Revista Brasileira de Computação Aplicada, July, 2022

DOI: 10.5335/rbca.v14i2.13487
Vol. 14, Nº 2, pp. 85–94
Homepage: seer.upf.br/index.php/rbca/index

ORIGINAL PAPER

Data Augmentation policies and heuristics effects over dataset imbalance for developing plant identification systems based on Deep Learning: A case study.

Luciano Araújo Dourado Filho[1], and Rodrigo Tripodi Calumby[1]

[1]University of Feira de Santana

*lucianoadfilho@ecomp.uefs.br; rtcalumby@uefs.br

Received: 2022-05-01. **Revised:** 2022-06-13. **Accepted:** 2022-07-20.

Abstract

Data augmentation (DA) is a widely known strategy for effectiveness improvement in computer vision models such as Deep Convolutional Neural Networks (DCNN). Although it enables improving model generalization by increasing data diversity, in this work we propose to investigate its effects with respect to two different sources of dataset imbalance (i.e., Content and Sampling imbalance) in a plant species recognition task. We systematically evaluated several techniques to generate the augmented datasets used to train the DCNN models that enabled a thorough investigation over the effects of DA in terms of imbalance attenuation. The results allowed inferring that data augmentation enables mitigating the negative effects related to underrepresentation mainly caused by the dataset imbalance.

Keywords: Data Augmentation; Deep Learning; Plant Recognition.

Resumo

O Data augmentation (DA) é uma estratégia amplamente conhecida para melhoria da eficácia em modelos de visão computacional, como Deep Convolutional Neural Networks (DCNN). Embora permita melhorar a generalização do modelo aumentando a diversidade de dados, neste trabalho propomos investigar seus efeitos em relação a duas fontes diferentes de desequilíbrio de conjunto de dados (ou seja, desequilíbrio de conteúdo e amostragem) em uma tarefa de reconhecimento de espécies de plantas. Avaliamos sistematicamente várias técnicas para gerar os conjuntos de dados aumentados usados para treinar os modelos DCNN que permitiram uma investigação completa sobre os efeitos da DA em termos de atenuação do desequilíbrio. Os resultados permitiram inferir que o aumento de dados permite mitigar os efeitos negativos relacionados à sub-representação causada principalmente pelo desequilíbrio do conjunto de dados.

Palavras-Chave: Aprendizagem profunda; Dados aumentados; Reconhecimento de plantas.

1 Introduction

For several years, machine learning algorithms for tackling computer vision problems involved handcrafted solutions for specific domains. In the context of plants, for instance, most studies regarded low-level features in single-organ (e.g., flower, fruit or leaves) tasks like detection, recognition, and segmentation (Chatfield et al., 2014). Such solutions have been continuously outperformed since *Deep Learning* (DL) architectures such as Deep Convolutional Neural Networks (DCNN) became more efficient and popular, after being considered infeasible for a long time.

The use of high-end parallel computing architectures associated to the automatic extraction of increasingly high-level characteristics from data (Lee et al., 2017) allowed to overcome important barriers previously associated to handcraft feature extraction algorithms (Pawara, Okafor, Surinta, Schomaker and Wiering, 2017; Chatfield et al., 2014; Mehdipour Ghazi

et al., 2017). Motivated by the capacity of achieving surprisingly high classification performance in several domains, this breakthrough motivated a notable endeavor of acquiring larger datasets to provide the massive amounts DL models often require.

Regarding plant species identification, the effort from botany experts and the interested community for scanning and cataloging thousands of specimens allowed building large repositories with data from several regions around the world through different sources (Kumar et al., 2012; Soltis, 2017; Beech et al., 2017; Goëau et al., 2019). This fostered the development of DCNN models capable of even surpassing human experts in species identification challenges (Goëau et al., 2018; Bonnet et al., 2018). Despite that, there are regions of the world for which there is still not enough data also with poor representativeness of species visual variations. This results in irregularity within datasets classes distributions and visual diversity (Buda et al., 2018; Graves et al., 2016; Picek et al., 2019) which may cause DCNN models to overfit and deteriorate its generalization ability for species with only a few sample images available.

These adversities are naturally expected due to difficulties behind the process of sample acquisition that also could be related to occurrence and incidence of the specimens in nature. There are examples of domains in which these issues are extraordinarily significant, to the point that the frequency of one class (e.g., a disease) being 1000 times less incident than another (e.g., healthy patient) (Buda et al., 2018). These challenges motivated researchers, specially in the field of plant recognition, to explore Data Augmentation (DA) to attenuate the lack of large amounts of samples and representative variations. DA allows to artificially extend a dataset through the application of label-preserving transformations over real samples. Therefore, its effectiveness in attenuating dataset imbalance, reducing overfitting and introducing invariance in DL models has being widely discussed (Shorten and Khoshgoftaar, 2019; Taylor and Nitschke, 2018; Cubuk et al., 2019; Mehdipour Ghazi et al., 2017).

In this context, DA has been shown an indispensable tool for achieving state-of-the-art classification performance (Mehdipour Ghazi et al., 2017; Sulc et al., 2018; Haupt et al., 2018). However, as previously discussed in Dourado Filho and Calumby (2021), DA has been mostly used in an ad-hoc way, through empirical heuristics. The concernings with this approach is that it can limit the optimization of model training and classification performance, even more so, considering increased proportion studies involving larger datasets. Another major concern is the suitability of the augmentation techniques with respect to every context of application (i.e., organ type, photo type, background, herbarium sheet, etc). Is comprehensible that exploring search spaces of astronomical magnitudes can still be computationally prohibitive, due to the complexity of the state-of-the-art DCNN models, although, adequate analysis and understanding of data augmentation impacts on tasks in different domains is essential to improve model optimization and performance gains.

