Institute for Infocomm Research
I²R

RETHINKING OF DEEP MODELS PARAMETERS WITH RESPECT TO DATA DISTRIBUTION

Shitala Prasad

Scientist, Visual Intelligence

Co-authors: Yiqun Li, Dongyun Lin, Dong Sheng, Oo Zaw Min

I2R / A*Star Singapore

Agenda

Introduction & Motivations	3-5 Page
Methodology	6-12 Page
Experimental Results	13-15 Page
Conclusion	16 Page

CREATING GROWTH, ENHANCING LIVES

Introduction and Motivation

The revolutionization of deep learning in field of computer vision has changed the way the visual learning takes place.

The performance of deep learning models are driven by various parameters but to tune all of them every time, for every given dataset, is a heuristic practice.

In this paper, unlike the common practice of decaying the learning rate, we propose a step-wise training strategy where the learning rate and the batch size are tuned based on the dataset size.

Introduction and Motivation

Deep learning has always such questions:

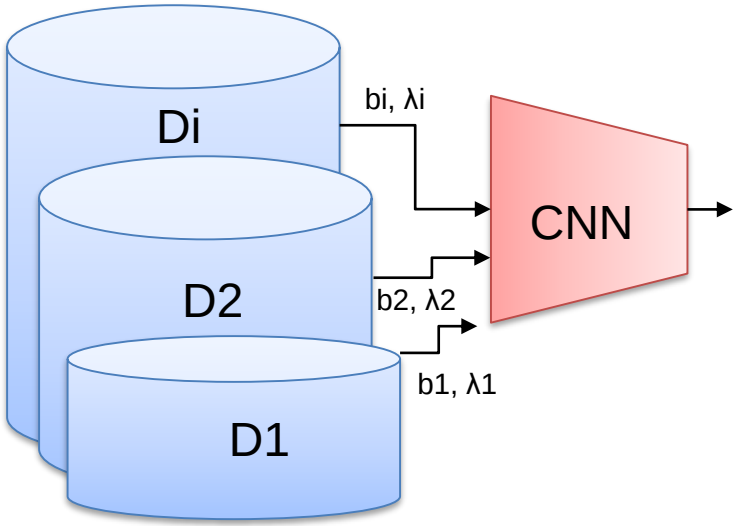
- *Can the learning performance be improved without additional data?*
- *Can accuracy be increased for the same architecture with the same dataset?*
- *Can data distribution strategy boosts the accuracy and reduces the training cost?*
- *Can it reduce network over-fitting?*

Introduction and Motivation

The proposed training strategy randomly splits the data into small subsets and incrementally train the classifier by updating the dataset with other subsets in a progressive step-wise fashion

- ✓ a simple **step-wise hyper-parameter tuning strategy** to boost the network classification performance **without using any additional data, consistently valued for several state-of-the-art** image classification **network** architectures and **reduces the overall training cost by ~40%**

