# Bluefield (KDE TUT) at LifeCLEF 2016 Plant Identification Task

Siang Thye Hang, Atsushi Tatsuma, Masaki Aono

Knowledge Data Engineering& Information Retrieval Laboratory (KDE Lab), Department of Computer Science and Engineering, Toyohashi University of Technology, Japan

hang@kde.cs.tut.ac.jp, tatsuma@cs.tut.ac.jp, aono@tut.jp

**Abstract.** In this paper, we propose an automatic approach for plant image identification. We enhanced the well-known VGG 16-layers Convolutional Neural Network model [1] by replacing the last pooling layer with a Spatial Pyramid Pooling layer [2]. Rectified Linear Units (ReLU) are also replaced with Parametric ReLUs [3]. The enhanced model is trained without external dataset. A post processing method is also proposed to reject irrelevant samples. We further improved identification performance using observation identity (ObservationId) provided in the dataset. Our methods showed outstanding performance in official evaluation results of the LifeCLEF 2016 Plant Identification Task.

Keywords: LifeCLEF, plant identification, deep learning, sample rejection.

### 1 Introduction

Nowadays, conservation of biodiversity is becoming an important duty for us. To achieve this, accurate knowledge is essential. However, even for professionals, identifying a species can be a very difficult task. Convolutional Neural Networks (CNNs) are leading the best performance in various image retrieval tasks such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [4] [5] [1] [6]. CNNs learn filters of filters through multiple convolution layers, enabling higher level of abstraction than hand crafted features such as Scale Invariant Feature Transform (SIFT).

In recent years, CNN has gained popularity in various image retrieval tasks including the LifeCLEF Plant Identification Task (PlantCLEF). Over the past years, participants of PlantCLEF have adopted to use CNN in their works [7]. In PlantCLEF 2016 [8] [9], in addition to the 1000 class identification task, a new problem is introduced. The evaluation set consists of not only native plant images but also images of potentially invasive plants or irrelevant objects such as a table or a computer keyboard.

Considering the new setting, we propose a post processing method to reject these samples. In the next section we propose to enhance the VGG 16-layers model with Spatial Pyramid Pooling and Parametric Rectified Linear Units. We detail our model

training strategy in Section 3, followed by irrelevant sample rejection algorithm in Section 4. Section 5 shows evaluation results and Section 6 concludes this paper.

### 2 Model Enhancement

In the following sections, a few enhancements to the VGG 16-layers model will be described.

#### 2.1 Spatial Pyramid Pooling

The original VGG 16-layers model requires input image of spatial size  $224 \times 224$ . On the other hand, the images in the PlantCLEF 2016 dataset are arbitrarily sized, as shown in Figure 1. Following the spatial size restriction in VGG 16-layers model, input images has to be either cropped or warped. Both of these methods may lead to information loss.

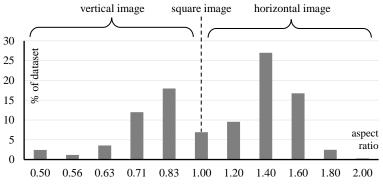


Fig. 1. Aspect ratio of images in the PlantCLEF 2016 dataset

To circumvent such restriction, we have replaced the last pooling layer (pool5) with a Spatial Pyramid Pooling (SPP) layer [2]. The last convolution layer (conv5\_3) produces feature map of 512 channels. With SPP, the feature map is spatially divided into  $1 \times 1$ ,  $2 \times 2$ ,  $4 \times 4$ , a total of 21 regions. Each region is then average pooled, producing a vector of fixed size  $21 \times 512 = 10752$ . Such conversion of arbitrarily sized feature map into a fixed size vector allows the model to accept input image of any size. Meanwhile, with such layer replacement, the number of parameters in the model is reduced from 138 million to 80 million.

#### 2.2 Parametric Rectified Linear Unit

We have also replaced all of the Rectified Linear Unit (ReLU) activations with a learnable version known as Parametric ReLU [3], which consistently outperforms ReLU in empirical experiments by Bing et al. [10]. Before training, we initialize the

learnable parameters with 0.05. Weight decay of the learnable parameters is disabled throughout the training process.

### 3 Model Training

The following strategies are used to train the enhanced model.

#### 3.1 Data preparation

The PlantCLEF 2016 dataset consists of 113204 images. 2048 images are used for validation purpose, while the remaining 111156 images are used to train the model.