Considering that, this work presents a set of analyzes from the perspective of some combination strategies of commonly used data augmentation techniques by relevant works in the literature. In this sense, the results obtained through the proposed experimental analyzes also represents our attempts to validate some of the main empirical strategies popularly employed. In addition to that, we raised some of the aspects identified as possibly decisive from the point of view of the performance improvement provided by DA, which allowed to conduct a more in-depth investigation regarding the underlying operating mechanisms that enables its effectiveness. Not only it allowed to obtain more insightful conclusions with regard to that but it corroborated with the process of validating some intuitions about the selection criteria behind these empirical strategies. In light of that we also attempt to consolidate the DA emergence as a promising alternative not only for introducing data variation and model generalization but also to overcome serious underrepresentation and extreme data imbalance.

Considering the above, this paper thoroughly discuss the investigation of performance improvements behind the utilization of Data Augmentation for training DCNNs to perform plant recognition. Following our preliminary work Dourado Filho and Calumby (2021), we investigate the impact of several DA approaches for plant species recognition while also introduce significant novel contributions, including:

i. An extended analysis that considers the recognition performance for every image sub-category corresponding to each plant organ or view (Leaf, Flower, Fruit, Stem and Entire plant).
ii. A comprehensive analysis of the data augmentation impacts considering multiple levels of representation for different sources of dataset imbalance (i.e., Sampling or Content: sub-type uneven distributions).

2 Related Work

Data Augmentation enables to artificially increase a dataset by obtaining transformed samples from the original ones Buda et al. (2018). This process enables improving DCNNs generalization power, through an invariance hardcoding procedure (Mehdipour Ghazi et al., 2017) that happens behind the process of training with images that presents variations of angle, position, light, brightness, alongside with the original corresponding versions. Consequently, the models are able to become more invariant to these adversities, increasing their capability of performing well for unseen data (Shorten and Khoshgoftaar, 2019; Mehdipour Ghazi et al., 2017).

For the image classification task, DA techniques are regarded as geometric when they cause modifications to the geometric constitution of the images (Fig. 1), or as photometric when changes performed in the color space (Taylor and Nitschke, 2018). Traditional geometric techniques includes: Crop, Flip (horizontally, vertically), Translate, Rotate, whereas photometric transformations mainly involves changes in brightness, light, color or saturation (Shorten and Khoshgoftaar, 2019). In terms of heuristics, these techniques (geometric or photometric) are usually applied individually or combined through

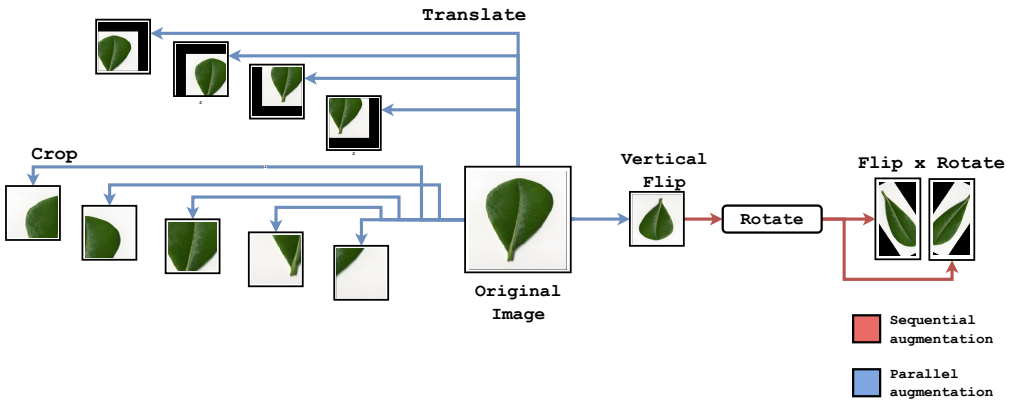

Figure 1: Illustration of geometric augmentation techniques.

sequential or parallel application procedures, referred as augmentation policies (illustrated in Fig. 1). In the *Sequential* modality, for example, samples are obtained from the chained application of multiple augmentation techniques over the same image. Fig. 1 illustrates the sequential heuristic with a Flip x Rotate example (highlighted in red). This process generates two rotated and vertically-flipped images from the application of Flip followed by Rotate. In turn, the *Parallel* approach yields augmentation policies from the independent application of two or more techniques over the image and inclusion of the obtained images to a unique set (illustrated in Fig. 1 in blue). Besides that, some researchers have also proposed to use more complex heuristics, such as through Deep-Learning-based methods (Wang et al., 2017).

In the context of plant recognition, the work of Pawara, Okafor, Schomaker and Wiering (2017) some DA techniques, mainly photometric ones, which included: Rotation, Blur, Contrast, Scaling, Illumination and Projection, combined through a parallel heuristic. Three datasets were considered: Folio (600 leaf images from 32 species), Swedish (1.000 images on a plain background of 15 Swedish tree species) and Agril Plant (3.000 images of 10 fruit species). Although promising results were demonstrated, the limited size and low complexity of the datasets used do not allow the findings to represent the effectiveness for plant recognition in the wild. More specifically, the perfectly balanced classes, high background homogeneity (except for Agril Plant) and the small number of species/specimens made the classification scenarios too simple and hardly realistic.

In Zhang et al. (2015) the authors proposed a CNN architecture for leaf classification and used the Flavia dataset (1907 images of 32 species). The model was trained with traditional DA techniques (Translate, Scale, Rotate, Contrast and Sharpening) randomly selected according to a desired augmentation factor (5x, 10x or 20x). In a similar way, the authors achieved a reasonable effectiveness improvement with DA, although, no significant differences were observed for the techniques assessed. Moreover, the low complexity of the dataset weakens the conclusions regarding the general effectiveness of the augmentation approaches. Similarly, in Pandian et al. (2019) the authors assessed many traditional DA techniques, including Flip, Rotate, Crop, color transformation, PCA and noise injection, as well as DL-based techniques (WGAN, DCGAN, neural style transfer). An imbalanced plant dataset was used with around 54 thousand specimens from 38 classes including healthy and diseased leaves. At the expense of higher complexity, processing and optimization time, the parallel application of DL-based approaches only slightly outperformed the parallel application of traditional ones. Additionally, the authors demonstrated that using all augmentation techniques at once lead to higher performance in contrast to the use in isolation.

