TAPAS: Train-less accuracy predictor for architecture searches

R. Istrate, F. Scheidegger, G. Mariani, P. Chatzidoukas, C. Bekas, C. Malossi
IBM Research – Zurich

June 5th, 2018 | Weekly Speaker Series : Measuring AI Progress with Cognitive Opentech

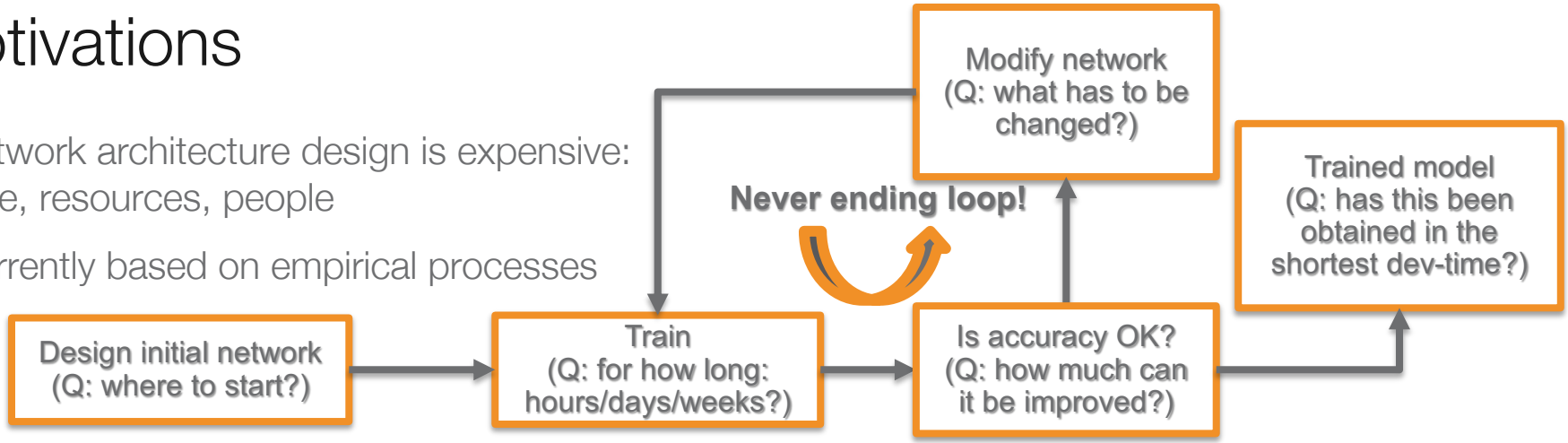
IBM

Table of content

- Motivations and state-of-the-art
- Methodology:
  - Train-less Accuracy Predictor for Architecture Search
  - Dataset characterization
  - Lifelong database of experiments
  - Train-less accuracy predictor
- Results
  - Accuracy prediction comparison with Peephole and Learning Curves Extrapolation
  - Performance comparison with Google Large Scale Evolution
- Conclusions and future works

Motivations

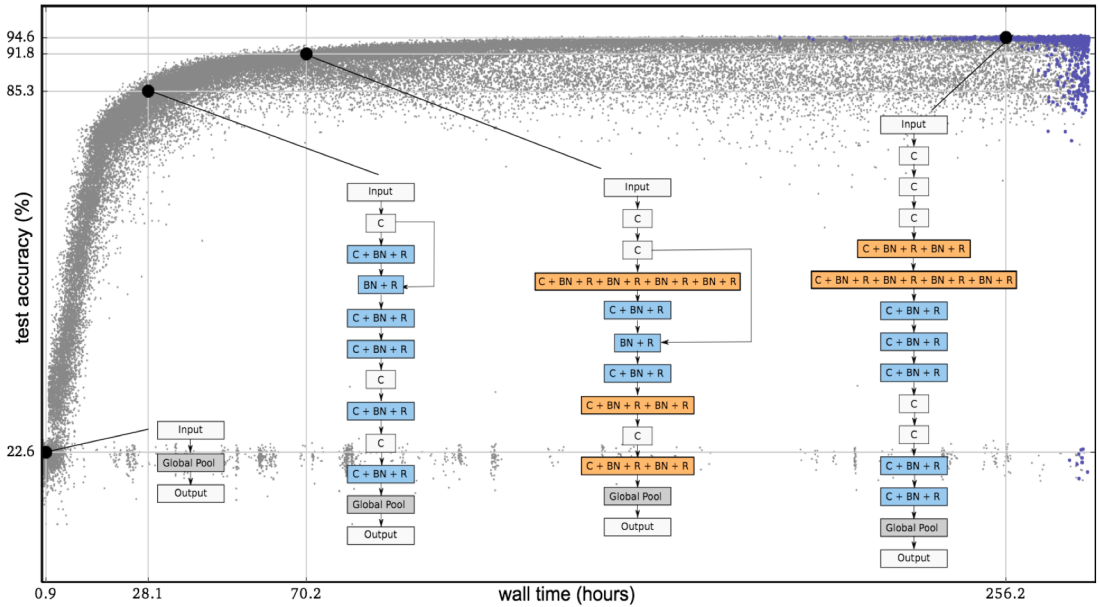
- Network architecture design is expensive: time, resources, people
- Currently based on empirical processes



