ANDANTE

AI FOR NEW DEVICES AND TECHNOLOGIES AT THE EDGE

Standard benchmarking for machine learning

Siavash A. Bigdeli

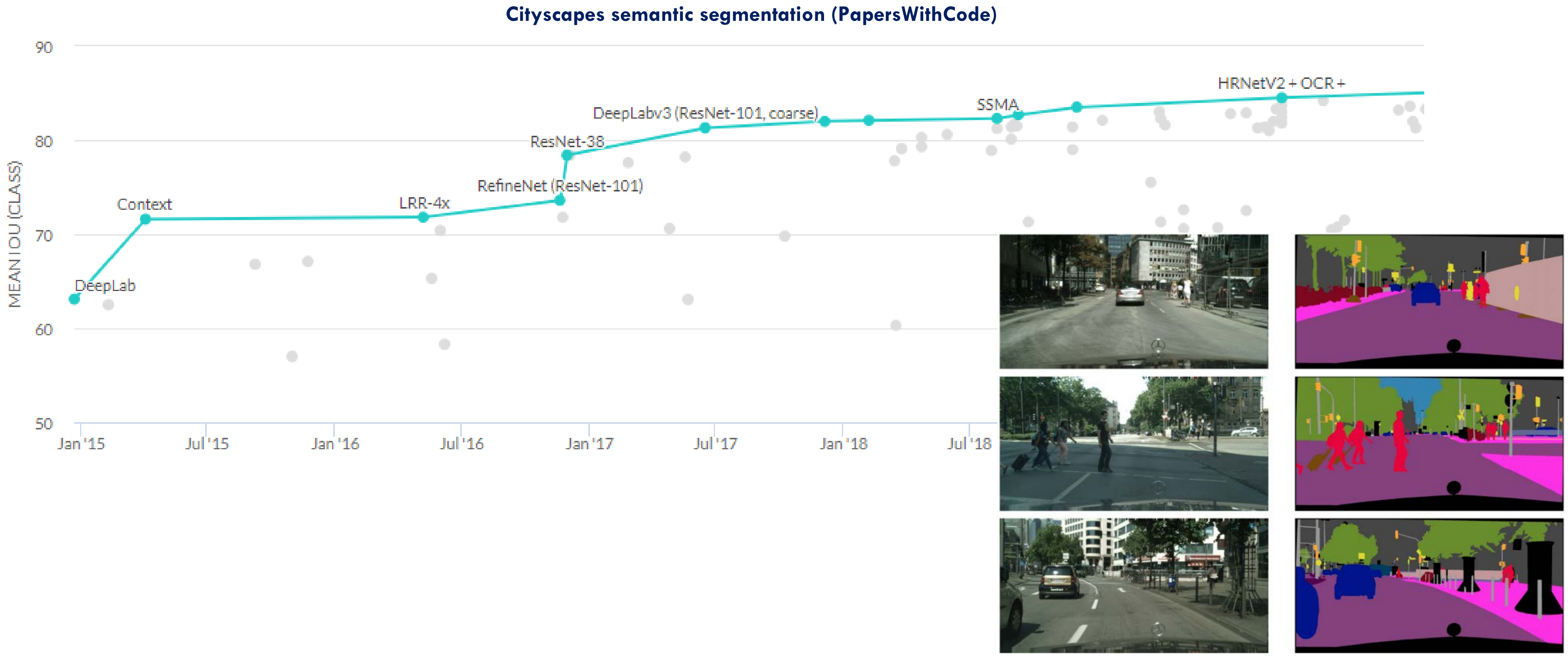
SAB@csem.ch

Benchmarks:

Key Performance Indicators for ML

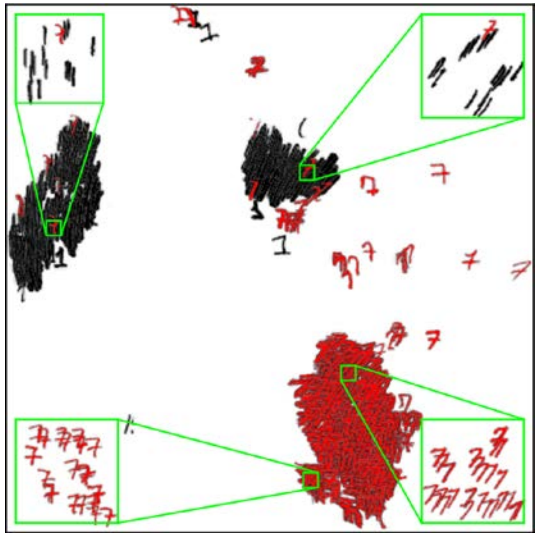
ACCURACY

- Top-k, AUC, IoU, confidence scales



ACCURACY

- Human level performance



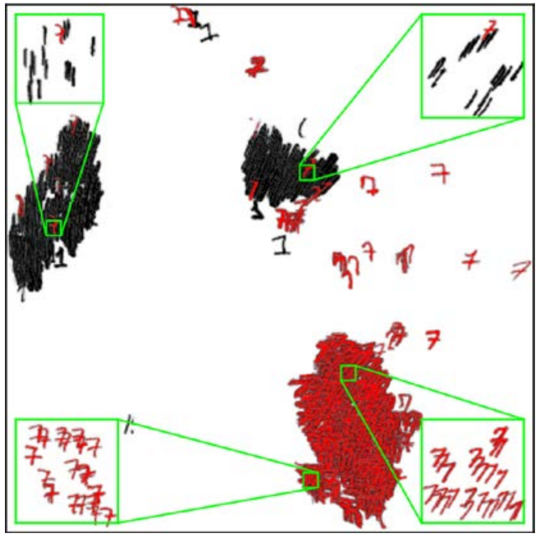
Portenier et al., VISIGRAPP 2018

ACCURACY

- Human level performance
- Bayes optimality bound [Theisen et al., arXiv 2021]

