# Tune It or Don't Use It: Benchmarking Data-Efficient Image Classification

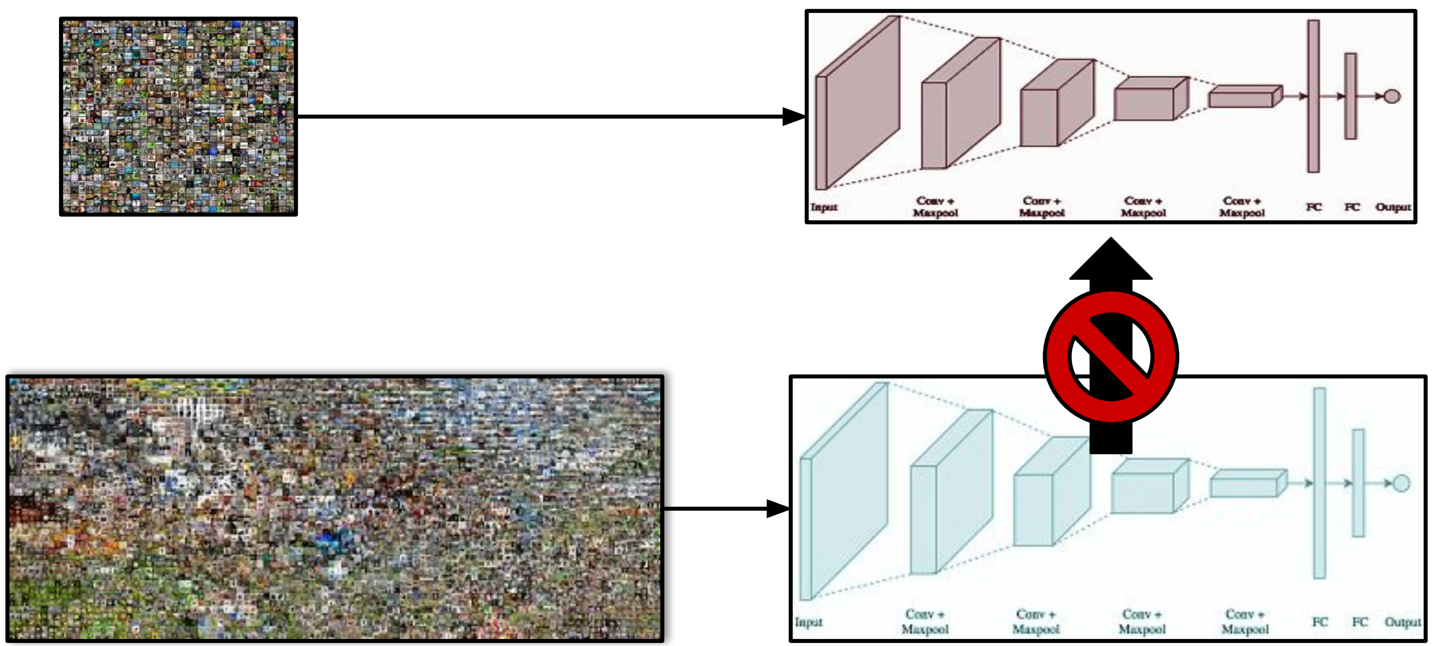
**L. Brigato**, B. Barz, L. Iocchi, and J. Denzler

*2nd Visual Inductive Priors for Data-Efficient Deep Learning Workshop*
*ICCV 2021*

# Context

**Data-Efficient Image Classification**

- Train classifiers from limited data - Few tens of samples per class (< 100, > 10)
- Do not employ pre-trained networks - NO Transfer Learning from large datasets