What it takes to automatize NN architecture discovery:

1. Preserve and re-use knowledge learned from previously experiments and models
2. Predict performance of architectures before training
3. Dynamically adapt to complexity of input dataset
4. Smart algorithms that perform large scale search, minimizing training

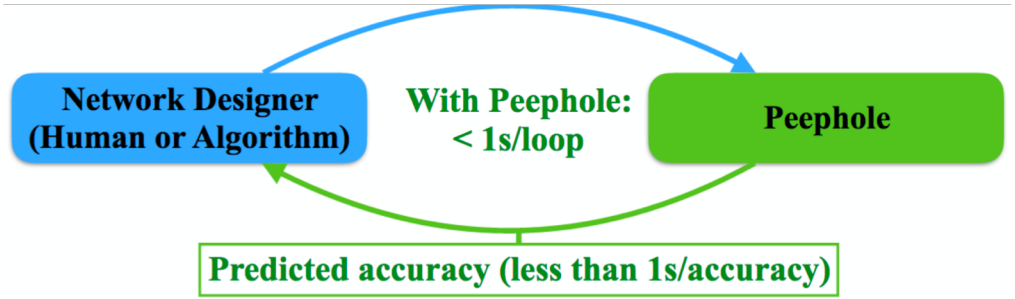
State-of-the-art: Google large scale evolution approach

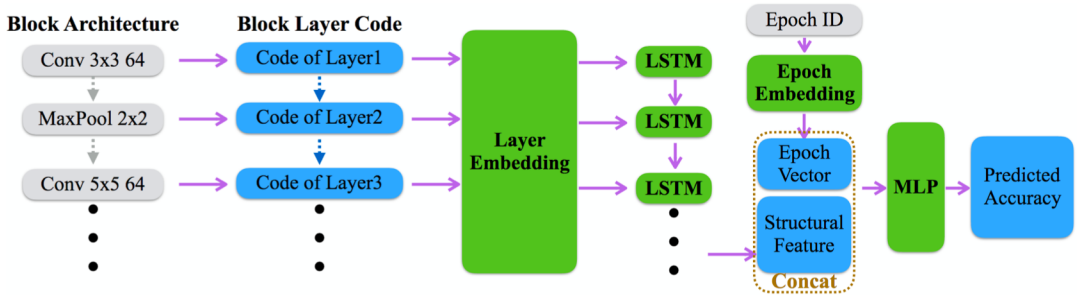


- ✓ Mutation algorithm to generate networks
- ✓ Architecture search over very large spaces
- − Expensive: 1000 individuals, 250 workers, 10 days of experiments for CIFAR-10 network
- − Would not scale when used on larger datasets

REF: Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc Le, Alex Kurakin, Large-Scale Evolution of Image Classifiers, 2017

State-of-the-art: Peephole, prediction before training



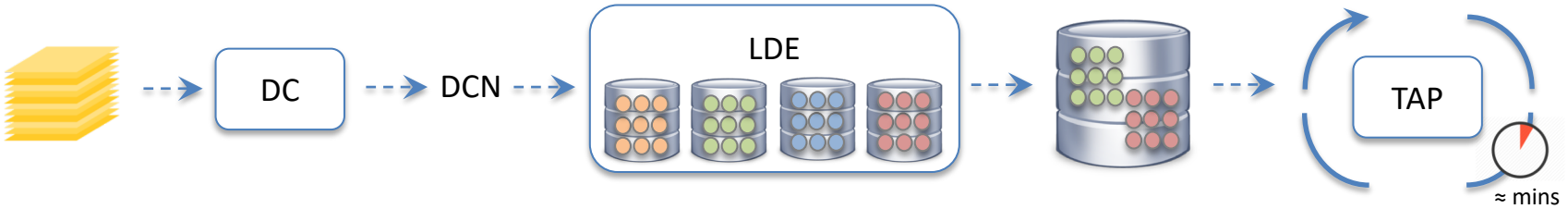


- ✓ Cheap: evaluates network performance without training
- ✓ Architecture search over large spaces at no cost

- ▬ Dataset specific
- ▬ Requires over 1000 networks trained on same dataset
- ▬ Generated networks have fixed convolutional structure

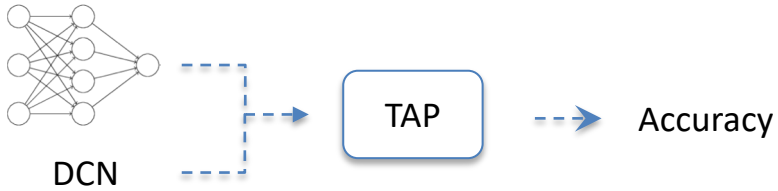
REF: Deng, Boyang, Junjie Yan, and Dahua Lin. "Peephole: Predicting network performance before training." arXiv preprint arXiv:1712.03351 (2017).

