IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-9, Sept. 2018 E-ISSN: 2347-2693

Enhancement of Image Classification through Data Augmentation using Machine Learning

Th. S. Kumar

Girijananda Chowdhury Institute of Management and Technology, Guwahati, India

Corresponding Author: thshantakumar@gmail.com, Tel.:+00-86380-48669

Available online at: www.ijcseonline.org

Accepted: 16/Sept/2018, Published: 30/Sept/2018

Abstract—Identification of plants species has become one of the challenges for image processing and machine learning. The need to find an efficient solution to such a problem is essential as medicinal plants and new plants' existence need to be studied. Most of the researches in identifying this plants species are based on color, shape and textures. This paper is based on these features with Data-Augmentation. Data-augmentation is an important technique in increasing the number of training dataset which further helps in increasing the prediction of classification. This paper uses machine learning algorithms in classifying the flower classes based on FLOWERS17 dataset. Data-augmentation is applied to the training dataset to enhance the prediction. It has been observed that Random Forest classifies flowers with an accuracy of 64% before data-augmentation and 94% after data-augmentation. This paper also shows that after increasing the number of classes from 17 to 21, the performance of Random Forest is consistent to 94%.

Keywords— Data Augmentation, Flower Recoginition, Image Processing, Machine Learning

I. Introduction

Flower identification from images still remains a challenge in image processing domain due to existence of variety of flowers in nature. Different models have been developed using deep learning. However, these models need a large amount of data as input. In many problems, this requirement of large dataset is difficult to fulfil and hence an alternative solution is required. A technique of identification of flower is seen in [1], where two input images, one for flower and other for leave, are required. This method also requires that a black cloth be placed behind the objects in order to recognise the object which is not convenient in real situations. Many of the plant species identification system like Pl@ntNet [2] and CLOVER [3] require the leaf as an input which is difficult as it requires expertised knowledge of flowers.

Reference [4] is an example that combines aspect ratio, eccentricity and Moving Median Center for identification purpose. The combination of shape, color and texture features with Zernike moments with Radial Basis Probabilistic neural networks is seen in [5]. Identification of flowers based on texture, color and shape features is seen in [6] and [7] which were tested on 103 labels. Combining domain knowledge of flowers and color clustering is seen in [8] which accurately identify flowers in images. The recent work of [9] which uses deep learning in identifying flower species have led to big success in terms of real time application using mobile phones.

The rest of the paper is organized as follows. Section II covers details of the global feature descriptor. Section III describes the different data-augmentation techniques used in this paper. Section IV highlights the different machine learning algorithms implemented. Section V gives the results and discussion. Section VI concludes the research work with future directions.

II. Global Feature Descriptors

A feature descriptor is a representation of an image that processes an image extracts useful information and removes inappropriate information in it. For the purpose of flower image identification the following attributes are most popularly used:

A. Color

'Color' happens to be the best feature when it comes to recognising flower species. The most commonly used color feature descriptor is the Color Histogram. It calculates the frequency of pixel intensities in a given image. It is obtained by counting the number of pixel in each of the RGB channel. The result thus obtained is plotted on 3 individual bar graphs. Certain statistics measure like mean could also be calculated to find the color distribution in the image. Color feature descriptor alone is not sufficient to identify flowers as two or more flowers may have the same colour and at the same time one species of flower may have different colors.

B. Texture

'Texture' is another feature used to identify flower species. It gives us uniformity of the combination of color and patterns in the image. Haralick et al. [10], described a number of texture extraction methods which were basically divided into two - structural and statistical. They introduced 14 statistical features based on "grey-tone spatial-dependence matrices" using which one could easily quantify images. The 14th feature is sometimes avoided due to the high computational overhead.

C. Shape

Another feature detector is 'shape' which is popularly used especially for natural objects. Two important shape detectors are Hu moments and Zernike moment in image processing, computer vision and related fields. Image moments compute the weighted average of the image pixel's intensities. They give useful information which describe about the image after segmentation like area (total intensity), its centroid and orientation. The Hu moments consist of seven moments – mean, variance, standard deviation, skew, kurtosis etc. forming a feature vector of size 7-d. Zernike moments introduced by Teague is another shape descriptor which is based on orthogonal functions.

D. Local Binary Patterns (LBPs)

Local Binary Patterns (LBPs) is texture classification and many other applications such as face detection, facial expression recognition, pedestrian detection etc. there are different variants of LBPs used in image processing, the simplest of which uses 3×3 window then processes it to extract an LBP code. It thresholds the central pixel of this 3×3 window with the surrounding pixels using window mean or window median. Once the LBP values are ready, a histogram with 256 bins is computed which is further normalised and concatenated. Different variations of this uniform pattern in LBP were also proposed to reduce the processing overhead.

