Improving Model Accuracy for Imbalanced Image Classification Tasks by Adding a Final Batch Normalization Layer: An Empirical Study

V. Kocaman, O. M. Shir, T. Baeck

Universiteit Leiden

Research Question

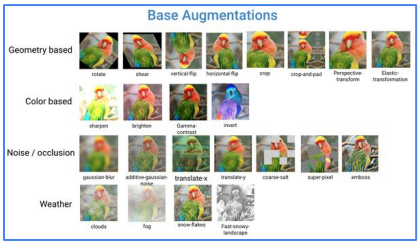
◆ **Learning from small samples in highly imbalanced image classification problems**

- In real world, especially in agriculture and healthcare, the anomalies are rare and it is usually expensive, time consuming or impossible to collect them.
- In this kind of highly imbalanced classification problems, DL frameworks favor majority classes over minority classes (generalisation power of DL).
- Under covariate shift (dataset shift), train and test set come from different distributions and model fails to predict samples which haven't seen during training.
- ***What is the most effective approach to enable learning of minority classes?***

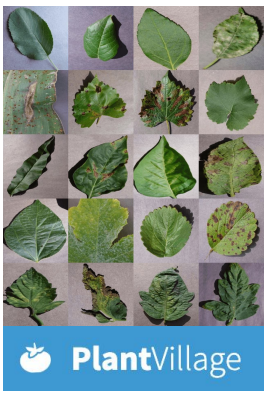
Possible solutions for learning from small samples in DL

- Get more data
- Transfer learning (fine tuning)
- Data augmentation
- Cosine Loss
- Going deeper and ensembling
- Autoencoders
- Prior knowledge (Domain adaptation)
- Class-balanced loss (CBL)
- ***Batch Normalization (BN)?***

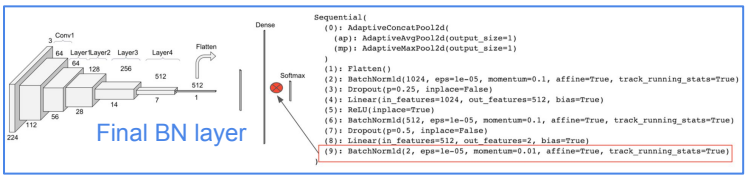
Base Augmentations



Batch Norm (BN) is a widely adopted technique that is designed to combat *internal covariate shift* and to enable faster and more stable training of DNNs. It is an operation added to the model before activation which normalizes the inputs and then applies learnable scale (γ) and shift (β) parameters to preserve model performance.



	Train set		Validation set		class size
	healthy	unhealthy	healthy	unhealthy	
Potato	121	1600	31	400	3
Peach	288	1838	72	459	2
Cherry	684	842	170	210	2
Grape	339	2912	84	727	4
Tomato	1272	13132	318	3284	10
Pepper	1181	797	295	200	2
Corn	16	2005	5	503	4
Orange	0	4405	0	1102	2
Blueberry	1202	0	300	0	2
Apple	1316	1220	329	306	4
Squash	0	1448	0	365	2
Soybean	4072	0	1018	0	2
Raspberry	297	0	74	0	2
Strawberry	364	886	92	222	2



Training set : 1% vs 99% Test set : %50 vs 50%

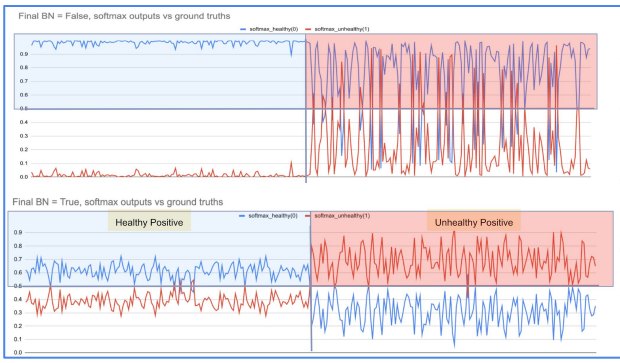
	Train set		Validation set		Test set	
	healthy	unhealthy	healthy	unhealthy	healthy	unhealthy
Appel	1000	10	150	7	150	150
Pepper	1000	10	150	7	150	150
Tomato	1000	10	150	7	150	150

TABLE I: Averaged **F1 test set** performance values over 10 runs, alongside BN's total improvement, using 10 epochs with VGG19, with/without BN and with Weighted Loss (WL) without BN.

plant	class	without final BN	with WL (no BN)	with final BN (no WL)	BN total improvement
Apple	Unhealthy	**0.2942**	**0.7947**	**0.9562**	**0.1615**
	Healthy	0.7075	0.8596	0.9577	0.0981
Pepper	Unhealthy	**0.7237**	**0.8939**	**0.9575**	**0.0636**
	Healthy	0.8229	0.9121	0.9558	0.0437
Tomato	Unhealthy	**0.5688**	**0.8671**	**0.9786**	**0.1115**
	Healthy	0.7708	0.9121	0.9780	0.0659

64 configurations: a final BN layer just before the output layer (BN), weighted cross-entropy loss according to class imbalance (WL), data augmentation (DA), mixup (MX), unfreezing or freezing (learnable vs pre- trained weights) the previous BN layers in ResNet34 (UF), and weight decay (WD).

Class	Config Id	Test set precision	Test set recall	Test set F1-score	Epoch	BN	DA	UF	WD
Unhealthy (class = 1)	**31**	0.9856	0.9133	**0.9481**	6	✓			
	23	0.9718	0.9200	0.9452	6	✓	✓		
	20	0.9926	0.8933	0.9404	7	✓	✓	✓	✓
Healthy (class = 0)	31	0.9193	0.9867	0.9518	6	✓			
	23	0.9241	0.9733	0.9481	6	✓	✓		
	20	0.9030	0.9933	0.9460	7	✓	✓	✓	✓



Upper right corner is the area that unhealthy samples are positive. When BN is True, softmax for unhealthy (red lines) are all in the area of being unhealthy. But when BN is False, most of the lines are blue (healthy) in the same area, showing that most of the samples in unhealthy area are predicted as healthy.

Conclusion

- Putting an additional BN layer just before the output layer has a considerable impact in terms of minimizing the training time and test error for minority classes in highly imbalanced datasets.
- Upon adding the final BN layer the F1 test score is increased from 0.2942 to 0.9562 for the unhealthy Apple minority class, from 0.7237 to 0.9575 for the unhealthy Pepper and from 0.5688 to 0.9786 for the unhealthy Tomato when WL is not used (all are averaged values over 10 runs)
- The highest gain in test F1 score for both classes (majority vs. minority) is achieved just by adding a final BN layer, resulting in a more than three-fold performance boost on some configurations.
- Trying to minimize validation and train losses may not be an optimal way of getting a high F1 test score for minority classes.
- Having a higher train and validation loss but high validation accuracy would lead to higher F1 test scores for minority classes in less time.
- The final BN layer in imbalanced classification problems has a calibration effect (*the probability associated with the predicted class label should reflect its ground truth correctness likelihood*). That is, the model might perform better even if it is not confident enough while making a prediction.
- Lower values in the softmax output may not necessarily indicate 'lower confidence level', leading to another discussion why softmax output may not serve as a good uncertainty measure for DNNs. A model can be uncertain in its predictions even when having a high softmax output