Train-Less Accuracy Predictor for Architecture Search



Three main components:

- Dataset Characterization (DC): rank dataset by difficulty
- Lifelong Database of Experiments (LDE): store experiments and grows over time
- Train-less Accuracy Predictor (TAP): predict performance of networks in real-time

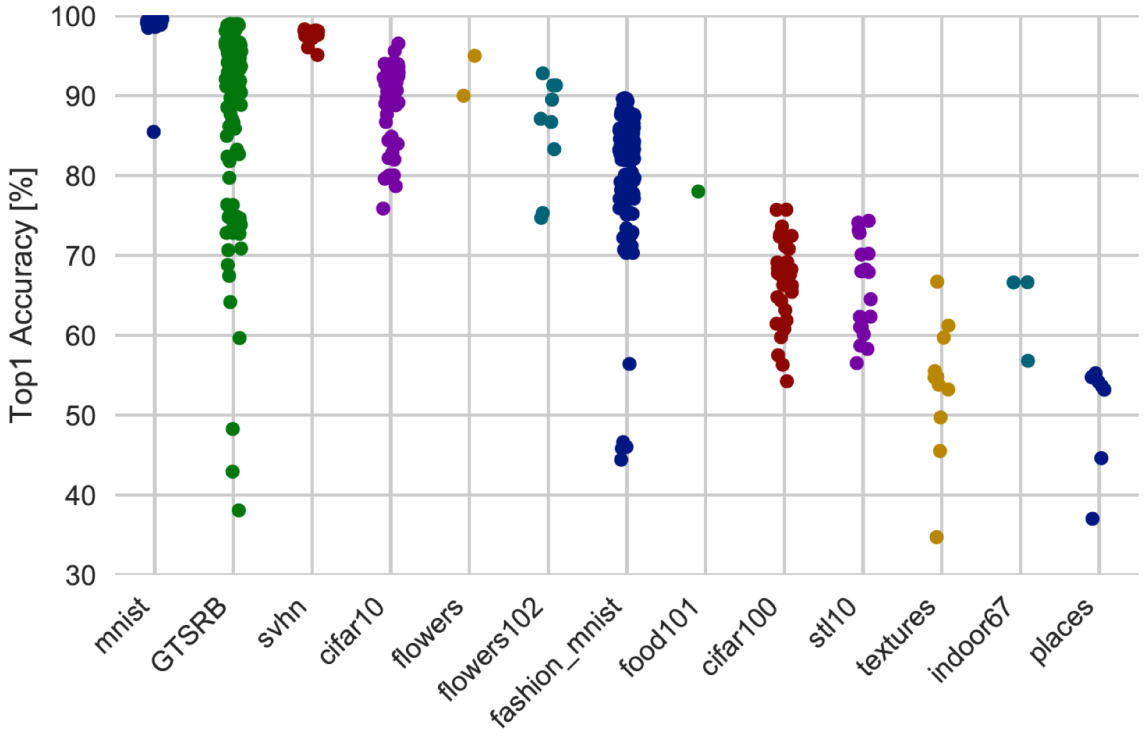


Dataset characterization from literature experiments

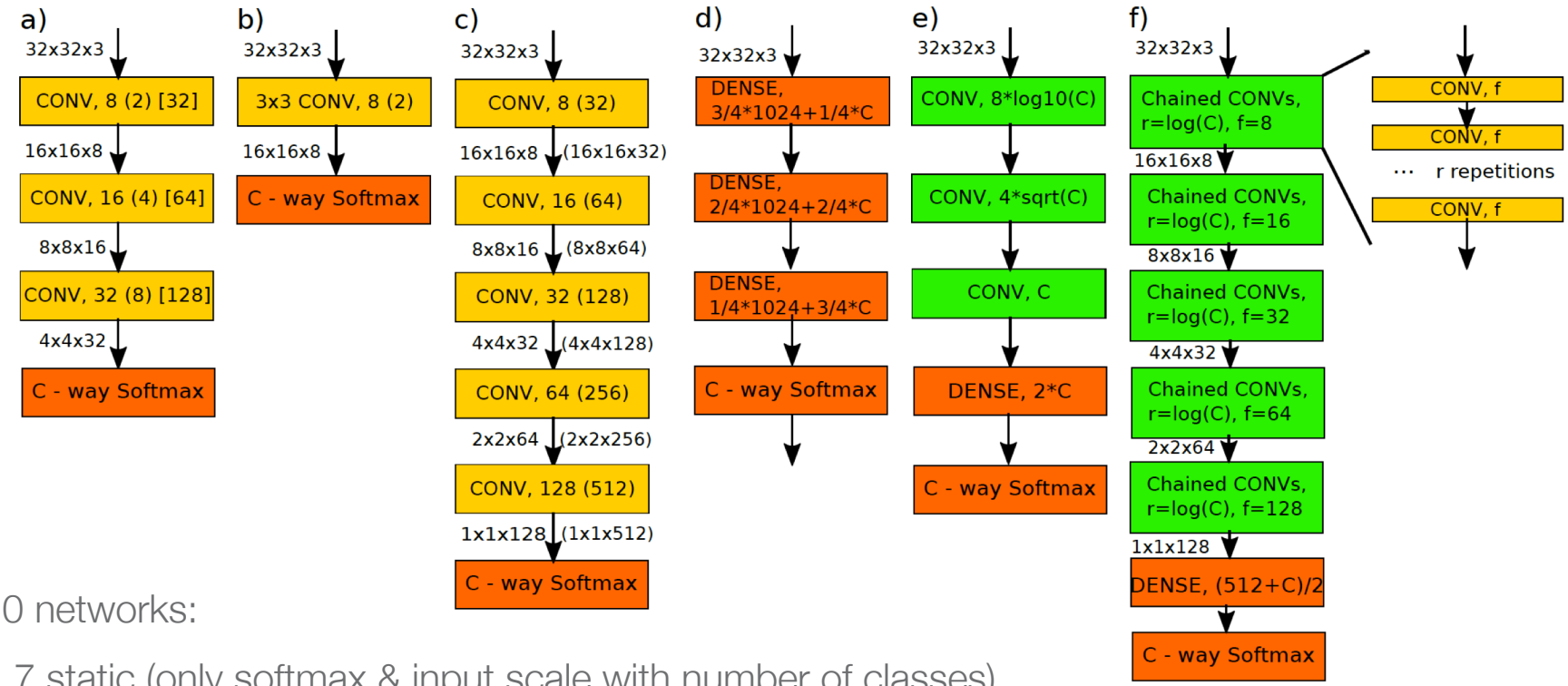
