Plant Classification based on Gated Recurrent Unit

Sue Han Lee[1], Yang Loong Chang[1], Chee Seng Chan[1], Joly Alexis[2], Pierre Bonnet[3][0000−0002−2828−4389], and Herve Goeau[3]

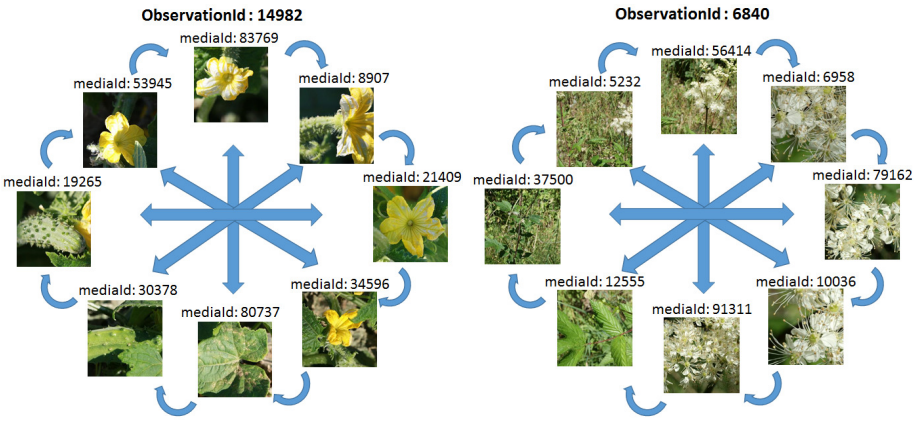
[1] Center of Image & Signal Processing, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia
{leesuehan,yangloong}@siswa.um.edu.my, cs.chan@um.edu.my
[2] INRIA, Montpellier, France
alexis.joly@inria.fr
[3] CIRAD-Amap, Montpellier, France
pierre.bonnet@cirad.fr, herve.goeau@inria.fr

Abstract. Classification of plants based on a multi-organ approach is very challenging due to the variability in shape and appearance in plant organs. Despite promising solutions built using convolutional neural network (CNN) for plant classification, the existing approaches do not consider the correspondence between different views captured of a plant. In fact, botanists usually observe and study simultaneously a plant from different vintage points, as a whole and also analyse different organs in order to disambiguate species. Driven by this insight, we introduce a new framework for plant structural learning using the recurrent neural network (RNN) approach. This novel approach supports classification based on a varying number of plant views composed of one or more organs of a plant, by optimizing the dependencies between them. We also present the qualitative results of our proposed models by visualizing the learned attention maps. To our knowledge, this is the first study to venture into such dependencies modeling and interpret the respective neural net for plant classification. Finally, we show that our proposed method outperforms the conventional CNN approach on the PlantClef2015 benchmark. The source code and models are available at https://github.com/cs-chan/Deep-Plant.

Keywords: Plant classification · Deep learning · Recurrent neural network.

1 Introduction

Plants are the backbone of all life on earth providing us with food and oxygen. A good understanding of plants is essential to help in identifying new or rare plant species in order to improve the drug industry, balance the ecosystem as well as the agricultural productivity and sustainability. Ever since LifeCLEF, one of the foremost visual image retrieval campaigns hosted a plant identification challenge, researchers have started to focus on automatic analysis of multiple



(a) (b)

Fig. 1: (a) and (b) represent examples of plant images taken from the plants tagged with ObservationID 14982 and 6840 respectively in PlantClef2015 dataset [15]. Different plant views captured of a plant exhibit correlated characteristic in their organ structures. Best viewed in color.

images exploiting different views of a plant capturing one or more organs. From year 2014, it has provided up to seven different plant views which are entire plant, branches, flower, leaf, leaf scan, fruit, and stem. Indeed, [14] has shown that combining different types of views in a query can increase the species identification rate. Previously, researchers [14, 20] consider that different images are independent from each other. A straightforward fusion scheme such as the mean of the categorical distributions predicted for each image is generally employed to combine the information contained in each image. However, in reality, different views are far from being independent because they correspond to multiple views of the same individual specimen. For example, as shown in Fig. 1, different plant views (or organs) captured of a plant exhibit correlated or overlapping characteristics in their organ structures, nonetheless these traits are distinctive between different plants. This information inevitably can be seen as one of the important cue to help differentiate species. Majority studies have used CNN to classify plant images [18, 21]. This approach however was designed to capture similar region-wise patterns within an image, thus disregarding the correlation between different plant views of a plant. In this work, we propose a new framework based on RNN [12] to model the dependencies between different plant views where the probability of a view is conditioned by the other views. Specifically, it takes in a varying number of plant view images composed of one or more organs, and optimizes the dependencies between them for species classification. Additionally, we introduce a Coarse-to-Fined Attention (*CFA*) module where it can locate the local regions that are highly voted by the RNN method in each plant view. Our contributions are summarized as follows:

