Deep Learning and SVM Classification for Plant Recognition in Content-Based Large Scale Image Retrieval

Bálint Pál Tóth, Márton Osváth, Dávid Papp, Gábor Szűcs

Department of Telecommunications and Media Informatics,
Budapest University of Technology and Economics
Magyar Tudósok krt. 2., H-1117, Budapest, Hungary,

{toth.b,pappd,szucs}@tmit.bme.hu, osvathmarton@gmail.com

Abstract. The PlantCLEF 2016 challenge focused on tree, herb and fern species identification based on different types of images. The aim of the task was to classify the plants in the images to species and to give a confidence score depicting the probability that a prediction is true. We elaborated different classification methods for this challenge. We applied dense SIFT for feature detection and description; and Gaussian Mixture Model based Fisher vector was calculated to represent an image with high-level descriptor. Fisher vectors were classified by a special SVM, the C-support vector classification algorithm with RBF (Radial Basis Function) kernel. Furthermore, we applied deep learning method to train convolutional neural network (CNN) for feature learning and fully-connected layers with softmax output for classification. We also combined these classifiers using the weighted average of their outputs. The final results show that the CNN achieved better result than the SVM, and the combined method slightly surpasses the CNN.

Keywords: deep learning, convolutional neural networks, GMM based Fisher vector, C-support vector classification

1 Introduction

Being able to identify the different species of plants growing in agricultural areas and to automatically detect the presence of invasive species is crucial. Identifying plants is usually a difficult task, sometimes for professionals (such as farmers or wood exploiters) as well. Using content-based image retrieval technologies is a promising possibility in this scenario. In order to solve it a challenge is announced in the LifeCLEF campaign [1].

The image-based plant identification task, briefly PlantCLEF 2016 [2] was focused on tree, herb and fern species identification based on different types of images. The number of species was 1000, and there were 7 viewpoints at the images: branch, leaf, scan (scan or scan-like pictures of leaf, briefly "LeafScan"), flower, fruit, stem, and entire views. The data were sample of the stream of the raw query images submitted by the users of the popu-

lar mobile application called Pl@ntNet (available on iPhone and Android), which accounts for several hundreds of thousands of active users submitting about ten thousands of query images daily.

The aim of the task was to classify images into the known categories (species), but the classification system had to be robust to unseen categories. It was a more difficult problem, because the test set contained images of species that were not in the training set (these are unseen categories). Besides the images contextual metadata (date, location, author and rating information) were also available.

2 Previous works

In the last two years there have been a number of successful deep learning, SVM and combined solutions in the LifeCLEF competition. In 2014 a combined system of convolutional neural nets and SVM won the challenge [3]. The CNNs had five convolutional layers, however their pure deep learning solution was outperformed with the combined systems. A part of our team participated in the same competition [4]. They used GMM based Fisher-vector for image representation, and SVM for classification. A different group, in the same year used the BoW model with OpponentColor SIFT descriptors and SVM, but their results were less convincing [5]. Also in 2014 another group used a pretrained Overfeat [6] network for feature learning, and the output of the fully-connected layer (before the softmax layer) was fed into a tree-based ensemble classifier [7]. However, other groups with SVM based solutions resulted better. In 2015 an Inception CNN model based network won the competition [8]. They have pretrained the model with ImageNet and fine-tuned with the PlantCLEF database. They used the combined output of five CNNs, that were fine-tuned with randomly selected parts of the database. Also in last year's competition a pretrained AlexNet was fine-tuned, which resulted the 4th place [9]. For fine-tuning they have reset the last (softmax) layer and they trained the last layer with relatively high learning rate (10) and the rest of the layers with a much lower learning rate (0.1).

We elaborated fully automatic methods (one by Fisher vectors and SVM, and another one by deep learning) for the classification of the images, and then they are taken in decreasing order based on reliability of classification decision; the next sections will present the details.