- Literature study on 14 datasets
- Some dataset are part of large competitions (more points)
- Some results are obtained with transfer learning

Key observation:

- Large variability per dataset
- There is a clear ranking



ProbeNets: networks for dataset characterization

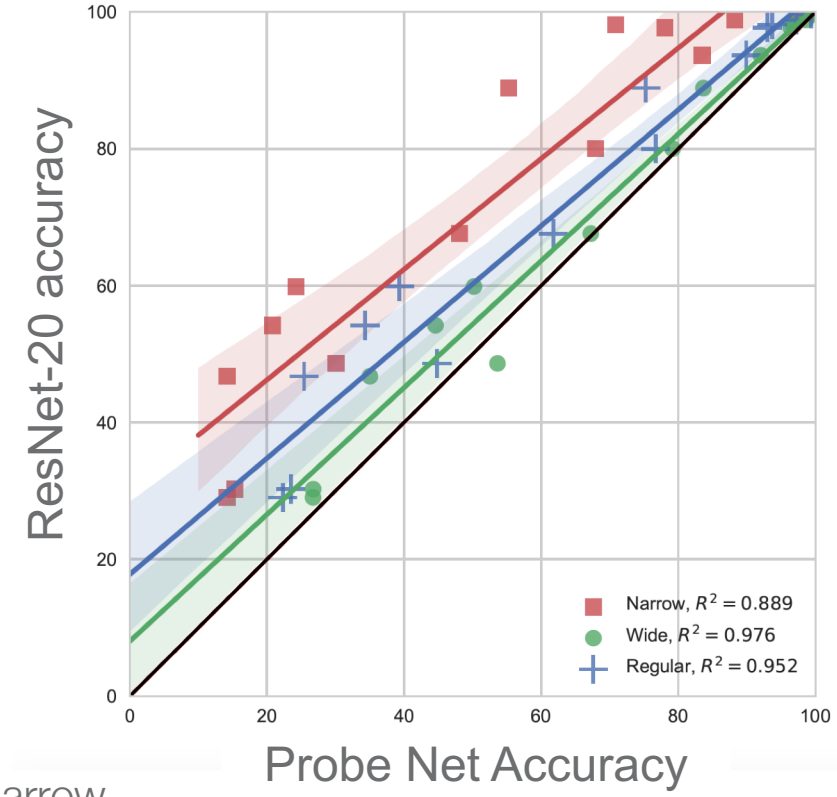


10 networks:

- 7 static (only softmax & input scale with number of classes)
- 3 dynamic (topology scales with number of classes)