Corpus	#classes	#samples	NLL	Bayes Error	SOTA Error [29]
MNIST	10	60,000	8.00e2	1.07e-4	1.6e-3 [3]
EMNIST (digits)	10	280,000	8.61e2	1.21e-3	5.7e-3 [27]
SVHN	10	73,257	4.65e3	7.58e-3	9.9e-3 [3]
Kuzushiji-MNIST	10	60,000	1.37e3	8.03e-3	6.6e-3 [11]
CIFAR-10	10	50,000	7.43e3	2.46e-2	3e-3 [10]
Fashion-MNIST	10	60,000	1.75e3	3.36e-2	3.09e-2 [32]
EMNIST (letters)	26	145,600	9.15e2	4.37e-2	4.12e-2 [15]
CIFAR-100	100	50,000	7.48e3	4.59e-2	3.92e-2 [10]
EMNIST (balanced)	47	131,600	9.45e2	9.47e-2	8.95e-2 [15]
EMNIST (bymerge)	47	814,255	8.53e2	1.00e-1	1.90e-1 [5]
EMNIST (byclass)	62	814,255	8.76e2	1.64e-1	2.40e-1 [5]

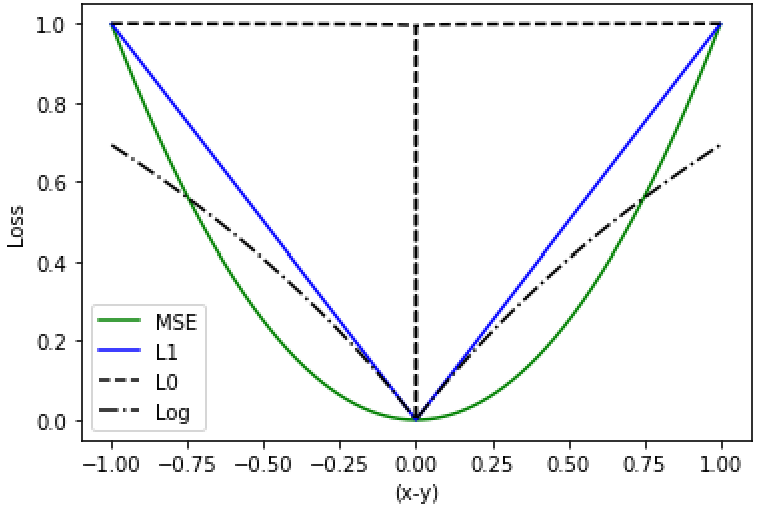
Theisen et al., arXiv 2021



Portenier et al., VISIGRAPP 2018

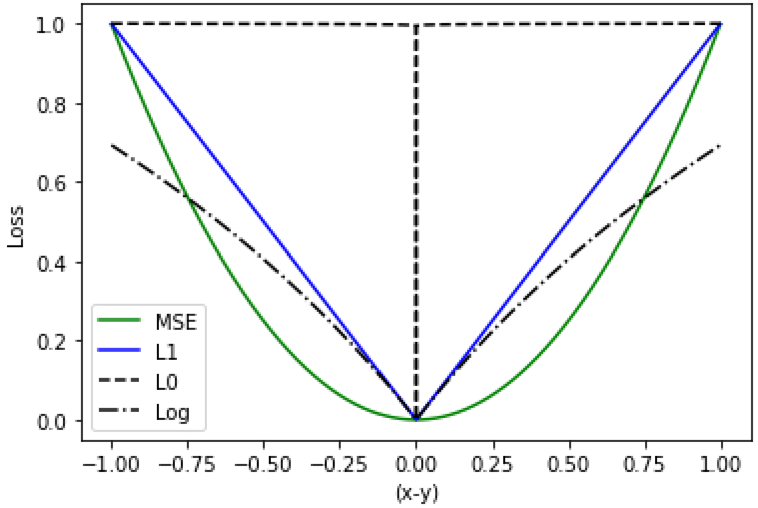
PRECISION

- Standard
  - MSE, L1, PSNR, SSIM, …

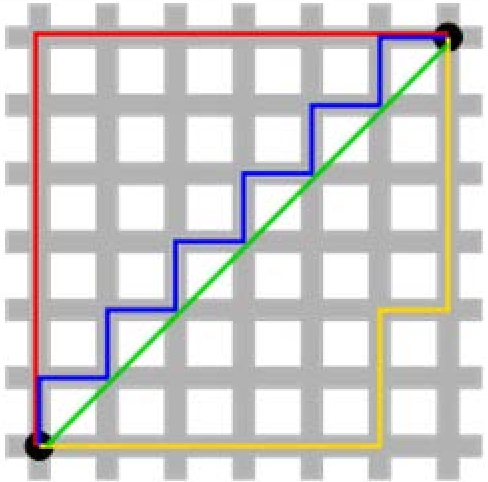


PRECISION

- Standard
  - MSE, L1, PSNR, SSIM, …
  - Similarity takes different meaning



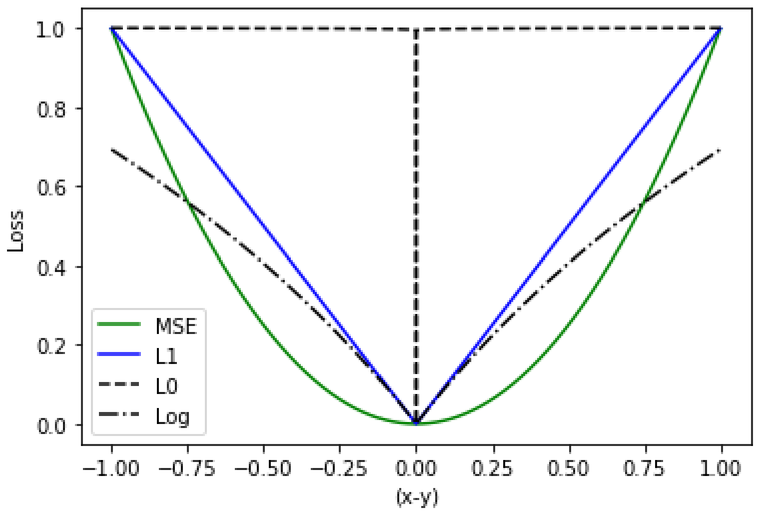
L2, L1, L1, L1

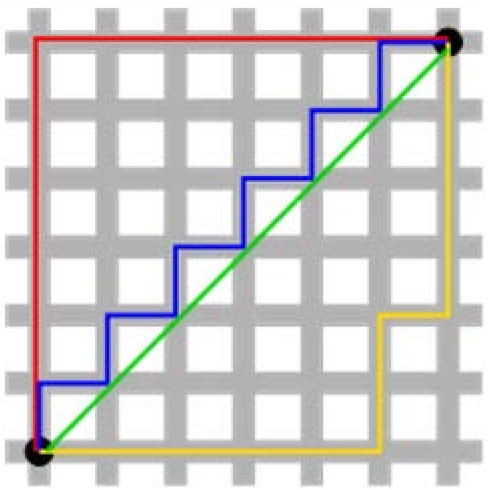


http://www.chioka.in/differences-between-l1-and-l2-as-loss-function-and-regularization/

PRECISION

- Standard
  - MSE, L1, PSNR, SSIM, …
  - Similarity takes different meaning
