Generalizing across the (in)visible spectrum

Ruchit Rawal [*] [1] [2] **Prabhu Pradhan** [*] [3]

Abstract

The effects of data disparity are well documented, and there are many proposed methods for handling such biases, however, most are either standalone, domain-specific or unfeasible. In our work Rawal & Pradhan (2020), we tried to evaluate parallel efforts for handling class-imbalance, focusing on complementary techniques, to present modular approaches that optimize performance for the minority classes. Here, we briefly expand on the apparently 'vague' effects of improving robustness for intrinsic generalization. We also shed light on standard metrics and their relations to the specific type of insights.

1. Introduction

Machine Learning is a tool driving many technologies across diverse sectors. However the fuel that drives this growth is data, and as is with every fuel it's not directly usable. A critical problem is class imbalance, both in supervised and unsupervised form of learning algorithms. A dataset can be treated as imbalanced if there is a noticeable mismatch between the target variable and other values. For example, medical-diagnostics data is conventionally biased towards the negative class (healthy samples are more numerous than the infected ones). Experimentally, high instability in performance has been observed in vanilla models when tested on imbalanced datasets (Buda et al., 2018).

Commonly, deep-net models are built to maximize predictive accuracy (ex. classification) but this metric is uneventful for the cases with limited labels, extreme classification etc. (Lipton & Steinhardt, 2019). This happens because the trained classifier focuses only on the most-numerous class (since it has a higher proportion) while remaining below-par on minority classes. This may prove catastrophic in critical use cases like medical diagnostics and self-driving cars where the rare instances are of utmost importance.

Presently, researchers tend to tackle the imbalance issues (either at input or intermediate pipeline) in its narrow context with domain-specific solutions. We present drawn-out insights on several techniques to mitigate data-imbalance problems. Our work[1] (Rawal & Pradhan, 2020) has following major contributions:

- We used a deep generative model for synthetic data augmentation of multi-spectral images. To the best of our knowledge, this specific area is still unexplored. We also show that certain spectral bands are better for particular tasks (here, vegetation area analysis).
- We show that a combination of Cyclic Learning Rate (CLR) (Smith, 2015) + Stochastic Weight Averaging (SWA) (Izmailov et al., 2018) is suitable for extreme imbalance scenarios.
- We further cemented the compatibility of LDAM: Label-distribution aware loss-function (Cao et al., 2019), which works better than crude re-sampling and can be further improved by using class-balanced loss (Cui et al., 2019).

The results for the same are presented in Table 1.

More specifically, in this paper we explore the following:

- We perform experiments on ResNet-18 architecture to crystallize the utility of the aforementioned strategies from a model-agnostic perspective.
- We analyse the effects of Augmix (Hendrycks et al., 2020) under high-imbalance setups in combination setting with SWA, CLR and LDAM. We observe that reliability isn't complimentary with high robustness.

The outline of this paper is as follows: section 2 explains our training data set (original + synthetic). section 3 introduces our modifications to the baseline neural-net model (including LDAM- loss function, Cyclic Learning Rate- CLR, and Stochastic Weight Averaging- SWA).

[*]Equal contribution [1]Netaji Subhas University of Technology (NSUT), New Delhi, India. [2]Internship (remote) at GCDSL, IISc Bangalore, India. [3]Max Planck Institute for Intelligent Systems, Tübingen, Germany. Correspondence to: RR <ruchitr.ec.17@nsit.net.in>, PP <ppradhan@tue.mpg.de>.

ICML 2020 Workshop on Extreme Classification. Copyright 2020 by the author(s).

[1] Paper link- https://arxiv.org/abs/2004.12344
Our codebase- `https://github.com/JARVVVIS/drought`.

section 4 presents new results with Augmix which is meant for robustness to perturbations. We discuss the results section 5 from a vantage point of generalization for imbalance, and finally conclude with some promising directions. We focus on imbalance-relevant metrics ex. Balanced Accuracy (Brodersen et al., 2010), F1 Score, ICV etc.[II]

