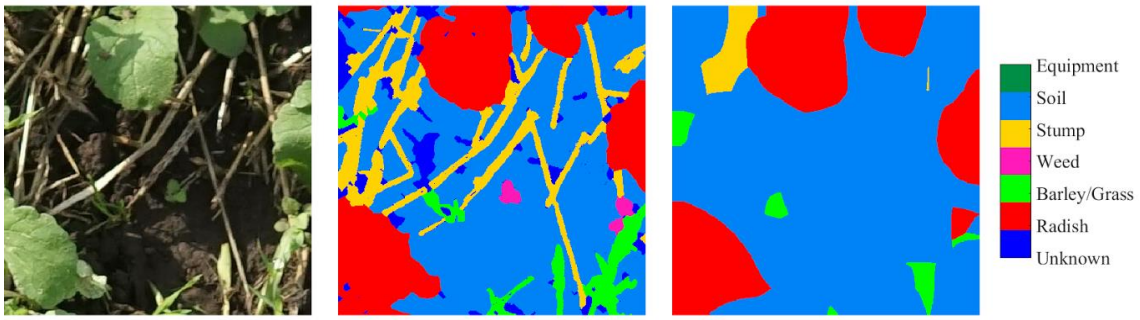

Figure 7. Example of an input image (left), ground truth (middle) and predicted pixel labels (right).

Passing the same test image through the network several times, but subject to the same 8 transformations used for the data argumentation prior to training, the network is to some extend independent on these transformations (figure 8). The coarse features such as the radish leafs, soil and areas with barley/grass were fairly consistent between the transformations. However, single straws were not detected and finer details of the objects varied greatly. This is particular evident for the stump and barley/grass areas. This is of course unfortunate, as it indicates that the orientation of the camera have an influence on the segmentation. However, these differences between the classifications can be combined to perform ensemble classification similar to Krizhevsky et al. (2012). To do so, we take each input image and perform the same transformations as for the data argumentation and feed the same input image through the network 8 times corresponding to each of the transformations. Then, the inverse transformation is applied to classified images and the final classification of the input image is found by averaging the 8 predictions. This increases both the pixel accuracy and the frequency weighted intersection over union by 1%-points for both the training and test sets.

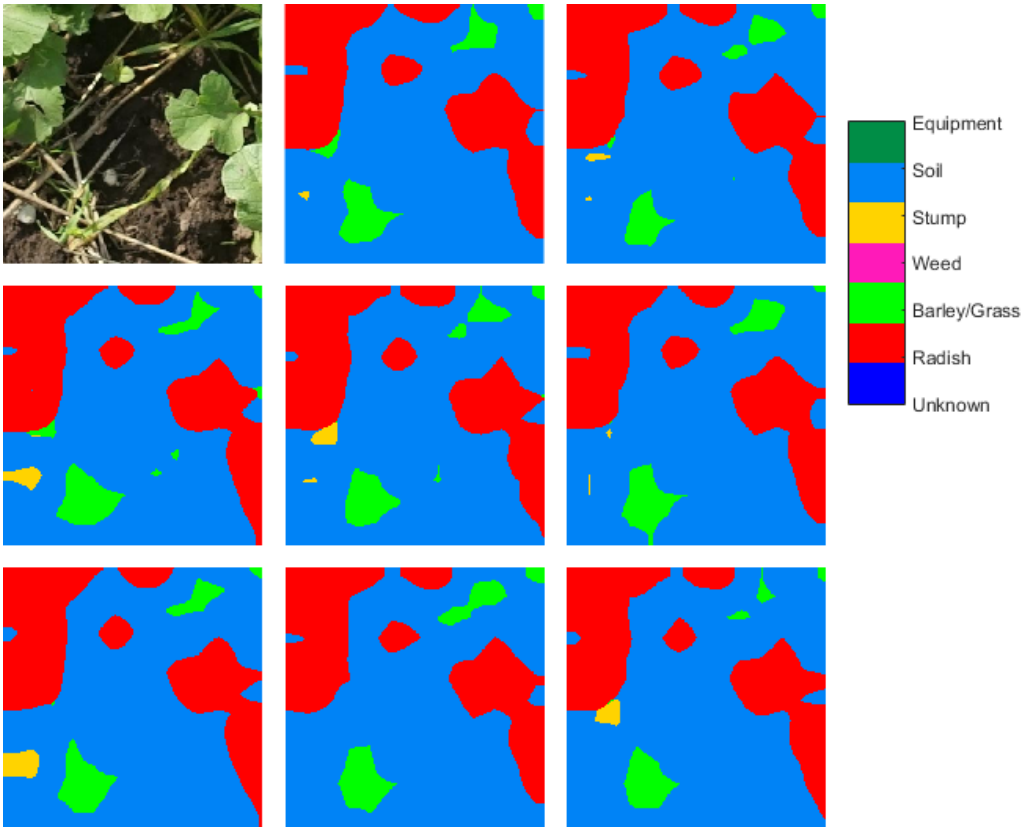

Figure 8. Example of input image (top left) and the resulting pixel-wise classification of the transformation used for data argumentation during training. The classified images have been transformed back after classification for better comparison.

4. Conclusions

This paper has explored the use of fully convolutional neural networks for semantic segmentation. These preliminary results show that the trained network models from the general semantic segmentation case can easily be used as a initialization point for fine-tuning on this special case. However, a smaller effective stride is necessary to achieve higher performance and classify the finer details on the input image. Further research and more data are required.

Acknowledgements

This research is part of the VIRKN-project and supported by The Danish Agrifish Agency (GUDP, “Grønt Udvikling- og DemonstrationsProgram”).

The authors thank the technical staff at Foulumgaard for their assistance in the data acquisition.

References

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. 2014. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. Iclr, 1–14.

Himstedt, M., Fricke, T., & Wachendorf, M. 2009. Determining the contribution of legumes in legume-grass mixtures using digital image analysis. Crop Science, 49 (October), 1910–1916.

Krizhevsky, A., Sutskever, I., & Hinton, G. 2012. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 1097–1105.

Long, J., Shelhamer, E., & Darrell, T. 2015. Fully Convolutional Networks for Semantic Segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440.

Mortensen, A.K., Gislum, R., Larsen, R. & Jørgensen, R. N., 2015. Estimation of above-ground dry matter and nitrogen uptake in catch crops using images acquired from an octocopter. In Precision Agriculture ’15. The Netherlands: Wageningen Academic Publishers, pp. 127–134.

Simonyan, K., & Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recoginition. Intl. Conf. on Learning Representations (ICLR), 1–14.