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

Veysel Kocaman | PhD Researcher

Universiteit Leiden
The Netherlands

ICPR2020

Co-authors :

Dr. Ofer M. Shir (Tel-Hai College, Israel)
Prof.Dr. Thomas Baeck (LIACS)

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

Veysel Kocaman
LIACS
Leiden University
Leiden, The Netherlands
v.kocaman@liacs.leidenuniv.nl

Ofer M. Shir
Computer Science Department
Tel-Hai College and Migal Institute,
Upper Galilee, Israel
ofersh@telhai.ac.il

Thomas Bäck
LIACS
Leiden University
Leiden, The Netherlands
t.h.w.baeck@liacs.leidenuniv.nl

***Abstract*—Some real-world domains, such as Agriculture and Healthcare, comprise early-stage disease indications whose recording constitutes a rare event, and yet, whose precise detection at that stage is critical. In this type of highly imbalanced classification problems, which encompass complex features, deep learning (DL) is much needed because of its strong detection capabilities. At the same time, DL is observed in practice to favor majority over minority classes and consequently suffer from inaccurate detection of the targeted early-stage indications. To simulate such scenarios, we artificially generate skewness (99% vs. 1%) for certain plant types out of the PlantVillage dataset as a basis for classification of scarce visual cues through transfer learning. By randomly and unevenly picking healthy and unhealthy samples from certain plant types to form a training set, we consider a base experiment as fine-tuning ResNet34 and VGG19 architectures and then testing the model performance on a balanced dataset of healthy and unhealthy images. We empirically observe that the initial F1 test score jumps from 0.29 to 0.95 for the minority class upon adding a final Batch Normalization (BN) layer just before the output layer in VGG19. We demonstrate that utilizing an additional BN layer before the output layer in modern CNN architectures has a considerable impact in terms of minimizing the training time and testing error for minority classes in highly imbalanced data sets. Moreover, when the final BN is employed, minimizing the loss function may not be the best way to assure a high F1 test score for minority classes in such problems. That is, the network might perform better even if it is not 'confident' enough while making a prediction; leading to another discussion about why softmax output is not a good uncertainty measure for DL models. We also report on the corroboration of these findings on the ISIC Skin Cancer as well as the Wall Crack datasets.**

crop diseases or premature malignant tumors in humans. The current work is rooted in Precision Agriculture, and particularly in Precision Crop Protection [1], where certain visual cues must be recognized with high accuracy in early stages of infectious diseases' development. A renowned use-case is the Potato Late Blight, with dramatic historical and economical impacts [2], whose early detection in field settings remains an open challenge (despite progress achieved in related learning tasks; see, e.g., [3]). The hard challenge stems from the actual nature of the visual cues (which resemble soil stains and are hardly distinguishable), but primarily from the fact that well-recording those early-stage indications is a rare event.

In recent years, reliable models capable of learning from small samples have been obtained through various approaches, such as autoencoders [4], fine tuning with transfer learning [5], data augmentation [6], cosine loss utilizing (replacing categorical cross entropy) [7], or prior knowledge [8]. Our primary research question is thus the following: *What is the most effective approach to enable learning of minority classes, and could Batch Normalization (BN) serve as one?*

A prominent effort towards plant diseases' detection is the PlantVillage Disease Classification Challenge. As part of the PlantVillage project, 54,306 images of 14 crop species with 26 diseases (including healthy, 38 classes in total) were made publicly available [9]. Following the release of the PlantVillage dataset, deep learning (DL) was intensively employed [10], with reported accuracy ranging from 85.53% to 99.34%. Without any

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.

