Image Classification for Soybean and Weeds Based on ViT

Jingxin Liang, Dong Wang* and Xufeng Ling

Shanghai Normal University Tianhua College AI School, No. 1661 North Shengxin Road, China
Email: jxliang999@126.com (Liang J X), 316467252@qq.com (Wang D)

Abstracts. In this paper, ViT deep neural network based on self-attention mechanism is used in classification for images of soybean and weeds. Firstly, the overall image is split into multiple tiles; with each tile regarded as a word, the whole image is regarded as a sentence, which can be used for image semantic recognition by natural language processing technology. We designed a ViT network with sequence length of 50, embedded dimension of 384, and self-attention module layers of 12. With soybean weed classification dataset, the network is trained, verified and tested. Experimental results showed that ViT network is superior in classification on dataset of soybean and weeds, with excellent generalization capability.

Keywords: Soybean and weeds, Classification, Self-Attention Mechanism, Deep Neural Network

1. Introduction

Weeds are one of the important reasons for soybean yield reduction. The most effective way to remove weeds is to spray herbicides. According to the type of weeds, spraying accurately with the corresponding herbicides is the key to improve the weeding efficiency, reduce the use of pesticides and build environment-friendly agriculture. By artificial intelligence image classification technology, weed types are identified to help improve the recognition rate, reduce labour intensity and improve efficiency, enabling weeder machine to remove weeds automatically. Researchers have proposed a variety of image classification algorithms, mainly including traditional machine-vision methods and deep neural network methods [1-7], to identify weed types.

Machine vision classification usually includes two steps: feature extraction and pattern recognition. Dongjian He et al. [2] proposed a multi-feature fusion method, using SVM (Support Vector Machine) to identify weeds by leaf shape, texture and other features. Chuanyuan Zhao et al. [3] transformed the multispectral image into lab color space, extracted a variety of leaf features, and used C4.5 algorithm to identify weeds. The above methods for weed recognition are based on feature extraction and classifier. Although good results have been achieved under specific conditions, there are still some problems such as high requirements for image quality, insufficient robustness, relatively low recognition rate and poor extensibility.

As an important field of machine learning, CNN (Convolutional Neural Network) can automatically learn the feature representations of the image data via large scale training without need for manually feature design. Various classic CNN image classification models, such as ResNet, VGGNet, have good recognition accuracy. With multiple convolution and aggregation layers, CNN can extract multi-level features of the whole or local region of an image. In recent years, CNN has been used to classify weeds and remarkable results are achieved. Xiangwu Deng et al. [4] proposed a rice seedling weed recognition method based on a combination of CNN and transfer learning, which
can transfer the parameters of the pre-trained CNN image classification model to the rice seedling weed recognition scenario, and realize the automatic recognition of rice seedling weeds under natural light and complex field background. To identify weeds in rape field automatically, Zhang Le et al. [5] proposed a weed identification method based on Faster R-CNN network, which showed obvious advantages in the object recognition of rape and weed. The accuracy of object recognition of rape and weed was 83.9%, while recall rate was 78.86% and F1 value was 81.30%. Gui Yue et al. [6] proposed a method of farmland weed classification and recognition based on CNN. The experimental results on farmland weed and rice datasets show that the network can effectively classify weeds and rice with accuracy above 85%.

Transformer has achieved great success in Natural Language Processing (NLP), such as BERT and GPT2 models. Recently, researchers have introduced transformer into the field of computer vision and achieved promising results [8-9]. At present, the application of transformer in three image subjects: classification (ViT), detection (DETR) and segmentation (SETR) has achieved excellent results.

2. Methods

In this paper, a method using vision transformer is designed to recognize soybean weeds. The processing steps are as follows: first, the input image is scaled to 224x224, and then divided into 196 tiles of size 16x16. We regard each tile as a word, and the whole image as a sentence of 196 words. Using the mature multi-layer transformer algorithm, the semantic recognition of this sentence is equivalent to semantic recognition of the corresponding image. The ViT network constructed in this paper is composed of: word embedding layer, positional encoding, transformer module (12 self-attention layers stacked one after another), dropout layer, fully-connected layer MLP (mapping high-dimension tensor to object classes). The ViT network are trained by a huge number of soybean weeds images to get a classifier with superior performance. The modules’ descriptions are as follows:

2.1. Embedding and Positional Encoding

Word embedding, which means a high-dimension word with one-hot encoding is mapped to a much lower-dimension tensor to achieve dimension reduction and feature extraction at the same time, is an import concept of NLP. In the ViT model, the image of size 224x224 is split into 196 tiles of 16x16 pixels. Each tile is mapped into a 384-dimension tensor through the embedding layer.

Positional encoding, which enables ViT model to learn image semantic information, mainly establishes the relative position relationship between image tiles. It should be noted that in the ViT model, the location coding is obtained by learning, and is different from the manual location coding in NLP.

2.2. Self-Attention

Self-attention is the basic module of transformer. It works out the relationship between each word of input sequence. After learning and synthesizing the overall information of the input sequence, self-attention is able to predict a single missing word from the rest of the sequence. With similar mode like above, self-supervised learning, such as cloze based on BERT and forward prediction based on GPT2, is realized.

