Very Deep Residual Networks with MaxOut for Plant Identification in the Wild

Milan Šulc, Dmytro Mishkin, and Jiří Matas

Center for Machine Perception, FEE, CTU in Prague
Technicka 2, 166 27 Prague 6, Czech Republic
{sulcmila,mishkdmy,matas}@cmp.felk.cvut.cz

Abstract. The paper presents our deep learning approach to automatic recognition of plant species from photos. We utilized a very deep 152-layer residual network [15] model pre-trained on ImageNet, replaced the original fully connected layer with two randomly initialized fully connected layers connected with maxout [13], and fine-tuned the network on the PlantCLEF 2016 training data. Bagging of 3 networks was used to further improve accuracy. With the proposed approach we scored among the top 3 teams in the PlantCLEF 2016 plant identification challenge.

1 Introduction

In the past, automatic recognition of plants was usually solved by requiring a photo of a specific plant organ, such as leaf [21,6,28], flower [24,1,23] or tree bark [6,29,3]. Moreover, a number of these systems and datasets placed further constraints on the input image, such as white background behind in a leaf image.

Recently, Deep Convolutional Neural Networks (CNNs) succeeded in a number of computer vision tasks, especially those related to complex recognition and detection of objects. This was also the case of plant recognition, where in PlantCLEF 2015 [9] the deep learning submissions [5,7,4,26] outperformed combinations of hand-crafted methods significantly (see Fig. 1).

Our submission to this year's challenge also employs a deep CNN pretrained on ImageNet, namely the very recent ResNet architecture [15], which allows to train very deep networks efficiently. The idea of ResNets is that each layer should not learn the whole feature space transformation, but only a residual correction to the previous layer. To bridge domain shift between a general-class network pretrained on ImageNet and a fine-grained task of plant classification, we added an additional maxout [13] layer on top of the pretrained network. This brought a noticeable improvement in accuracy in our validation experiments.

The rest of this paper is organized as follows: Section 2 gives a quick overview of the PlantCLEF 2016 plant identification task. The proposed method is described in Section 3 and the preliminary results and validation in Section 4. The results on the test set are discussed in Section 5. Finally, Section 6 contains the conclusions and the discussion for future work.

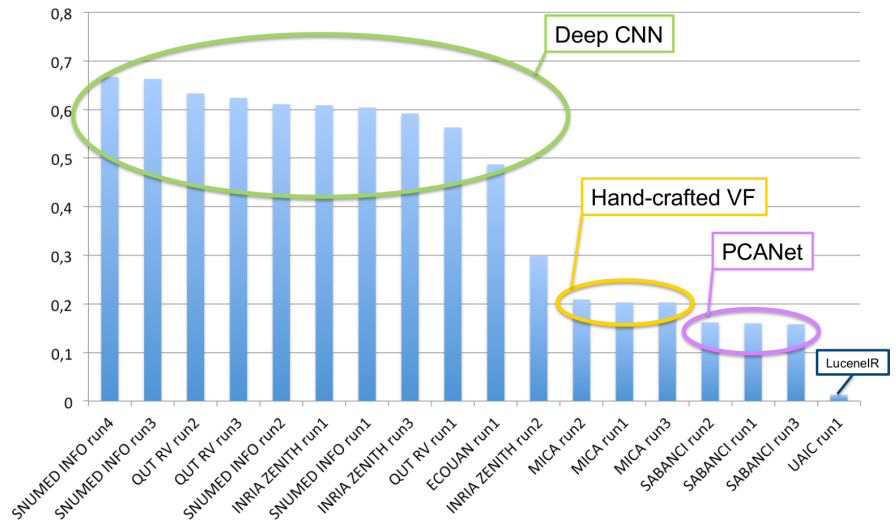

Fig. 1. Results of the previous year's (PlantCLEF2015) main task. Source: The LifeCLEF 2015 Plant Identification Task [9] presentation.