As reported in Mehdipour Ghazi et al. (2017), the combination by sequential and parallel application of vertical flip, rotation, and scaling techniques through an image patch extraction pipeline, allowed to achieve state-of-the-art performance in a multi-organ, large-scale plant classification task involving about 1,000 species. The authors observed a decrease of overfitting and the improvement of the benefits of fine-tuning. They also showed that an 80-fold augmentation outperformed a 10-fold by roughly 6% in accuracy. This illustrates how large the augmentation factor usually is to allow reasonable improvements as well as how some augmentation heuristics can generate large datasets that increase training costs.

In general, large-scale DA studies demonstrated its important role towards developing real-world plant recognition systems. Nevertheless, most studies were not successful on determining general techniques or combining heuristics, mainly due to the low complexity of the data or the limited amount of DA techniques assessed. Hence, this work aims at assessing several DA techniques through different combination heuristics and a representative scenario of large-scale plant recognition in terms of class imbalance, visual image heterogeneity (multiple plant organs), number of species and specimens.

3 Experimental setup

The characteristics of the classification task and data used are decisive for robust DA studies involving DCNN (Pawara,

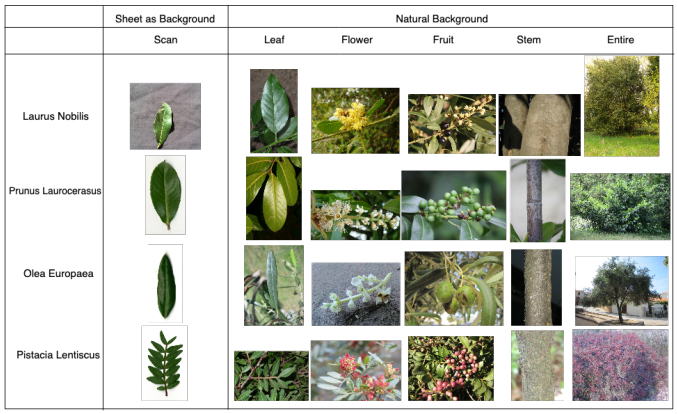

Figure 2: Samples of images with multiple organs of different species divided by the categories "Sheet as Background" and "Natural Background".

Okafor, Schomaker and Wiering, 2017). While small or too simple datasets weakens in-depth analysis of real-world challenges, mainly due to the limited amount of samples and/or classes and its irregular distribution large-scale or more complex datasets closer to real-world scenarios makes comprehensive systematic analysis infeasible, specially considering the multiple possible variations and combinations of DA techniques.

In order to enable large experimentation, considering the depicted challanges, the analysis in this work was conducted over the PlantCLEF2013 (PC2013) dataset (Goëau et al., 2013), a dataset considered as mid-range, in terms of scale and complexity. It presents 26,077 images of 250 plant species from the French flora, from which 5092 (20%) images were separated for testing purposes. Introducing additional complexity, the dataset includes images of multiple plant organs or views considering different perspectives, resulting in two image types and 6 sub-types as illustrated in Fig. 2. The two main types are: *Sheet as Background* (a homogeneous surface of uniform white background) representing 42% of the total images (11,031 samples) and *Natural Background*, with 15,046 samples representing 58% of total images, composed of natural photographs captured outdoors from different perspectives of different organs. An additional challenge imposed by this type of data regards the high intra-class variations and inter-class similarity, which significantly increase the difficulty of class generalization and discriminative feature learning.

For the task of plant species recognition, transfer learning based on the ResNet He et al. (2016) DCNN architecture was performed considering pre-trained weights from the Imagenet Russakovsky et al. (2015) dataset. The ResNet is a highly recognized effective network architecture that achieved the first place at the ILSVRC 2015 classification task with 3.57% error on the ImageNet test set (Russakovsky et al., 2015). Considering that our experiments were conducted upon the performance comparison between the network trained without augmentation (**baseline**) versus the model performance upon training with each proposed augmentation policy based on a corresponding combining heuristic (Individual, Sequential or Parallel, as illustrated in Fig. 1). Similarly Dourado Filho and Calumby (2021) in this work the ResNet50 model was considered for evaluation whereas in the effectiveness assessment step considering the independent test set, the model performance for each sample sub-type was taken into consideration, in order to account for a more factored analysis as proposed.

This way, for test set images, the same data augmentation policies as in the training phase were applied. Considering the recognition is performed for 1+N images (1 original + N augmented versions), the class prediction is performed according to Softmax of the average class scores. All the DA policies assessed, the resulting amount images, and augmentation factors are presented in Table 1. For the DA techniques, the following configurations were considered: Translate (4 different directions with offsets equivalent to 20% of the image width over the horizontal axis and 20% of the height over the vertical axis with black pixel padding); Rotate (30 degrees clockwise and counterclockwise); Crop (four corner patches and a central crop with 50% of the image size).

Table 1: DA policies and number of images. Augmentation factor: size in relation to the original dataset (Bold).

Dataset	Train+Validation	Test Set
Original Dataset	20.985	5092
Flip	41.970	10.184 (**2x**)
Rotate	62.955	15.276 (**3x**)
Flip x Rotate	62.955	15.276 (**3x**)
Flip + Rotate	83.940	20.368 (**4x**)
Translate	104.925	25.460 (**5x**)
Crop	125.910	30.552 (**6x**)
Flip x Crop	125.910	30.552 (**6x**)
Flip + Translate	125.910	30.552 (**6x**)
Flip + Crop	146.895	35.644 (**7x**)
Translate + Rotate	146.895	35.644 (**7x**)
Crop + Rotate	167.880	40.736 (**8x**)
Translate x Rotate	188.865	45.828 (**9x**)
Translate + Crop	209.850	50.920 (**10x**)
Translate x Crop	440.685	106.932 (**20x**)

The evaluation was conducted with a stratified random sampling protocol to keep class proportion in training and validation sets. More specifically, we considered 80% of the training data for model construction and the 20% for validation, followed by the testing with the independent held-out test set. To ensure comparability, instead of performing individualized optimizations, all models were trained with the same configurations, for a fixed number of epochs (75) and amount of trainable weights. The Categorical Crossentropy loss function and the Accuracy measure were considered in the training phase, whereas the Micro-F1 measure was computed to account for class imbalance in the test phase. For weight update and optimization the Adam optimizer (Kingma and Ba, 2017) with a batch size of 64 was proposed. First and Second Moment exponential decays and Epsilon were set to default (0.9, 0.99 and 10^{-7}, respectively), and the

learning rate was set to 2×10^{-6}.