With queries as Q, index keys as K and value as V, the input sequence X is mapped to Q, K and V by multiplying X with matrix Wq, Wk and Wv; output self-attention as:

$$Z = \text{softmax}\left(\frac{QK^T}{\sqrt{d_q}}\right)V$$

It should be noted that Q, K and V is obtained through training. In order to get multiple relationships at different positions in the sequence, multi-head attention is introduced, and each head corresponds to a relationship. In this paper, eight heads are used, as shown in figure 1.
2.3. **Transformer Model**

Transformer is built on several self-attention modules. In this paper, the transformer is composed of 12 self-attention modules. The dimension of the embedded word is set to 384, while one word is divided into 8 heads and dimension of each head is 96. The experimental results show that the lower-level self-attention modules represent the basic features of the sequence, while the higher-level ones represent more complex relationships. The diagram of transformer is shown in Figure 2.

![Figure 1. Multi-Head Attention Diagram.](image)

![Figure 2. Diagram of Transformer Encoder.](image)

2.4. **Vision Transformer (ViT)**

Vision transformer (ViT) directly applies pure transformer architecture to a series of image tiles for classification. The main process is as follows:

1. Image to tiles: the image is resized into a fixed size (e.g., 224x224), and then split into fixed size tiles (e.g., 16x16)

2. Word embedding: concatenated tiles, multiplied with an embedding matrix, are fed into transformer

3. Training: build a pre-trained model through a huge amount of training. At present, there are some pre-trained models, such as B_16, B_16_imagenet1k, B_32, B_32_imagenet1k, L_16_imagenet1k, L_32, L_32_imagenet1k.

4. Application: based on the pre-trained model, we perform transfer learning, whose work is mainly about the fine-tuning of the last full-connection layer, on image of soybean weeds. Following the front-end pre-training model as a feature extraction network, which outputs the image semantic features, a full-connection classifier is used to perform the image classifications. The system diagram of ViT is shown in Figure 3 as illustration of descriptions above.
3. Experiments and Analysis

3.1. Dataset Selection
In this paper, we use the open dataset of Weed Detection in Soybean Crops, which contains 10896 RGB images of soybean and three kinds of weeds. 60% of all images were randomly selected as the training dataset, 20% as the verification set, and the rest as the test set. Each image is resized to the same size of 224*224.

3.2. Data Augmentation
The training dataset is augmented by random resized clipping and random horizontal flipping. Finally, the tensor is normalized to balance the gray-scale level and dynamic range of the image.

3.3. Parameters and Training
The parameters of the model: batch size as 64, final epoch number as 500, optimizer as Adm, and the learning rate as 3e-5.

(a) loss for training and validation.  
(b) validation accuracy.

Figure 4. ViT train and validation.
The loss curve of the model for training and validation dataset is shown in figure 4(a), and the accuracy curve for validation dataset is shown in figure 4(b). The accuracy of the model for the validation dataset reaches an extremely high score as 99.8%.

3.4. Comparisons
We compared the ViT model with ResNet and VGG on the Pytorth platform. The results are shown in table 1.

| Margin         | ViT      | ResNet   |
|----------------|----------|----------|
| Accuracy (%)   | 99.8     | 97.1     |
| Epochs         | 500      | 160      |
| Parameter Scale (M) | 21.4     | 25.6     |

Compared with CNN models in comparative scale, ViT model has higher accuracy, whereas the disadvantages are the longer time for training, as shown in figure 5.

![Figure 5. ResNet train and validation.](image)

4. Conclusion
In this paper, a ViT classification network is constructed, trained and verified using soybean weed dataset. The experimental results showed that the accuracy of ViT classification model is significantly higher than that of ResNet or VGG models based on CNN, which indicates that ViT model has promising value in the field weed recognition application. The future work is to optimize the model on more datasets, and integrate it into the actual weeder machine.

References
[1] Ren Q H and Yang B H 2020 Application of image processing technology in weed recognition in field Journal of Chinese Agricultural Mechanization 025(06).
[2] He D J, Qiao Y L, Li P, Gao Zh, Li H Y and Tang J L 2013 Weed recognition based on SVM-DS multi-feature fusion Transactions of the Chinese Society for Agricultural Machinery 34(02).
[3] Zhao Ch Y, He D J and Qiao Y J 2013 Transactions of the Chinese Society of Agriculture Engineering 29(02).
[4] Deng X W, Ma X, Qi L, Deng G X, Liang S and Jin J 2021 Identification of weeds in rice seedling stage based on convolution neural network and transfer learning Agricultural Mechanization Research Journal 43(10).

[5] Zhang L, Jin X, Fu L Y and Li Sh W 2020 Recognition method for weeds in rapeseed field based on faster R-CNN deep network Laser & Optoelectronics Progress 57(02).

[6] Gui Y, Fan G H, Tang B B, Zhang Y H and Wu Y Zh 2019 Classification and recognition of weeds and rice based on convolution neural network Journal of Luoyang Institute of Science and Technology(Natural Science Edition) 29(01).

[7] Zhang Y Ch 2019 Research on rapid identification of weeds based on convolutional neural network Journal of Anhui Agricultural Science 47(14).

[8] Alexey DosoViTskiy, Lucas Beyer, Alexander Kolesnikov, et al 2020 An image is worth 16X16 words: Transformers for image recognition at scale DOI: arxiv-2010.11929

[9] Ablavatski A, Lu Sh J and Cai J F 2017 Enriched deep recurrent visual attention model for multiple object recognition Proceedings of 2017 IEEE Winter Conference on Applications of Computer Vision (WACV) DOI: 10.1109/WACV.2017.113.