Improving Image Classifiers for Small Datasets by Learning Rate Adaptations

S Mishra, T Yamasaki, H Imaizumi

MVA 2019

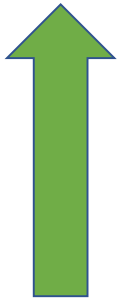
Oral Session 3

May 29, 2019

Scope

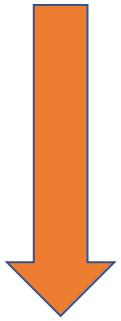
Rapid improvements in image classification tasks

- Larger better & detailed datasets
- Faster hardware resources
- Better architectures



However (the ugly truth)!

- More iterations to SOTA
- Longer train time
- Higher costs
- Small dataset reliability low



Scope

Deployment costs can adversely impact individuals or smaller groups.

SOLUTION?

- Organic combination of proven techniques, field tested on benchmark datasets.
- Optimization by learning rate (v) adaptations.
- Transfer modus-operandi to smaller, untested data.
- Ensure repeatability.

Baseline

- Multi-class classification on CIFAR-10
- Test candidate architectures of increasing size/complexity

  Resnet-34, ResNet-50, ResNet-101, ResNet-152

  DenseNet161
- Baseline Performance

  5:1 split, Early stopping, lower LR restarts

  BCE with logits loss

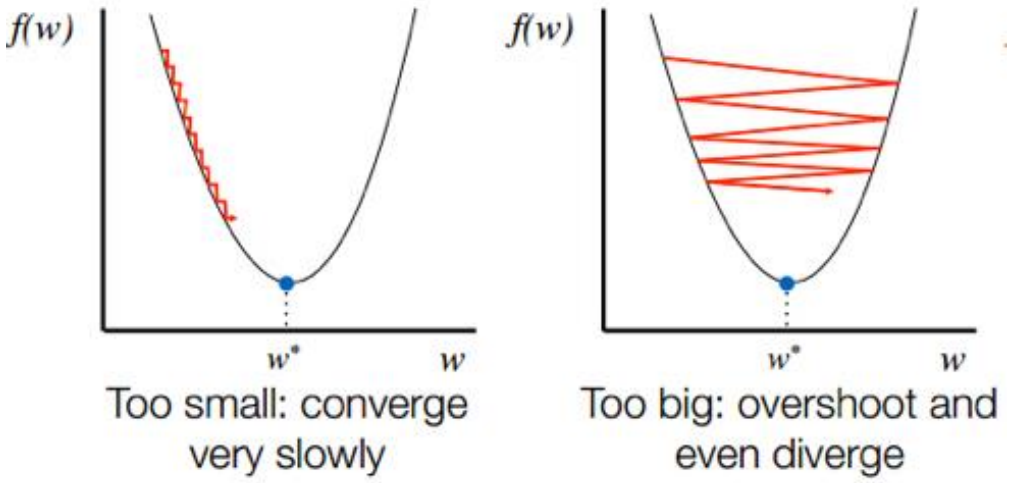
  Train to 90%+ validation accuracy mark

Baseline

Architecture	Accuracy (Top-1)	Time (s)
ResNet 34	90.36%	17,757
ResNet-50	90.54%	34,039
ResNet-101	90.71%	60,639
ResNet-152	90.68%	91,888
DenseNet-161	93.02%	54,628