- Data driven
  - Perceptual, Gram loss, …
  - FID, KDE, KL-divergence, Average Log-likelihood





L2, L1, L1, L1

http://www.chioka.in/differences-between-l1-and-l2-as-loss-function-and-regularization/

PRECISION

- NVIDIA Benchmark for loss functions
  - 6 training / 9 test losses
  - Denoising+demosaicing
  - JPEG de-blocking
  - Super-resolution

Denoising + demosaicking			Training cost function					
Image quality metric	Noisy	$BM3D$	ℓ_2	ℓ_1	SSIM_5	SSIM_9	MS-SSIM	Mix
$1000 \cdot \ell_2$	1.65	0.45	0.56	0.43	0.58	0.61	0.55	**0.41**
PSNR	28.24	34.05	33.18	34.42	33.15	32.98	33.29	**34.61**
$1000 \cdot \ell_1$	27.36	14.14	15.90	13.47	15.90	16.33	15.99	**13.19**
SSIM	0.8075	0.9479	0.9346	0.9535	0.9500	0.9495	0.9536	**0.9564**
MS-SSIM	0.8965	0.9719	0.9636	0.9745	0.9721	0.9718	0.9741	**0.9757**
IW-SSIM	0.8673	0.9597	0.9473	0.9619	0.9587	0.9582	0.9617	**0.9636**
GMSD	0.1229	0.0441	0.0490	0.0434	0.0452	0.0467	0.0437	**0.0401**
FSIM	0.9439	0.9744	0.9716	0.9775	0.9764	0.9759	0.9782	**0.9795**
FSIM_c	0.9381	0.9737	0.9706	0.9767	0.9752	0.9746	0.9769	**0.9788**

Zhao et al., TCI 2015

GENERALIZATION

- Overfitting and memorization

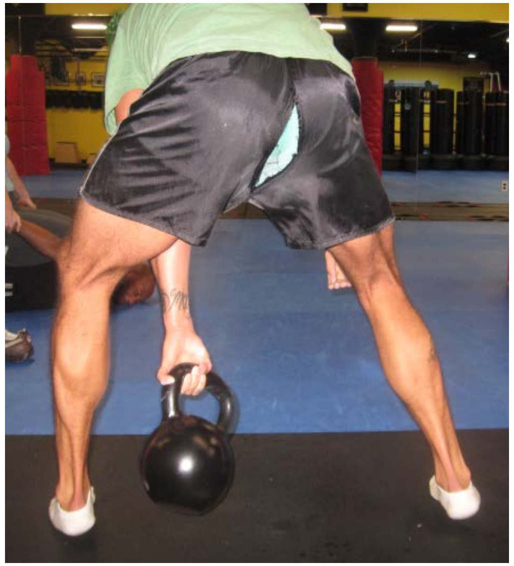
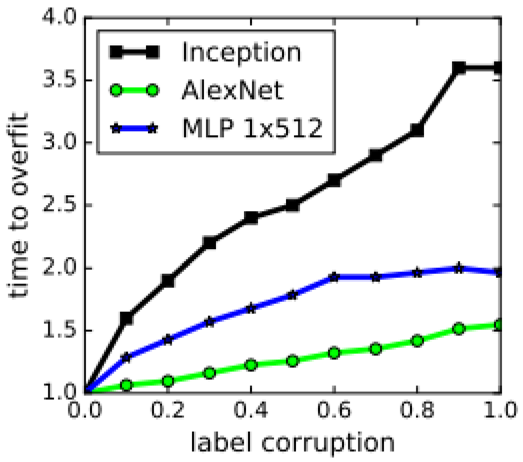