1. We propose a RNN based approach to model different plant views capturing one or more organs of plant for species classification.
2. We introduce a *CFA* module that provides a better visual understanding on the local features emphasized by the RNN method in plant views dependencies modeling.
3. Our proposed model achieves a better performance compared to the conventional CNN approach on PlantClef2015 benchmark.

2 Related works

Plant identification. Over the past few years, researchers have worked on recognizing plant species using solely a single plant organ. A majority of the studies have utilized leaves to identify species. Leaf characters such as shape, texture, and venation are the most generally used features to distinguish leaves of different species [18]. To fit better with a real scenario where people generally try to identify a plant by observing several plant organs or a similar organ from different viewpoints, researchers in computer vision have focused on designing an automated plant classification system to identify multi-organ plant images. Earliest attempts [27, 33, 11] in general, adopt organ-specific features for discrimination. Specifically, they first group the images of plants into their respective organ categories. Then, based on each organ category, organ-specific features are extracted using feature engineering approaches. Ever since, DL has been proved extremely high recognition capabilities in dealing with very large datasets, [10] proposed using an end-to-end CNN to replace those hand-crafted feature extractors. They introduced organ-specific CNN models where each model is trained on dedicated plant organs. There are also researchers [24, 6] focused on using CNN to learn generic features of plants, irrespective of their organ information. Lately, [21] showed that using the HGO-CNN which incorporates both the organ-specific and generic features could provide the best result in the LifeClef2015 challenge. Despite promising results obtained using CNN based approach, the representation learned focuses only on the information contained in each image, but fails to capture the high-level semantics corresponding to the interaction between different plant views (organs). Henceforth, this work moves beyond existing practice, venturing into a new alternative to solve this problem.

RNN based classification. The RNN has always been used to process sequential data such as language translation [17, 30] and action recognition [22, 29]. Lately, CNN and RNN have been employed to combine information, integrating the domain of computer vision and natural language processing [28, 9, 34, 32, 31]. Despite using RNN to model complex structures of video or language, a few publications have showed the capability of RNN in processing variable length of fixed-sized data in a sequential manner though data originally is not in a form of sequences. For example, it has been actively explored in segmentation [23, 25], scene labeling [4, 26], object recognition [2, 3] as well as image generation [13]. In such case, RNN is used to model the dependencies between pixels or regions within an image. In our work, we formulate RNN to the contrary, to

learn the structure of an object based on its different views which do not have a form of sequences. We introduce a probabilistic model to process different plant views captured of a plant where each state variable is conditioned upon all other states, and not only its previous ones.

3 Approach

Notations We denote the plant view images as $\mathbf{I_t} \in \{\mathbf{I_1}, \mathbf{I_2}, \cdots, \mathbf{I_T}\}$ where $t = 1, \cdots, T$ are the states corresponding to the indices of plant view images of the same plant. Each $\mathbf{I_t}$ is associated with a species annotation $(\mathbf{I_t}, r_t)$ where r_t is a one hot vector with only the species label set as positive. For each plant view image, we extract its convolutional features from a CNN model, $\boldsymbol{\delta_t} \in \{\boldsymbol{\delta}_1, \boldsymbol{\delta_2}, \cdots, \boldsymbol{\delta_T}\}$, $\boldsymbol{\delta_t} \in \mathbb{R}^{H \times W \times C}$ where H, W and C are the height, width and number of channels of feature maps.