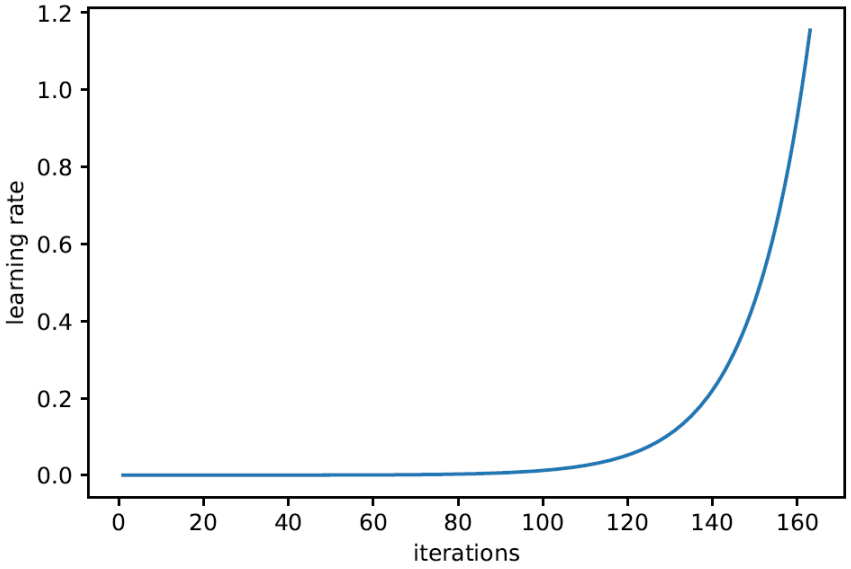
Learning Rate range-test

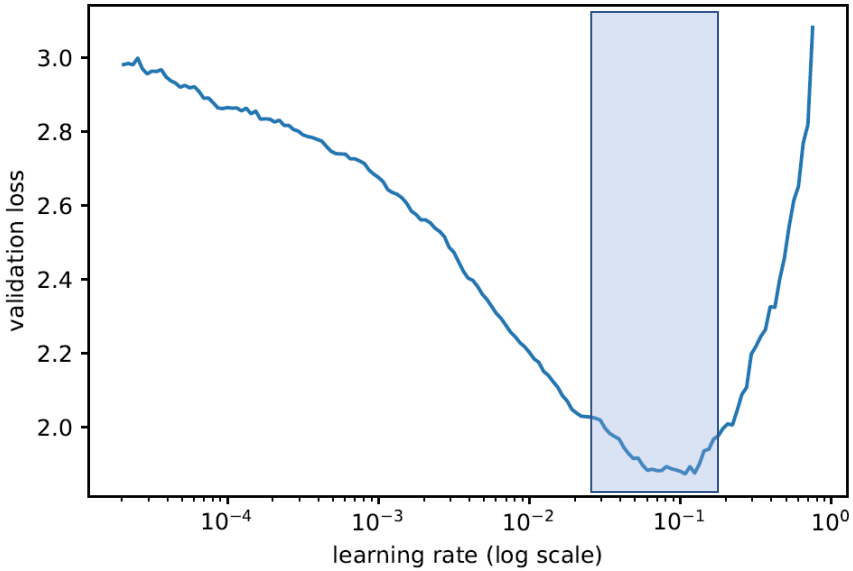
Learning rate under-exploited by Monotonic change.



Courtesy:
Cyclical Learning rates for training NN, L. Smith [2017]
Deep Learning, S. Verma et al. 2018

Learning Rate range-test





Steadily increase the LR and observe the Cross entropy loss

Test several mini-batches to see a point of inflexion

Courtesy:

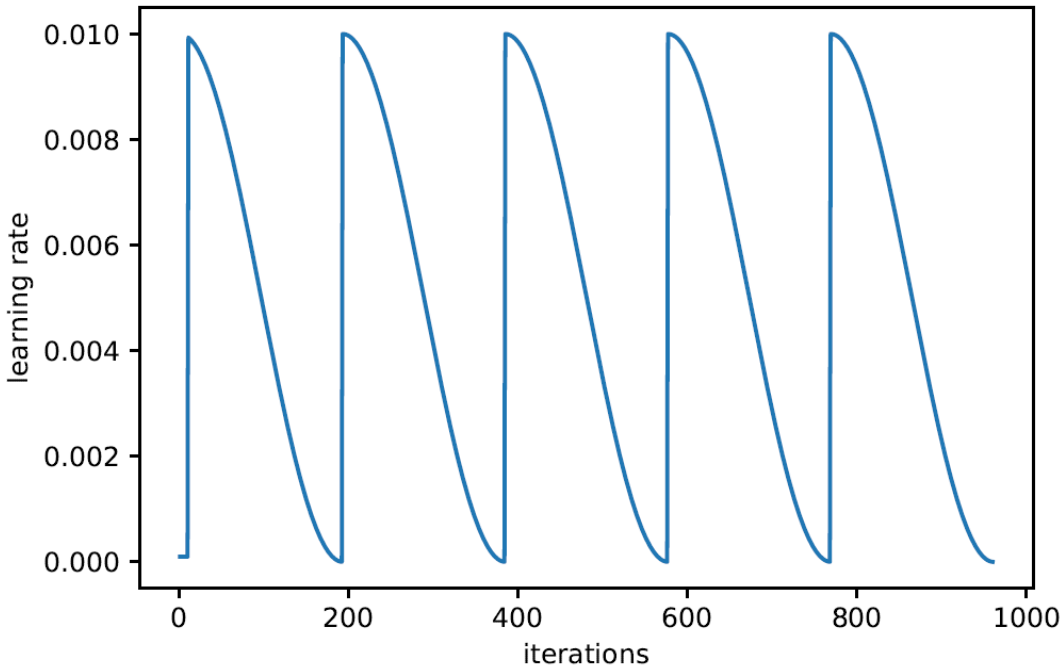
Cyclical Learning rates for training NN, L. Smith [2017]