3 Classification by Fisher vector and SVM

The first part of the classification was the representation of each image based on visual content. Following the general trend, we applied BoW (Bag-of-Words) model [10–12] for this purpose. This consists of three steps: (i) fea-

ture detection, (ii) feature description, (iii) image description as usual phases in computer vision and we solved these steps similarly to our previous work [3].

For feature detection and description we used the SIFT (Scale Invariant Feature Transform) algorithm [13] with dense keypoint sampling. After that, we performed PCA (Principal Component Analysis) [14, 15] to reduce the dimensions of the descriptor vectors from 128 to 80. Finally, we encoded the low-level descriptor vectors to get GMM (Gaussian Mixture Model) [16, 17] based Fisher-vectors [16, 18]. These vectors were the final representations (image descriptor) of the images.

For the classification subtask we used a variation of SVM (Support Vector Machine), the C-SVC (C-support vector classification) [19, 20] with RBF (Radial Basis Function) kernel. We applied the one-against-all technique to extend SVM for multi-class classification. Furthermore, a validation set were used to optimize the two hyperparameters (C from C-SVC and γ from RBF kernel).

The results of SVM classification was submitted as '**BME TMIT Run2**'.

4 Classification by deep learning

Nowadays state-of-the-art image recognition and classification solutions generally use deep learning methodology. Deep convolutional neural networks are able to learn the descriptive features of the image database in many abstraction levels. Convolutional neural networks raised a lot of interests in 2012, when a team led by Geoffrey Hinton and Alex Krizhevsky won the ImageNet Large Scale Visual Recognition Competition [21] by a large margin [22]. This model is often referred to as AlexNet. AlexNet consists of five convolutional layers, from which the first, second and fifth are followed by max-pooling layers. This part is responsible for feature learning. The second part of AlexNet includes three fully-connected layers with an output layer of 1000 softmax neurons for classification.

In the data preparation phase we applied cropping, scaling and normalization. Hence we only cropped the center of the images along the shorter dimension and scaled it down to the network's input dimension, which is 224x224 pixels. Finally, we normalized the red, green and blue color channels individually to zero mean and unit variance. An example of the resulting image is shown in Figure 1.

Figure 1. Example of an image before (left) and after cropping and normalization (right).

For training the PlantCLEF 2015 database we used a modified version of AlexNet [22]. We changed the ReLU activation functions to parametric ReLUs (PReLUs) [23]. Furthermore, we applied batch normalization [24] before the max-pooling layers of AlexNet. The block diagram of the proposed convolutional network is shown in Figure 2.

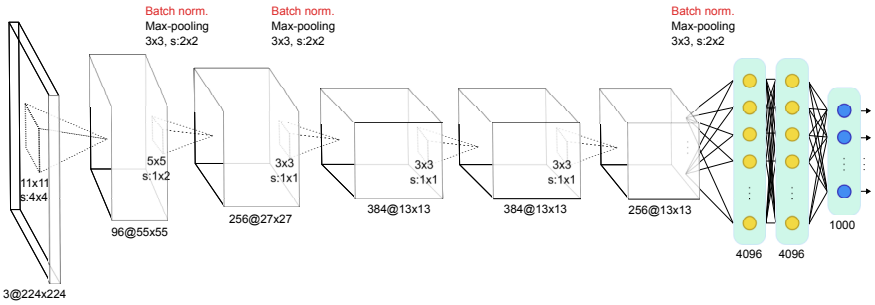

Figure 2. The block diagram of the proposed convolutional network, which is a modification of AlexNet [16]. (A@BxB refers to A number of planes with size BxB. The CxC, s: DxE refers to CxC kernel size with DxE stride.)

For optimizer we chose AdaDelta [25], which is a great tool for adaptively adjusting the learning rate. Negative log-likelihood criterion was used for multi-class classification purposes. We performed hyperparameter optimization with manual grid search in terms of batch size. If there wasn't improvement in the global correct rates in 100 epochs the training was stopped. According to the results that are shown in Table 1, 130 was chosen as the batch size. The loss and the average correct classification rates within rows of the confusion matrix are shown in Figure 3.