2 The PlantCLEF 2016 Plant Identification Task

The goal in the PlantCLEF challenge is to automatically identify plant species from photos of different nature. This years challenge [10,19] addresses the task as an open-set or open-world recognition problem [27,2]: the test data contains distractors of unseen categories and the metric for the classification is the mean average precision (mAP), considering each class of the training set as a query. This poses a requirement on robustness to unseen categories.

The training data consist of the training and test set of PlantCLEF 2015, which were fully annotated with the correct taxonomic species as well as other meta-data such as type of view (*Leaf*, *LeafScan*, *Flower*, *Fruit*, *Stem*, *Branch*, *Entire*), date of acquisition, author ID or its GPS coordinates (if available). The training set consists of 113,205 pictures of herb, tree and fern specimen belonging to 1,000 species. The evaluated metric is the mean average precision – mAP. The test set contains 8000 images, including the above mentioned distractor images. The official results on the test (evaluated by the organizers) are discussed in Section 5.

3 The Proposed Approach: Very Deep Residual Maxout Networks

Recently, the very deep residual networks of He et al. [15] gained a lot of attention after achieving the best results in both the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) 2015 and the COCO (Common Objects

in Context) 2015 Detection Challenge. The residual learning framework allows to efficiently train networks that are substantially deeper than the previously used CNN architectures. We used the pre-trained models on ImageNet publicly available[1] for Caffe [18].

To further improve the classification accuracy, we made a small change in the network architecture: the additional fully-connected layer with 512 neurons was added on top of the network, right before the softmax classifier. The activation function in the new layer is maxout [13] with 4 linear pieces. Dropout with a ratio of 50% is applied after the maxout layer and before the classifier. The final layer is a 1000-way softmax classifier corresponding to the number plant species needed to be recognized. Glorot [8] initialization was used for these two layers.

Thereafter, we have finetuned the network for 150,000 iterations in submissions "CMP Run 1" and "CMP Run 2" and for 370,000 iterations in "CMP Run 3", all with the following parameters:

- The learning rate was set to 10^{-3} and lowered by factor of 10 each 100 000 iterations.
- The momentum was set to 0.9, weight decay to $2 \cdot 10^{-4}$. r
- The effective batch size was set to 28 (either computed at once on NVIDIA Titan X, or split into more batches using Caffe's *iter_size* parameter when used on lower-memory GPUs).
- A horizontal mirroring of input images was performed during training.

Beyond fine-tuning the network, we performed bagging, inspired by its impact on the PlantCLEF 2015 results, where an interesting margin was gained with bagging of 5 networks by Choi [5]. Due to computational limits at training time, we only used bagging of 3 networks, although we believe that using a higher number of more diverse networks would further improve the accuracy. The voting was done by taking species-wise maximum of output probabilities.

We have also experimented with training Support Vector Machines (SVMs) on L2 normalized outputs of the last pooling layer "kernelized" using an approximate χ^2 feature map [30]. Platt's probabilistic output was used [22,25] to obtain comparable results with One-vs-All arrangement of the binary classifiers.

4 Preliminary Results and Validation

For our preliminary experiments and validation, we used the PlantCLEF 2016 training set, which is the union of previous year's (PlantCLEF 2015) training and test sets. Our validation process was threefold:

First, we validated networks trained on the PlantCLEF 2015 training set by evaluating them on the PlantCLEF 2015 test set using the previous year's score metric [9]. The goal of this phase was to validate that the ResNet-152 architecture is superior to the GoogleNet networks finetuned by the winners of the PlantCLEF 2015 challenge. While the best submission without bagging achieved

[1] `https://github.com/KaimingHe/deep-residual-networks`

a score of 59.4% on image observations in the PlantCLEF 2015 challenge, a finetuned ResNet-152 (without maxout) scored 62.1%. Applying SVMs on top of the last pooling layer pushed the score further to 62.5%, and training different SVMs for different groups of view types (always one for *Leaf* and *LeafScan*, second for *Flower* and *Fruit*, and third for *Stem*, *Branch* and *Entire*) gives another small improvement to 62.7%. Note that the meta information about view type is also available at test-time.