SGD-R

Avoid monotonicity

Cycle LR v by Cosine scheduling function

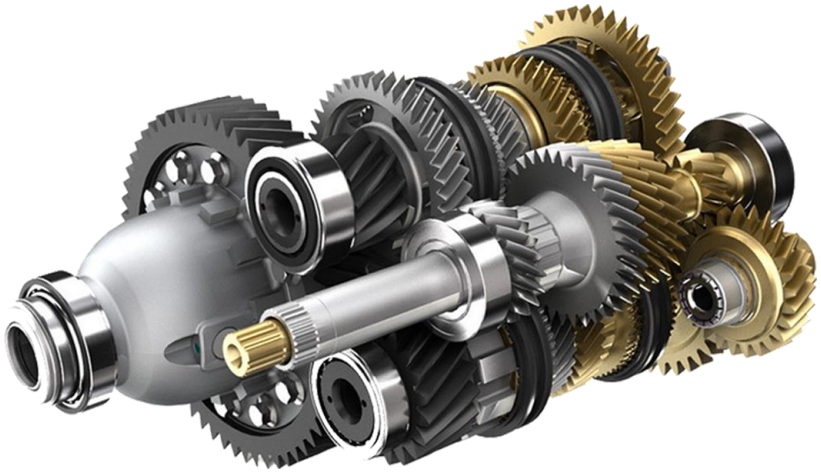
$$v(t) = \frac{1}{2}\left(1 + v\, cos\left(\frac{t\pi}{T}\right)\right) + \varepsilon$$



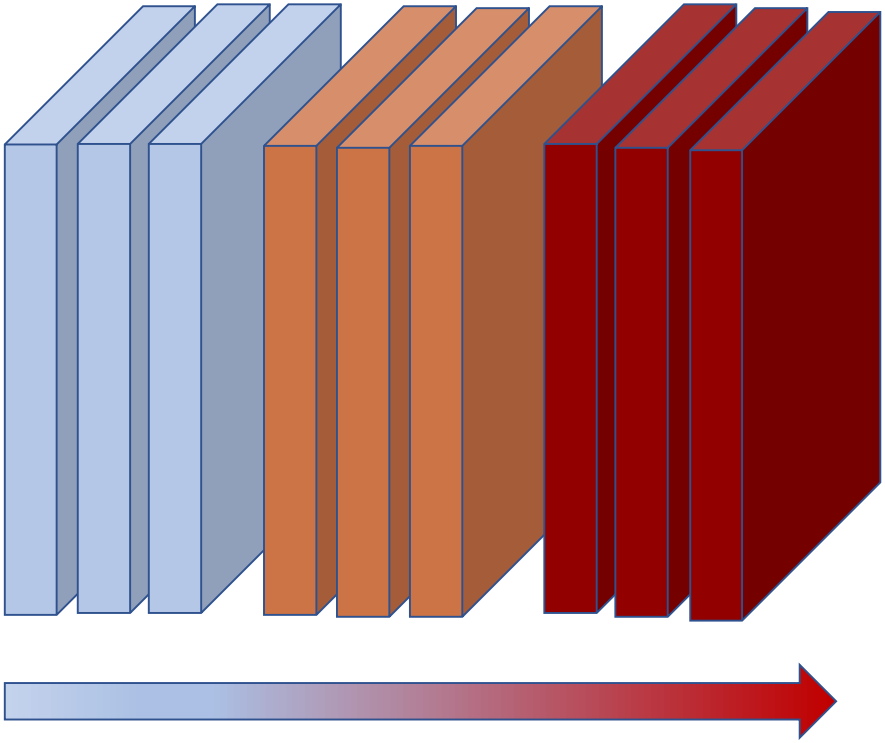
Perform initial coarse fit by tuning the last (or last few) FC layer

Courtesy:
SGD with Warm restarts, Loschilov [2017]

Differential learning



Gear-box need not spin all gears equally!



