Rethinking of Deep Models Parameters with Respect to Data Distribution

Shitala Prasad, Yiqun Li, Dongyun Lin, Dong Sheng, Oo Zaw Min

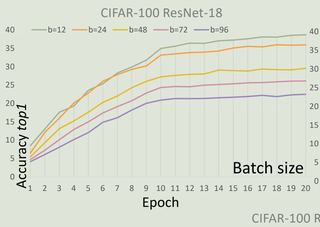
Visual Intelligence, Institute for Infocomm Research (I2R), A*Star Singapore

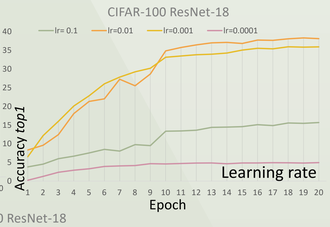
Introduction

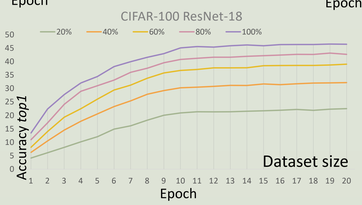
The revolutionization of deep learning (DL) [1] in field of computer vision (CV) has changed the way the visual learning takes place. Because of which today transfer learning is very much feasible even with small datasets, by simply sharing the useful knowledge from one domain to another. A well-established pre-trained model helps to achieve decent performance by fine-tuning to the target training dataset.

Challenges:

- Can the learning performance be improved without additional data?
- Can accuracy be increased for the same architecture with the same dataset?
- Can data distribution strategy boosts the accuracy and reduces the training cost?
- Can it reduce network over-fitting?







Objective

Inspired by the teacher student ideas, in this paper we showed that adjusting data distribution while training can lead to a significant performance improvement without any additional data contribution. Then, we showed that this strategy actually works for various different state-of-the-art image classification architectures and different types of datasets.

- A simple step-wise hyper-parameter tuning strategy to boost the network classification performance without using any additional data
- Evaluated the proposed strategy to be consistently valued for several state-of-the-art image classification network architectures
- Compared with the baseline training, the proposed strategy significantly achieves uprise in top-1 and top-5 classification accuracies on CIFAR-100 [20], Birdsnap [21], Food-101 [22] and COVID-19 mask-nomask datasets for different networks
- The proposed step-wise training reduces the overall training cost by ≈40%

COVID-19 mask-nomask Dataset: 1. **covid mask images:** https://www.kaggle.com/danielferrazcampos/face-mask-images
2. **mask dataset:** https://www.kaggle.com/ahmetfurkandemr/mask-datasets-v1
3. **COVID19 mask image dataset:** https://github.com/UniversalDataTool/coronavirus-mask-image-dataset

Methodology

Here, we describe and derive the data distribution problem in image classification, then study different training approaches and finally propose our new training strategy to optimize the learning.

A. Problem Definition

Let's say N be a CNN model where it's i layer is defined as $Y_i = F_i(X_i)$ such that Y_i is the output, F_i is the operator and X_i is the input to the i layer with tensor shape (b, h_i, w_i, C_i). Here, (h_i, w_i) is the spatial dimension, C_i is the channel number and b is the batch size. If model N has α parameters then to minimize the loss, L(α) is defined as:

$$L(\alpha) = \frac{1}{M}\sum_{m=1}^{M} L_m(\alpha)$$

where M is the training set size and $L_m(\alpha)$ denotes the loss for the m training sample.

If the training set M is split into several small subsets, *i.e.*,$M = \{m_1, m_2, \dots m_n\}$ and they are gradually added for training the network N , then the new updated loss function $L'(\alpha)$ can be defined as:

$$L'(\alpha) = \frac{1}{\sum_{i=1}^{d} M_i}\sum_{m=1}^{\sum_{i=1}^{d} M_i} L_m(\alpha)$$

where Mi is the subset of M *s.t.* $M_i = (m_1 + m_2 + \dots m_i)$.

B. Training Approach

We observed that scaling up the dataset with an optimal batch size and learning decay gives better performance. Hence, it is important to obtain an optimal set of (M, b, λ).

The noise scale can be equated as:

$$g_i \approx \frac{\lambda' M_i}{b_i} \approx \lambda_i M_i$$

where i represents the progressive update during the training.