The hardware we used for training were a NVidia GTX 970 (4 GB) and a NVidia Titan X (12 GB) GPU cards hosted in two i7 servers with 32 GB RAM. Ubuntu 14.04 with Cuda 7.5 and cuDNN 4.0 was used as general software architecture. For data preparation, training and evaluating deep neural networks the Torch7 [26] deep learning framework was used. For calculat-

ing mean average precision (MAP) values the sklearn Python package was used.

The results of CNN classification was submitted as '**BME TMIT Run1**'.

Table 1. Batch size optimization. The global correct rates correspond to the ratio of the overall correct classification in the confusion matrix. 'Early stopped #epochs' refers to the number of epoch when early stopping was applied.

Batch size	Train global correct [%]	Validation global correct [%]	Early stopped #epochs
20	96.19	32.98	189
30	89.57	29.26	138
40	93.96	26.54	199
50	92.99	25.32	209
60	68.35	24.48	124
80	89.33	23.52	226
100	87.79	31.63	102
130	90.28	37.76	148
150	84.5	20.66	286
200	47.31	20.96	208

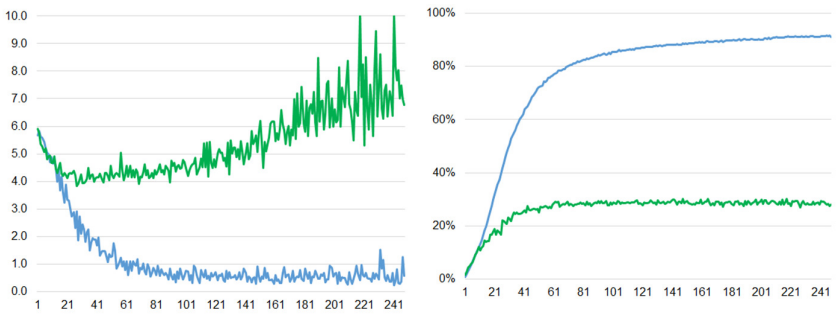

Figure 3. Loss (left) and average correct rows in the confusion matrix (right) during training (blue: train, green: validation). The horizontal axis corresponds the number of epochs.

5 Combination of the classifiers

We combined the outputs of the classifiers. Based on our preliminary testing the weighted average of the outputs were used for creating the '**BME TMIT Run4**'. Besides two classifiers described above a third one was constructed using metadata. After normalizing and cleaning the data we calculated new metadatas (e.g. the season) in order to get more informative variables. Next we applied Random Forest (RF), and we measured the accuracy. The

metadata based classification had the lowest MAP value, SVM and CNN gave much better results. Therefore, we chose the following weight parameters: 0.1 for metadata, 0.3 for SVM and 0.6 for CNN.

Furthermore, we built an aggregated model to detect unseen categories by attempting to filter out the images with unknown classes. We measured the largest distance among fisher vectors of training images; in the case if a test image's fisher vector is farther from all training fisher vectors then this distance, we reject that particular image (as outlier). As well as, images with very low overall (0.3) decision values were also rejected. As a result, only the remaining test images were included in '**BME TMIT Run3**'.

6 Evaluation

We trained 1 CNN and 7 SVM classifiers (one for each viewpoint) and we conducted a preliminary testing on the PlantCLEF 2015 test data, and measured the MAP (Mean Average Precision) values. The results of the preliminary evaluation can be seen in the first row of Table 2.