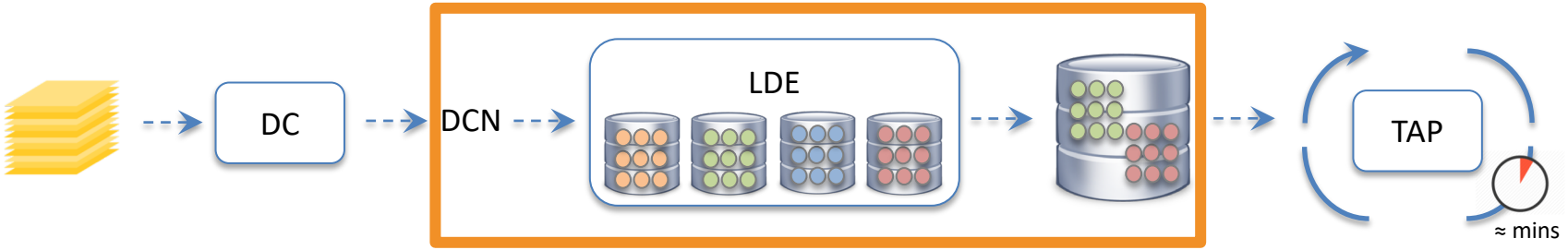
ProbeNets: cheap, quick, and accurate

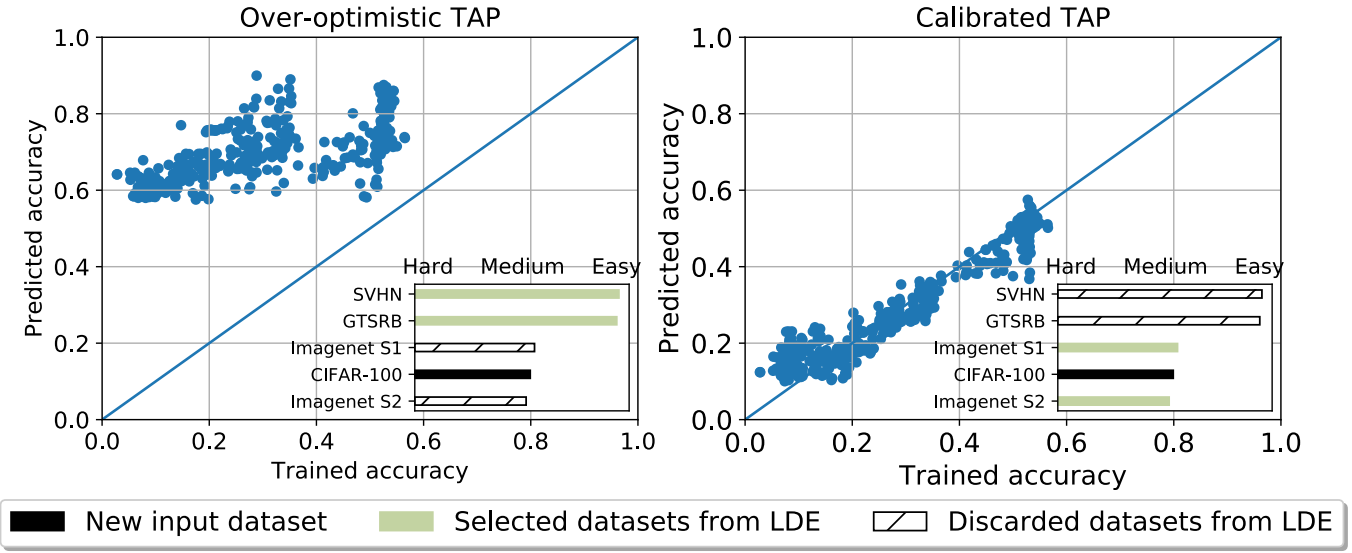
Probe Net	$C = 10$		$C = 100$	
	OPs	Weights	OPs	Weights
Regular	0.81M	11K	0.86M	57.5K
Narrow	0.09M	2K	0.10M	13K
Wide	10.34M	114K	10.52M	299K
Shallow	0.24M	21K	0.42M	205K
Shallow norm.	0.06M	5K	0.10M	51K
Deep	1.40M	100K	1.41M	112K
Deep norm.	19.76M	1576K	19.81M	1622K
MLPs	2.90M	2908K	3.10M	3107K
Kernel depth	0.53M	6K	4.56M	384K
Length	1.41M	118K	4.39M	338K
ResNet-20	40.55M	271K	40.56M	277K