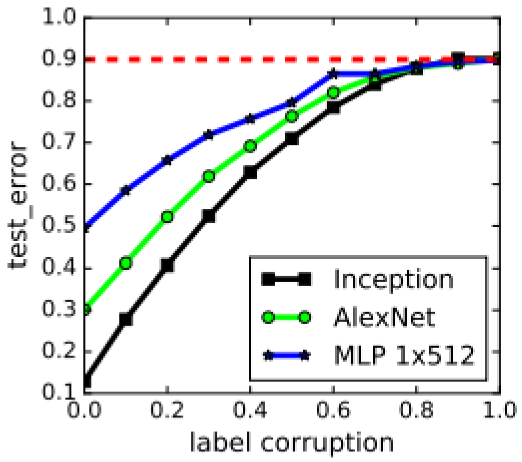
Image from: Diary of the Unexpected Housewife: Oh Damn...I Ripped My Pants (wheredmyjobgo.blogspot.com)

GENERALIZATION

- Overfitting and memorization
- Learning random labels vs. capacity [Zhang et al., ICLR 2017]



convergence slowdown



generalization error growth

GENERALIZATION

- Overfitting and memorization
- Learning random labels vs. capacity
- Gap between the test and train / Bias

Natural Language Inference

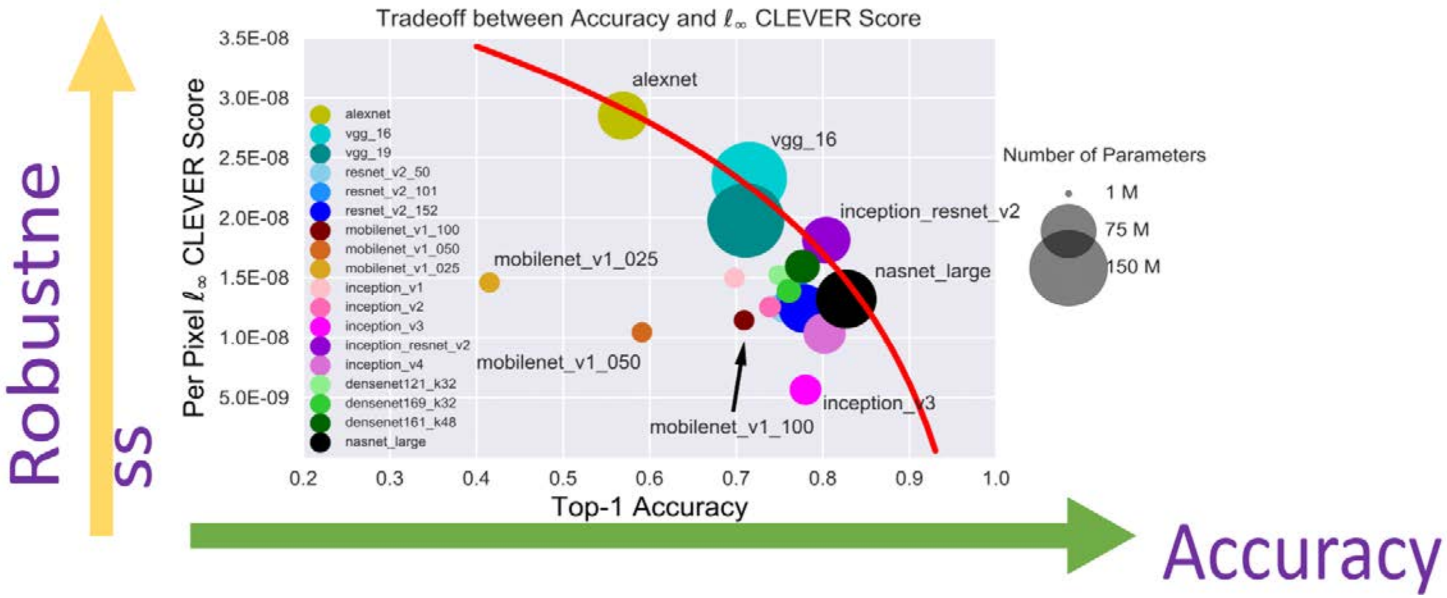
Premise	Label	Hypothesis
An older and younger man smiling.	neutral	Two men are smiling and laughing at the cats playing on the floor.