E. Histogram of Oriented Gradients (HOG)

HOD is another global descriptor which uses the concept of gradient orientations in a particular section of the image. This technique differs from others in that it is computed on a dense grid of uniformly spaced cells and uses overlapping contrast normalisation.

F. Segmentation

Segmentation happens to be the first step in image processing transforming a given image to high-level image description. It partitions an input image into distinct regions with similar features. This is done in order to remove the unwanted background while only the foreground information is extracted. Grabcut segmentation algorithm is one important algorithm used in computer vision.

III. DATA AUGMENTATION

This section describes the different data-augmentation techniques used in order to increase the number of training dataset for every image. The different techniques are:

A. Image noise

Image noise is simply the random variation of brightness in an image. It is an undesirable result from the original image which is sometimes incomprehensible. Noise can be introduced in an image using simple linear function such as the one given in (1) or with a multiplication of non-linear function.

$$f(i,i) = s(i,j) + n(i+j) \tag{1}$$

where *f(i,j)* is the new image, *s(i,j)* is the original image and *n(i,j)* is the noise

B. Flipped

A flipped image is an image obtained by a mirror-reversal of the original across the axes. Both flipping across the horizon and vertical can be used depending on the application.

C. Rotate

Tilting of image is one technique used in data augmentation in the domain of computer vision. This paper uses any random angle between [-25^0, 25^0].

D. Contrast

The images in RGB are first mapped to an HSV color map. A random factor is used to multiply and alter the S and V components and then the image is converted back to RGB.

E. Blur

One technique popularly used in image processing to blur image is the Gaussian blur or Gaussian smoothing. This paper uses a kernel size of $2\lceil\sigma\rceil+1$, where σ stands for standard deviation of the distribution which is any random number between 2 and 8.

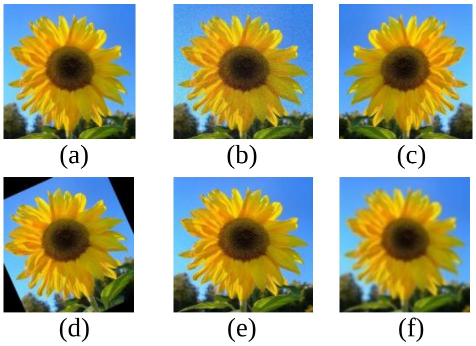


Figure 1. Example of a augmented dataset. (a) original, (b) with noise, (c) flippled, (d) rotate, (e) contrast, (f) blur

An example of the dataset generated using data-augmentation with these techniques is given in Figure 1.

IV. Machine Learning Algorithms

The following machine learning algorithms are used in order to select the most efficient algorithm for image classification:

A. Logistic Regression

Logistic regression is normally used to predict the probability of a binary outcome. However, the binary logistic regression can be generalised to more dependent variables. Categorical variables are modelled by multinomial logistic regression while ordinal logistic regression is used for ordinal variables. The prediction can have one or more predictors, both numerical and categorical [11].

B. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA), is a generalisation of Fisher's linear discriminant which is used in pattern recognition and machine learning in finding linear combination of certain features which separates the classes. The result of LDA may be used for dimensionality reduction before classification which is done in order to avoid overfitting and to reduce computational overhead.

C. K Nearest Neighbors

The k-nearest neighbors (kNN) is the simplest and the oldest classifier. It classifies each of the test data by the majority of the k-nearest neighbors in the training set. A suitable distance metric is chosen to find the distance between objects. This distance metric thus selected influences the overall performance of the algorithm. Euclidean distance is used in most of the cases. However it does not exploit on any statistical information that can be extracted from a large training set [12].

D. Decision Tree

A decision trees classifier repetitively divides the training dataset into subparts. In the process, it selects the attributes that contributes maximum information. It uses entropy or the measure of impurity, which is given by:

$$H = -\sum p(x)log_{p}(x) \tag{2}$$

Where, *p(x)* is the probability of item x.

E. Random Forest

Random forest builds a forest of decision trees. Bagging method is used in most of the cases. The principle of bagging method is that when learning models are combined the overall result increases. In the process of creating the forest, it adds additional randomness to the model. It searches for the best feature from a set of random features.

F. Naïve Baysian classifier

This classifier is based on Bayes' Theorem. It assumes that there is independence among predictors, that is, the presence of a feature in a label is not related to the presence of other feature.

$$R = \frac{P(i \mid X)}{P(j \mid X)} = \frac{P(i)P(X \mid i)}{P(i) \mid P(X \mid j)} = \frac{P(i)\prod P(X_r \mid i)}{P(j)\prod P(X_r \mid j)} \tag{3}$$

Equation (3) is used to compare the two probabilities and then the larger probability is taken to be the predicted class label.