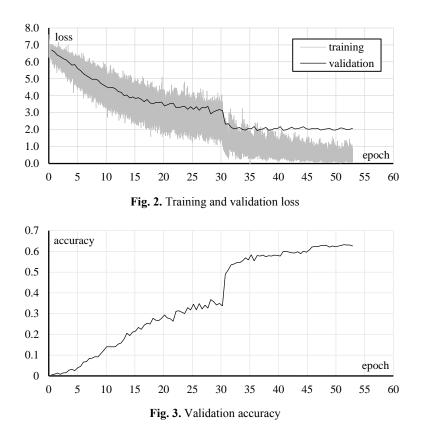
We have augmented the training images as follows: while preserving aspect ratio, images are resized such that the shorter side becomes 224 and 336. Random cropping to  $224 \times 224$  and random flipping by y-axis are also applied to the resized images. For evaluation process, images are resized such that the shorter side becomes 224, and no cropping or flipping is applied. Cropping is not required during evaluation as the enhanced model accepts image of any size.

#### 3.2 Image Mean

As the enhanced model is trained from scratch i.e. without external resources, image mean is computed from the training set. Excluding LeafScan images (with high RGB values due to bright background), the computed mean values of red, green and blue channels are 105, 111, and 79 respectively. Images augmented in Section 3.1 are sub-tracted with these mean values before being used for training.

#### 3.3 Training

We use the Caffe framework to train the enhanced model. To train such a deep model, we use Xavier's method [11] to initialize the weights of the convolution and fully connected layers. We train the model using Stochastic Gradient Descent with momentum 0.9, learning rate 0.01 through 0.0001, and batch size 50. The learning rate is multiplied by 0.1 when the validation accuracy stops improving. As the result, we have trained with learning rate 0.01 for 30 epochs, 0.001 for 15 epochs and finally 0.0001 for 8 epochs. Figure 2 shows the normalized softmax loss of both training and validation. Figure 3 shows the validation accuracy of the first prediction.



### 4 Irrelevant Sample Rejection

The next section discusses the One-Versus-All method's limitation to reject irrelevant samples in multiclass classification. To overcome such problem, an algorithm is proposed in Section 4.2.

#### 4.1 Multiclass Classification with One Versus All Method

Assume a binary classifier F of class C mapping an input x into a score F(x). F can be optimized such that F(x) > 0 when  $x \in C$ , and F(x) < 0 when  $x \notin C$ .

In the case of multiclass classification, the One-Versus-All method can be used to classify N classes  $C = [C_1, ..., C_N]$  using N binary classifiers  $F = [F_1, ..., F_N]$ . As shown in (1),  $x \in C_i$  when  $F_i(x)$  is the highest among all of the binary classifiers.

$$x \in C_i$$
 where  $i = \arg \max_{k \in N} F_k(x)$  (1)

With (1), it is guaranteed a class will be assigned to x, regardless how high or low the scores produced by each classifier. Figure 4's first example (from left) is ideal, as there is only one strong positive score, thus we can confidently predict that input  $\alpha$  in

Figure 4 belongs to the first class. Meanwhile, second and third examples have either weak or no positive score. With (1), input  $\beta$  is classified to the third class, while input  $\gamma$  is classified to the first class, despite that all of the scores are rather low or even negative.

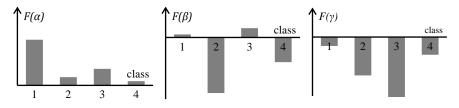


Fig. 4. An example of class scores (N = 4) based on three inputs  $\alpha$ ,  $\beta$ ,  $\gamma$ 

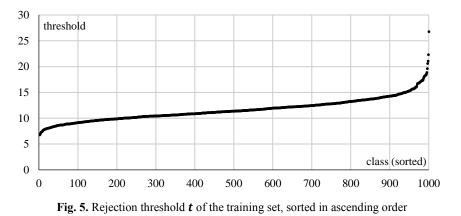
Based on the scenario above, we realized that there are cases that a sample should be rejected by all of the binary classifiers. To identify such irrelevant sample, an algorithm is elaborated in the next section.