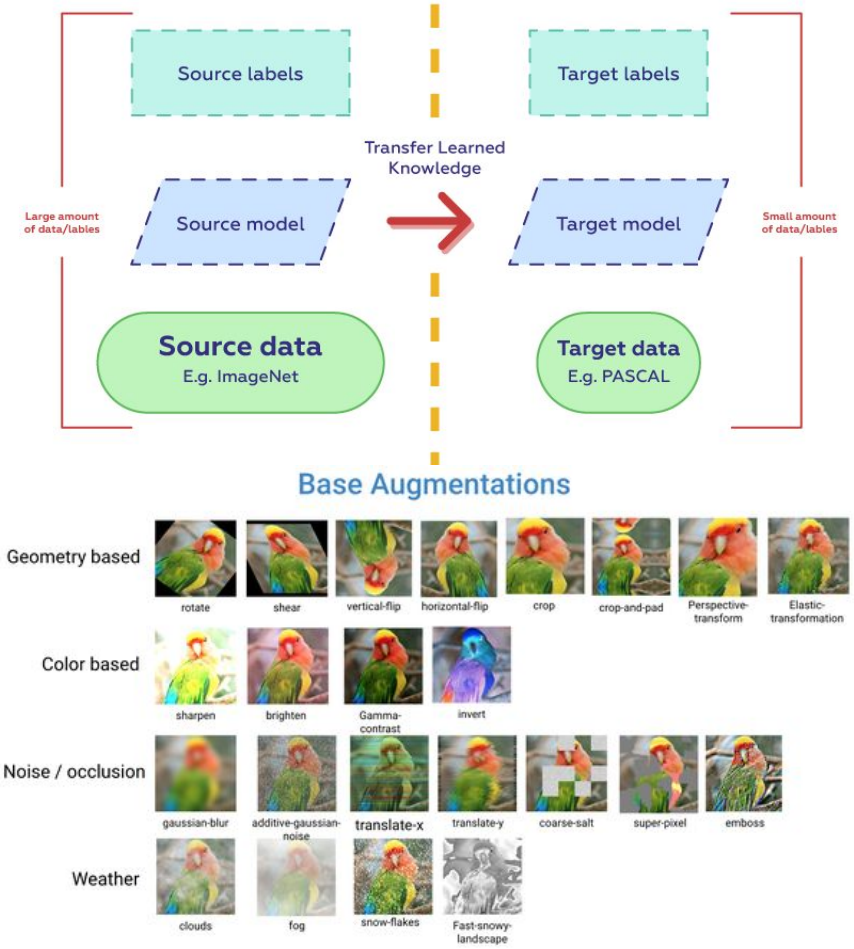
- In this kind of highly imbalanced classification problems, DL frameworks favor majority classes over minority classes (generalisation power of DL).

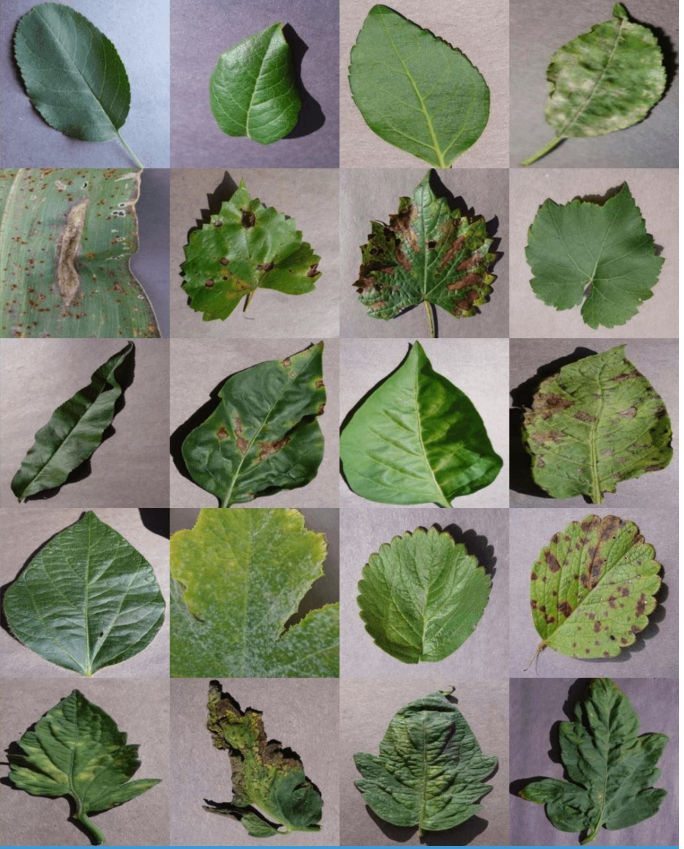
- Under covariate shift (dataset shift), train and test set come from different distributions and model fails to predict samples which haven't seen during training.