Architecture It is known that human brain processes information iteratively, where it keeps the current state in an internal memory and uses it to infer future observation, capturing the potential relationships between them [8]. Driven by this insight, we build a new plant classification framework upon the RNN based approach, which can hold and relate different structural information of a plant. Moreover, it is versatile to an arbitrary number of plant images. In this work, the Gated Recurrent Unit (GRU) [7], one of the gating mechanism in RNNs, is adopted for a more light-weight and simple network structure. The activation $\mathbf{h_t}$ is a linear interpolation between the previous activation $\mathbf{h_{t-1}}$ and the current candidate activation $\tilde{\mathbf{h}}_\mathbf{t}$: $\mathbf{h_t} = (1 - \mathbf{z_t})\mathbf{h_{t-1}} + \mathbf{z_t}\tilde{\mathbf{h}}_\mathbf{t}$ where $\mathbf{z_t}$ is the update gate that decides how much of the previous state should be kept around. The $\mathbf{z_t}$ is computed as $\mathbf{z_t} = \sigma(\mathbf{W_{z1}x_t} + \mathbf{W_{z2}h_{t-1}})$. The candidate activation $\tilde{\mathbf{h}}_\mathbf{t}$ which is processed with a hyperbolic tangent is formulated as follows: $\tilde{\mathbf{h}}_\mathbf{t} = \tanh(\mathbf{W_{h1}x_t} + \mathbf{W_{h2}}(\mathbf{v_t} \odot \mathbf{h_{t-1}}))$ where $\mathbf{v_t}$ is the reset gate that determines to which extent the new input should be combined with the previous state and $\odot$ is an element-wise multiplication operator. The $\mathbf{v_t}$ is formulated as follows: $\mathbf{v_t} = \sigma(\mathbf{W_{v1}x_t} + \mathbf{W_{v2}h_{t-1}})$. The activations of both gates are element-wise logistic sigmoid functions σ. It maps $\mathbf{v_t}$ and $\mathbf{z_t}$ in between 0 to 1. All the $\mathbf{W}$ matrices are trained parameters. The network is fed by the current input vector $\mathbf{x_t}$, while all the $\mathbf{W}$ matrices are trained parameters.

Attention (*attn*) The attention module is used to reduce the dimensionality of convolutional features in order to ease the computational burden of a network [5]. The attention map $\boldsymbol{\lambda_t}$ controls the contribution of convolutional features at the t-th state. Larger value in $\boldsymbol{\lambda_t}$ indicates higher importance. The term $\boldsymbol{\epsilon_t}$ introduced as the weighted average of convolutional features that is dependent on the previous activation $\mathbf{h_{t-1}}$ and convolutional features $\boldsymbol{\delta_t}$. The attention function $g : \boldsymbol{\delta_t}, \mathbf{h_{t-1}} \rightarrow \boldsymbol{\epsilon_t}$ is defined as follows:

$$\boldsymbol{\zeta_t} = \{\tanh(\boldsymbol{\delta_t}\mathbf{W_\delta} + \mathbf{h_{t-1}}\mathbf{W_h})\}\mathbf{W_a} \tag{1}$$

$$\boldsymbol{\lambda_t} = softmax(\boldsymbol{\zeta_t}) \tag{2}$$

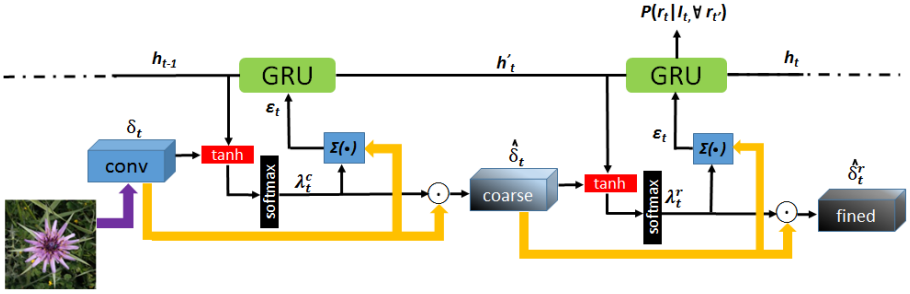


Fig. 2: The proposed Coarse-to-Fined Attention module. Best viewed in color.

$$\boldsymbol{\epsilon_t} = \sum_{i,j} \boldsymbol{\lambda_{t,ij}} \boldsymbol{\delta_{t,ij}} \tag{3}$$

where the embedding matrices $\mathbf{W_\delta} \in \mathbb{R}^{C\times C}$, $\mathbf{W_h} \in \mathbb{R}^{E\times C}$, $\mathbf{W_a} \in \mathbb{R}^{p\times 1}$, E is the dimensionality of GRU cell, $p = H \times W$ and $\boldsymbol{\delta_{t,ij}}$ denotes convolutional feature at location $(i, j) \in p$.