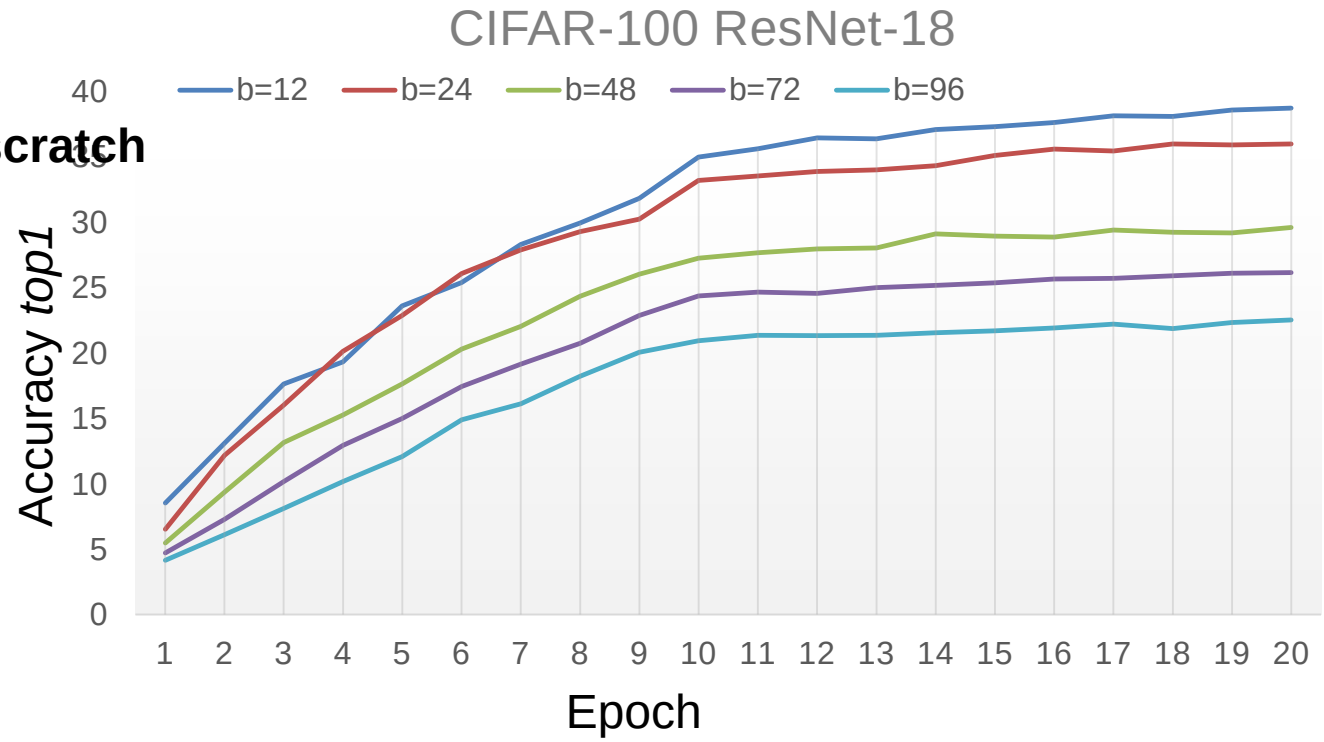
Methodology



Standard Baseline Approach

In fine-tuning deep models, mainly:

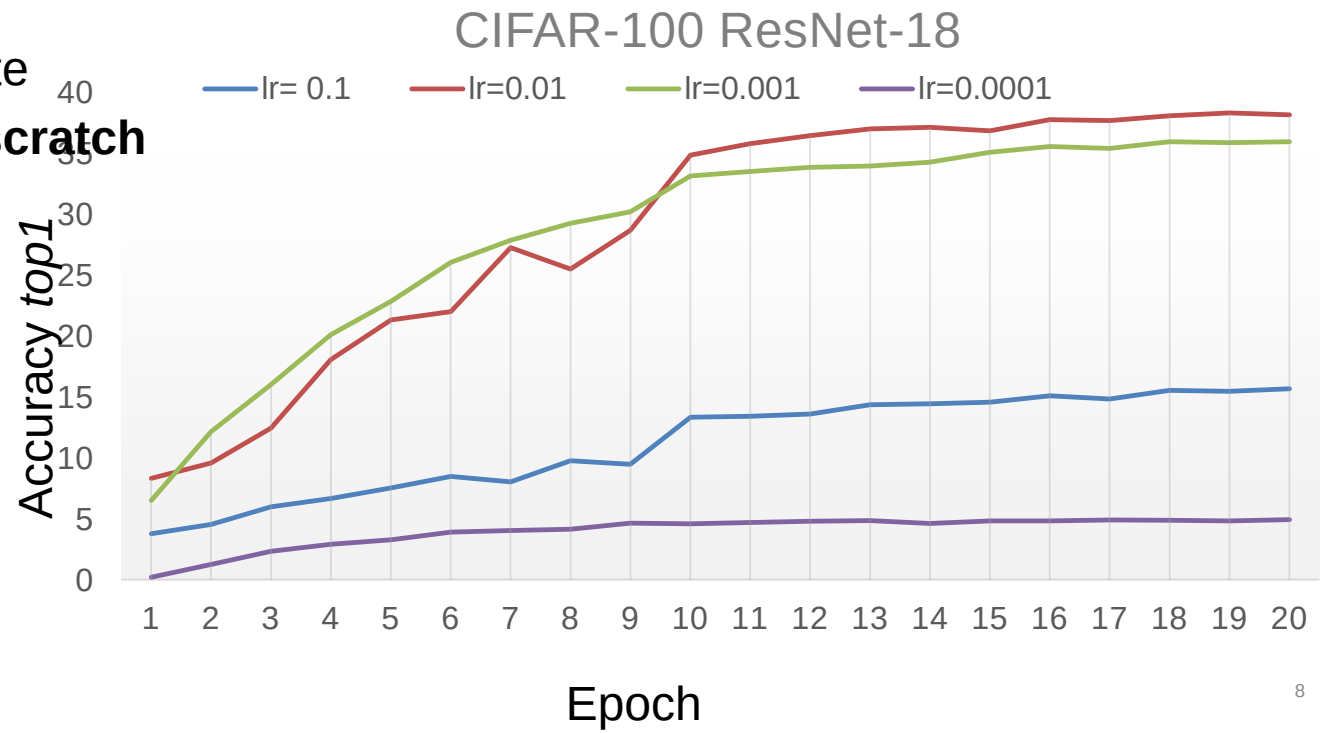
- Batch size
- Train from **scratch**



Standard Baseline Approach

In fine-tuning deep models, mainly:

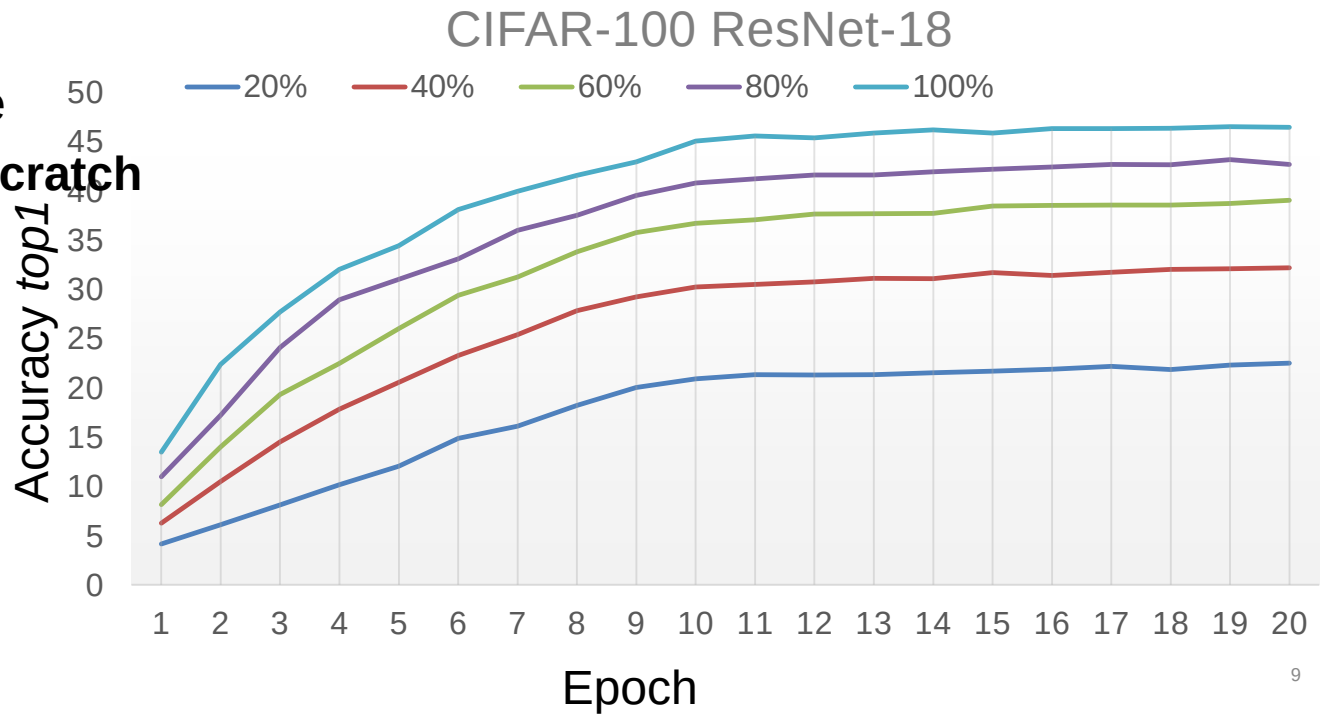
- Learning rate
- Train from **scratch**