- ***What is the most effective approach to enable learning of minority classes?***

Research Question

- **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)?***



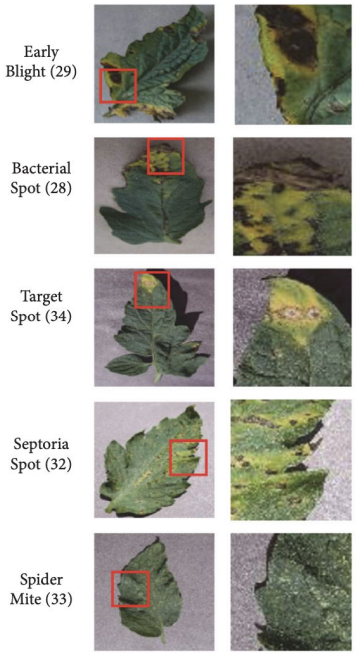


PlantVillage

Experiments with PlantVillage Dataset

	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

54,306 images of 14 crop species with 26 diseases (38 classes in total)

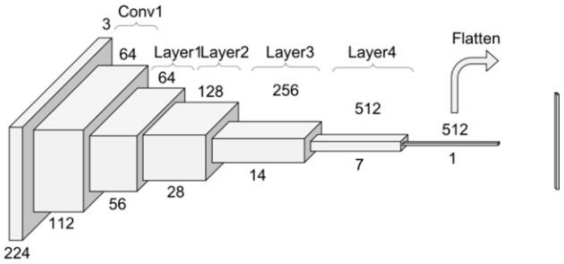


Experiments with ResNet34 and VGG19

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

- treated **unhealthy** class as a **minority class**
- the ratio of unhealthy images is 1% in train set while it is evenly distributed (balanced) in test set.
- a balanced dataset for test set to test the model performance on unseen images that are in an equal number for both classes.
- 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.




```
Sequential(
  (0): AdaptiveConcatPool2d(
    (ap): AdaptiveAvgPool2d(output_size=1)
    (mp): AdaptiveMaxPool2d(output_size=1)
  )
  (1): Flatten()
  (2): BatchNorm1d(1024, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (3): Dropout(p=0.25, inplace=False)
  (4): Linear(in_features=1024, out_features=512, bias=True)
  (5): ReLU(inplace=True)
  (6): BatchNorm1d(512, eps=1e-05, momentum=0.1, affine=True, track_running_stats=True)
  (7): Dropout(p=0.5, inplace=False)
  (8): Linear(in_features=512, out_features=2, bias=True)
  (9): BatchNorm1d(2, eps=1e-05, momentum=0.01, affine=True, track_running_stats=True)
)
```

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

Experiments with ResNet34 and VGG19

TABLE II: Best performance metrics over the Apple dataset under various configurations using ResNet34.

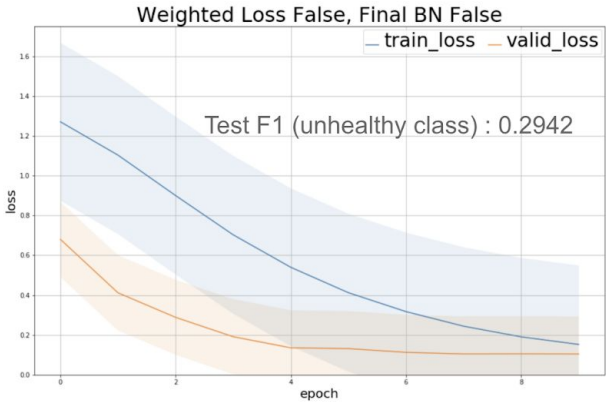
Class	Config Id	Test set precision	Test set recall	Test set F1-score	Epoch	BN	DA	UF	WD
Unhealthy (class = 1)	**31**	0.9856	0.9133	**0.9481**	6	✓			
	23	0.9718	0.9200	0.9452	6	✓	✓		
	20	0.9926	0.8933	0.9404	7	✓	✓	✓	✓
Healthy (class = 0)	31	0.9193	0.9867	0.9518	6	✓			
	23	0.9241	0.9733	0.9481	6	✓	✓		
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