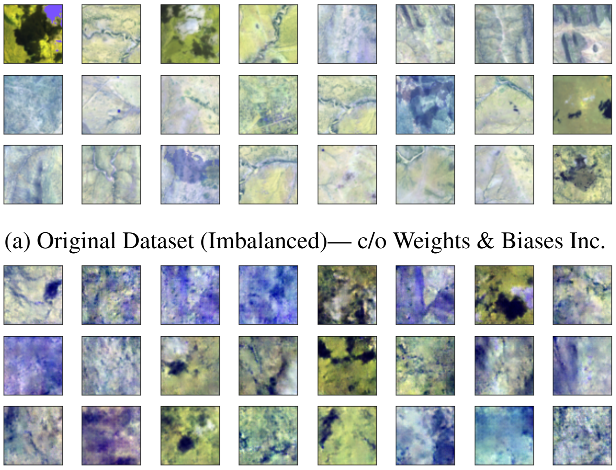
2. Training Data

2.1. Original Data set

The expert-labelled, multi-spectral satellite dataset (Hobbs & Svetlichnaya, 2020) is highly imbalanced (roughly 60% of the data gathered is of class 0, classes 1 and 2 have 15% each, and the remaining 10% is class 3). The model can erroneously achieve 60% accuracy just by predicting 0 every time. However, such high mis-classification is very problematic since these algorithms will be deployed in high-stake real-world settings. We would like to make dense predictions no matter the location of the pixel, since there is high amount of sparsity in the labels. Hence, we need to train a model that is satisfactorily robust to complex samples and generalizes well on all the inherent classes i.e. independent-&-identically-distributed (i.i.d) samples. We focus on striking a pragmatic balance. Also, at the time of writing the official Test Data had still not been released, thus for all intents and purposes, we treat the Validation Data as the Test Data i.e. unseen during the training time.

2.2. Generating Synthetic Images

Standard data augmentation has been used as a go-to technique for enhancing generalizability. Generative adversarial networks offer a novel method for data augmentation (Sandfort et al., 2019), but have still not been adopted by either the earth observation or remote sensing community. We use DC-GAN (Radford et al., 2016), which employs deep convolutional neural networks for both the Generator (G) and Discriminator (D), to generate synthetic images for the low represented classes as a form of data-augmentation to equalize the number of samples of each class. Following best practices, we only operate on a subset of bands (6-5-2 aka Agriculture)[III], also, it is easier to visually gauge the images this way than when all the bands are combined.



(a) Original Dataset (Imbalanced)— c/o Weights & Biases Inc.

(b) GAN generated images (only minority classes)

Figure 1. Sample images — Final dataset = (a) + (b)

3. Improvements via Network Architecture

Ever since winning the 2015 ILSVRC (Russakovsky et al., 2015) challenge ResNet (He et al., 2016) has inspired a family of deep convolutional neural networks. The skip connections in ResNet allow one to build deep networks (up to 1000 layers) while still keeping them optimizable,

Model	Training Details	Test Acc	Balanced Acc	Intra-Class Variance
ResNet-50	Learning Rate : 1e-3 Loss : Cross-Entropy	0.7465	0.6123	0.4510
ResNet-50	Learning Rate : 1e-3 Sampler : $\propto 1/n_j$ Loss : Cross-Entropy	0.7184	0.6467	0.2757
ResNet-50	CLR : Yes Sampler : $\propto 1/n_j$ Loss : LDAM+DRW	0.7022	0.2076	0.6406
Efficient-Net B4	Learning Rate : 1e-3 Loss : Cross-Entropy	0.7630	0.6380	0.4443
Efficient-Net B4	CLR : Yes Sampler : $\propto 1/n_j$ Loss : LDAM+DRW	0.7196	0.6779	0.2185
Efficient-Net B4	SWA + CLR : Yes Sampler : $\propto 1/n_j$ Loss : LDAM+DRW	0.7292	0.6740	0.2417
Efficient-Net B4	Bands: 6,5,2 SWA + CLR : Yes Sampler : $\propto 1/n_j$ Loss : LDAM+DRW	0.7441	0.6955	0.2115
$ Efficient-Net B4	Learning Rate : 1e-3 Sampler : $\propto 1/n_j$ Loss : Cross-Entropy	0.67	0.6803	0.0942
$ Efficient-Net B4	SWA + CLR : Yes Sampler : $\propto 1/n_j$ Loss : LDAM+DRW	0.70	0.6569	0.1967