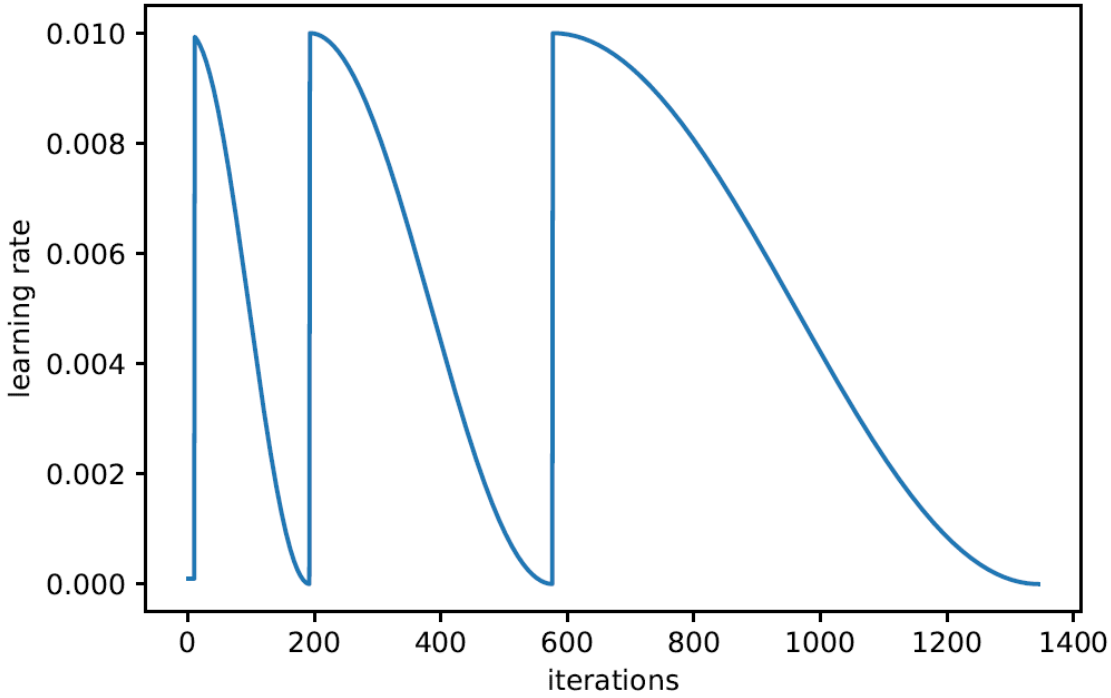
Reduce computational overhead by assigning different learning rates.

Courtesy:
J Howard, T. Parr [2018]

SGD-R + CLM + DLR

Cycle Length Multiply by integral powers of 2.

Cover larger # epochs per anneal cycle



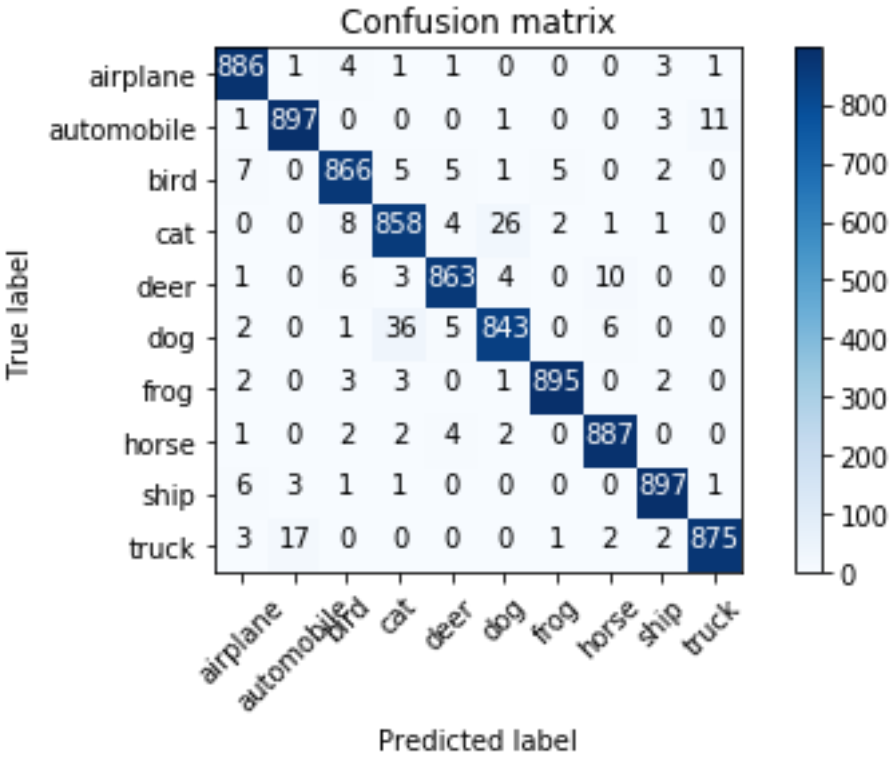
Perform tighter fit over all layers

Courtesy:
SGD with Warm restarts, Loschilov [2017]