Standard Baseline Approach

In fine-tuning deep models, other than batch size and learning rate, dataset size is also a key factor:

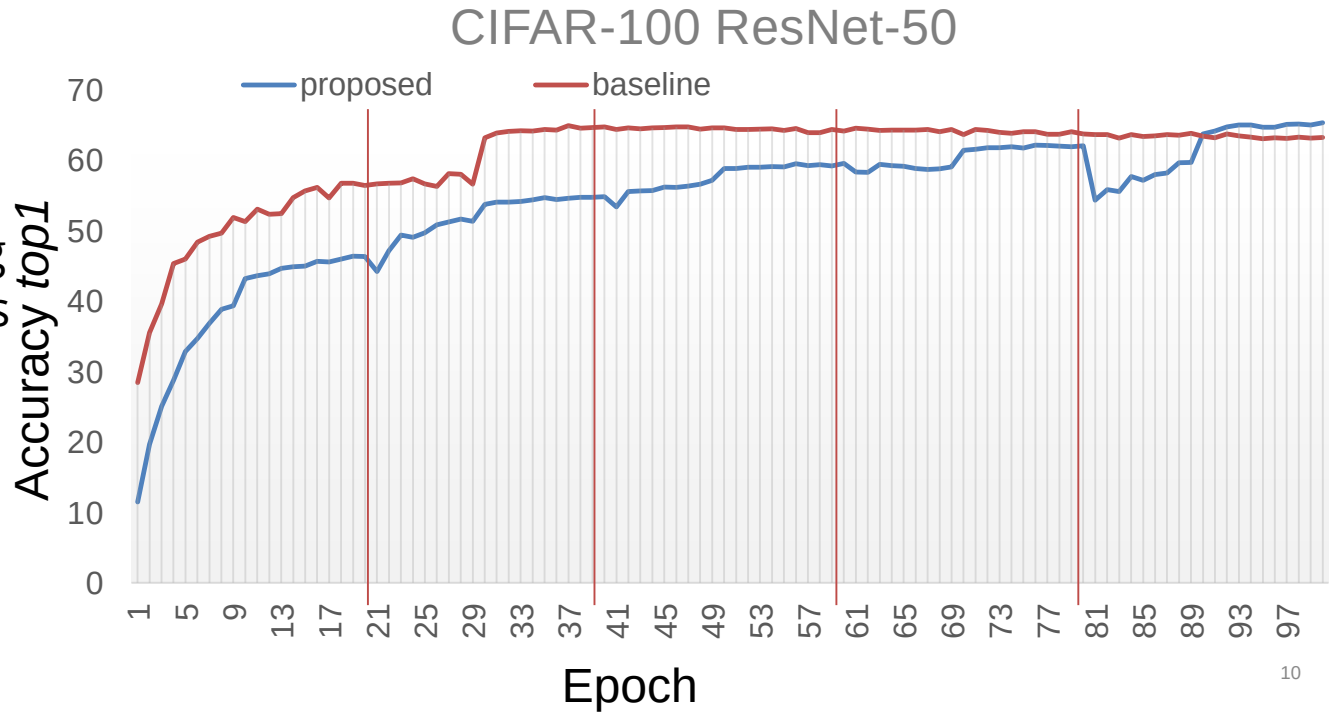
- Dataset size
- Train from **scratch**



