The extraction of maize lodging regions in UAV images using deep fully convolutional neural network

Ergong Zheng	extsuperscript{1*}, Yingfang Tian	extsuperscript{2}, Zhaohong Xu	extsuperscript{1} and Tao Chen	extsuperscript{3}

	extsuperscript{1}Aviation University of Air Force, Changchun, Jilin, 130022, China

	extsuperscript{2}Henan Academy of Agricultural Sciences, Zhengzhou, Henan, 450002, China

	extsuperscript{3}Jilin Branch, The People’s Insurance Company (Group) of China Limited, Changchun, Jilin, 130000, China

*Corresponding author’s e-mail: 330983114@qq.com

Abstract. This paper presents a method for automatic extraction of maize lodging regions in unmanned aerial vehicle (UAV) visible images. A deep fully convolutional neural network is used to achieve the purpose. Firstly, images are collected by using UAV remote sensing, and then the network is trained and validated on a total of more than 20000 labelled images, and finally, the network is able to accurately extract lodging regions. Experimental results demonstrate the effectiveness of the method and F1 score reaches 89.5% on test dataset. This study makes a major contribution to the assessment of maize lodging disaster.

1. Introduction

Maize lodging is a common and frequent natural disaster. Meanwhile, agricultural insurance is an effective means to resist risks and reduce the economic losses of farmers. After the maize lodging disaster, insurance companies need to send employees to investigate the disaster. However, manual investigation is laborious. UAV remote sensing meets this demand, which is a convenient way to shoot images of the maize field. The main disadvantage of current methods is that the identification of maize lodging regions in the UAV images still needs manual visual interpretation.

Since 2012, deep convolutional neural network has achieved great success in many fields such as image classification\cite{1}, object detection\cite{2}, and image semantic segmentation\cite{3}. However, little work has been devoted to agricultural UAV remote sensing by using methods of deep convolutional neural network. The extraction of maize lodging regions in UAV images can be regarded as an image semantic segmentation problem. Thus, this paper presents an image segmentation model for extracting maize lodging regions in UAV images, which is a fully convolutional neural network and is trained on a total of more than 20000 labelled images. Experimental results show that the proposed method can be served as a valuable tool of investigation and assessment of maize lodging disaster.

2. Material and methodology

In order to train the image segmentation model to be able to identify maize lodging regions, a large number of training images are required. The study site is located in the Sansheng Village, Mishazi Town, Dehui County, Changchun City, Jilin Province, China. On September 14, 2016, about 2000 images with size of 5456×3632 pixels were acquired by using a quad-rotor UAV at a flight relative height of 170 m and 4 air routes. The ground sampling distance is no more than 5 cm/pixel. 316 images from the first air route are selected for construction of training and validation dataset, and the maize
lodging regions have been annotated manually. The original resolution images are cropped with 50% overlapping into images with size of 512×512 pixels, and images without maize lodging are discarded. Totally, 20662 images are obtained, and 20% of them are randomly split into validation dataset. At last, the training dataset consists of 16530 images, and the validation dataset consists of 4132 images. The test dataset consists of 100 images from the fourth air route with size of 5456×3632 pixels. Image samples from the test dataset are shown in figure 1.

Figure 1. Two samples from the test dataset. Notice the maize lodging regions in the images

2.1 Network architecture

The proposed fully convolutional neural network consists of three parts: an encoder, a decoder and an attention block, as illustrated in figure 2. The encoder is based on pre-trained EfficientNet-B5, a member of Efficient Nets family by Google, which are designed by neural architecture search and achieve much better accuracy and efficiency than previous convolutional neural networks [4]. To construct the encoder, the final pooling layer and fully connected layer have been removed. Previous studies have reported the effectiveness of transfer learning for neural networks training [5]. Using a pre-trained encoder is beneficial for the network to converge faster and get better weights. The decoder is the same as the decoder module proposed in Link Net [6], which doubles the size of the feature maps to output an image whose size coincides with the input image and where each pixel corresponds to a probability of lodging. The larger the pixel values in the output image are, the more likely they are lodging regions. Afterward, the lodging regions can be easily extracted by
thresholding the probability map. Recent evidence suggests that self-attention mechanism can be flexibly embedded in convolutional neural networks to improve the accuracy with little extra computational cost. In this paper, in addition to the channel attention module proposed in [7], the spatial attention module is used to construct our attention block. Furthermore, unlike previous work that embeds attention modules in encoder or decoder [8], our attention modules are embedded in skip connection between encoder and decoder to make full use of multi-scale semantic information from encoder to refine the segmentation results.