4 Results

The classification models were trained using the Individual, Parallel and Sequential heuristics. The overall performance presented in Table 2 demonstrates that the models trained over the augmented datasets yielded expressively superior results in comparison to the baseline, with relative increases from 0.37% up to roughly 55% in Micro-F1. In terms of heuristics, the Parallel policies generally outperformed the Sequential and Individual ones. The overall worst performing Parallel policy (Flip + Rotate, Micro-F1=0.4255) allowed results similar to the best performing Sequential policy (Flip x Crop, Micro-F1=0.4271).

The models trained with data augmented by the Sequential heuristics presented consistently inferior performance to the ones trained with the best Individual policies. Such inferior results may be a consequence of the Sequential application of multiple geometric policies which increases the chances of violation of the label-preservation principle of data augmentation.

For instance, techniques that are inherently prone to promote more intense distortions, when combined through Sequential heuristics may end up producing policies with higher chances of enhancing irrelevant image regions (e.g., background) or specific regions that does not necessarily include relevant features. The Translate x Crop policy results in poor effectiveness even though Translate and Crop techniques composes the best performing Individual policies. The Individual geometric procedures performed by the Translate and Crop techniques, when applied sequentially (Translate x Crop) are more likely to deteriorate the visual information correspondent to the label associated with the generated image. In contrast to that, the same Sequential association but with Flip and Crop techniques, that conduct less intense visual distortions, resulted in the best performing policy (Flip x Crop) for this type of heuristic.

Considering the number of resulting images (Table 1), our findings demonstrate that the DA policies that yielded better results were not necessarily the ones of higher augmentation factors. For instance, the *Translate x Crop* policy increased the original dataset to over 440.000 images (20x factor) and yet allowed only 0.37% gain in relation to the baseline. In contrast, the single application of Flip (2x factor) allowed an improvement of roughly 28% while the the best performing policy (Translate + Crop) with a 10x factor resulted in a 55% improvement.

These results enabled inferring the effects of data augmentation combining heuristics over the performance of the models trained over the derived policies. Despite the augmentation factor has also being considered as a relevant aspect of investigation, it was able to verify that deeper analyzes should be necessary to enable more assertive conclusions with respect to data augmentation effectiveness. In light of this a more factored performance analysis of the models in terms of individual organ (image sub-type) was conducted and is further presented in Section 4.1. Moreover the different amounts of image sub-types is not only a characteristic of the whole data set, but is also present in the set of images of each plant species. This motivated the investigation of how decisive DA can be in overcoming different sources of data imbalance as investigated and discussed in Section 4.2.

Table 2: ResNet50 results: Combination heuristics and Augmentation Policies (Micro-F1 on test set).

Baseline	
Augmentation	**Micro-F1**
None (Original data)	0.3177
Individual Application	
Crop	**0.4636**
Translate	0.4550
Flip	0.4094
Rotate	0.4049
Sequential Application	
Flip x Crop	**0.4271**
Translate x Rotate	0.4010
Flip x Rotate	0.3800
Translate x Crop	0.3189
Parallel Application	
Translate + Crop	**0.4919**
Flip + Crop	0.4672
Crop + Rotate	0.4573
Translate + Rotate	0.4522
Flip + Translate	0.4518
Flip + Rotate	0.4255

Table 3: ResNet50 results: Heuristics, Augmentation Policies and performance for the image sub-types (Accuracy on test set).

Baseline							
Policy	Scanned Leaf	Flower	Fruit	Stem	Leaf	Entire	Overall
Original Dataset	0.5608	0.2368	0.2442	0.3206	0.2924	0.1051	0.3177
Individual Application							
Crop	0.5624	**0.5182**	**0.3980**	**0.4049**	**0.4037**	**0.2809**	**0.4636**
Translate	**0.6368**	0.4655	0.3557	0.3702	0.3746	**0.2809**	0.4550
Flip	0.5960	0.4201	0.3038	0.3487	0.3759	0.1902	0.4094
Rotate	0.5488	0.4306	0.3250	0.3652	0.3582	0.2132	0.4049
Sequential Application							
Flip x Crop	**0.5320**	**0.4906**	**0.3538**	**0.4016**	**0.3658**	0.2449	**0.4271**
Translate x Rotate	0.5072	0.4282	0.3153	0.3851	0.3341	**0.2521**	0.4010
Flip x Rotate	0.5192	0.3998	0.3019	0.3438	0.3506	0.1916	0.3800
Translate x Crop	0.3600	0.4168	0.3076	0.2363	0.2341	0.2247	0.3189
Parallel							
Translate + Crop	**0.6280**	**0.5498**	**0.4500**	0.4016	0.4101	**0.3472**	**0.4919**
Flip + Crop	0.5832	0.5296	0.4288	**0.4264**	**0.4151**	0.2708	0.4672
Crop + Rotate	0.5440	0.5482	0.4442	0.3603	0.4012	0.2982	0.4573
Translate + Rotate	0.6160	0.4882	0.3750	0.3570	0.3949	0.2997	0.4522
Flip + Translate	0.6200	0.4801	0.3692	0.3834	0.3784	0.3040	0.4518
Flip + Rotate	0.5720	0.4533	0.3403	0.3884	0.3822	0.2579	0.4255

4.1 Single-Organ Analysis

As described in Section 3, the dataset is composed by homogeneous background (42%) and natural background (58%) images. These can be divided in 6 image sub-types (see Fig. 2) according to the corresponding plant organs or views: Scanned Leaves (42%), Leaf (16%), Flower (18%), Fruit (8%), Stem (8%) and Entire plant(8%). In this context, classification performance can be affected by many aspects including the uneven training data distribution, which could deteriorate the model performance for some image sub-types. Moreover, for different organs, the suitability of the selected augmentation policy and the specific classification rates must be considered. Therefore, given that the training

data imbalance and the inherent classification challenges for different organs play a key role on the resulting performance, the impact of each augmentation policy was evaluated for each image sub-type.