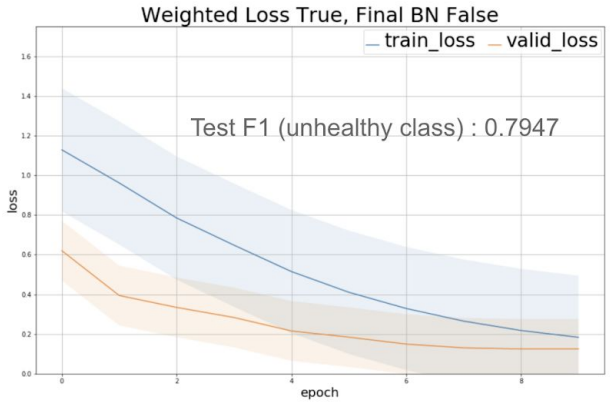
64 configurations: a final BN layer just before the output layer (BN), weighted cross-entropy loss according to class imbalance (WL), data augmentation (DA), mixup (MX), unfreezing or freezing (learnable vs pre- trained weights) the previous BN layers in ResNet34 (UF), and weight decay (WD)

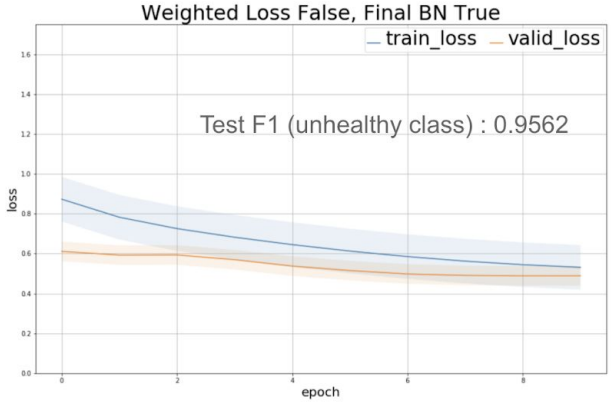
TABLE III: Best (top three) and worst (bottom three) performing configurations (F1 measure, for the unhealthy/minority class) when using ResNet34 for the Apple, Pepper and Tomato datasets.

Config Id	Apple	Pepper	Tomato	Average	BN	DA	MX	UF	WD
29	0.9332	0.9700	0.9866	**0.9633**	✓			✓	
28	0.9332	0.9566	0.9966	0.9622	✓			✓	✓
31	0.9499	0.9433	0.9833	**0.9588**	✓				
49	0.6804	0.5080	0.5339	0.5741		✓	✓	✓	
48	0.5288	0.5638	0.5638	0.5521		✓	✓	✓	✓
50	0.6377	0.5027	0.4528	0.5311		✓	✓		✓