Table 1. **Core results from Rawal & Pradhan (2020).**
Legends: **$** = GAN-Augmented Dataset

[II] Bal. Acc : The arithmetic mean of the TPR (True Positive Rate) and TNR (True Negative Rate).
ICV (Intra-Class Variance): $\sum_{i=0}^{4}(acc - class_i)^2$ (where acc denotes the validation accuracy and $class_i$ denotes the accuracy of the i^{th} class for a given experiment)

[III] LANDSAT Agriculture- It is a spectral combination of SWIR-1 (6), near-infrared (5) and blue (2). The short-wave and near infrared allows this combination to be used for crop monitoring.

Model	Training Details	Test Acc	Bal. Acc	F-1 Score	Recall				Precision			
					0	1	2	3	0	1	2	3
(1) ResNet-18	Loss : Cross-Entropy	0.7575	0.6470	0.6595	0.8968	0.4883	0.5631	0.6397	0.8376	0.5909	0.6054	0.6685
(2) ResNet-18	Bands: 6,5,2 Loss : Cross-Entropy	**0.7589**	0.6474	0.6600	0.9019	0.4784	0.5529	0.6565	0.8361	0.6091	0.5972	0.6677
(3) ResNet-18 #	Bands: 6,5,2 Loss : Cross-Entropy	0.7561	0.6029	0.6365	0.9517	0.3883	0.4614	0.6103	0.7867	0.6782	0.6701	0.6819
(4) ResNet-18	Bands: 6,5,2 Loss : LDAM+DRW	0.7170	0.6694	0.6389	0.7852	0.5731	0.6007	0.7185	0.8960	0.4987	0.5154	0.5620
(5) ResNet-18 #	Bands: 6,5,2 Loss : LDAM+DRW	0.7005	0.6496	0.6179	0.7700	0.5450	0.6037	0.6796	0.8932	0.5065	0.4709	0.5222
(6) ResNet-18	Bands: 6,5,2 SWA + CLR : Yes Loss : LDAM+DRW	0.7564	**0.6744**	0.6680	0.8665	0.5275	0.5906	0.7132	0.8734	0.5677	0.5672	0.6436
(7) ResNet-18 #	Bands: 6,5,2 SWA + CLR : Yes Loss : LDAM+DRW	0.7579	0.6721	**0.6685**	0.8702	0.5199	0.6061	0.6922	0.8673	0.5691	0.5906	0.6373

Table 2. **Comparison on Metrics**. Legends: # = AugMix, Bal. Acc = Balanced Accuracy

He et. al (He et al., 2016) demonstrated that even for fixed baseline architecture increase in depth almost always leads to increased accuracy.

Tan & Le (2019) argued that to we need to harmoniously scale a model across all the dimensions (scale,depth,width) rather than focusing on one dimension (for e.g. width), and proposed a compound scaling method for the same. In the same work they also proposed a new baseline "Efficient-Net" by leveraging Neural Architecture Search (Zoph & Le, 2017) to optimize for both accuracy and FLOPS.

We briefly discuss the modular improvements below:

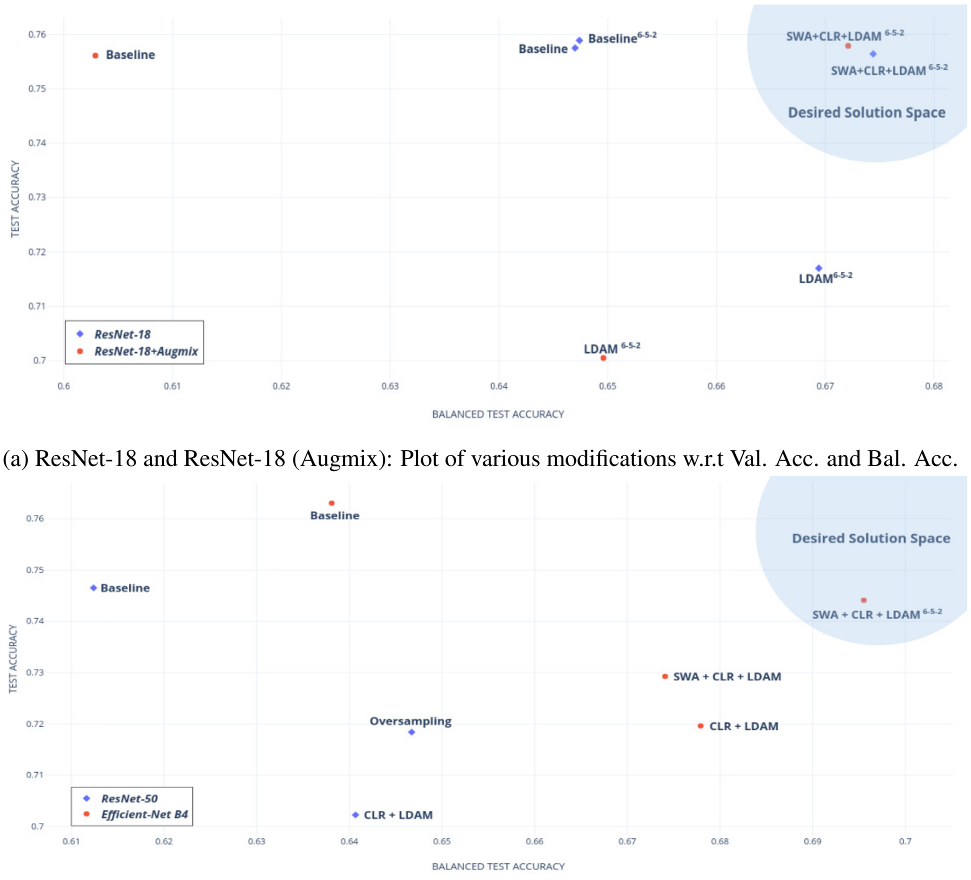
- **Label Distribution Aware loss function** (LDAM): It regularizes the minority class more aggresively, facilitating improved generalization on minority class without affecting the majority class performance.
- **Cyclical Learning Rates** (CLR) : A learning rate routine which oscillates between a range of values. The occasional high learning rates during the cycle helps in getting out of sharp-minima quickly as well as expedite traversal when stuck on saddle-points.
- **Stochastic Weight Averaging** (SWA) : (Garipov et al., 2018) showed that using CLR with stochastic gradient descent traversed on the periphery of the optimal weights but never quite reached it's center. SWA assists in reaching these optimal weights by averaging in weight domain at different snapshots of time.

In Rawal & Pradhan (2020), we evaluated ResNet-50 and Efficient-Net B4 under various combination settings in order to deal with skewed data problems. In this work we present additional set of experiments on ResNet-18 architecture. We observe SWA + CLR + LDAM yields the best solution (Figure 2) for two highly distinct architectures i.e. ResNet-18 and Efficient-Net B4, illustrating the viability of these modular improvements from a model-agnostic purview.

4. Augmix

Data augmentation can be understood as an explicit regularization technique wherein we attempt to increase our input space by augmenting the original dataset with domain specific transformations of the input data. Researchers have tried to train robust Deep Neural Networks by training on corrupted images however, recent work by (Geirhos et al., 2018) has substantially shown that Deep Neural Networks tend to memorize specific distortions shown during training and fail to generalize on unseen distortions.

Augmix (Hendrycks et al., 2020) is a highly modular data augmentation technique which facilitates training of robust and accurate models with high uncertainty estimates. An augmentation chain consists of layered transformations which are stochastically sampled from a predefined pool of augmentations. The final augmented image consists of a weighted sum across various augmentation chains and the original image. The length of chains, mixing weights, severity of transformations are also sampled stochastically leading to diverse augmented images. Augmix also employs a Jason Shannon Divergence consistency loss over the origi-



(a) ResNet-18 and ResNet-18 (Augmix): Plot of various modifications w.r.t Val. Acc. and Bal. Acc.

(b) Efficient-Net B4 and ResNet-50: Plot of various modifications w.r.t Val. Acc. and Bal. Acc.

Figure 2. ResNet-18 and Efficient-Net B4 both show progressive improvement in Balanced Validation Accuracy when equipped with proposed methodologies to deal with imbalanced data. *The top-right section is the desirable solution space*

nal image and it's augmented variants posterior distribution in addition to the standard cross-entropy loss to enforce smoother responses.