Coarse-to-Fined Attention (*CFA*) Using the aforementioned attention mechanism (Eq. 1-3), the GRU decodes species prediction based on global image features attained from a CNN. The attention mechanism trained by such global image features might not be able to infer the discriminative local features of plant structures. To gain a better visual understanding on which part of a plant view image is mostly emphasized by the RNN based approach, a better localization of the attention map is inevitably necessary. To this end, we refine the attention map acquired in each state t by proposing *CFA* module as shown in Fig. 2. Basically, the convolutional feature $\boldsymbol{\delta_t}$ is first processed to obtain a coarse attention map $\boldsymbol{\lambda_t^c}$. The $\boldsymbol{\lambda_t^c}$ is then element-wise multiplied with the $\boldsymbol{\delta_t}$ to produced a masked convolutional feature $\hat{\boldsymbol{\delta}}_\mathbf{t}$ which is to be fed to the following GRU. The attention mechanism at the later stage is therefore trained to look for pertinent features from this refined image feature $\hat{\boldsymbol{\delta}}_\mathbf{t}$ and identify the best local features. With the use of the refined attention map produced as $\boldsymbol{\lambda_t^r}$, we can eventually locate these local features in each plant view.

Training Contrary to modeling video or language data where variable number of inputs are conditioned upon their previous states $P(r_t|\mathbf{I_t}, r_1, \cdots, r_{t-1})$, in our case, it is logical to condition the inputs upon all other states information for the plant structural modeling, $P(r_t|\mathbf{I_t}, \{r_d\}_{d\neq t})$. The reason is that, states in our context are analogous to the collections of different plant views captured from a similar plant, so the relationships between these states are interrelated. Henceforth, to tackle this challenge, we design in such a way that it would be able to iteratively classify images of a plant while conjointly operate on all of its related instances. In particular, we build a bidirectional states modeling mechanism where the forward neuron activations $\overrightarrow{\mathbf{h}}$ models $P_{fw} = P(r_t|\mathbf{I_t}, r_1, \cdots, r_{t-1})$ and the backward neuron activations $\overleftarrow{\mathbf{h}}$ models $P_{bw} = P(r_t|\mathbf{I_t}, r_{t+1}, \cdots, r_T)$.

Then, we put in correspondence between both neurons for every state and train them upon the respective species classes. In this manner, each state t can be considered as condition upon the collections of the related plant images from states $1, \cdots, t-1, t+1, \cdots, T$. In order to correlate between both states, the output activations of the forward and backward GRU are cascaded as follows: $\mathbf{h_t} = [\overrightarrow{\mathbf{h}}_\mathbf{t}, \overleftarrow{\mathbf{h}}_\mathbf{t}]$. Then, we multiply $\mathbf{h_t}$ with a class embedding matrix, $\mathbf{W_{em}}$, which is $\mathbf{s(I_t)} = \mathbf{W_{em}h_t}$ before normalizing it with a softmax function: $P(r_t|\mathbf{I_t}, \{r_d\}_{d \neq t}) = \frac{e^{s_r(\mathbf{I_t})}}{\sum_{m=1}^{M} e^{s_m(\mathbf{I_t})}}$ where M and r stand for the total number of classes and the target class respectively. We perform the softmax operation for every state t preceding the computation of the overall cross entropy function: $L_{psn} = \frac{1}{T} \sum_{t=1}^{T} L_t$, where $L_t = -logP(r_t|\mathbf{I_t}, \{r_d\}_{d \neq t})$.

4 Datasets and Evaluation metrics

Dataset. The PlantClef2015 dataset [15] was used. It has 1000 plant species classes. Training and testing data comprise 91759 and 21446 images respectively. Each image is associated with a single organ type (branch, entire, flower, fruit, leaf, stem or leaf scan).

Evaluation metrics. We employ the *observation* and *image-centered* scores [15] to evaluate the model's performance. The purpose of the observation score is to evaluate the ability of a model to predict the correct species labels for all the users. To this end, the observation score is the mean of the average classification rate per user as defined: $S_{obs} = \frac{1}{U} \sum_{u=1}^{U} \frac{1}{P_u} \sum_{p=1}^{P_u} S_{u,p}$ where U represents the number of users, P_u is the number of individual plants observed by the u-th user, and $S_{u,p}$ is the score between 0 and 1 as the inverse of the rank of the correct species (for the p-th plant observed by the u-th user). Each query observation is composed of multiple images. To compute $S_{u,p}$, we adopt the Borda count (BD) to combine the scores of multiple images: $BD = \frac{1}{n} \sum_{k=1}^{n} score_k$ where n is the total number of images per query observation and *score* is the softmax output score, which describes the ranking of the species.