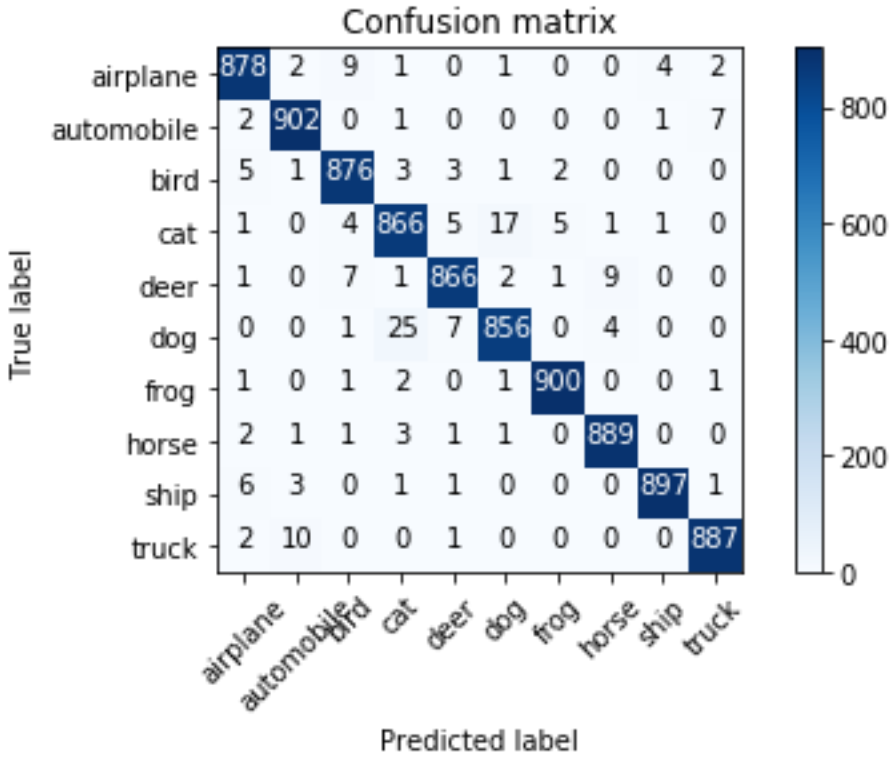
Results

Architecture	Accuracy (Top-1)	Time (s)	η
ResNet 34	96.84%	9,565	1.84
ResNet-50	96.82%	11,817	2.88
ResNet-101	97.61%	6,673	9.09
ResNet-152	97.78%	9,012	10.2
DenseNet-161	97.15%	7,195	7.59