Train data	Test data	Test accuracy	Δ	Model
SNLI	**SNLI**	**86.1**		**BiLSTM-max** (our baseline)
SNLI	**SNLI**	**86.6**		**HBMP** (Talman et al., 2018)
SNLI	**SNLI**	**88.0**		**ESIM** (Chen et al., 2017)
SNLI	**SNLI**	**88.6**		**KIM** (Chen et al., 2018)
SNLI	**SNLI**	**88.6**		**ESIM + ELMo** (Peters et al., 2018)
SNLI	**SNLI**	**90.4**		**BERT-base** (Devlin et al., 2019)
SNLI	MultiNLI-m	55.7*	-30.4	BiLSTM-max
SNLI	MultiNLI-m	56.3*	-30.3	HBMP
SNLI	MultiNLI-m	59.2*	-28.8	ESIM
SNLI	MultiNLI-m	61.7*	-26.9	KIM
SNLI	MultiNLI-m	64.2*	-24.4	ESIM + ELMo
SNLI	MultiNLI-m	75.5*	-14.9	BERT-base
SNLI	SICK	54.5	-31.6	BiLSTM-max
SNLI	SICK	53.1	-33.5	HBMP
SNLI	SICK	54.3	-33.7	ESIM
SNLI	SICK	55.8	-32.8	KIM
SNLI	SICK	56.7	-31.9	ESIM + ELMo
SNLI	SICK	56.9	-33.5	BERT-base
MultiNLI	**MultiNLI-m**	**73.1***		**BiLSTM-max**
MultiNLI	**MultiNLI-m**	**73.2***		**HBMP**
MultiNLI	**MultiNLI-m**	**76.8***		**ESIM**
MultiNLI	**MultiNLI-m**	**77.3***		**KIM**
MultiNLI	**MultiNLI-m**	**80.2***		**ESIM + ELMo**
MultiNLI	**MultiNLI-m**	**84.0***		**BERT-base**
MultiNLI	SNLI	63.8	-9.3	BiLSTM-max
MultiNLI	SNLI	65.3	-7.9	HBMP
MultiNLI	SNLI	66.4	-10.4	ESIM
MultiNLI	SNLI	68.5	-8.8	KIM
MultiNLI	SNLI	69.1	-11.1	ESIM + ELMo
MultiNLI	SNLI	80.4	-3.6	BERT-base
MultiNLI	SICK	54.1	-19.0	BiLSTM-max
MultiNLI	SICK	54.1	-19.1	HBMP
MultiNLI	SICK	47.9	-28.9	ESIM
MultiNLI	SICK	50.9	-26.4	KIM
MultiNLI	SICK	51.4	-28.8	ESIM + ELMo
MultiNLI	SICK	55.0	-29.0	BERT-base
SNLI + MultiNLI	**SNLI**	**86.1**		**BiLSTM-max**
SNLI + MultiNLI	**SNLI**	**86.1**		**HBMP**
SNLI + MultiNLI	**SNLI**	**87.5**		**ESIM**
SNLI + MultiNLI	**SNLI**	**86.2**		**KIM**
SNLI + MultiNLI	**SNLI**	**88.8**		**ESIM + ELMo**
SNLI + MultiNLI	**SNLI**	**90.6**		**BERT-base**
SNLI + MultiNLI	SICK	54.5	-31.6	BiLSTM-max
SNLI + MultiNLI	SICK	55.0	-31.1	HBMP
SNLI + MultiNLI	SICK	54.5	-33.0	ESIM
SNLI + MultiNLI	SICK	54.6	-31.6	KIM
SNLI + MultiNLI	SICK	57.1	-31.7	ESIM + ELMo
SNLI + MultiNLI	SICK	59.1	-31.5	BERT-base

Aarne and Chatzikyriakidis, arXiv 2018

GENERALIZATION (ROBUSTNESS)

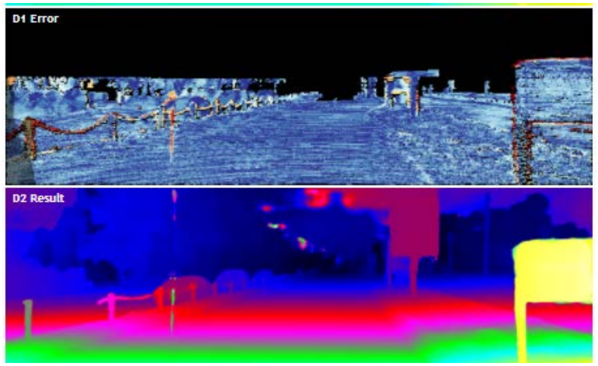
- Domain adaptation
- Perturbation
- Adversarial attacks

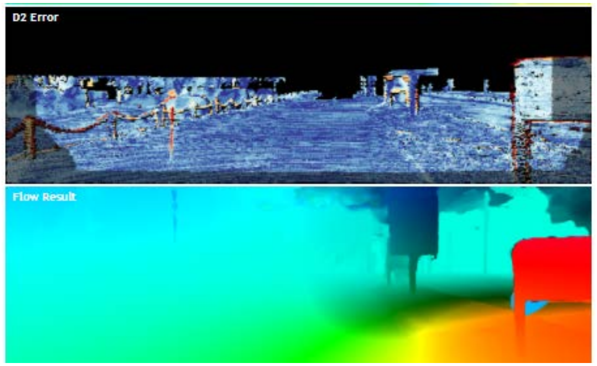


IBM CLEVER, image from: AI Tradeoff: Accuracy or Robustness? - EE Times Europe

CHARACTERIZATION

KITTI scene flow evaluation [The KITTI Vision Benchmark Suite (cvlibs.net)]