2.2 Network losses

The sum of two loss functions is used for optimizing the network weights during the training phase. These losses are a classification loss and a shape-constraints loss. Since image segmentation task can be considered as a pixel binary classification problem, we can use the common loss function for binary classification tasks - binary cross entropy that is defined as:

$$\text{Loss}_{\text{BCE}} = -\frac{1}{n} \sum_{i=1}^{n} (y_i \log \hat{y}_i + (1 - y_i) \log (1 - \hat{y}_i))$$  \hspace{1cm} (1)

where \( y_i \) is a binary label of the corresponding pixel \( i \) in the labelled image and \( \hat{y}_i \) is the predicted probability for the pixel in the output image of neural network.

The shape-constraints loss is the Jaccard loss that is defined as:

$$\text{Loss}_{\text{Jaccard}} = -\log \left( \frac{1}{n} \sum_{i=1}^{n} \left( \frac{y_i \hat{y}_i}{y_i + \hat{y}_i - y_i \hat{y}_i} \right) \right)$$  \hspace{1cm} (2)

The Jaccard loss can be interpreted as maximization of the intersection between masks and corresponding predictions. The final loss function is obtained by combing equation (1) and equation (2) as following:

$$\text{Loss} = \alpha \cdot \text{Loss}_{\text{BCE}} + (1 - \alpha) \cdot \text{Loss}_{\text{Jaccard}}$$  \hspace{1cm} (3)

where \( \alpha = 0.7 \) in order to balance the values of two loss functions just like the work in [9].

2.3 Evaluation

To test the efficacy of the network, we choose precision, recall and \( F_1 \) score as evaluation metrics. In this study, the foreground pixels are pixels in maize lodging regions. The precision indicates the proportion of true positives relative to all predicted foreground pixels. The recall indicates the proportion of true positives relative to all annotated foreground pixels. The \( F_1 \) score, as shown in equation (4), is a harmonic mean of precision and recall. The higher values of these metrics indicate the better performance of the neural network model.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$  \hspace{1cm} (4)

3. Results

Since the original test images are too large to be fed into the neural network at one time, an overlap-tile strategy is employed for seamless segmentation of arbitrary large images. Small areas in the binary segmentation images are filtered out by morphological post-processing. Finally, the performance of extraction of maize lodging regions was tested on 100 test images. This yielded a precision of 91.5\% and a recall of 87.5\%. The average \( F_1 \) score on test dataset is 89.5\%. Figure 3 shows the extraction results of lodging regions for two samples in figure 1. We can see that the proposed method is capable of extracting both strip-like and surface-like regions of maize lodging.
4. Conclusion
In this paper, we present a method for extracting regions of maize lodging in UAV images using deep fully convolutional neural network. The network has been trained by an end-to-end manner with thousands of annotated images. Experimental results show that the proposed algorithm is able to reliably extract maize lodging regions in UAV images. This study makes an attempt on combination of agricultural UAV remote sensing and deep neural network, and has potential application in investigation and assessment of maize lodging disaster. Due to the lack of lodging images of maize indifferent growth stages, although prediction ability of neural network model is available, the generalization ability needs further study in future.

Acknowledgments
This study was supported by the science and technology projects in Jilin Province Department of Education under grant No.2016-515.

References
[1] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. In: Conference and Workshop on Neural Information Processing Systems. Lake Tahoe. pp. 1097-1105.
[2] Redmon, J., Divvala, S., Girshick, R., et al. (2016) You Only Look Once: Unified, Real-time Object Detection. In: IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas. pp. 779-788.
[3] Long, J., Shelhamer, E., Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. Boston. pp. 3431-3440.
[4] Tan M., Le Q.V. (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv preprint arXiv:1905.11946.
[5] Ruber S. (2018) Transfer learning-machine learning’s next frontier. http://ruder.io/transfer-learning.
[6] Chaurasia A., Culurciello E. (2017) LinkNet: Exploiting Encoder Representations for Efficient Semantic Segmentation. In: IEEE Visual Communications and Image Processing. St. Petersburg. pp. 518-521.
[7] Hu J., Shen L., Sun G., et al. (2018) Squeeze-and-Excitation Networks. In: IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City. pp. 7132-7141.
[8] Fu J., Liu J., Tian H., et al. (2019) Dual Attention Network for Scene Segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. Long Beach. pp. 3146-3154.
[9] Iglovikov V., Seferbekov S., Buslaev A., et al. (2018) TernausNetV2: Fully Convolutional Network for Instance Segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City. pp. 233-237.