### 4.2 Irrelevant Sample Rejection Algorithm

After training the enhanced model as mentioned in Section 3, a matrix of raw (i.e. before softmax normalization) scores  $S_{M \times N}$  of N classes and M training samples are extracted from the classifier layer (fc8). Incorrect predictions are omitted. With only correctly predicted class scores  $S_{M' \times N}(M' \le M)$ , rejection threshold  $t = [t_1, ..., t_N]$  of each class is computed from the class wise minima, as shown in (2).

$$t_i = \min_{k \in M'} S_{k,i} \tag{2}$$

Figure 5 shows the threshold t of each class based on the PlantCLEF 2016 training set (N = 1000), sorted in ascending order.



During evaluation, any sample with score lower than t for all of the classes will be rejected as irrelevant. With the evaluation set, 195 out of 8000 images are rejected. Some of the images are shown in Figure 6.



Fig. 6. Subset of rejected samples with threshold t

### 4.3 Taking Validation Set into Account

The threshold t obtained in Section 4.2 is solely based on the training set. Due to factors such as overfitting, there is a possibility that lower threshold can be acquired from the validation set. Figure 7 shows threshold obtained from the validation set, corresponding to the sorted classes shown in Figure 5.

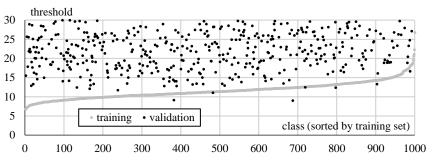


Fig. 7. Rejection threshold of training and validation set

Although we expect lower thresholds from the validation set, as shown in Figure 7, majority of the thresholds are higher than that of the training set. This is due to the fact that there is too little sample  $(2048 \ll M)$  in the validation set. Thus, only lower thresholds are considered. Ratios of each of these (seven) thresholds to its corresponding training set thresholds are computed, and then averaged into Q. As detailed in (3), the denominators are the thresholds of the training set, while the numerators are of the validation set. The values in (3) are based on Figure 7.

$$Q = \frac{1}{7} \left( \frac{9.1}{10.8} + \frac{11.1}{11.3} + \frac{9.0}{12.4} + \frac{12.5}{12.6} + \frac{12.4}{13.3} + \frac{13.3}{14.3} + \frac{16.7}{18.0} \right) \approx 0.91$$
(3)

Q is then multiplied to the threshold t of the training set, as shown in (4), before using it to reject samples during evaluation.

$$\boldsymbol{t}' = \boldsymbol{Q}\boldsymbol{t} \tag{4}$$

Applying t' to the evaluation set, only 69 samples are rejected as it is lower compared to the original t. Some of the images are shown in Figure 8.



Fig. 8. Subset of rejected samples with threshold t'

#### 4.4 Observation Based Identification

Identification based on a single image may be insufficient. In the PlantCLEF2016 dataset, images based on the same observation share a unique ObservationId. One ObservationId may be assigned with multiple images of different organs. Specifically images of flower or fruit which have much characteristic features, their existence in an observation often improves identification performance.

To further improve the identification performance, after rejecting samples as explained in Section 4.2 and 4.3, we sum the raw (i.e. before softmax normalization) class scores of images with the same ObservationId. The summed scores are then softmax normalized. As the result, images with the same ObservationId share the same normalized scores.

# 5 Evaluation

In our evaluation experiments, the enhanced CNN model is used to extract class scores. Four different post processing methods are then applied to the extracted class scores, detailed as follows. As mentioned in Section 4.2 and 4.3, t is the rejection threshold obtained from training set, while t' is the rejection threshold obtained by considering both training and validation sets.

- Run 1: Sample rejection with t, identification based on single sample
- Run 2: Sample rejection with t', identification based on single sample
- Run 3: Sample rejection with t, observation based identification
- Run 4: Sample rejection with t', observation based identification

In our run files (Bluefield), we provide scores up to top 30 classes, and rejected samples are entirely excluded from our run files. Evaluation results compared with other participants are summarized in Table 1 and Figure 9.