Datasets

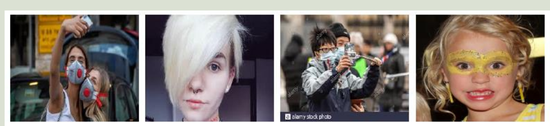
For training and testing, we used two different types of datasets: interclass CIFAR-100 dataset and intraclass Birdsnap and Food-101 datasets.

CIFAR is a natural object dataset of 100 classes with 50k training samples and 10k testing samples.

Birdsnap is the second dataset used in this paper which includes 500 North American bird species collected from Flicker. It has total 49,829 images out of which 2443 images are used for testing.

Food-101 is the third dataset used in this paper which consists of 101 food categories with total 101k training and testing images. Among 1000 images per food category, 750 are involved in training and 250 is used for testing.

We introduce a new COVID-19 mask-nomask dataset with 2,509 train images and 539 test images which is a combination of three smaller datasets for masked and non-masked face classification.



Results

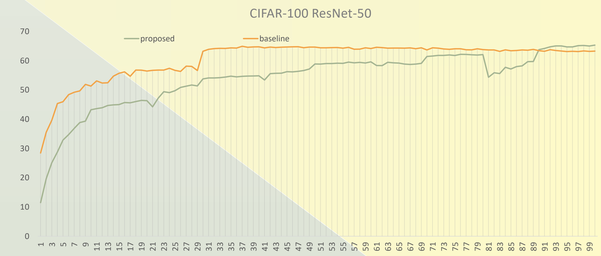
TABLE I: ResNet-50 performance analysis on CIFAR-100 dataset. Letters A, B, C, D, and O respectively represents 20%, 40%, 60%, 80% and 100% training sets of CIFAR-100, as discussed in Section III-A. Note: all the trainings are from the scratch and the **bold** is the best results.

CIFAR-100 Dataset	ResNet-50		
	top-1	*top-5*	Epochs
with set O (baseline)	64.9400	88.6200	100
with set A	35.7900	66.3300	20
with set B	46.7300	77.4700	20
with set C	52.5200	81.9900	20
with set D	55.2900	84.3500	20
with set O	57.2800	85.8100	20
with set A $\xrightarrow{2b}$ B $\xrightarrow{3b}$ C $\xrightarrow{4b}$ D $\xrightarrow{b}$ O	**66.3400**	**90.7200**	20 *each*

TABLE II: Performance analysis of ResNeXt-50 and Res2NeXt-50 on CIFAR-100 dataset.

CIFAR-100 Dataset	ResNeXt-50		
	top-1	*top-5*	Epochs
with set O (baseline)	65.5800	88.6700	100
with set A $\xrightarrow{2b}$ B $\xrightarrow{3b}$ C $\xrightarrow{4b}$ D $\xrightarrow{b}$ O	**66.8000**	**90.1900**	20 *each*

CIFAR-100 Dataset	Res2NeXt-50		
	top-1	*top-5*	Epochs
with set O (baseline)	66.4100	88.7400	100
with set A $\xrightarrow{2b}$ B $\xrightarrow{3b}$ C $\xrightarrow{4b}$ D $\xrightarrow{b}$ O	**66.9800**	**90.5900**	20 *each*



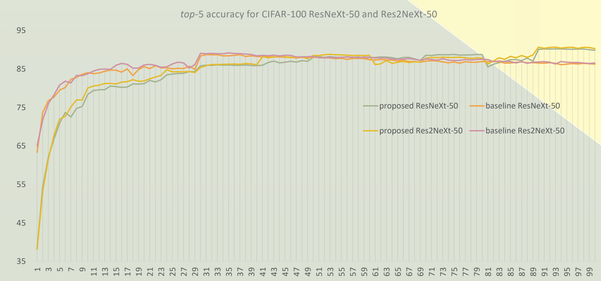


TABLE III: Performance analysis of Res2Net-50 and Res2NeXt-50 on Birdsnap dataset.