# Motivation

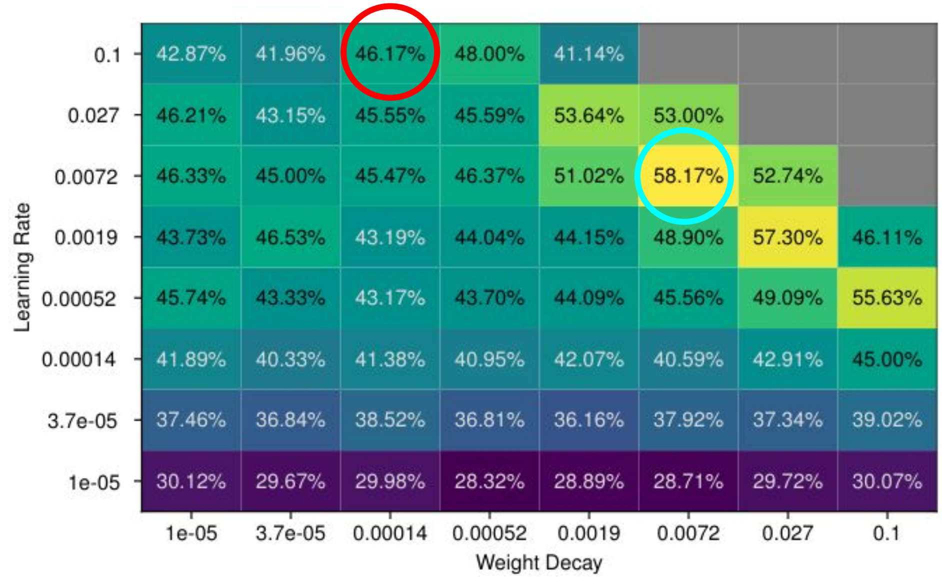
**Lack of a common benchmark**

- Subsampled datasets without canonical splits
- Overfitting on natural images (e.g. CIFAR10) which generally do not have data-deficiency issues

**Lack of reliable comparisons**

- Neglecting the existing state of the art
- Using untuned baselines

Example on 1% of CIFAR10:
Default *lr* and *wd* provide sub-optimal performance ( ~12 pp)



# Contribution

**Propose 1st benchmark for data-efficient image classification**

- Composed of 6 different datasets covering different data types and domains
- Follow a strict evaluation pipeline
- Re-evaluate 8 state-of-the-art methods along with the baseline

# Findings

**Well-tuned cross-entropy is a strong baseline**

- Cross-entropy ranks second in our benchmark only behind one specialized method
- Hyper-parameter optimization (HPO) is fundamental to avoid biased evaluations

**Published baselines are underperforming**

- Our cross-entropy baseline beats published ones by up to ~18 pp on similar setups

**A strong baseline is obtainable with:**

- Very small batch sizes
- Weight decay tuned ad-hoc for each dataset
- Small learning rates

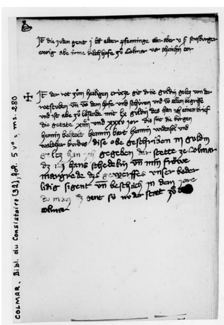
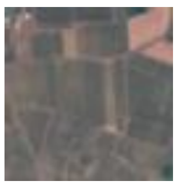
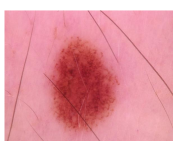
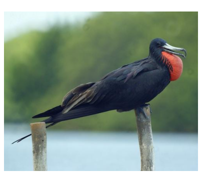
# Datasets

**Comprehensive set concerning data domains and data types**

- From popular academic datasets (e.g. ImageNet) to handwriting recognition
- From RGB, to multispectral, to grayscale images

**Datasets dimension**

- Mostly training sets with 50 images per class


| Dataset | Classes | Imgs/Class | #Trainval | #Test | Problem Domain | Data Type |
|---|---|---|---|---|---|---|
| ImageNet-1k [24] | 1,000 | 50 | 50,000 | 50,000 | Natural Images | RGB |
| ciFAIR-10 [15, 2] | 10 | 50 | 500 | 10,000 | Natural Images | RGB (32x32) |
| CUB [33] | 200 | 30 | 5,994 | 5,794 | Fine-Grained | RGB |
| EuroSAT [12] | 10 | 50 | 500 | 19,500 | Remote Sensing | Multispectral |
| ISIC 2018 [7] | 7 | 80 | 560 | 1,944 | Medical | RGB |
| CLaMM [26] | 12 | 50 | 600 | 2,000 | Handwriting | Grayscale |

# Methods

**Re-evaluation of 8 approaches published between 2017 and 2021**

- Selected approaches originally tested on small/subsampled versions of popular datasets
- Using publicly available code when available

| |
|---|
| Cross-Entropy Baseline |
| Deep Hybrid Networks [21, 22] |
| OLÉ [16] |
| Grad-$\ell_2$ Penalty [3] |
| Cosine Loss [1] |
| Cosine Loss + Cross-Entropy [1] |
| Harmonic Networks [31, 32] |
| Full Convolution [13] |
| Dual Selective Kernel Networks [27] |
| T-vMF Similarity [14] |

**Approaches of different type**

- Loss-based (e.g. Cosine loss, T-vMF similarity)
- Geometric priors (e.g. Deep Hybrid Networks, Harmonic Networks)
- Architecture-based (e.g. Full Convolution, DSK networks,)

SAPIENZA
UNIVERSITÀ DI ROMA

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

# Experimental Setup

**Data pre-processing**

- Channel-wise normalization of input images (mean and std)
- Dataset-specific data augmentation (e.g. h/v flipping, scale/shift augmentation)