Experiments with ResNet34 and VGG19







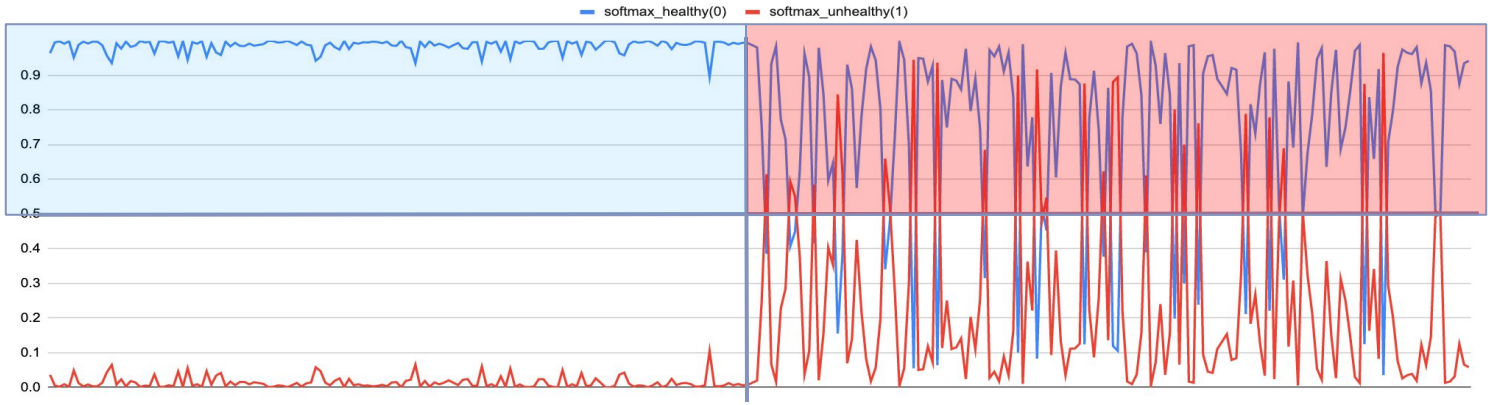
Without final BN layer	With final BN layer
0.1082	0.5108
0.1464	0.6369
0.1999	0.6082
0.2725	0.6866
0.3338	0.7032

TABLE IV: Softmax output values (representing class probabilities) for five sample images of unhealthy plants. Left column: Without final BN layer, softmax output values for unhealthy, resulting in a wrong classification in each case. Right column: With final BN layer, softmax output value for unhealthy, resulting in correct but less "confident" classifications.

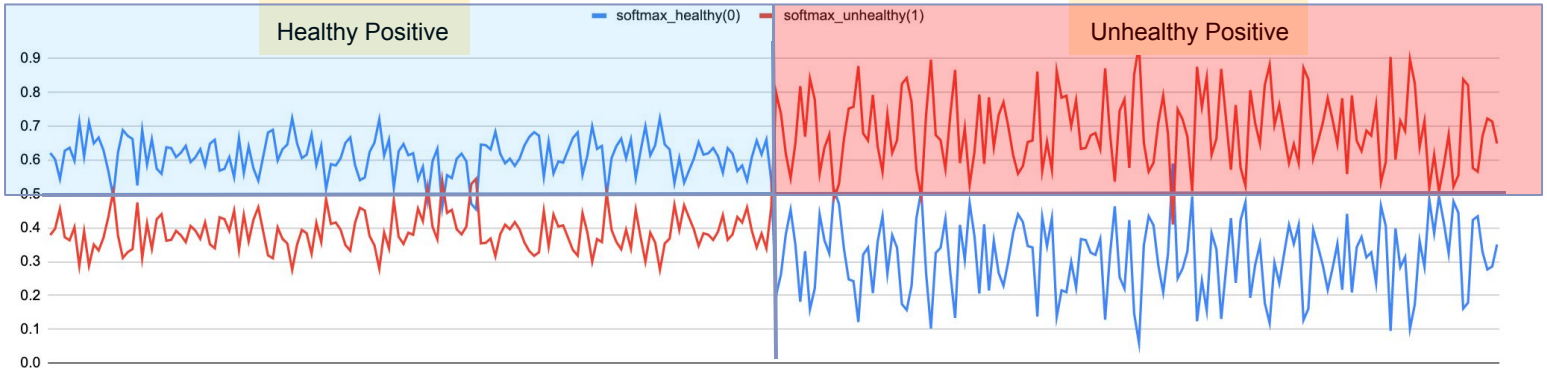
the network might perform better even if it is not 'confident' enough while making a prediction; leading to another discussion about why softmax output is not a good uncertainty measure for DL models.

Experiments with ResNet34 and VGG19

Final BN = False, softmax outputs vs ground truths



Final BN = True, softmax outputs vs ground truths



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

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

Veysel Kocaman | PhD Researcher

Universiteit Leiden
The Netherlands

ICPR2020

Co-authors :

Dr. Ofer M. Shir (Tel-Hai College, Israel)
Prof.Dr. Thomas Baeck (LIACS)