Table 3 presents the results for all combining heuristics and each policy performance with respect to every image sub-type. In terms of top performing heuristic, similar result was achieved for all image sub-types. In general, the Parallel heuristic allowed the best results, while the Sequential heuristic was not able to outperform the best Individual policies. The Parallel heuristic yielded the top performing policies for all image sub-types, except for the Scanned Leaf which was best classified with the Individual application of the Translate policy.

Considering the top-performing policies, the best results for each image sub-type was allowed by different policies. While Translate + Crop allowed the best results for Flower, Fruit and Entire, the Flip + Crop was the best policy for Stem and Leaf. Nevertheless, regarding all image sub-types, the parallel Translate + Crop was the overall best policy (Micro-F1=0.4919), given it was significantly superior for some sub-types while also quite similar to the top performing policies for the others.

More specifically, the results demonstrate that, for Scanned Leaf images, the most frequent sub-type (42% of the dataset), even the baseline (no DA) was capable of presenting reasonable performance with Micro-F1 over 0.50, while the performance of the less frequent sub-types was below 0.33. Nevertheless, Scanned Leaf images with homogeneous background may be the less challenging image sub-type in comparison with the other sub-types with natural background. On the other hand, the result for equally frequent sub-types, such as Stem and Entire, were quite different, with the entire sub-type representing a harder classification task. Such differences are consistently noticed regardless of the DA policies and heuristics.

Fig. 3 presents for each image sub-type the relative gains of the best augmentation policies of each heuristic in relation to the best Individual policy for the corresponding sub-type (baseline).

No composed augmentation policy was able to outperform the best Individual policy for Scanned Leaf sub-type. In this case, considering the baseline (Translate), the data generated by the additional application of the Crop operation (Translate + Crop) demonstrated to slightly inferiorize effectiveness. That result was possibly influenced by irrelevant information that the corner crops have may have introduced in the model by mostly covering white background regions and reducing the actual leaf area depicted (as illustrated in Fig. 1). Similarly, Sequential policies achieved almost none or even negative gains in relation to the baseline. Finally, the Parallel policies achieved significant gains from roughly 3% up to roughly 24% (for the entire sub-type). These results (Fig. 3) indicate the potential of generalization of the DA techniques with respect to the mixed organ classification and also in the analysis for specific image sub-types. More specifically, the best results were achieved through parallel combination involving Crop, Translate and Flip operations.

Despite promising, the smaller amount of samples of some image sub-types may have prevented the learning of better representations. For instance, the although Translate + Crop policy only achieved Micro-F1=0.34, enabled leveraging the performance for the entire sub-type by **230%** in contrast to the baseline. The expressive gains for the less represented sub-types suggest the potential of Data Augmentation techniques in attenuating the problems for underrepresented types of organs.

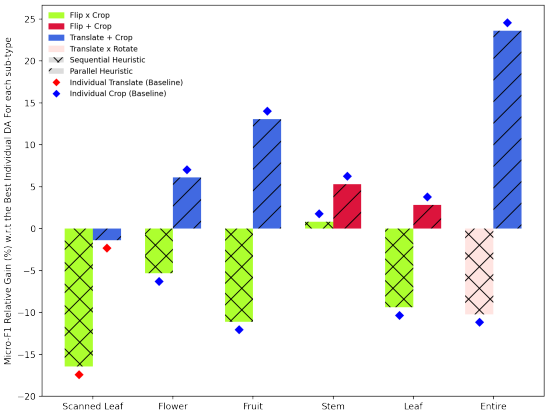

Figure 3: Micro-F1 gain (%) w.r.t best Individual DA policies for each Heuristic and correspondent sub-types.

4.2 Data Imbalance

A statistical analysis of the data revealed two important types of imbalance. Specifically, the dataset presents different amounts of images for each species and also an irregular intra-class distribution for a given species regarding the image sub-types. The number of samples of each plant species (classes) ranges in a broad interval from 11 to 260 images. This source of imbalance was previously referred as **Sampling imbalance** in Seeland et al. (2019), and widely known for causing low accuracy on under-represented classes. Another source of imbalance worth of investigation regards how well represented a class is in terms of the possible image sub-types and how balanced the sub-types are. This **Content imbalance** (Seeland et al., 2019), is also an obstacle that can result in classification biases for species with underrepresented sub-types. The DA impact in relation to these problems are investigated in Sections 4.2.1 and 4.2.2.

4.2.1 *Inter-class - Sampling imbalance Analysis*

The class size distribution is presented in Fig. 4 (outliers removed). While the number of samples per class ranges from 11 to 260, the median class size is 45. The distribution also shows that 25% of the classes presents more than 120 samples. In contrast, 50% of the classes has less than 45 samples and a considerable 25% of the classes have less than 25 samples. These amount of samples are quite below the estimated 100-500 images per species necessary to

learn accurate visual representations for precise species recognition (Carranza-Rojas et al., 2017; Seeland et al., 2019). Precisely, X classes (70% of the dataset) have less than 100 samples, which imposes a significant challenge for the learning of effective and generalized classification models.

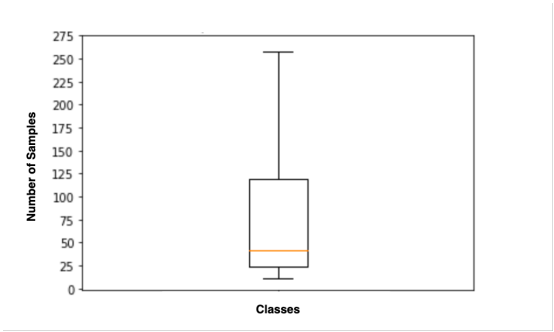

Figure 4: Distribution from the number of samples by class for the original dataset.