Next, the image-centered score evaluates the ability of a system to provide the correct species labels based on a single plant observation. It calculates the average classification rate for each individual plant defined as: $S_{img} = \frac{1}{U} \sum_{u=1}^{U} \frac{1}{P_u} \sum_{p=1}^{P_u} \frac{1}{N_{u,p}} \sum_{n=1}^{N_{u,p}} S_{u,p,n}$ where U and P_u are explained earlier in the text, $N_{u,p}$ is the number of pictures taken from the p-th plant observed by the u-th user and $S_{u,p,n}$ is the score between 0 and 1 equal to the inverse of the rank of the correct species (for the n-th picture taken from the p-th plant observed by the u-th user). We compute the rank of the correct species based on its softmax scores. Besides S_{obs} and S_{img}, we also compute the top-1 classification result to infer the robustness of the system: $Acc = T_r/T_s$ where T_r is the number of true species prediction and T_s represents total number of testing data.

Table 1: Performance comparison between the E-CNN [20] and the GRU architecture.

Method	Acc	S_{img}	S_{obs}
E-CNN [19, 20]	0.635	0.710	**0.737**
GRU (*conv*7) + *attn*	0.669	0.709	0.718
GRU (*conv*5_3) + *CFA*	0.662	0.711	0.723
GRU (*conv*5_3) + *attn*	**0.686**	**0.718**	0.726

5 Experiments

We firstly group the training and testing images into their respective observation ID. Note that, each observation ID consists of T number of plant images captured from a p-th plant observed by a u-th user. By doing so, we have 27907 and 13887 observation IDs for training and testing respectively. Next, we apply the multi-scale image generation process proposed in [19] on these images. For each plant image, we extract its image representation using the enhanced HGO-CNN (E-CNN) [19, 20]. We train the architecture based on random sequence, disregarding the order of the plant images fed into the network. This is driven by our understanding that botanists usually observe and study a plant from different vintage points simultaneously, as a whole and also analyse different organs, and this is done without specific order. We test the performance of GRU architecture using different levels of image abstraction representation. We use *conv*5_3 and *conv*7 features extracted from the last convolutional layer of generic and species layer of E-CNN [20] respectively. The GRU architecture is trained using the Tensorflow library [1]. We use the ADAM optimizer [16] with the parameters $\alpha = 1e-08$, $\beta 1 = 0.9$ and $\beta 2 = 0.999$. We applied L_2 weight decay with penalty multiplier set to 1×10^{-4}, and dropout ratio set to 0.5, respectively. We set the learning rate to 1×10^{-3}, and, reduce it to 1×10^{-4} when the training performance stops improving. Mini-batch size is set to 15.

5.1 Performance Evaluation

In Table 1, we compare the performance of the GRU architecture with the E-CNN baseline [19, 20]. It can be seen that using the GRU with *conv*5_3 input layer, achieved the highest top-1 accuracy of 0.686, outperforms the previous E-CNN [19, 20]. However, we found that its S_{img} and S_{obs} do not seem to have much improvement. We then explore the cause and observe that most of the misclassifications occur when there is only one testing image per observation ID. Table 2 shows that there is a total of 9905 testing images that fall in category A, which is nearly 47% of the testing set. The GRU performs noticeably better in category B than A (top-1 accuracy of 0.754 compared to 0.607), while E-CNN [19, 20] performs almost equally in all cases for category A and B (top-1 accuracy of 0.634 and 0.637). This can be deduced from the characteristic of both E-CNN and GRU based models used in this context. To recognize a plant

Table 2: Comparison of top-1 classification accuracy for different categories of observation ID. Note that, Category A = a single image per observation ID; Category B = number of images ≥ 2 per observation ID

Category	A	B
Total number of training images for each category	11690	80069
Total number of testing images for each category	9905	11541
E-CNN [19, 20]	**0.634**	0.637
GRU (*conv*5_3) + *attn*	0.607	**0.754**

Table 3: Classification performance comparison of each content based on S_{img}.