Table 1. Validation of the finetuned ResNet-152 using the PlantCLEF 2015 [9] score. *LR* denotes the learning rate, *it.* denotes the number of iterations, *sepSVM* is the set of SVM classifiers, each trained only on images of certain types (leaves, flowers & fruits, stem & branch & entire).

Method	*Score*
LR = 0.001, 70K it.	58.7%
LR = 0.001, step size 100K, 150K it.	62.1%
LR = 0.001, step size 100K, 150K it. + SVM	62.5%
LR = 0.001, step size 100K, 150K it. + sepSVM	**62.7%**

Second, we tested networks finetuned on PlantCLEF 2015 training data by evaluating them on the test set of PlantCLEF 2015 with the mAP metric chosen for PlantCLEF 2016. This step demonstrated the difference between the metrics used in PlantCLEF 2015 and PlantCLEF 2016.

While deploying the SVMs slightly improved the 2015 score, using the CNN SoftMax output worked better for the 2016 metric. This is probably due to better comparability of SoftMax outputs among different samples, which was not important for the 2015 score, but which is crucial for the mAP used in PlantCLEF 2016. Finetuning ResNet-152 for 150,000 iterations lead to 52.2% mAP, while finetuning for 130,000 iterations with maxout lead to 56.75% mAP, proving that using maxout improves the accuracy significantly. The training set size is big enough to finetune networks for a larger number of iterations without overfitting, which brings additional 0.5% of mAP points, see Table 2. The test-time 10-view image augmentation [20] (4 corner-crops, center crop and their mirrored versions) added only 0.2% of mAP, unlike in the ImageNet competition. Evaluating the network in a fully convolutional style on images scaled to 448 pixels (in longer dimension) surprisingly decreased the accuracy.

Third and lastly, we performed bagging on the selected pipeline by dividing the 2016 training set into three folds and finetuning 3 networks, each using a different fold for validation and the remaining two folds for finetuning. The only goal of this validation was to check, that finetuning of all 3 networks converged to a meaningful solution.

Table 2. Validation of the finetuned ResNet-152 using the PlantCLEF 2016 [10] mAP score. *LR* denotes the learning rate, *it.* denotes the number of iterations, *sepSVM* is the set of SVM classifiers, each trained only on images of certain types (leaves, flowers & fruits, stem & branch & entire).

Method	*mAP*
ResNet-50 (150K it.)	50.3%
ResNet-152 (150K it.)	52.2%
ResNet-152 (150K it.) + SVM	51.8%
ResNet-152 (150K it.) + sepSVM	50.6%
ResNet-152 + maxout (130K it.)	56.8%
ResNet-152 + maxout (130K it.) + 10-view test aug. [20]	56.9%
ResNet-152 + maxout (130K it.) + fully convolutional eval.	55.9%
ResNet-152 + maxout (370K it.)	**57.3**%

5 Results on the Test Set

Our primary submission - the bagging of 3 deep residual (ResNet-152) networks finetuned with maxout - scored with 71.0% mAP among the top 3 teams in the challenge and among top 7 runs, loosing by 3.2 to the leader. This submission is denoted as CMP Run 1 in Figure 2.

The second submission (CMP Run 2), achieving mAP 64.4%, was generated using only one of the three residual networks utilized in CMP Run 1. The difference between the first two submissions is 6.6%, underlining the importance of bagging.

The third submission (CMP Run 3), achieving mAP 63.9%, was generated by a network pretrained only on the LifeCLEF 2015 training set, for a higher number of iterations.