Fig. 5 presents for each class, the contrastive results between the DA and the original dataset (baseline). To better visualize and understand the performance variation in light of the class size, the classes were grouped according to the size distribution quartiles as presented in Fig. 4. For simplicity, the **small-size classes** (1st and 2nd quartiles) were merged into a single category ranging from **11 to 45 samples**, which represented a concise interval. In turn, the remaining quartiles were used to define **medium and large-size classes**, ranging from **46-115 and 116-270** samples, respectively. The dashed diagonal lines represent equal performance between the methods.

Fig. 5-a depicts the distribution of the comparative results of the overall best DA in relation to the baseline. Following the overall results presented in Table 2, this detailed analysis shows that the overall best DA (parallel) allowed superior effectiveness for the vast majority of the classes. Additionally, the distributions in Fig. 5-a shows that the most significant results (far from the diagonal) were achieved for the smaller classes (in green).

Similarly, Fig. 5-b presents the comparative results of the worst performing DA method against the baseline. Once again the greater improvements were achieved for the smaller classes. Moreover, while such DA method was generally equivalent to the baseline in terms of Micro-F1, the results shows that for some classes the DA deteriorated the performance in comparison to the baseline, more noticeably for the larger classes (in red).

Similarly to Fig. 5-a and b, Fig. 5-c and d presents a comparative of the baseline in relation to the other heuristics (Individual and Sequential). Once again, the best results were achieved for the smaller classes also with a lesser deterioration of the performance for the bigger classes.

Finally, Fig. 5 highlighted the expressive improvements allowed by the DA, specially for the classes with the worst original baseline results or small sample sets. These findings demonstrate the DA was successful in attenuating the problems related to data imbalance and lack of representativeness of some species.

4.2.2 *Intra-class - Content imbalance*

Considering the intra-class content imbalance, i.e., uneven intra-class image sub-type distributions, we investigated the impact of the utilization of Data Augmentation in light of the imbalance degrees. To represent the imbalance degree of the internal sub-type distribution of the classes (species), the sub-type Entropy within the classes were computed for the training dataset according to Eq. (1). The Entropy (H) with respect to a random variable (X) can be defined in terms of the average level of surprise inherent to its possible n outcomes, given $x_1, x_2, \ldots, x_n$, possible outcomes of X which may occur with probability $P(x_1), P(x_2), \ldots, P(x_n)$.

$$H(X) = -\sum_{i=1}^{n} P(x_i)\log(P(x_i)) \quad (1)$$

With Eq. (1), the Entropy level for a given species (S) is computed by averaging the product between the outcome probability and its correspondent surprise, considering the six possible sub-types within S. Therefore, the Entropy is used to quantify the level of internal imbalance of each species in the dataset, which enabled to investigate the impact of data augmentation over this source of imbalance[1].

The resulting entropy values were grouped according to three sub-intervals to represent the different entropy levels, specifically: **Small Entropy Classes** ([0.4,0.6]), **Medium Entropy Classes** (]0.6,0.8]) and **High Entropy Classes** (]0.8,1]). It allowed to analyze the relationship between the imbalance degree and the corresponding performance improvements provided by the utilization of data augmentation. For that we related the entropy degree (small/medium/high) and the Delta-F1, which corresponds to the test-set Micro-F1 difference, between the model trained with best DA policy (**Crop + Translate**) and the model trained with the **Original dataset** (no augmentation) policy.

Imbalance	Average Entropy	Average Delta-F1
High	0.5390 ± 0.0531	0.2528 ± 0.2485
Medium	0.7088 ± 0.0520	0.2958 ± 0.2279
Low	0.8663 ± 0.0496	0.1821 ± 0.1549

Table 4: Micro (Content) Imbalance Results

Table 4 presents the average improvements provided by Data Augmentation with respect to the defined levels of imbalance. The low entropy deviation means that each group includes the species with similar entropy (imbalance). Despite the high Delta-F1 variability (high

[1] In order to deal with the possible absence of samples for a given sub-type, Laplace Smoothing (with default K=1) was used to avoid $P(x_i) = 0$. Min-max normalization was used to scale the Entropy between 0 and 1.