Method	Branch	Entire	Flower	Fruit	Leaf	LeafScan	Stem
E-CNN [19, 20]	0.564	0.573	0.801	0.657	0.666	0.759	0.384
GRU (*conv*5_3) + *attn*	**0.650**	**0.643**	**0.823**	**0.709**	**0.729**	**0.790**	**0.546**
Gain (%)	+15.2	+12.2	+2.7	+7.9	+9.5	+4.1	+42.2

image, the E-CNN based model is trained to find similar patterns on all different subfields of an image, while the GRU based model is trained to look for higher level features, modeling the dependencies between a series of images. Next, we noticed that the number of training samples in category A is significantly less than category B. Such imbalanced training set might be another factor that affects the performance of the GRU in predicting species for category A. Based on these findings, we therefore deduce that the poor performance of the GRU based model is most likely due to the inadequate samples of plants given one observation ID. Besides, we found that using GRU + *CFA* module, the S_{img} and S_{obs} are 0.711 and 0.723 respectively, which are comparable to the *attn* module but the top-1 accuracy on the other hand is only 0.662. This is probably due to the absence of global information when the network is explicitly forced to focus on local regions of plant structures. Moreover, using the GRU with the *conv*5_3 as the input layer is proven to be better compared to the *conv*7. We attribute this performance difference to *conv*5_3 features being more generic compared to *conv*7, as we note that there is a transition from generic to class specific features within the CNNs. Hence, the generic features are more versatile when re-purposed for a new task. Additionally, training the GRU with generic features does not make any explicit use of the organ tags, which inevitably reduces the complexity in model training.

5.2 Detailed Scores for Each Plant Organ

In this section, we analyse the classification performance for each of the organ based on the image-centered score, S_{img}. We observe that the GRU model essentially improved the recognition performance of each organ, especially the 'stem' organ. As shown in Table 3, the improvement gained is 42.2% which is considerably significant compared to other organs. This is due to the fact that the stem organ has the least number of images in category A compared to other organs.

Table 4: Percentage of testing images that fall under category A for each organ category (%)

Branch	Entire	Flower	Fruit	Leaf	LeafScan	Stem
56.49	68.17	64.81	50.98	33.59	64.23	25.77

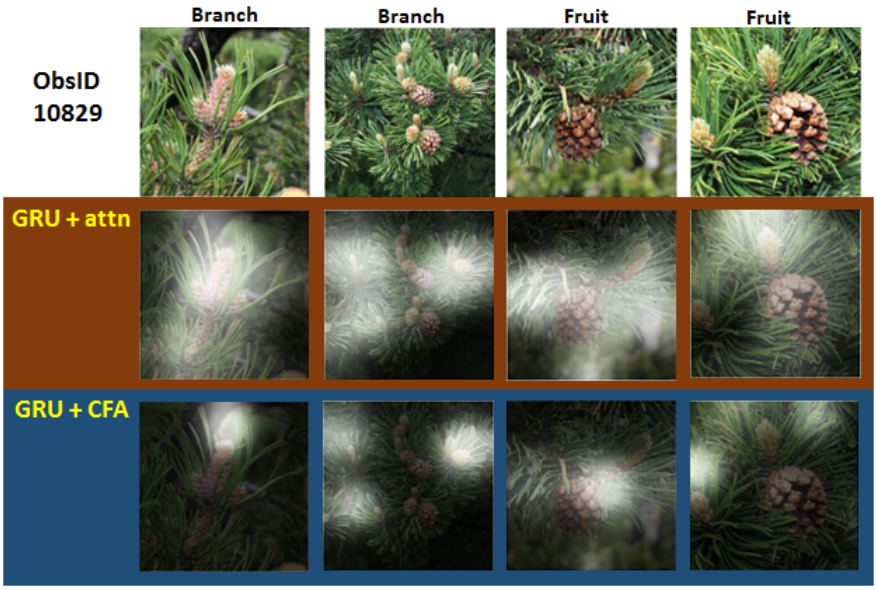
That is the majority of stem images co-exists with other plant images in one observation ID. For this reason, we can see that although the stem organ is considered as the least informative one compared to other organs, using the RNN method, we can successfully boost its classification performance. Besides, note that, although improvement gained for the 'flower' is not as high as the 'stem' organ, its performance is the highest for the overall plant views. This shows that flower is the most effective organ to identify plant species.