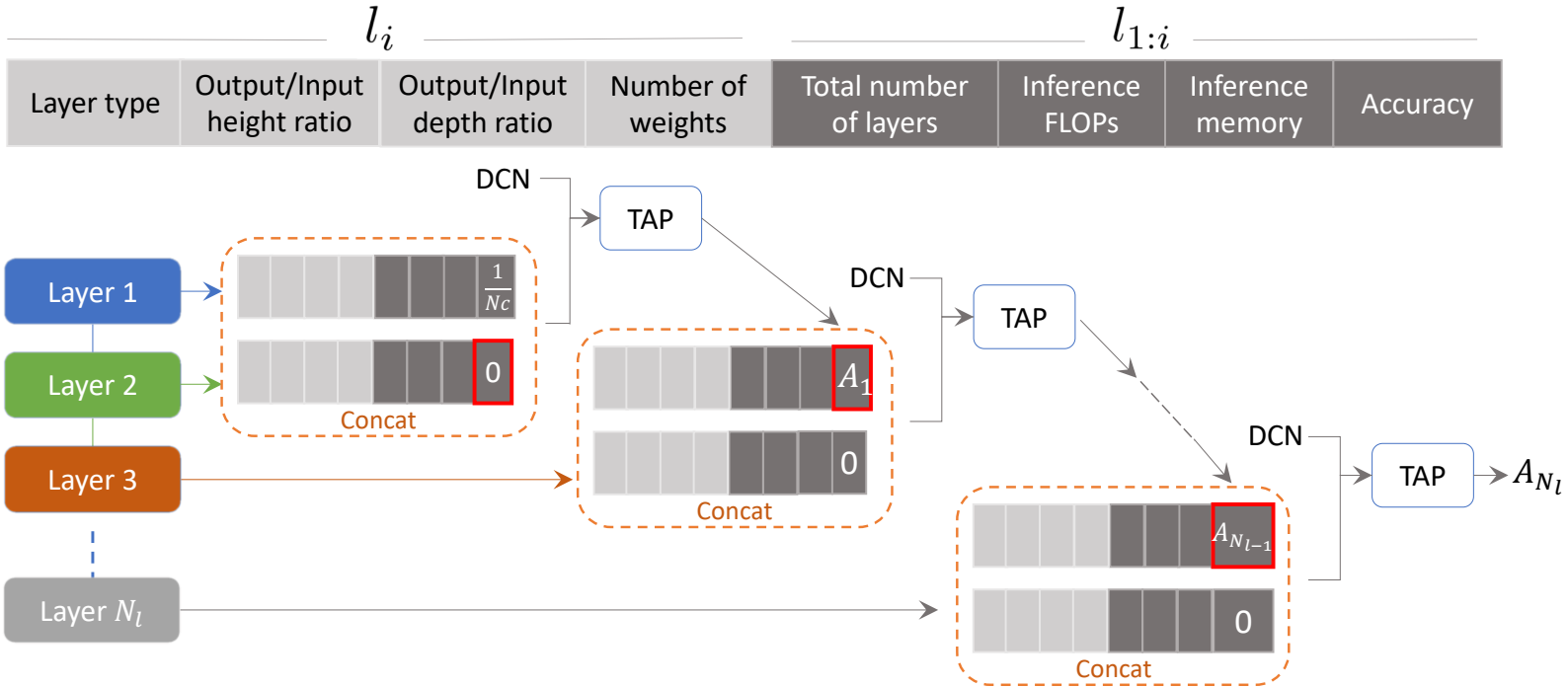
- Good performance compared to ResNet-20
- Cost reduced up to 50x for Regular, and 400x for Narrow

Lifelong database of experiments (LDE)





TAP layer-by-layer workflow



- Network prediction build on the intermediate evaluation of all sub-networks
- Incremental training approach used to populate LDE and obtain intermediate accuracies

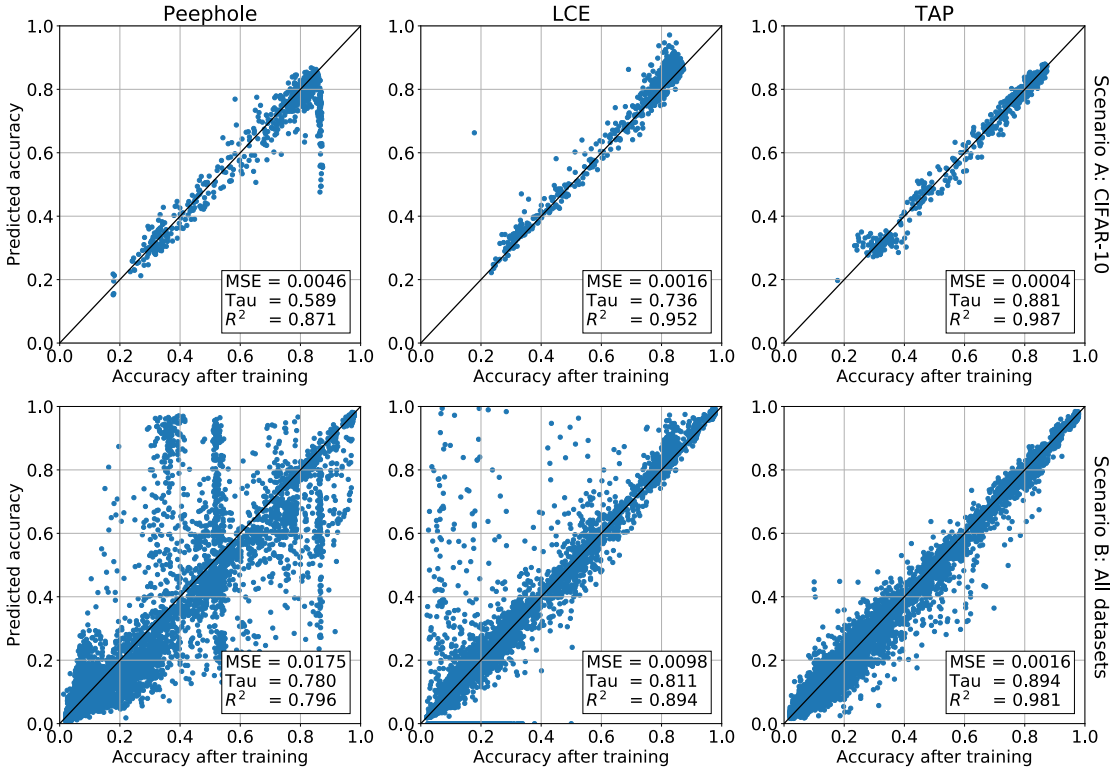
Accuracy prediction comparison with Peephole and LCE