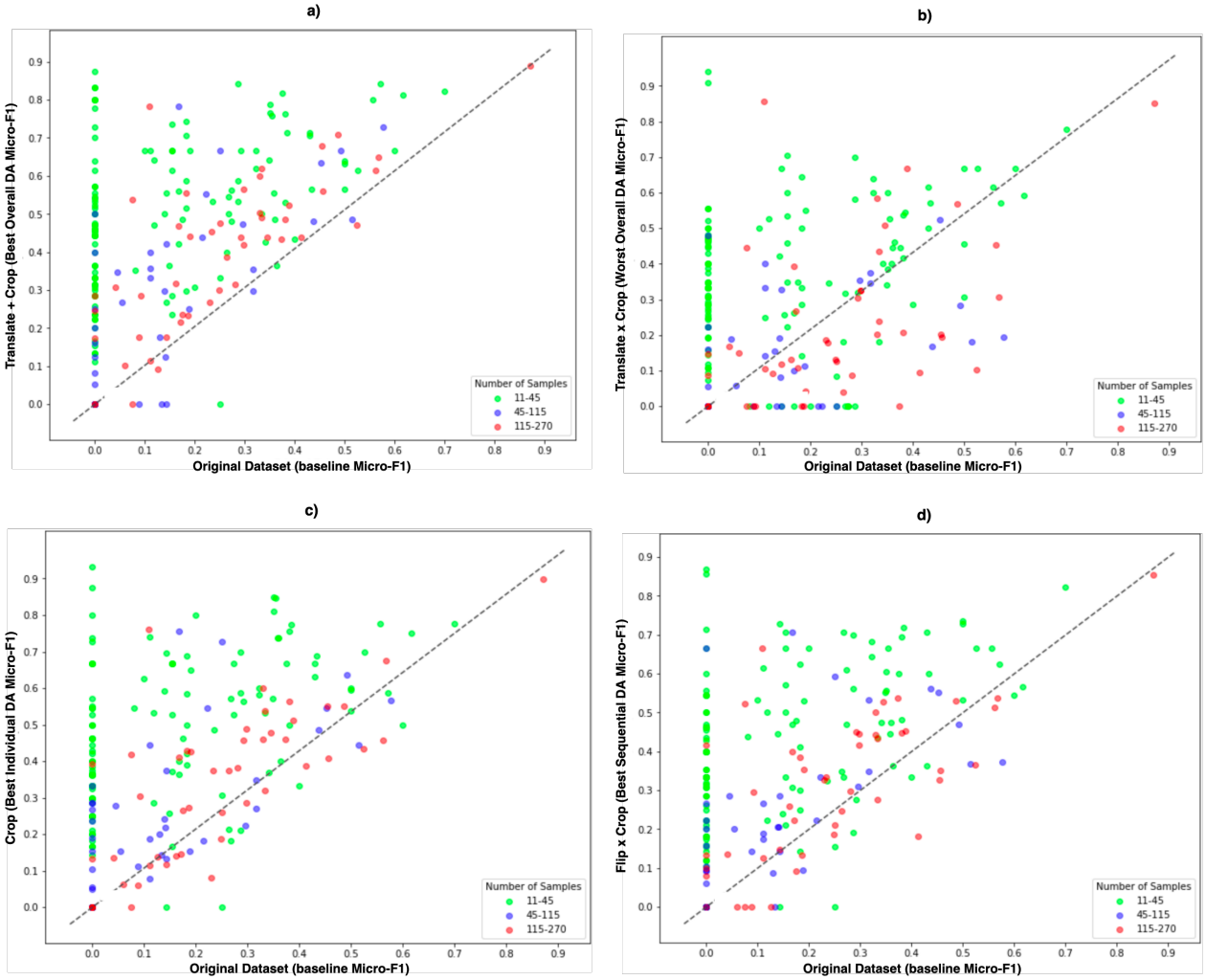

Figure 5: a & b - Micro-F1 (**DA**) x Micro-F1 (**Baseline**) performance for the Crop + Translate and Crop x Translate policies. **c & d** - Micro-F1 (**DA**) x Micro-F1 (**Baseline**) for the Crop and Flip x Crop policies.

standard deviation), the average Delta-F1 was positive.

Besides the Delta-F1 for the species with low imbalance was positive, the performance improvement was even higher for the species with medium and high imbalance. It suggests, the DA allowed performance improvements for all levels of imbalance, but more significantly for the classes with higher internal distribution irregularity, attenuating poor data representativeness even for extremely imbalanced situations.

As demonstrated in Table 4, the model trained with the best DA allowed an average performance improvement of roughly 0.25 per class for the more irregular classes (Low Entropy). These results represented a substantial **39% increase** in relation to the High Entropy classes that obtained average **0.18 Micro-F1** improvement per class.

Considering that the classes that most benefited from the DA were the most irregular ones (Low and Medium Entropy), which represents together roughly **75% of the dataset**, these findings are aligned with the hypothesis that data augmentation acts more significantly in poorly distributed classes, enabling leveraging performance more considerably in comparison with better distributed ones.

5 Conclusion

In this work, the performance improvement provided by the utilization of Data Augmentation for training DCNN models to perform plant species recognition from images was thoroughly investigated in terms of several aspects including image sub-type and dataset imbalance. This results demonstrated that DA acted significantly by leveraging classification performance hence enabling models to learn more accurate visual representations, mostly over underrepresented classes. Besides that, these findings emphasized how promising Data Augmentation could be for attenuating class imbalance and its potential effectiveness for application in small-size or underrepresented classes, in case of computational resources limitations. Otherwise integral or even contextual (dynamic) utilization of techniques

suitable for specific image sub-types should also provide better performance improvement.

We believe that the findings and analyzes presented can represent some aspects that designers of DCNN-based plant recognition systems should consider for developing more rigorous applications when facing imbalanced datasets. Furthermore, the dataset imbalance analyzes showed that despite promising, Data Augmentation demonstrates potential for enhancement in terms of intra-class content imbalance. In this sense we believe that random heuristics instead of deterministic approaches may lead to more substantial performance improvements, therefore, future work addressing this topic may be prospective to provide more insightful analysis with this respect.

Acknowledgments

This work was partially supported by the National Council for Scientific and Technological Development (CNPq grant no. 124989/2021-7) and a Quadro® P6000 GPU donation by NVIDIA™ Corporation.

References

Beech, E., Rivers, M., Oldfield, S. and Smith, P. (2017). Globaltreesearch: The first complete global database of tree species and country distributions, *Journal of Sustainable Forestry* **36**(5): 454–489. Available at https://www.tandfonline.com/doi/full/10.1080/10549811.2017.1310049.

Bonnet, P., Goëau, H., Hang, S. T., Lasseck, M., Šulc, M., Malécot, V., Jauzein, P., Melet, J.-C., You, C. and Joly, A. (2018). Plant identification: experts vs. machines in the era of deep learning, *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, Springer, pp. 131–149. Available at https://hal.archives-ouvertes.fr/hal-01913277/document.

Buda, M., Maki, A. and Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks, *Neural Networks* **106**: 249 – 259. Available at https://doi.org/10.1016/j.neunet.2018.07.011.

Carranza-Rojas, J., A.J. Joly, A., Bonnet, P., H.G. Goëau, H. and Mata-Montero, E. (2017). Automated herbarium specimen identification using deep learning, *Biodiversity Information Science and Standards* **1**: e20302. https://doi.org/10.3897/tdwgproceedings.1.20302.

Chatfield, K., Simonyan, K., Vedaldi, A. and Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets, *arXiv preprint arXiv:1405.3531* . Available at https://arxiv.org/pdf/1405.3531.pdf.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V. and Le, Q. V. (2019). Autoaugment: Learning augmentation strategies from data, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 113–123. Available at https://openaccess.thecvf.com/content_CVPR_2019/papers/Cubuk_AutoAugment_Learning_Augmentation_Strategies_From_Data_CVPR_2019_paper.pdf.