5.3 Qualitative Analysis

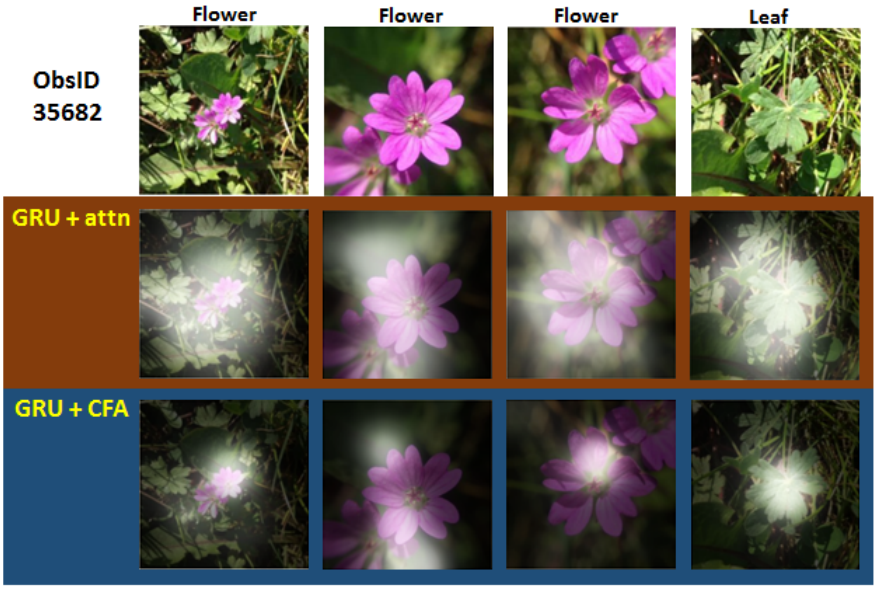
Contrary to CNN, RNN learns the high-level structural features of a plant by modeling the dependencies between different plant views. Besides quantitative analysis, we go deeper into exploring, analyzing and understanding the learned features by using both, the *attn* and the *CFA* modules. Fig. 3 shows the visualisation results of the GRU(*conv*5_3) + *attn* and the GRU(*conv*5_3) + *CFA*. It is noticed that, using the *attn* module, the highly activated regions mostly fall on the holistic plant structures. Hence, we deduce that the GRU(*conv*5_3) + *attn* is able to locate the pertinent foreground regions that are analogous to the plant structures. On the other hand, using the *CFA* module to recurrently refine the attention regions can precisely locate the discriminative local regions of plant structures, which are voted the most by the RNN method. Based on the visualisation results in Fig. 3, we can notice that the refined features are focused on the boundary of the flower's petals as well as the center of the compound leaflets, radiating from the tip of the petiole. This shows that the *CFA* can provide more localized attention that emphasizes the most distinctive local regions rather than the holistic plant structures. These insights therefore provide us with a better visual understanding from the global to the local perspective of image representation learned through the RNN in modeling plant views correlation.

6 Conclusion

We presented a novel plant classification framework based on RNN approach where it supports classification based on a varying number of plant views composed of one or more organs of a plant, by optimizing the dependencies between them. Experiments on the PlantClef 2015 benchmark showed that modeling the higher level features of plant views interaction can essentially improve the classification performance, especially for the less distinctive 'stem' organ. Furthermore, with the help of the proposed *CFA* module, we can achieve better insights of the discriminative subparts of the plant structures which are voted the most by the RNN approach for species classification.



(a)



(b)