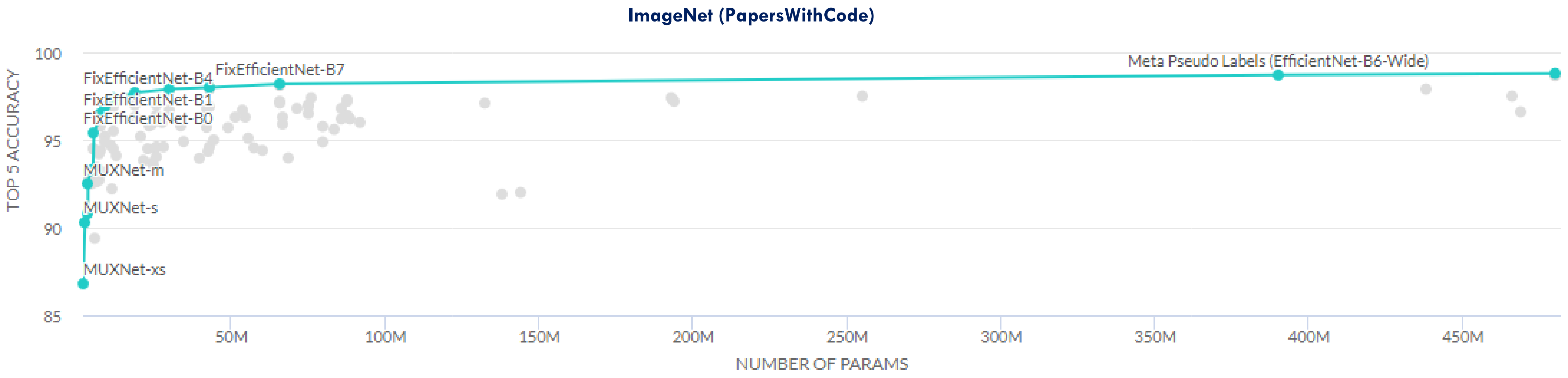
12 KPIs

Error	D1-bg	D1-fg	D1-all	D2-bg	D2-fg	D2-all	Fl-bg	Fl-fg	Fl-all	SF-bg	SF-fg	SF-all
All / All	1.89	4.45	2.02	2.12	4.96	2.26	3.20	4.09	3.24	3.55	5.94	3.66
All / Est	1.89	4.45	2.02	2.12	4.96	2.26	3.20	4.09	3.24	3.55	5.94	3.66
Noc / All	1.70	4.45	1.84	1.85	4.96	2.03	2.37	4.09	2.47	2.72	5.94	2.92
Noc / Est	1.70	4.45	1.84	1.85	4.96	2.03	2.37	4.09	2.47	2.72	5.94	2.92

4 sub-sets

NUMBER OF PARAMETERS

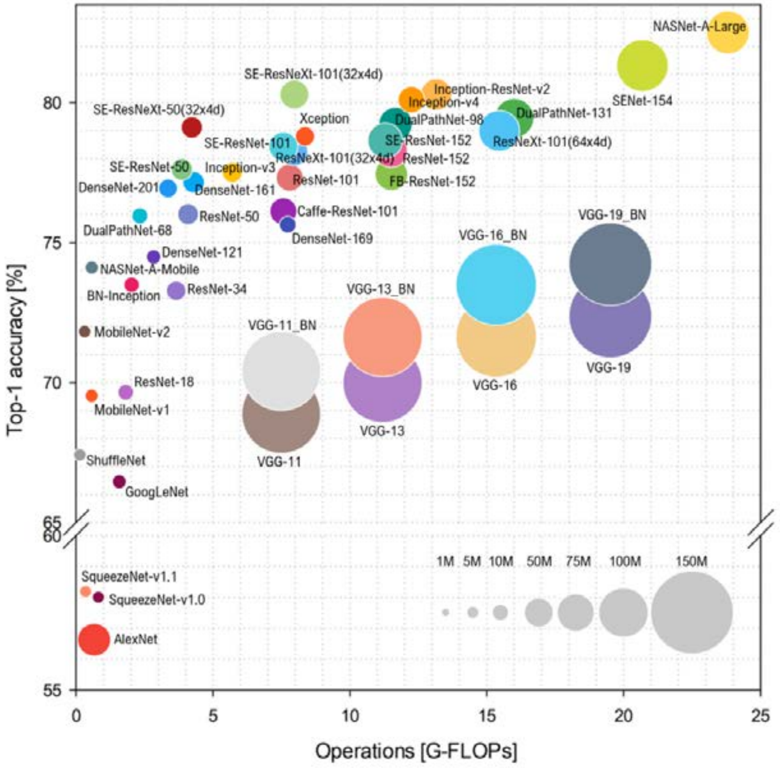
- Smaller the better (Occam's razor)
- Memory footprint
- Storage/transfer/flash



COMPUTATION AND SPEED

- Calculations
  - FLOPs
  - MADDs/MACs
  - Sometimes viewed in expected performance (early exit, attention, …)
  - #spikes

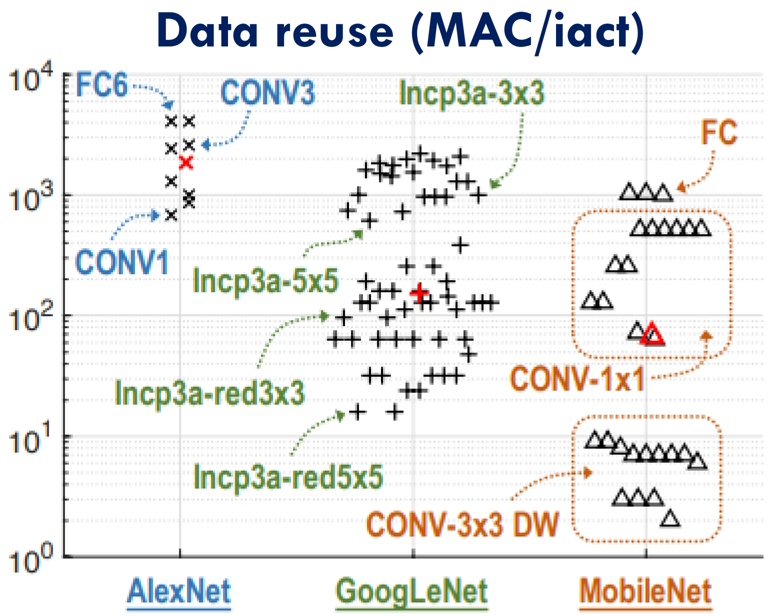
MLPerf inference [Reddi et al., ISCA 2020]