Dourado Filho, L. A. and Calumby, R. T. (2021). Experimental evaluation of data augmentation heuristics for plant identification systems based on deep learning, *Anais do XIII Congresso Brasileiro de Agroinformática*, SBC, pp. 136–143. Available at https://sol.sbc.org.br/index.php/sbiagro/article/download/18384/18217.

Goëau, H., Bonnet, P. and Joly, A. (2018). Overview of expertlifeclef 2018: how far automated identification systems are from the best experts? Available at https://hal.archives-ouvertes.fr/hal-01913244/file/expert.pdf.

Goëau, H., Bonnet, P. and Joly, A. (2019). Overview of LifeCLEF Plant Identification Task 2019: diving into Data Deficient Tropical Countries, *CLEF 2019 - Conference and Labs of the Evaluation Forum*, Vol. 2380 of *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, Linda Cappellato and Nicola Ferro and David E. Losada and Henning Müller, CEUR, Lugano, Switzerland, pp. 1–13. Available at https://hal.umontpellier.fr/hal-02283184/file/Goeau_etal_CLEF_Lugano_2019_paper_247.pdf.

Goëau, H., Joly, A., Bonnet, P., Bakic, V., Barthélémy, D., Boujemaa, N. and Molino, J.-F. (2013). The imageCLEF Plant Identification Task 2013, *Proceedings of the 2nd ACM International Workshop on Multimedia Analysis for Ecological Data* (i): 23–28. dx.doi.org/10.1145/2509896.2509902.

Graves, S. J., Asner, G. P., Martin, R. E., Anderson, C. B., Colgan, M. S., Kalantari, L. and Bohlman, S. A. (2016). Tree species abundance predictions in a tropical agricultural landscape with a supervised classification model and imbalanced data, *Remote Sensing* **8**(2): 161. Available at https://mdpi-res.com/d_attachment/remotesensing/remotesensing-08-00161/article_deploy/remotesensing-08-00161-v2.pdf?version=1456799441.

Haupt, J., Kahl, S., Kowerko, D. and Eibl, M. (2018). Large-scale plant classification using deep convolutional neural networks., *CLEF (Working Notes)*. Available at http://ceur-ws.org/Vol-2125/paper_92.pdf.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778. Available at https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf.

Kingma, D. P. and Ba, J. (2017). Adam: A method for stochastic optimization. Available at https://arxiv.org/pdf/1412.6980.pdf.

Kumar, N., Belhumeur, P. N., Biswas, A., Jacobs, D. W., Kress, W. J., Lopez, I. and Soares, J. V. B. (2012). Leafsnap: A computer vision system for automatic plant species identification, *The 12th European Conference on Computer Vision (ECCV)*. Available at https://link.springer.com/content/pdf/10.1007/978-3-642-33709-3_36.pdf.

Lee, S. H., Chan, C. S., Mayo, S. J. and Remagnino, P. (2017). How deep learning extracts and learns leaf features for plant classification, *Pattern Recognition* **71**: 1–13. Available at https://eprints.kingston.ac.uk/id/eprint/38201/1/Remagnino-P-38201-AAM.pdf.

Mehdipour Ghazi, M., Yanikoglu, B. and Aptoula, E. (2017). Plant identification using deep neural networks via optimization of transfer learning parameters, *Neurocomputing* . Available at https://www.sciencedirect.com/science/article/abs/pii/S0925231217300498.

Pandian, J. A., Geetharamani, G. and Annette, B. (2019). Data augmentation on plant leaf disease image dataset using image manipulation and deep learning techniques, *2019 IEEE 9th International Conference on Advanced Computing (IACC)*, IEEE, pp. 199–204. Available at https://ieeexplore.ieee.org/document/8971580.

Pawara, P., Okafor, E., Schomaker, L. and Wiering, M. (2017). Data augmentation for plant classification, *International Conference on Advanced Concepts for Intelligent Vision Systems*, Springer, pp. 615–626. Available at https://link.springer.com/chapter/10.1007/978-3-319-70353-4_52.

Pawara, P., Okafor, E., Surinta, O., Schomaker, L. and Wiering, M. (2017). Comparing local descriptors and bags of visual words to deep convolutional neural networks for plant recognition., *ICPRAM* **479**: 486. Available at https://www.scitepress.org/Papers/2017/61962/61962.pdf.

Picek, L., Šulc, M. and Matas, J. (2019). Recognition of the amazonian flora by inception networks with test-time class prior estimation, *CLEF (Working Notes)* . Available at http://ceur-ws.org/Vol-2380/paper_108.pdf.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision (IJCV)* **115**(3): 211–252. dx.doi.org/10.1007/s11263-015-0816-y.

Seeland, M., Rzanny, M., Boho, D., Wäldchen, J. and Mäder, P. (2019). Image-based classification of plant genus and family for trained and untrained plant species, *BMC bioinformatics* **20**(1): 1–13. Available at https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2474-x.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning, *Journal of Big Data* **6**(1): 60. Available at https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0.

Soltis, P. S. (2017). Digitization of herbaria enables novel research, *American journal of botany* **104**(9): 1281–1284. Available at https://bsapubs.onlinelibrary.wiley.com/doi/pdfdirect/10.3732/ajb.1700281.

Sulc, M., Picek, L. and Matas, J. (2018). Plant recognition by inception networks with test-time class prior estimation., *CLEF (Working Notes)*. Available at http://ceur-ws.org/Vol-2125/paper_152.pdf.

Taylor, L. and Nitschke, G. (2018). Improving deep learning with generic data augmentation, *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, pp. 1542–1547. Available at https://arxiv.org/pdf/1708.06020.pdf.

Wang, J., Perez, L. et al. (2017). The effectiveness of data augmentation in image classification using deep learning, *Convolutional Neural Networks Vis. Recognit* **11**. Available at https://arxiv.org/pdf/1712.04621.pdf.

Zhang, C., Zhou, P., Li, C. and Liu, L. (2015). A convolutional neural network for leaves recognition using data augmentation, *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, IEEE, pp. 2143–2150. Available at https://ieeexplore.ieee.org/abstract/document/7363364.