6 Conclusions

Our results confirm the suitability of very deep convolutional networks for plant recognition in the wild, allowing to use a unified end-to-end pipeline for recognition of different plant organs as well as overall views in an uncontrolled environment. Significant improvements, compared to the most successful approaches from the previous year, were achieved by deploying a very deep (152-layer) residual network and placing a maxout layer in front of the classifier.

6.1 Future Work

Very recent improvements to the ResNet architecture, such as deeper networks trained with stochastic depth [17] or the new residual unit proposed based on the latest study of identity mappings [16], would be the next steps for future work in plant recognition using deep residual networks.

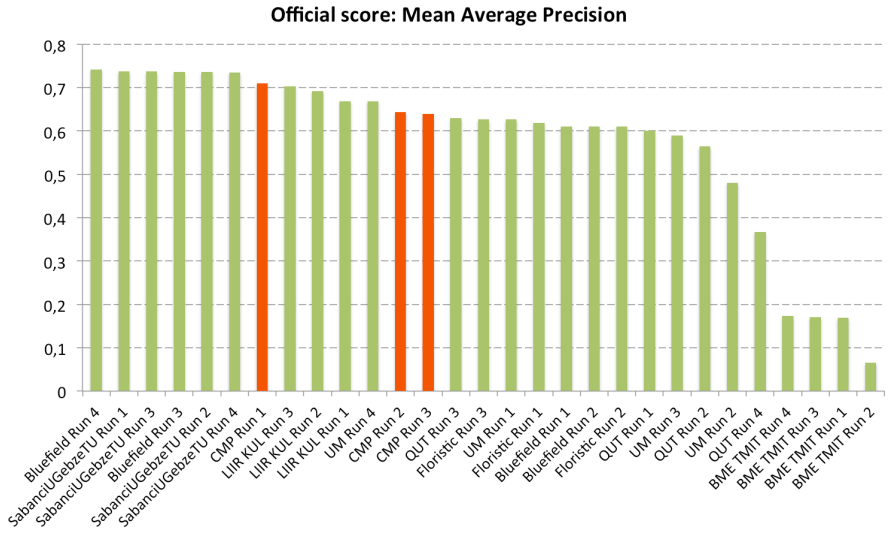

Fig. 2. Results of the main task in PlantCLEF2016 [10]. CMP results are in orange, our primary submission (CMP Run 1) scored among the 3 best performing teams.

The second open question is what is the best way of learning for domain adaptation. We believe that the results can be further significantly improved without any change in the CNN architecture.

One of the attractive applications of plant recognition are mobile apps with automatic field guides [21,12,11]. Compared to the standard automatic plant recognition approaches, deploying a CNN allows to process different types of images using exactly the same pipeline. However, since our networks are very deep and have a significant memory footprint (241MB), they should be optimized in terms of speed, memory and energy efficiency prior to utilization on mobile devices. The Deep Compression of Han et al. [14] is a promising direction for this optimization.

Acknowledgments

Milan Šulc and Dmytro Mishkin were supported by the CTU student grant SGS15/155/OHK3/2T/13, Jiří Matas by the Czech Science Foundation Project GAČR P103/12/G084. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

References

1. Angelova, A., Zhu, S., Lin, Y.: Image segmentation for large-scale subcategory flower recognition. In: Applications of Computer Vision (WACV), 2013 IEEE