**Architecture and optimizer**

- Wide ResNet 16-8 / ResNet-50 for ciFAIR10 / other datasets
- SGD with momentum (0.9) and cosine learning rate schedule

**HPO**

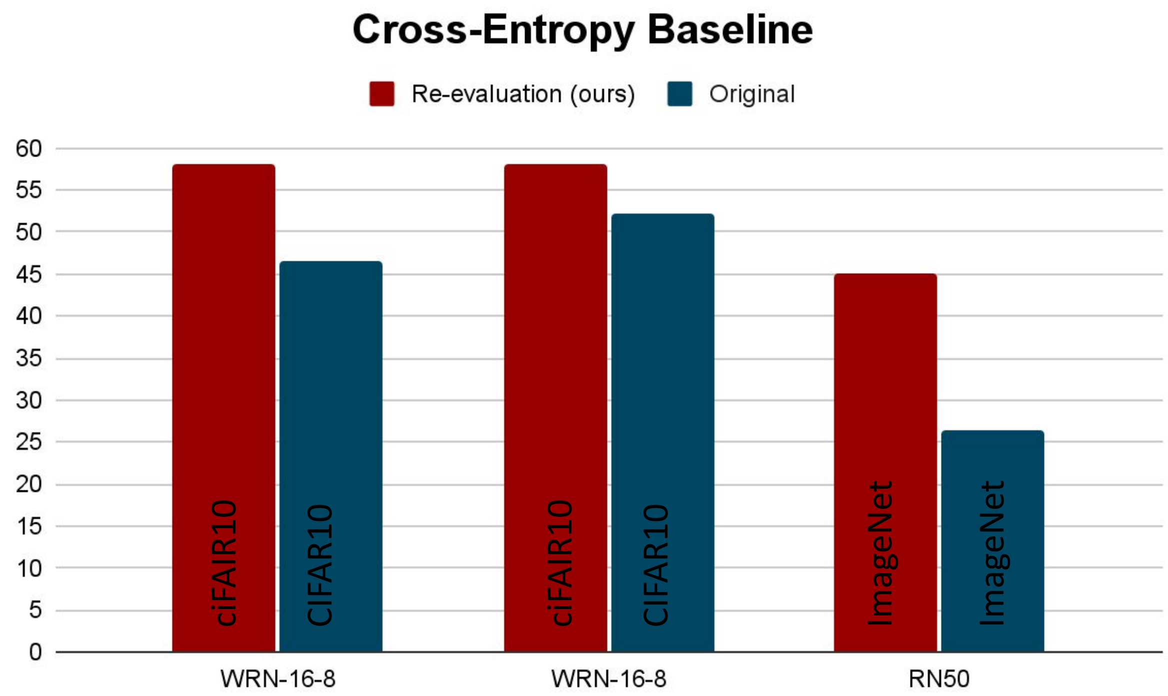
- Asynchronous HyperBand with Successive Halving (ASHA) algorithm
- We tuned *lr*, *wd*, *bs,* and method-specific additional hyper-parameters
- 40% of training set used for validation

| Hyper-Parameter | ImageNet | ciFAIR-10 | CUB | EuroSAT | ISIC 2018 | CLaMM |
|---|---|---|---|---|---|---|
| Learning Rate | | | `loguniform`(1e-4, 0.1) | | | |
| Weight Decay | | | `loguniform`(1e-5, 0.1) | | | |
| Batch Size | {8, 16, 32} | {10, 25, 50} | {8, 16, 32} | {10, 25, 50} | {8, 16, 32} | {8, 16, 32} |
| Epochs | 200 (500) | 500 | 200 | 500 | 500 | 500 |
| HPO Trials | 100 | 250 | 100 | 250 | 100 | 100 |
| Grace Period | 10 | 50 | 10 | 25 | 25 | 25 |

# Results

## Comparison with published methods on similar settings

- Re-evaluations of data-efficient methods similar to published values ( -0.5 pp, -1.9 pp, +5.5 pp).
- Re-evaluations of baselines much better than published values ( +11.7 pp, +6.0 pp, +18.6 pp)