G. Support Vector Machine

A support vector machine (SVM) is another popular a discriminative classifier used for classification. It constructs a hyperplane or set of hyperplanes in a high-dimensional space. It can be used for both the tasks of classification and regression. When the distance between the nearest training-data point of any class or margins is large enough, it is said to have a good separation. Larger the margin lower is the generalisation error of the classifier [13].

V. Experiments and Results

FLOWERS17, having 17 species of flowers, from Visual Geometry group at University of Oxford is a benchmark dataset for image processing domain. The dataset has been extended by incorporating with a collection of 4 more labels and is named as FLOWERS21. Each class of flowers has 80 images. Thus the numbers of images are 17×80=1360 and 21×80=1680 respectively. The five data-augmentation techniques - Image noise, Flipped, Rotate, Contrast and Blur are applied to these datasets and thus the number of dataset has been increased to 17×80×6=8,160 and 21×80×6=10,080 respectively including the original image.

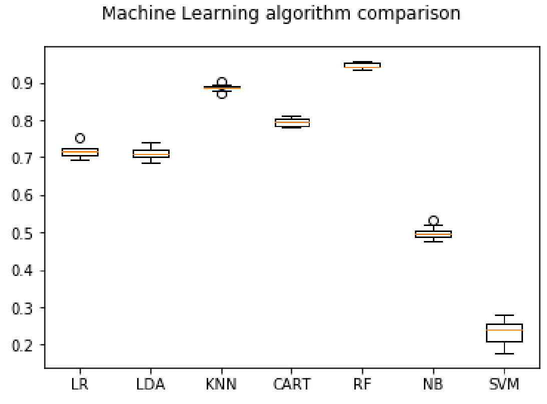


Figure 2. Comparision of algorithms after Data-Augmentation

The experiment uses K-Fold cross validation with K=10. Thus both the datasets FLOWERS17 and FLOWERS21 are divided into 90-10 splits for training and testing uniquely over each round up to 10 times. Then the succeeding process is divided into three parts. First, the features from the training datasets are extracted using color histogram, Haralick texture and Hu moments. Second, the training dataset is trained using machine learning techniques – Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Decision Tree (CART), Random Forest (RF), Naïve Baysian Classifier (NB), Support Vector Machine (SVM). Third, the test dataset is given to the models for prediction. The process is repeated for both the datasets FLOWERS17 and FLOWERS21 before and after data-augmentation. Then the accuracy of the model's prediction is evaluated. Figure 2 gives the results of the comparison of the seven machine learning algorithms.

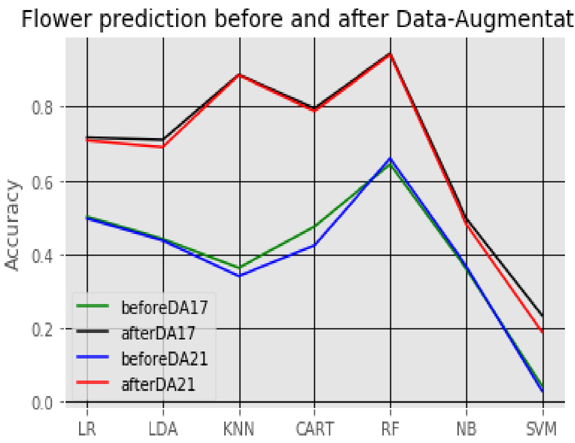
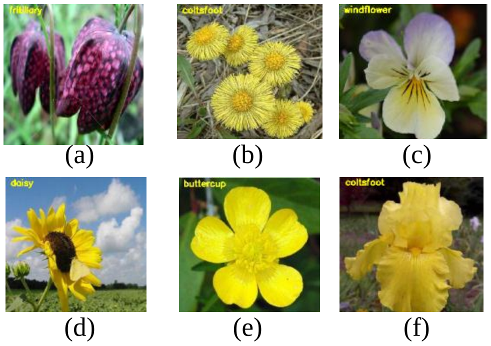


Figure 3. Accuracy comparison of the 7 machine learning techniques

It has been observed that the prediction is good after data-augmentation in both the two datasets as can be seen in Figure 3. Random Forest is found to give an accuracy of 0.64% and 0.66% in the two datasets before data-augmentation. After data-augmentation it gave an accuracy of 0.94% and 0.94% in the two datasets respectively. It may be noted that Support Vector Machine gave a poor accuracy of 0.04% and 0.02% before data-augmentation and 0.23% and 0.19% after data-augmentation for the two datasets respectively.