Fig. 3: Visualisation of the activation maps generated by the GRU(*conv*5_3) + *attn* and GRU(*conv*5_3) + *CFA* for plant samples tagged with observation ID (a)10829 and (b) 35682 in PlantCLef 2015 dataset. It can be seen that the CFA module can refine the attention regions to locate the most distinctive local regions rather than the holistic plant structures. Best viewed in color.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: OSDI. vol. 16, pp. 265–283 (2016)
2. Ba, J., Mnih, V., Kavukcuoglu, K.: Multiple object recognition with visual attention. arXiv preprint arXiv:1412.7755 (2014)
3. Bell, S., Lawrence Zitnick, C., Bala, K., Girshick, R.: Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In: CVPR. pp. 2874–2883 (2016)
4. Byeon, W., Breuel, T.M., Raue, F., Liwicki, M.: Scene labeling with lstm recurrent neural networks. In: CVPR. pp. 3547–3555 (2015)
5. Cho, K., Courville, A., Bengio, Y.: Describing multimedia content using attention-based encoder-decoder networks. IEEE Transactions on Multimedia **17**(11), 1875–1886 (2015)
6. Choi, S.: Plant identification with deep convolutional neural network: Snumedinfo at lifeclef plant identification task 2015. In: CLEF (Working Notes) (2015)
7. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
8. Clark, A.: Whatever next? predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences **36**(3), 181–204 (2013)
9. Fu, K., Jin, J., Cui, R., Sha, F., Zhang, C.: Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2321–2334 (2017)
10. Ge, Z., McCool, C., Sanderson, C., Corke, P.: Content specific feature learning for fine-grained plant classification. In: CLEF (Working Notes) (2015)
11. Goëau, H., Joly, A., Yahiaoui, I., Bakić, V., Verroust-Blondet, A., Bonnet, P., Barthélémy, D., Boujemaa, N., Molino, J.F.: Plantnet participation at lifeclef2014 plant identification task. In: CLEF2014 Working Notes. Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 724–737. CEUR-WS (2014)
12. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: Acoustics, speech and signal processing (ICASSP), 2013 IEEE International Conference on. pp. 6645–6649 (2013)
13. Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., Wierstra, D.: Draw: A recurrent neural network for image generation. arXiv preprint arXiv:1502.04623 (2015)
14. Joly, A., Goëau, H., Bonnet, P., Bakić, V., Barbe, J., Selmi, S., Yahiaoui, I., Carré, J., Mouysset, E., Molino, J.F., et al.: Interactive plant identification based on social image data. Ecological Informatics **23**, 22–34 (2014)
15. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Planqué, R., Rauber, A., Palazzo, S., Fisher, B., et al.: Lifeclef 2015: multimedia life species identification challenges. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction, pp. 462–483. Springer (2015)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
17. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R.: Ask me anything: Dynamic memory networks for natural language processing. In: ICML. pp. 1378–1387 (2016)

18. Lee, S.H., Chan, C.S., Mayo, S.J., Remagnino, P.: How deep learning extracts and learns leaf features for plant classification. Pattern Recognition **71**, 1–13 (2017)
19. Lee, S.H., Chan, C.S., Remagnino, P.: Multi-organ plant classification based on convolutional and recurrent neural networks. IEEE Transactions on Image Processing **27**(9), 4287–4301 (2018)
20. Lee, S.H., Chang, Y.L., Chan, C.S.: Lifeclef 2017 plant identification challenge: Classifying plants using generic-organ correlation features. Working Notes of CLEF **2017** (2017)
21. Lee, S.H., Chang, Y.L., Chan, C.S., Remagnino, P.: Hgo-cnn: Hybrid generic-organ convolutional neural network for multi-organ plant classification. pp. 4462–4466. ICIP (2017)
22. Ng, J.Y.H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: CVPR. pp. 4694–4702 (2015)
23. Ren, M., Zemel, R.S.: End-to-end instance segmentation with recurrent attention. arXiv preprint arXiv:1605.09410 (2017)
24. Reyes, A.K., Caicedo, J.C., Camargo, J.E.: Fine-tuning deep convolutional networks for plant recognition. In: CLEF (Working Notes) (2015)
25. Romera-Paredes, B., Torr, P.H.S.: Recurrent instance segmentation. In: ECCV. pp. 312–329. Springer (2016)
26. Shuai, B., Zuo, Z., Wang, B., Wang, G.: Dag-recurrent neural networks for scene labeling. In: CVPR. pp. 3620–3629 (2016)
27. Szűcs, G., Papp, D., Lovas, D.: Viewpoints combined classification, method in image-based plant identification task. In: CLEF (Working Notes). vol. 1180, pp. 763–770 (2014)
28. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4), 652–663 (2017)
29. Wu, Z., Jiang, Y.G., Wang, X., Ye, H., Xue, X.: Multi-stream multi-class fusion of deep networks for video classification. In: Proceedings of the 2016 ACM on Multimedia Conference. pp. 791–800 (2016)
30. Xiong, C., Merity, S., Socher, R.: Dynamic memory networks for visual and textual question answering. In: ICML. pp. 2397–2406 (2016)
31. Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: ECCV. pp. 451–466. Springer (2016)
32. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: CVPR. pp. 21–29 (2016)
33. Yanikoglu, B., Tolga, Y., Tirkaz, C., FuenCaglartes, E.: Sabanci-okan system at lifeclef 2014 plant identification competition. In: CLEF (Working Notes) (2014)
34. Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W.: Video paragraph captioning using hierarchical recurrent neural networks. In: CVPR. pp. 4584–4593 (2016)