# Results

**Data-efficient image classification benchmark**

- Harmonic networks is the champion of our benchmark with 68.70% average accuracy
- The baseline ranks second with a close 67.90%
- T-vMF, DSK Networks, Cosine+Xe loss, and OLE follow with ~64/65%

| Method | ImageNet | ciFAIR-10 | CUB | EuroSAT | ISIC 2018 | CLaMM | Average |
|---|---|---|---|---|---|---|---|
| Cross-Entropy Baseline | 44.97 | **58.22** | 71.44 | 90.27 | 67.19 | **75.34** | 67.90 |
| Deep Hybrid Networks [21, 22] | 38.69 | 54.21 | 52.54 | 91.15 | 59.64 | 65.74 | 60.33 |
| OLÉ [16] | 43.05 | 54.92 | 63.32 | 89.29 | 62.89 | 71.42 | 64.15 |
| Grad-$\ell_2$ Penalty [3] | 25.21 | 51.03 | 51.94 | 79.33 | 60.21 | 65.10 | 55.47 |
| Cosine Loss [1] | 37.22 | 52.39 | 66.94 | 88.53 | 62.42 | 68.89 | 62.73 |
| Cosine Loss + Cross-Entropy [1] | 44.39 | 51.74 | 70.80 | 88.77 | 64.52 | 69.29 | 64.92 |
| Harmonic Networks [31, 32] | **46.36** | 56.50 | **72.26** | **92.09** | **70.42** | *74.59* | **68.70** |
| Full Convolution [13] | 36.58 | 55.00 | 64.90 | 90.82 | 61.70 | 63.33 | 62.06 |
| Dual Selective Kernel Networks [27] | 45.21 | 54.06 | 71.02 | 91.25 | 64.78 | 61.51 | 64.64 |
| T-vMF Similarity [14] | 42.79 | *57.50* | 67.43 | 88.53 | 65.37 | 66.40 | 64.67 |

# Qualitative picture of computational demand

## Total run times for the baseline

- 4 Nvidia V100 GPUs 32GB
- Around 6 days to evaluate the baseline on the full benchmark

| Dataset | HPO (#hrs) | Final Training - 10 reps (#hrs) |
|---|---|---|
| ciFAIR10 | 3.7 | 1.2 |
| EuroSAT | 3.3 | 1.4 |
| ISIC 2018 | 4.5 | 3.3 |
| ClaMM | 4.8 | 3.5 |
| CUB | 6.5 | 6.3 |
| ImageNet | 63 | 37 |
| **Total** | **85.8** | **52.7** |

# Conclusions

**Introduced the 1st benchmark for data-efficient image classification**

- Comprehensive with respect to data domains and types
- Dataset splits publicly available: **https://github.com/cvjena/deic**
- Possible to extend with new splits/datasets

**Proposed a fair evaluation pipeline to re-evaluate existing approaches**

- Common experimental setup (e.g. architecture, optimizer etc.)
- Careful tuning of hyper-parameters

**Cross-entropy has been largely undervalued**

- Ranks second only behind Harmonic Networks
- Beats published baselines by large margins

**Our benchmark and training procedure allows to:**

- Compare with the existing state of the art
- Avoid misleading comparison with weak baselines
- Set the stage for future developments of the research field

# Context

**Data-Efficient Image Classification**

- Increase in number of publications concerning this research area
- Organization of workshops and challenges
- tagliare

**Deep Learning on Small Datasets without Pre-Training using Cosine Loss**

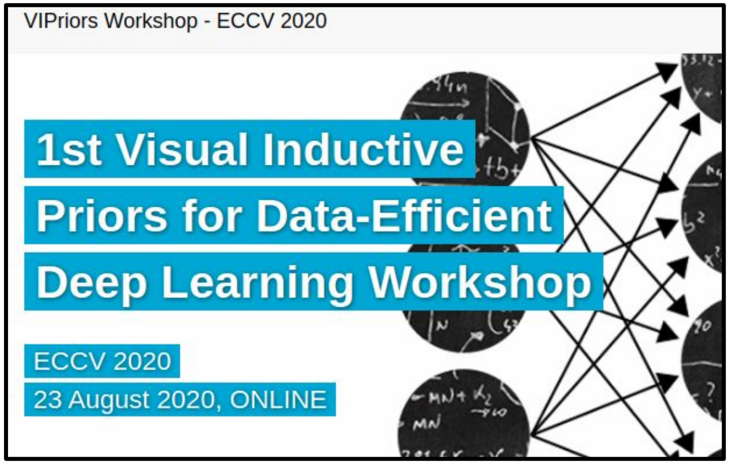

Björn Barz Joachim Denzler

Friedrich Schiller University Jena, Germany
Computer Vision Group


A Close Look at Deep Learning with Small Data


Lorenzo Brigato
Dpt. of Computer, Control, and Management Engineering
Sapienza University of Rome

Luca Iocchi
Dpt. of Computer, Control, and Management Engineering
Sapienza University of Rome

# Results

**Tuned hyper-parameters (baseline)**

- Small batch sizes, high *wd* and small *lr*
- *Lr* and *wd* appear to be negatively correlated *r = -0.58* (after taking the log of both)