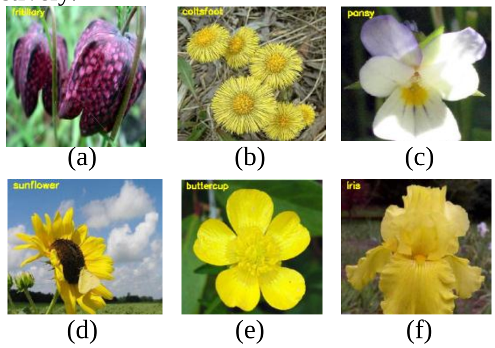


(a) (b) (c)

(d) (e) (f)

Figure 4. Prediction before data-augmentation. (a) fritillary (b) coltsfoot (c) windflower [wrong] (d) daisy [wrong](e) buttercup (f) coltsfoot [wrong]

Random Forest being the best technique is chosen to predict some images randomly. The results of this prediction before and after data-augmentation are given in Figure 4 and Figure 5 respectively.



(a) (b) (c)

(d) (e) (f)

Figure 5. Prediction after data-augmentation. (a) fritillary (b) coltsfoot (c) pansy (d) sunflower (e) buttercup (f) iris

VI. Conclusion And Future Scope

This paper demonstrates that by applying Data-Augmentation to the given dataset, the accuracy level increases. This is true even after increasing the number of class labels. Random Forest is found to give an accuracy of 64% and 94% when applied to FLOWERS17 before and after Data-Augmentation. Further when the number of class is increased to 21 Random Forest gives 66% and 94% accuracy with FLOWERS21 before and after Data-Augmentation. On the contrary Support Vector Machine gives a very poor accuracy. To handle a millions of flower and plant species around the world, a more robust method is required so that the user can take an image of any parts of a plant – leaves, fruits, branches and the system could identify if the plant can be used for medicinal purposes.

References

[1] T. Saitoh and T. Kaneko, "Automatic recognition of wild Flowers", Pattern Recognition, Proceedings. 15th International Conference on, vol.2, no., pp.507-510 vol.2, 2000.

[2] D. Barthelemy. "The pl@ntnet project: A computational plant identification and collaborative information system", Technical report, XIII World Forestry Congress, 2009.

[3] Y. Nam, E. Hwang, and D. Kim, "Clover: A mobile content-based leaf image retrieval system", In Digital Libraries: Implementing Strategies and Sharing Experiences, Lecture Notes in Computer Science, pages 139-148, 2005.

[4] J.-X. Du, X.-F. Wang and G.-J. Zhang, "Leaf shape based plant species recognition", Applied Mathematics and Computation, vol. 185, 2007.

[5] H. Kulkarni, H. M. Rai, K. A. Jahagirdar and P. S. Upparamani, "A Leaf Recognition Technique for Plant Classification Using RBPNN and Zernike Moments", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 1, pp. 984-988, 2013.

[6] M.E. Nilsback and A. Zisserman, "A Visual Vocabulary for Flower Classification", Computer Vision and Pattern Recognition, IEEE Computer Society Conference on. Vol.2, 2006.

[7] M.E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes", Indian Conference on Computer Vision, Graphics and Image Processing. pp. 722-729, 2008.

[8] S. Fadzilah, M.A.Salahuddin, and S.A. Yusof, "Digital Image Classification for Malaysian Blooming Flower", Computational Intelligence, Modelling and Simulation (CIMSiM), IEEE, 2010.

[9] I. Gogul and V.S. Kumar, "Flower Species Recognition System using Convolution Neural Networks and Transfer Learning", 4th International Conference on Signal Processing, Communications and Networking (ICSCN -2017), March 16–18, 2017, Chennai, India

[10] R.M. Haralick, K. Shanmugam, I.H. Dinstein, "Textural Features for Image Classification", IEEE Transactions on Systems, Man and Cybernetics, Vol.SMC-3, No. 6, November 1973, pp.610-621, 1973.

[11] A.B. Walker, S.H., Duncan, DB (1967). "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54 (1/2): 167–178. doi:10.2307/2333860. JSTOR 2333860

[12] C. Domeniconi, D. Gunopulos, J. Peng, "Large margin nearest neighbor classifiers" in IEEE Transactions on Neural Networks, 2005. https://doi.org/10.1109/TNN.2005.849821

[13] J. Han and M. Kamber, Data Mining: Concepts and Techniques, The Morgan Kaufmann Series, 2006

Author Profile

Dr. Th. Shanta Kumar pursued Bachelor of Science (Honours) from Manipur University, Manipur in 1995, Master in Computer Application from Jorhat Engineering College, Assam in 2000 and Ph.D. from Himachal Pradesh University, Himachal Pradesh in 2012. He is currently working as Associate Professor and Head of the Department, Department of Computer Science and Engineering, Girijananda Chowdhury Institute of Management and Technology, Assam. He has published more than 13 papers in Journals and Conferences. He also has contributed chapters to books. His main research work focuses on Data Mining Algorithms, Machine Learning and Image Processing. He is a life member of International Association of Computer Science and Information Technology (IACSIT), International Association of Engineers (IAENG) and Science and Engineering Institute (SCIEI). He has more than 17 years of teaching experience and 12 years of Research Experience.