Owing to several sources of randomness while generating the augmented images, Augmix facilitates robust training rather than 'rote' memorization. We employ Augmix in a high-imbalance setup to evaluate whether the diverse augmentations lead to better generalization on all the classes.

5. Results

Table 2 shows a comparative overview of various methodologies. We observe that all the 3 vanilla models (Table 2 - (1),(2),(3)) fall prey to overfitting owing to high class-imbalance, (with Augmix being the worst affected) w.r.t Balanced Accuracy & F1 Score.

Employing LDAM+DRW (Table 2 - (4),(5) w.r.t. (2),(3)) leads to better Balanced Accuracy and Recall (for rare classes), however we see a drop in F-1 score owing to poor precision. Here also, Augmix performs slightly worse than the relevant parent model.

We observe that using SWA+CLR+LDAM (Table 2) has the best performance as it leads to better Bal. Acc. as well as F-1 score while maintaining the overall accuracy. This result is highly desirable since the model has high precision as well as high recall for all the inherent classes. Augmix variant marginally outperforms the parent model on F-1 score.

The experiments on Augmix (Table 2 - (3),(5),(7)) indicate that robustness to unseen,complex corruptions doesn't necessarily ensure generalization on all the inherent classes, rather, we observe Augmix tends to deteriorates the model's performance on imbalance focused metrics, while maintaining similar overall accuracy (see Figure 2).

Summary

We observe that improvements in robustness does not necessarily translate to better 'intrinsic' generalization (especially in case of high imbalance or long-tail scenarios).

An additional takeway is that vanilla metrics (ex. Test Accuracy) are superficial and should not be trusted for ascertaining model reliability when dealing with real-world data.

The diverse set of methodologies discussed in this paper facilitate training of Deep-Networks equipped to handle adverse effects of imbalanced data. We concretely demonstrate the effectiveness of such technique-blends across different architectures. Also, fusion of complimentary methods provides overall improvement, and allows high flexibility to optimize for domain-specific applications.

References

Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M. The balanced accuracy and its posterior distribution. *20th International Conference on Pattern Recognition (ICPR)*, pp. 3121–3124, 2010. doi: https://doi.org/10.1109/ICPR.2010.764.

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

Cao, K., Wei, C., Gaidon, A., Aréchiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. J. Class-balanced loss based on effective number of samples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9260–9269, 2019.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. Generalisation in humans and deep neural networks. In *NeurIPS*, 2018.

He, K., Zhang, X., Ren, S., and Sun., J. Deep residual learning for image recognition. In *Proc.. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, USA, June 2016. doi: 10.1109/CVPR.2016.90.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. *ArXiv*, abs/1912.02781, 2020.

Hobbs, A. and Svetlichnaya, S. Satellite-based prediction of forage conditions for livestock in northern kenya, 2020. ICLR 2020 Workshop on Computer Vision for Agriculture (CV4A).

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. P., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

Lipton, Z. C. and Steinhardt, J. Troubling trends in machine learning scholarship. *Queue*, 17(1):80:45–80:77, February 2019. ISSN 1542-7730. doi: 10.1145/3317287.3328534.

Pradhan, P., Sarkar, M., and Ghose, D. Smarter prototyping for neural learning. In *Neural Information Processing Systems (NeurIPS) Workshop*. OpenReview, 2019. URL https://openreview.net/forum?id=H1l7mN6AwH. ML-Retrospectives @ NeurIPS'19.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations, (ICLR)*, 2016.

Rawal, R. and Pradhan, P. Climate adaptation: Reliably predicting from imbalanced satellite data. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. arXiv 2004.12344.

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. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Sandfort, V., Yan, K., Pickhardt, P. J., and Summers, R. M. Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks. In *Scientific Reports*, 2019.

Smith, L. N. Cyclical learning rates for training neural networks. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 464–472, 2015. doi: 10.1109/WACV.2017.58.

Tan, M. and Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning (ICML)*, 2019. URL http://proceedings.mlr.press/v97/tan19a.html.

Zoph, B. and Le, Q. V. Neural architecture search with reinforcement learning. In *International Conference on Learning Representations, (ICLR)*, volume abs/1611.01578, 2017. URL https://openreview.net/forum?id=r1Ue8Hcxg.