Workshop on. pp. 39–45. IEEE (2013)
2. Bendale, A., Boult, T.: Towards open world recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1893–1902 (2015)
3. Boudra, S., Yahiaoui, I., Behloul, A.: A comparison of multi-scale local binary pattern variants for bark image retrieval. In: Advanced Concepts for Intelligent Vision Systems. pp. 764–775. Springer (2015)
4. Champ, J., Lorieul, T., Servajean, M., Joly, A.: A comparative study of fine-grained classification methods in the context of the lifeclef plant identification challenge 2015. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015. CEUR-WS (2015)
5. Choi, S.: Plant identification with deep convolutional neural network: Snumedinfo at lifeclef plant identification task 2015. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015. CEUR-WS (2015)
6. Fiel, S., Sablatnig, R.: Automated identification of tree species from images of the bark, leaves and needles. In: Proc. of 16th Computer Vision Winter Workshop. pp. 1–6. Mitterberg, Austria (2011)
7. Ge, Z., McCool, C., Sanderson, C., Corke, P.: Content specific feature learning for fine-grained plant classification. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015. CEUR-WS (2015)
8. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: International conference on artificial intelligence and statistics. pp. 249–256 (2010)
9. Goëau, H., Bonnet, P., Joly, A.: Lifeclef plant identification task 2015. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015. CEUR-WS (2015)
10. Goëau, H., Bonnet, P., Joly, A.: Plant identification in an open-world (lifeclef 2016). In: CLEF working notes 2016 (2016)
11. Goëau, H., Bonnet, P., Joly, A., Affouard, A., Bakic, V., Barbe, J., Dufour, S., Selmi, S., Yahiaoui, I., Vignau, C., et al.: Pl@ ntnet mobile 2014: Android port and new features. In: Proceedings of international conference on multimedia retrieval. p. 527. ACM (2014)
12. Goëau, H., Bonnet, P., Joly, A., Bakić, V., Barbe, J., Yahiaoui, I., Selmi, S., Carré, J., Barthélémy, D., Boujemaa, N., et al.: Pl@ ntnet mobile app. In: Proceedings of the 21st ACM international conference on Multimedia. pp. 423–424. ACM (2013)
13. Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y.: Maxout networks. arXiv preprint arXiv:1302.4389 (2013)
14. Han, S., Mao, H., Dally, W.J.: Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. ArXiv e-prints (Oct 2015)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385 (2015)
16. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. arXiv preprint arXiv:1603.05027 (2016)
17. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.: Deep networks with stochastic depth. arXiv preprint arXiv:1603.09382 (2016)
18. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)

19. Joly, A., Goëau, H., Glotin, H., Spampinato, C., Bonnet, P., Vellinga, W.P., Champ, J., Planqué, R., Palazzo, S., Müller, H.: Lifeclef 2016: multimedia life species identification challenges. In: Proceedings of CLEF 2016 (2016)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
21. Kumar, N., Belhumeur, P.N., Biswas, A., Jacobs, D.W., Kress, W.J., Lopez, I.C., Soares, J.V.: Leafsnap: A computer vision system for automatic plant species identification. In: Computer Vision–ECCV 2012, pp. 502–516. Springer (2012)
22. Lin, H.T., Lin, C.J., Weng, R.C.: A note on platt's probabilistic outputs for support vector machines. Machine learning 68(3) (2007)
23. Mattos, A.B., Herrmann, R.G., Shigeno, K.K., Feris, R.S.: Flower classification for a citizen science mobile app. In: Proceedings of International Conference on Multimedia Retrieval. p. 532. ACM (2014)
24. Nilsback, M.E., Zisserman, A.: An automatic visual flora-segmentation and classification of flower images. Oxford University (2009)
25. Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers 10(3) (1999)
26. Reyes, A.K., Caicedo, J.C., Camargo, J.E.: Fine-tuning deep convolutional networks for plant recognition. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015. CEUR-WS (2015)
27. Scheirer, W.J., Jain, L.P., Boult, T.E.: Probability models for open set recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36(11), 2317–2324 (2014)
28. Sulc, M., Matas, J.: Texture-based leaf identification. In: Computer Vision-ECCV 2014 Workshops. pp. 185–200. Springer (2014)
29. Sulc, M., Matas, J.: Kernel-mapped histograms of multi-scale lbps for tree bark recognition. In: Image and Vision Computing New Zealand (IVCNZ), 2013 28th International Conference of. pp. 82–87. IEEE (2013)
30. Vedaldi, A., Zisserman, A.: Efficient additive kernels via explicit feature maps. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(3), 480–492 (2012)