Scenario A (first row):

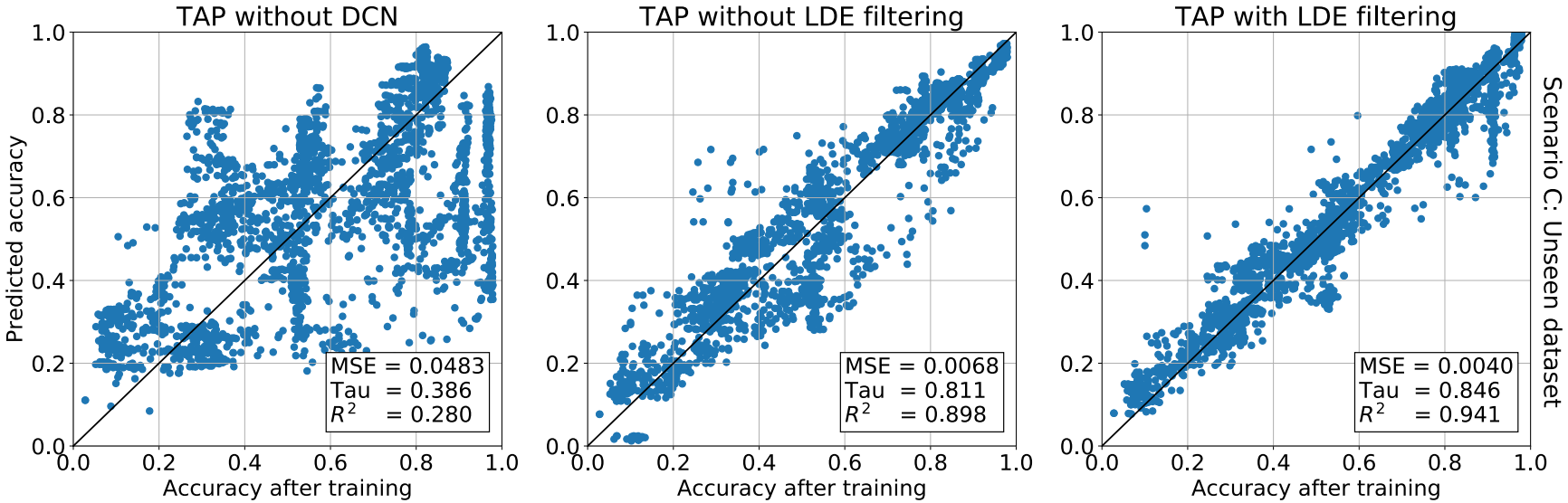
- Input dataset: CIFAR-10
- 1 dataset in LDE: CIFAR-10
- 90% cross-validation

Scenario B (second row):

- Input dataset: CIFAR-10
- 20 dataset in LDE (including CIFAR-10)
- 90% cross-validation