Stay tuned for Simon's overview of TinyMLPerf

COMPUTATION AND SPEED

- Calculations
  - FLOPs
  - MADDs/MACs
  - Sometimes viewed in expected performance (early exit, attention, …)
  - #spikes
- Inference speed
  - Parallelizability
  - Complexity
  - Acceleration (memory, ops)



Eyeriss 2 [Chen et al., CAS 2019]

CONVERGENCE

- Time/complexity (e.g. RNNs)
- Optimality (e.g. in GraphCuts)
- Stability (e.g. training GANs)
- Error bounds

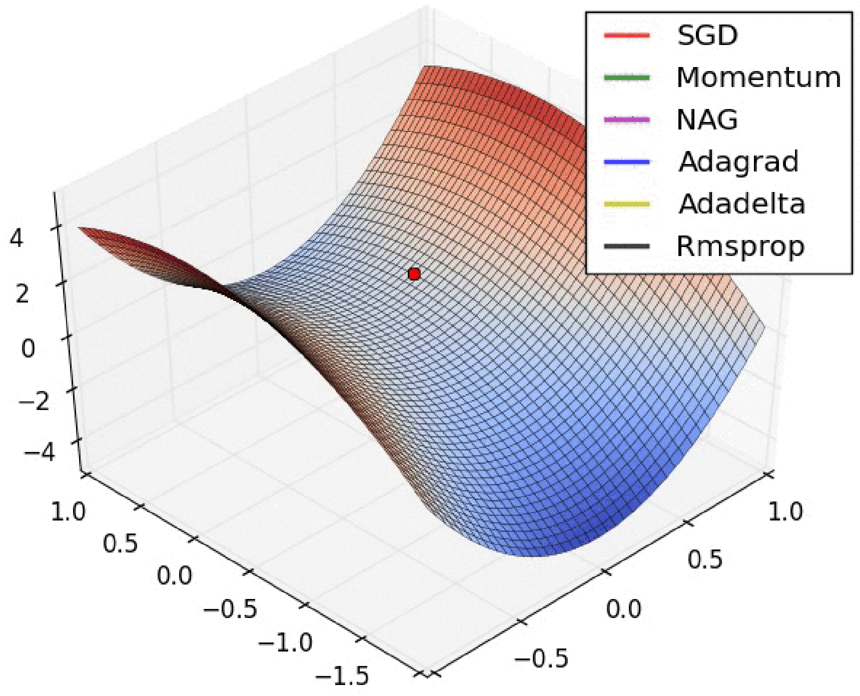
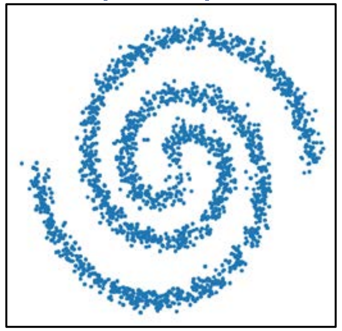


Image from: https://ruder.io/content/images/2016/09/saddle_point_evaluation_optimizers.gif

INTERPRETABILITY

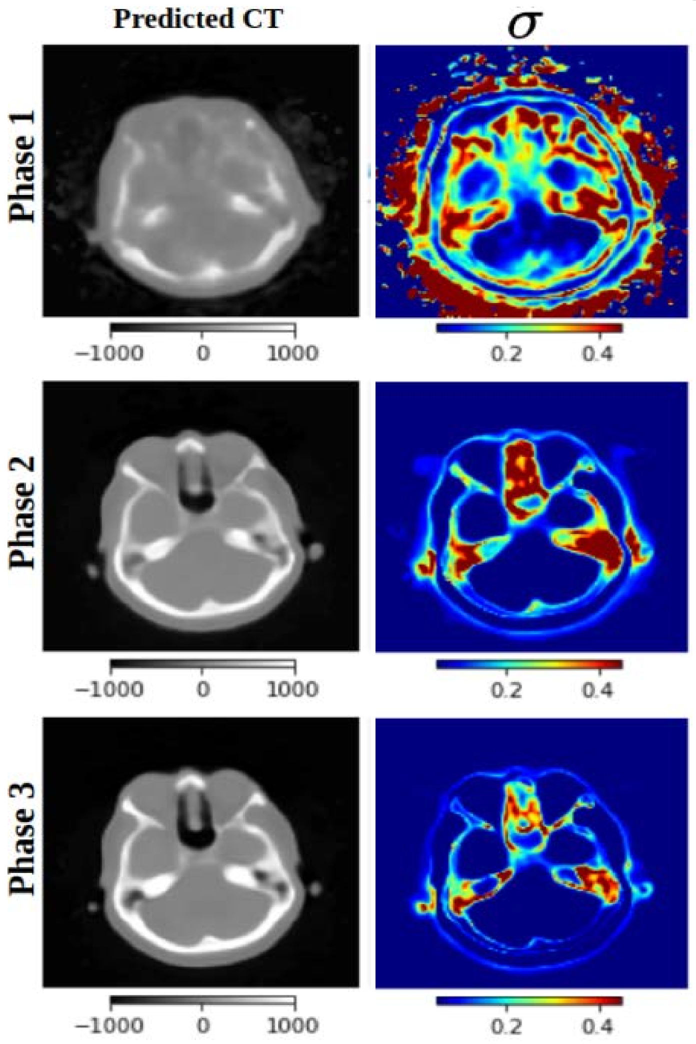
- Explicit
  - Dataset (balanced, toy)
  - Loss/Objective
  - Architecture (e.g. normalizing flows)