Results



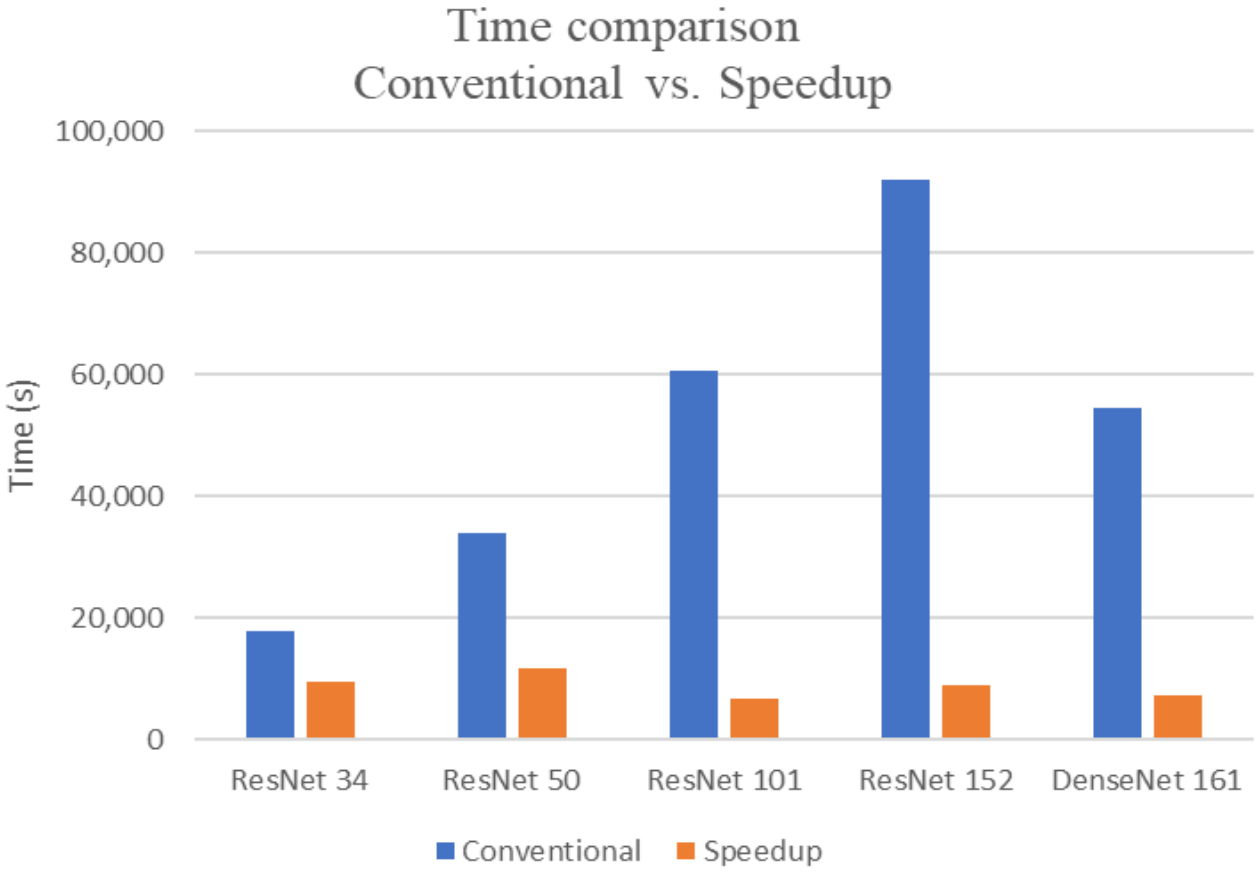
DenseNet 161



ResNet 152

** Appendix*

Results



Higher dividends when architecture size grows larger.
Possible by offsetting the computation overhead by DLR

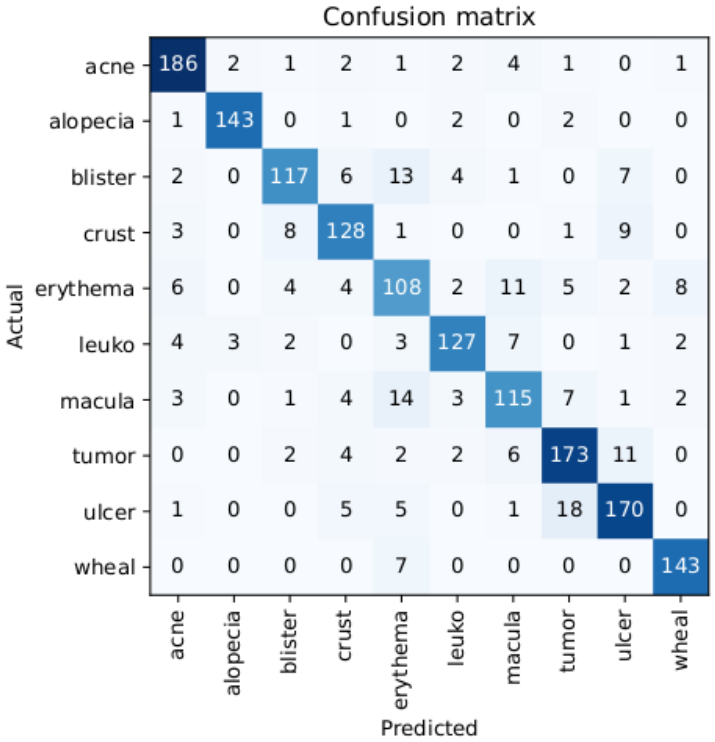
Application

exMedio

Dermoscopic detection as diagnostic aid:

- 10 common classes & homogenous
- Requires rapid retrain
- Severe class imbalance

- >10% acc. improvement
- 3.1x - 5.7x in time
- $ 6.9 (old) vs $ 0.7 (new)



Application - Samples

Acne

Alopecia

Crust

Tumor