Some of the test images had no viewpoint attribute at all, and some of them were marked as 'Other'. Therefore, an image with known viewpoint was classified with the appropriate SVM classifier (with same viewpoint), and decision values of this classifier were used in the runfile as predictions. For testing an image with unknown viewpoint we constructed a classifier using the weighted average of the decision values coming from all trained classifiers. At the estimation of weight parameters we took the "goodness" of different viewpoint classifiers into consideration: LeafScan: 0.3, Leaf: 0.15, Flower: 0.15, Fruit: 0.15, Stem: 0.15, Branch: 0.05, Entire: 0.05.

In Table 2, we presented multiple MAP (Mean Average Precision) values based on the size of the sorted lists associated with the classes. The results of the post-testing (i.e. test was after the run submission) of SVM and CNN show that the best MAP value is reached when only the 10 most probable predictions are involved in the calculations. Unfortunately, we analyzed the influence of MAP calculations (second and third rows of this table belong to post-testing) only after the official results were released. For the competition we submitted predictions of all the 1000 classes for every test image - that caused worse MAP value.

The evaluation was executed on the PlantCLEF 2015 test data, and we calculated the predictions on 2016 test data. It is important to note that these runfiles contained the total results, which means that we gave all decision values for each ClassId for each MediaId.

Table 2. MAP values of the SVM and CNN network with different length of the sorted lists.

Length of the sorted lists	SVM	CNN
total	0.1365	0.0592
top 100	0.1774	0.1852
top 10	0.2062	0.2671

7 Official results

In the official evaluation MAP was used for measurement of goodness of the image classification, considering each class Ci of the training set as a query. In this query evaluation all predictions with ClassId=Ci in the runfile were extracted, and ranked by decreasing probability and the average precision (AP) was computed for that class. The MAP is mean of these AP values. To evaluate more specifically the targeted usage scenario consisting in detecting invasive species, a secondary MAP was computed by considering as queries only a subset of the species that belongs to a blacklist of invasive species. Recognition system was expected to be robust to unseen categories by automatically detecting the numerous false positives classification hits. The official results can be seen in Table 3, where the first two columns contain the unseen categories, while the last column ignores them.

Table 3. Official results of the 4 submitted runs.

Runs	Official score MAP	MAP restricted to a blacklist of (potentially) invasive species	MAP ignoring unknown classes and queries
BME TMIT Run1	0.169	0.125	0.196
BME TMIT Run2	0.066	0.128	0.101
BME TMIT Run3	0.17	0.125	0.197
BME TMIT Run4	0.174	0.144	0.213

8 Conclusion

We elaborated different classification methods for image-based plant identification task. We applied dense SIFT for feature detection and description; and Gaussian Mixture Model based Fisher vector was calculated to represent an image with high-level descriptor. The chosen classifier was the C-support

vector classification algorithm with RBF (Radial Basis Function) kernel, and we optimized two hyperparameters (C from C-SVC and γ from RBF kernel) by a grid search with two-dimensional grid.

We also used convolutional neural networks for the task. The images were normalized to zero mean and unit variance, they were also cropped and scaled down. We used a modified version of the AlexNet model, and we performed batch size optimization. With the winning batch size the deep learning method achieved considerably higher MAP score than SVM.

We constructed a classifier by combining the decisions values of the metadata, SVM and CNN classification methods. The weights of these techniques were determined based on the preliminary tests. Furthermore, a novel approach was used for rejecting the outlier test images (i.e. images with unseen categories), this approach used the information coming from both distance measurement of Fisher vectors and CNN.

It should be noted that our team was formed at the middle of March and we started working on the project in April. According to the investigation of MAP calculations, we must admit that if we would have optimized the length of the sorted lists for the MAP calculations, we might have achieved better results in the official competition.

Acknowledgement

Bálint Pál Tóth gratefully acknowledges the support of NVIDIA Corporation with the donation of an NVidia Titan X GPU used for his research.