Proposed Training Approach

Batch size + learning rate + dataset size:
Train from **scratch**

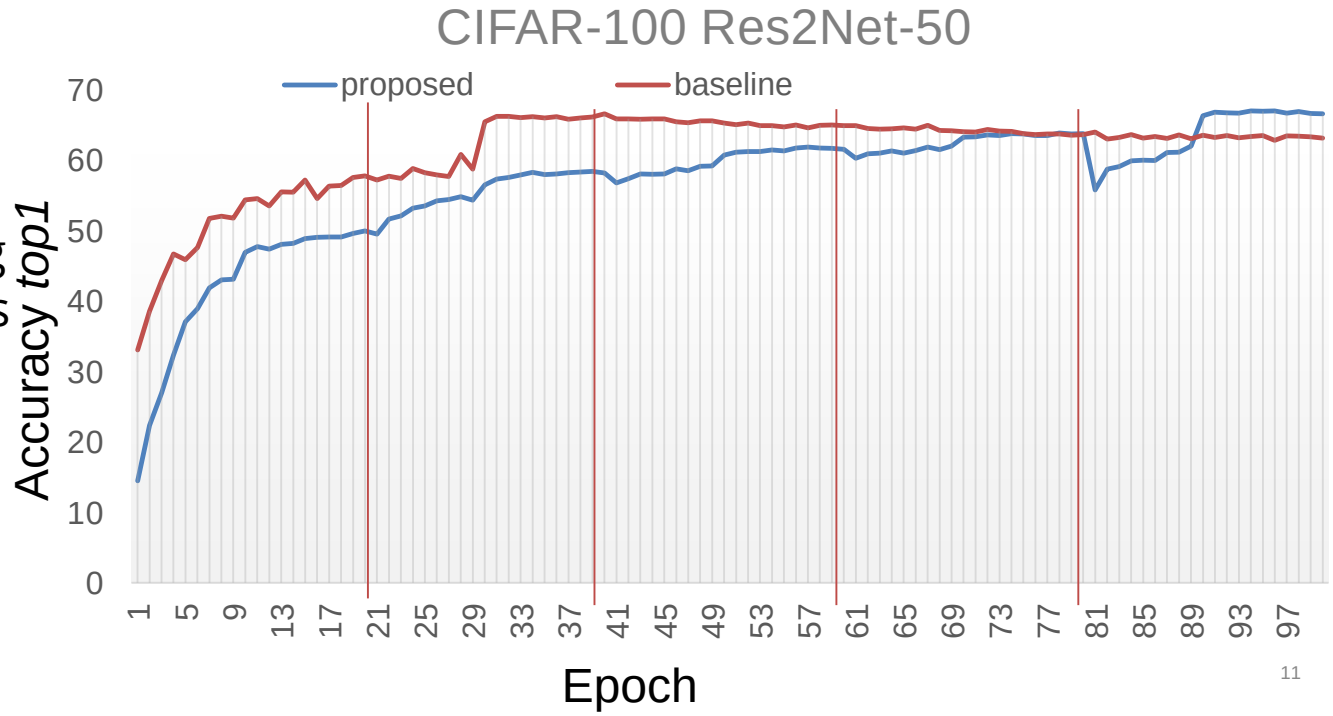
Our proposed step-wise training strategy improves accuracy and reduces computational cost.