Two spirals toy dataset



INTERPRETABILITY

- Explicit
  - Dataset (balanced, toy)
  - Loss/Objective
  - Architecture (e.g. normalizing flows)
- Implicit
  - Precision / confidence



[Upadhyay et al., arXiv 2021]

INTERPRETABILITY

- Explicit
  - Dataset (balanced, toy)
  - Loss/Objective
  - Architecture (e.g. normalizing flows)
- Implicit
  - Precision / confidence
  - Decision boundaries / saliency

Images from: Explainable AI Demos (fraunhofer.de)

Qualitative / Simulations [Ribeiro et al. SIGKDD 2016]

User-comparison [Lundberg and Lee Neurips 2017], Simulatability [Hase and Bansal, ACL 2020]

REPRODUCIBILITY

- Standard/Interpretable datasets
  - CIFAR, MNIST, PascalVOC, …

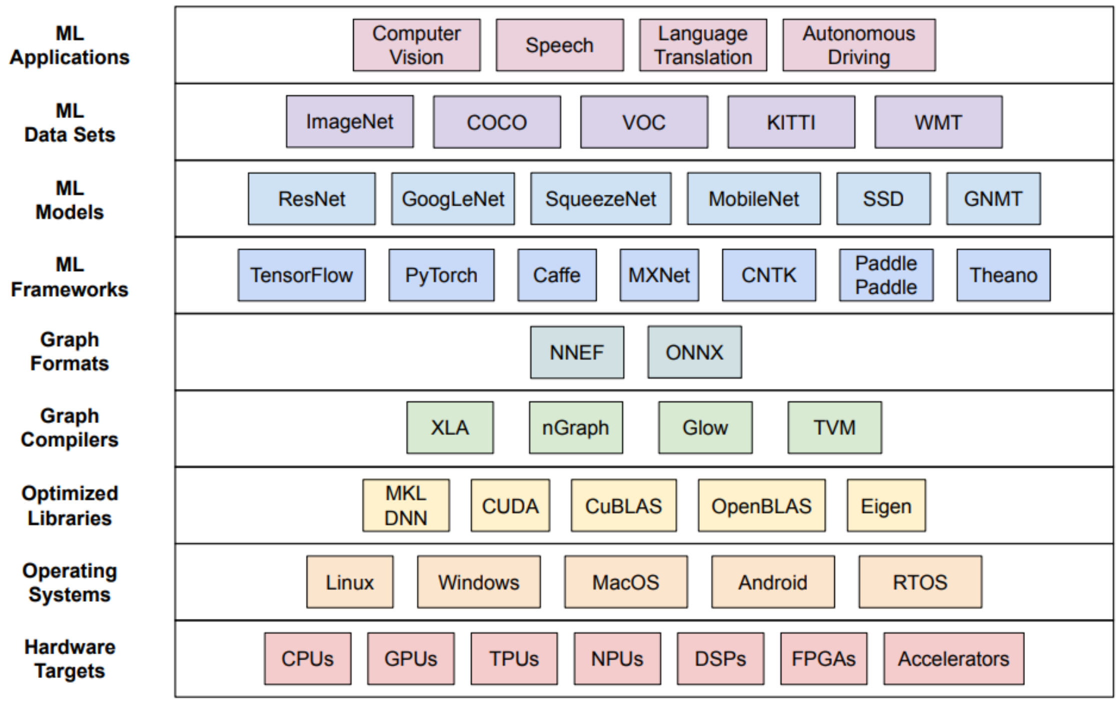
REPRODUCIBILITY

- Standard/Interpretable datasets
  - CIFAR, MNIST, PascalVOC, …
- Easy of installation and testing
  - Conventional packages like TF
  - Minimum requirements
  - Popular data interfaces (ONNX, png, …)
  - Initialization/optimization

Image from: Validation of Machine Learning Libraries (johner-institute.com)

REPRODUCIBILITY

- Standard/Interpretable datasets
  - CIFAR, MNIST, PascalVOC, …
- Easy of installation and testing
  - Conventional packages like TF
  - Minimum requirements
  - Popular data interfaces (ONNX, png, …)
  - Initialization/optimization
- Deployment
  - Platforms
  - Scalability



Reddi et al. Arxiv 2020

SUMMARY

- KPIs:
  - precision/accuracy, generalization/memorization, robustness, parameters/activation size, convergence time/optimality/reproducibility, interpretability
- Datasets:
  - Interpretability, bias, toy sets, standard, exhaustive
- Data-driven metrics:
  - FID, ALL, expected FLOPs, Perceptual

AIBench vs. MLPerf [Fei et al., ISPASS 2021]