References

1. 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, Proceedings of CLEF 2016 (2016)
2. Goëau, H., Bonnet, P., Joly, A. Plant identification in an open-world (LifeCLEF 2016), CLEF working notes 2016 (2016)
3. Chen, Q., Abedini, M., Garnavi, R. and Liang, X.: IBM Research Australia at LifeCLEF2014: Plant Identification Task. In CLEF (Working Notes), pp. 693-704. (2014)
4. Szűcs, G., Papp, D., Lovas, D., Viewpoints Combined Classification Method in Image-based Plant Identification Task In: Cappellato L., Ferro N., Halvey M., Kraaij W. (eds) Working Notes for CLEF 2014 Conference. Sheffield, Great Britain, September 15-18, pp. 763-770. Vol. 1180. (2014)
5. Issolah, M., Lingrand, D., & Precioso, F. Plant Species Recognition using Bag-Of-Words with SVM classifier in the Context of the LifeCLEF Challenge In: Cappellato L., Ferro N., Halvey M., Kraaij W. (eds) Working Notes for CLEF 2014 Conference. Sheffield, Great Britain, September 15-18, pp. 738-746. Vol. 1180. (2014)

6. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun. Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
7. Sünderhauf, N., McCool, C., Upcroft, B. and Perez, T.: Fine-Grained Plant Classification Using Convolutional Neural Networks for Feature Extraction. In CLEF (Working Notes), pp. 756-762. (2014)
8. Sungbin C.: Plant identification with deep convolutional neural network: SNUMedinfo at LifeCLEF plant identification task 2015, In Working notes of CLEF 2015 conference. 2015.
9. Reyes, A. K., Caicedo, J. C., & Camargo, J. E. : Fine-tuning deep convolutional networks for plant recognition. In Working notes of CLEF 2015 conference. (2015)
10. Fei-Fei, L., Fergus, R., & A. Torralba, A.: Recognizing and Learning Object Categories, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), (2007)
11. Chatfield, K., Lempitsky, V., Vedaldi A. and Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods, British Machine Vision Conference, pp. 76.1-76.12. (2011)
12. Lazebnik, S., Schmid, C. and Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, Vol. 2, pp. 2169-2178. (2006)
13. Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, Vol. 60, No 2., pp. 91-110. (2004)
14. Abdi H., Williams L. J.: Principal Component Analysis, Wiley Interdisciplinary Reviews: Computational Statistics, Vol 2. No. 4, pp. 433-459. (2010)
15. Ke, Y., & Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors, In Computer Vision and Pattern Recognition, CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, Vol. 2, pp. II-506. (2004)
16. Reynolds D. A.: Gaussian Mixture Models, Encyclopedia of Biometric Recognition, Springer, February, pp. 659-663. (2009)
17. Tomasi C.: Estimating gaussian mixture densities with EM: A tutorial, (Tech. rep., Duke University); Chinese Journal of Electron Devices, pp, 15-18. (2004)
18. Perronnin, F., Dance, C.: Fisher kernel on visual vocabularies for image categorization, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), (2007)
19. Boser, B., Guyon, I., Vapnik, V.: A Training Algorithm for Optimal Margin Classifier, Proc. of the 5th Annual ACM Workshop on Computational Learning Theory, pp. 144-152. (1992)
20. Cortes, C., Vapnik, V.: Support-vector networks, Machine Learning, Vol. 20, No. 3, pp. 273-297. (1995)
21. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Hunag, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., & Fei-Fei, L. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), 211-252. (2015)
22. 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)

23. He, K., Zhang, X., Ren, S., & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. arXiv preprint arXiv:1502.01852 (2015)
24. Ioffe, S. and Szegedy, C. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." arXiv preprint arXiv:1502.03167 (2015)
25. Zeiler, M. D. "ADADELTA: an adaptive learning rate method." arXiv preprint arXiv:1212.5701 (2012)
26. Collobert, R., Kavukcuoglu, K., & Farabet, C. Torch7: A matlab-like environment for machine learning. In BigLearn, NIPS Workshop (No. EPFL-CONF-192376). (2011)