Proposed Training Approach

Batch size + learning rate + dataset size:
Train from **scratch**

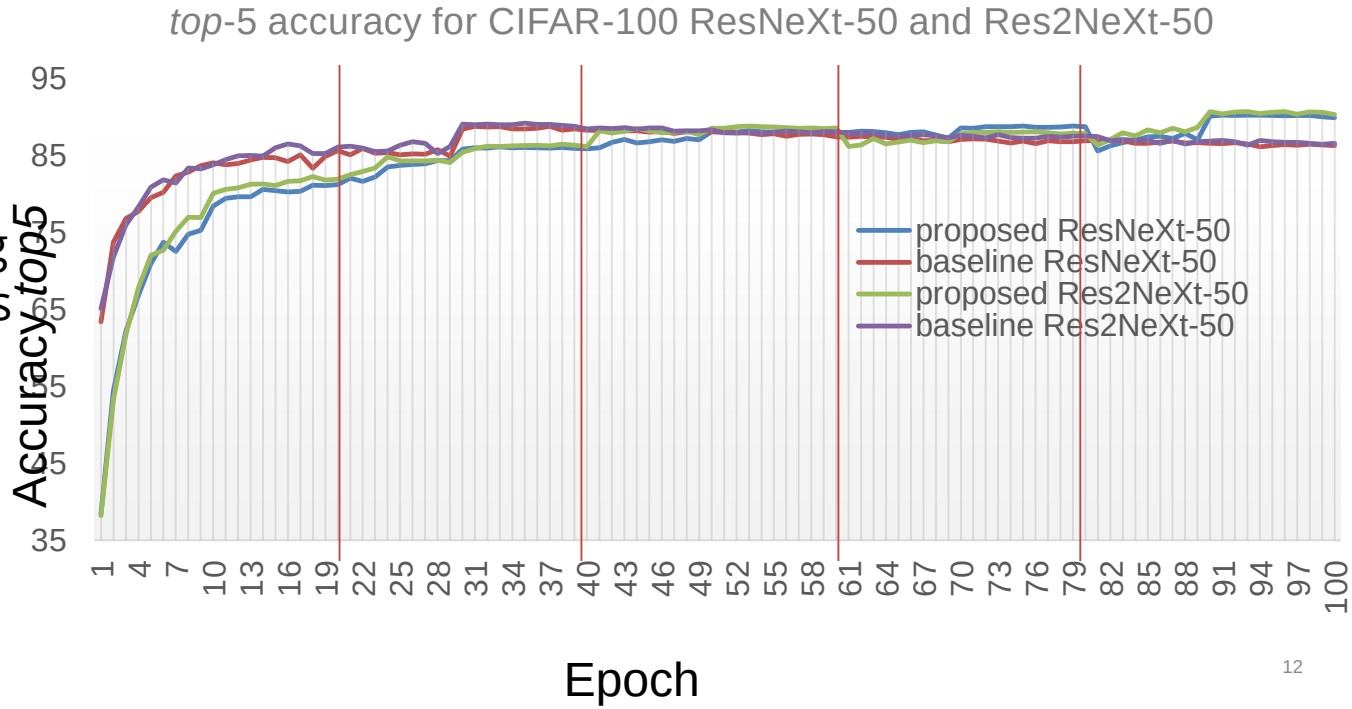
Our proposed step-wise training strategy improves accuracy and reduces computational cost.



Proposed Training Approach

Batch size + learning rate + dataset size:
Train from **scratch**

Our proposed step-wise training strategy improves accuracy and reduces computational cost.



Experimental Results

Datasets

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*

Food-101 Dataset	Res2Net-50				Epochs	Weight Shared
	Pre-trained=False		Pre-trained=True			
	top-1	*top-5*	*top-1*	*top-5*		
with set O (baseline)	53.5584	81.0297	85.7347	96.8832	100	✓
with set A	12.0832	32.2733	75.5881	93.2198	20	
with set B	15.7109	38.2733	80.2851	95.1723	20	
with set C	17.0257	40.8119	82.0371	95.8020	20	×
with set D	18.2574	42.7287	83.9247	96.6059	20	
with set O	37.8614	67.5485	86.2574	97.1406	20	
with set A	13.0218	33.4020	75.5881	93.2198		
$\xrightarrow{2b}$ B	21.8970	48.4634	78.7247	94.3842	20 each	✓
$\xrightarrow{3b}$ C	30.2851	59.0416	80.6337	95.0416		
$\xrightarrow{4b}$ D	38.2772	66.9782	81.8614	95.5762		
$\xrightarrow{b}$ O	**54.9901**	**81.5010**	**86.7089**	**97.2554**		

CREATING GROWTH, ENHANCING LIVES

Datasets

Clubbed together: 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

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**	



COVID-19 mask-nomask dataset. Details: train (1050 mask, 1459 no-mask) and test (264 mask, 375 no-mask) images.

Take back

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

CREATING GROWTH, ENHANCING LIVES

Institute for Infocomm Research
I²R

THANK YOU

www.a-star.edu.sg