Birdsnap Dataset	Res2Net-50		
	top-1	*top-5*	Epochs
with set O (baseline)	61.7901	83.3790	100
with set A $\xrightarrow{2b}$ B $\xrightarrow{3b}$ C $\xrightarrow{4b}$ D $\xrightarrow{b}$ O	**62.1391**	**85.150**	20 *each*

Birdsnap Dataset	Res2NeXt-50		
	top-1	*top-5*	Epochs
with set O (baseline)	61.9900	84.0025	100
with set A $\xrightarrow{2b}$ B $\xrightarrow{3b}$ C $\xrightarrow{4b}$ D $\xrightarrow{b}$ O	**62.4611**	**84.9875**	20 *each*

Ablation Study

COVID-19_mask-nomask Dataset	Res2Net-50	
	top-1	Method
with set O (baseline) (*b=16*)	98.90	$\mathcal{G} \approx \frac{\lambda_i M_i}{b_i}$
with set A $\xrightarrow{2b}$ B $\xrightarrow{3b}$ C $\xrightarrow{4b}$ D $\xrightarrow{b}$ O	97.50 $\xrightarrow{2b}$ 98.75 $\xrightarrow{3b}$ 99.06 $\xrightarrow{4b}$ 98.90 $\xrightarrow{5b}$ 99.67	
with set A $\xrightarrow{2b}$ B $\xrightarrow{3b}$ C $\xrightarrow{4b}$ D $\xrightarrow{b}$ O	97.50 $\xrightarrow{2b}$ 98.75 $\xrightarrow{3b}$ 99.06 $\xrightarrow{4b}$ 98.90 $\xrightarrow{b}$ **99.84**	
with set O (baseline) (*b=24*)	99.53	
with set A $\xrightarrow{2b}$ B $\xrightarrow{3b}$ C $\xrightarrow{4b}$ D $\xrightarrow{b}$ O	97.65 $\xrightarrow{2b}$ 97.18 $\xrightarrow{3b}$ 97.34 $\xrightarrow{4b}$ 97.50 $\xrightarrow{b}$ **100**	
with set O (baseline) (*b=24*)	**99.53**	$\mathcal{G} \approx \frac{\lambda_i M_i}{b}$
with set A $\xrightarrow{2b}$ B $\xrightarrow{3b}$ C $\xrightarrow{4b}$ D $\xrightarrow{b}$ O	97.65 $\xrightarrow{2b}$ 97.65 $\xrightarrow{3b}$ 97.81 $\xrightarrow{4b}$ 98.28 $\xrightarrow{b}$ 97.97	
with set O (baseline) (*b=24*)	99.53	$\mathcal{G} \approx \frac{\lambda M_i}{b_i}$
with set A $\xrightarrow{2b}$ B $\xrightarrow{3b}$ C $\xrightarrow{4b}$ D $\xrightarrow{b}$ O	97.65 $\xrightarrow{2b}$ 97.50 $\xrightarrow{3b}$ 98.12 $\xrightarrow{4b}$ 98.75 $\xrightarrow{b}$ **99.69**	
with set O (baseline) (*b=24*)	99.53	$\mathcal{G} \approx \frac{\lambda M_i}{b}$
with set A $\xrightarrow{2b}$ B $\xrightarrow{3b}$ C $\xrightarrow{4b}$ D $\xrightarrow{b}$ O	97.66 $\xrightarrow{2b}$ 98.44 $\xrightarrow{3b}$ 99.22 $\xrightarrow{4b}$ 99.37 $\xrightarrow{b}$ **99.69**	

TABLE V: Performance analysis of Res2Net-50 on Food-101 dataset with varying *b*.

Food-101 Dataset	Res2Net-50					
	Pre-trained=True				Epochs	Weight Shared
	b=12		*b=16*			
	top-1	*top-5*	*top-1*	*top-5*		
with set O (baseline)	84.6931	96.5901	85.2356	96.7050	100	✓
with set A	75.2832	92.9386	76.0119	93.1248	20 each	✓
$\xrightarrow{2b}$ B	81.5089	95.5129	81.5327	95.2752		
$\xrightarrow{3b}$ C	83.3307	96.2059	83.7109	9.61069		
$\xrightarrow{4b}$ D	84.1624	96.4792	84.4554	96.4356		
$\xrightarrow{b}$ O	**85.9168**	**97.1129**	**86.2653**	**96.9822**		

Conclusion

The M_i trained network weights are uses as the new initializer for Mi+1 subset training that boost the learning curve without saturating

Analyzes a close interrelation between M, b and λ and propose a step-wise training to up rise the performance instead of traditional baseline training without performing any change in the network architecture

The proposed stepwise training reduces the risk of over-fitting by adopting different b and also reduces the training cost by 40%

In future, we would like to explore other aspects of CV such as object detection and segmentation where annotation is the biggest challenge

Acknowledgement

Special thanks to our department and I2R for supporting this research work.