Run	Official score MAP	MAP restricted to potentially invasive species	MAP ignoring unknown classes and queries
Bluefield Run 4	0.742	0.717	0.827
SabanciUGebzeTU Run 1	0.738	0.704	0.806
SabanciUGebzeTU Run 3	0.737	0.703	0.807
Bluefield Run 3	0.736	0.718	0.820
SabanciUGebzeTU Run 2	0.736	0.683	0.807
SabanciUGebzeTU Run 4	0.735	0.695	0.802
CMP Run 1	0.710	0.653	0.790
LIIR KUL Run 3	0.703	0.674	0.761
LIIR KUL Run 2	0.692	0.667	0.744
LIIR KUL Run 1	0.669	0.652	0.708
UM Run 4	0.669	0.598	0.742
CMP Run 2	0.644	0.564	0.729
CMP Run 3	0.639	0.590	0.723
QUT Run 3	0.629	0.610	0.696
Floristic Run 3	0.627	0.533	0.693
UM Run 1	0.627	0.537	0.700
Floristic Run 1	0.619	0.541	0.694
Bluefield Run 1	0.611	0.600	0.692
Bluefield Run 2	0.611	0.600	0.693
Floristic Run 2	0.611	0.538	0.681
QUT Run 1	0.601	0.563	0.672
UM Run 3	0.589	0.509	0.652
QUT Run 2	0.564	0.562	0.641
UM Run 2	0.481	0.446	0.552
QUT Run 4	0.367	0.359	0.378
BME TMIT Run 4	0.174	0.144	0.213
BME TMIT Run 3	0.170	0.125	0.197
BME TMIT Run 1	0.169	0.125	0.196
BME TMIT Run 2	0.066	0.128	0.101

Table 1. Evaluation results sorted by official score MAP (MAP: Mean Average Precision)

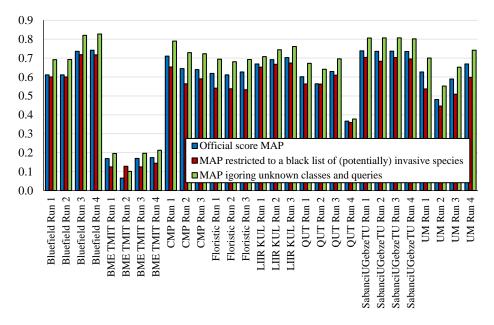


Fig. 9. Evaluation results sorted by run name in alphabetical order

Bluefield Run 4 yields the highest official score among all of the participants. Usage of ObservationId shows significant improvement. On the other hand, rejection threshold t' that takes validation set into account shows a slight improvement in official MAP.

### 6 Conclusion

In this paper, we described our approach to PlantCLEF 2016, focusing on model enhancements, data augmentations, and an irrelevant sample rejection strategy. We still leave some rooms for improvements, which are itemized as follows:

- As mentioned in Section 3.1, images for training process are resized to two scales. We should consider applying random scaling instead of a constant number of (two) scales. Other than scaling, random rotation should be applied as well.
- In Section 4.2 and 4.3, instead of taking minima as rejection threshold, we should consider using mean and standard deviation to obtain a more stable threshold.
- The PlantCLEF 2016 dataset includes rich metadata information such as Genus, Family, Date, Longitude, Latitude, Location and Content (organ type). However, in our work, only ClassId (class label) and ObservationId are used. More metadata information should be utilized to obtain more accurate identification performance.

### Acknowledgement

We would like to thank MEXT KAKENHI, Grant-in-Aid for Challenging Exploratory Research, Grant Number 15K12027 for partial support of our work.

## References

- 1. K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *CVPR*, 2014.
- 2. K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in *TPAMI*, 2015.
- 3. K. He, X. Zhang, S. Ren and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *CVPR*, 2015.
- 4. A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional," in *NIPS*, 2012.
- C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2014.
- 6. K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *CVPR*, 2015.
- 7. H. Goeau, P. Bonnet and A. Joly, "LifeCLEF Plant Identification Task," in *CLEF*, 2015.
- A. Joly, H. Goëau, H. Glotin, C. Spampinato, P. Bonnet, W.-P. Vellinga, J. Champ, R. Planqué, S. Palazzo and H. Müller, "LifeCLEF 2016: Multimedia Life Species Identification Challenges," in *LifeCLEF*, 2016.
- 9. H. Goëau, P. Bonnet and A. Joly, "Plant Identification In An Open-World (LifeCLEF 2016)," in *PlantCLEF*, 2016.
- 10. B. Xu, N. Wang, T. Chen and M. Li, "Empirical Evaluation of Rectified Activations in Convolution," in *CVPR*, 2015.
- 11. X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*, 2010.