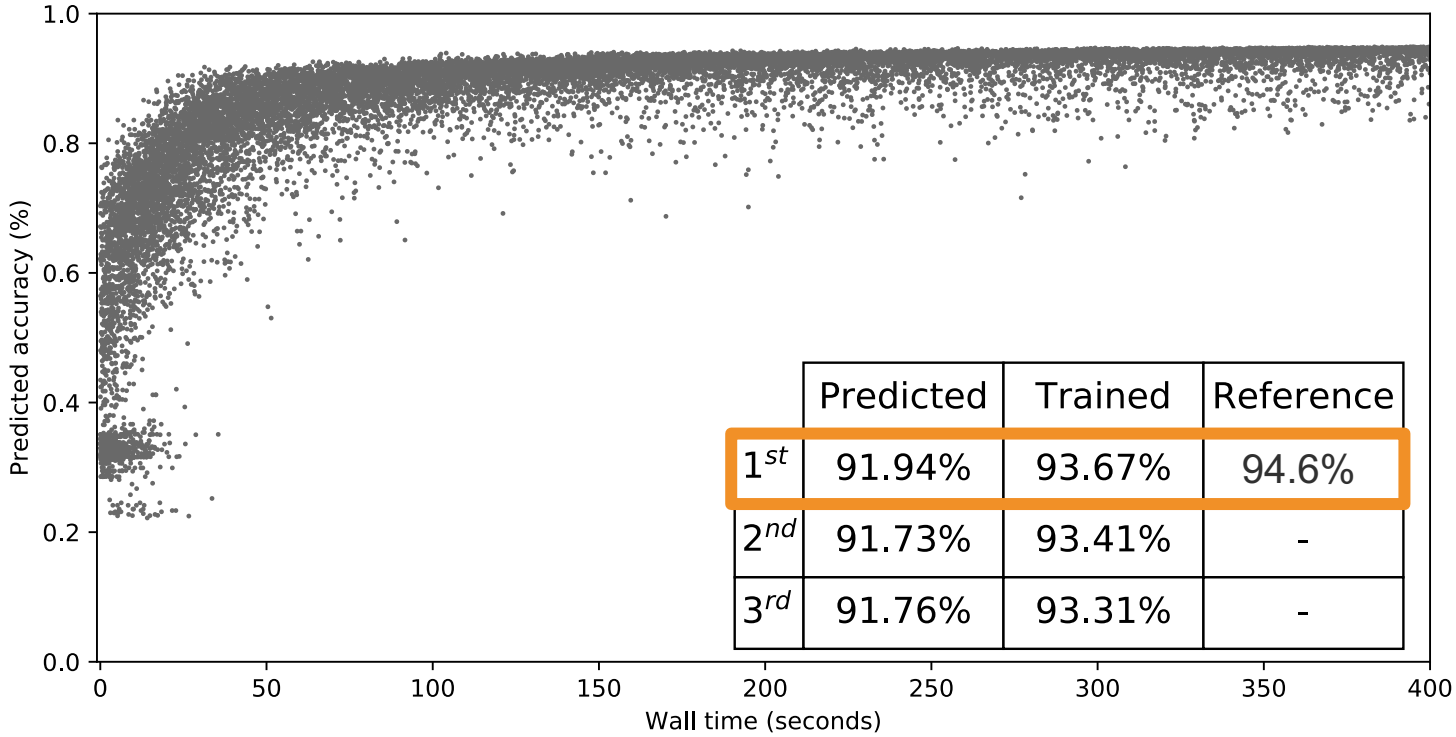
TAP on unseen datasets: effect of LDE filtering and DCN



Scenario C:

- 11 leave-one-out cross-validations
- Input dataset: one of the eleven available
- 19 dataset in LDE (10 real not including the input one + 9 sub-sampled from Imagenet)

Performance comparison: Google large scale evolution



	Predicted	Trained	Reference
1^{st}	91.94%	93.67%	94.6%
2^{nd}	91.73%	93.41%	-
3^{rd}	91.76%	93.31%	-

Comparison of resources utilization:

- Google: 256h on 250 workers (many GPUs)
- TAPAS 400 s on 1 GPU (without training)

Conclusions and future works

Framework features summary:

- Dataset agnostic
- Leverage experience from previously trained networks (continuous learning)
- Real-time prediction that scales with resources
- Can be used in many architecture search algorithms

Near future works:

- Extension to other type of DL problems: object detection, scene labeling NLP, etc.
- Extension to other type of DL dataset: video, text, audio signal, etc.
- Full framework for architecture search on IBM Cloud (under development)
- Full framework delivery on OpenPOWER

References

- R. Istrate, F. Scheidegger, G. Mariani, P. Chatzidoukas, D. Nikolopoulos, C. Bekas, A. C. I. Malossi. TAPAS: Train-less Accuracy Predictor for Architecture Search, In Preparation, 2018
- R. Istrate, A. C. I. Malossi, C. Bekas and D. Nikolopoulos. Incremental Training of Deep Convolutional Neural Networks. In Proceedings of AutoML at ECML-PKDD, Skopje, Macedonia, https://arxiv.org/abs/1803.10232, 2017
- F. Scheidegger, R. Istrate, G. Mariani, L. Benini, C. Bekas, A. C. I. Malossi. Efficient Image Dataset Classification Difficulty Estimation for Predicting Deep-Learning Accuracy, https://arxiv.org/abs/1803.09588, 2018
- G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, C. Malossi. BAGAN: Data Augmentation with Balancing GAN, https://arxiv.org/abs/1803.09655, 2018

THANK YOU