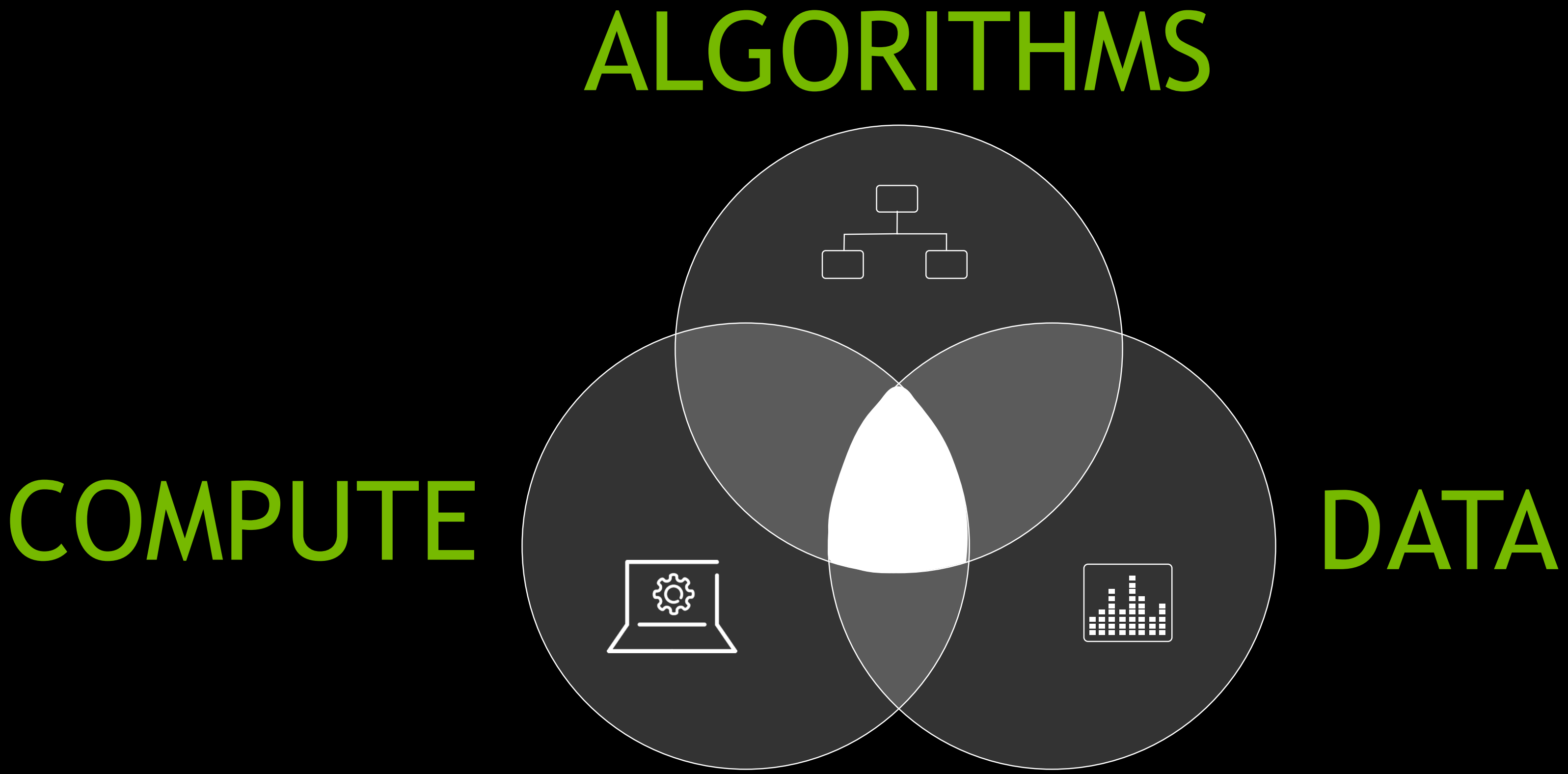
Anima Anandkumar

Caltech

NVIDIA

BEYOND BLACK-BOXES: INFUSING STRUCTURES INTO DEEP LEARNING

TRINITY OF AI/ML



DEEP LEARNING IS DATA-HUNGRY

STRUCTURE-INFUSED LEARNING

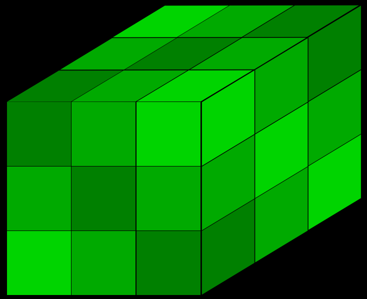
Learning = Data + Priors

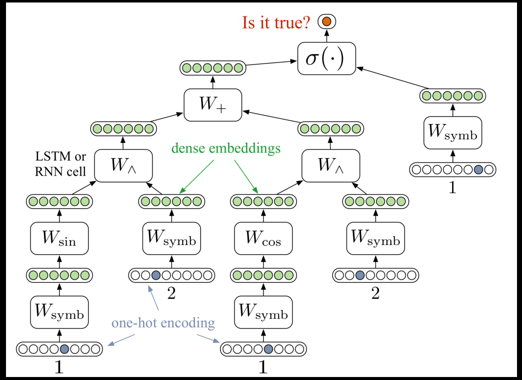
USE OF PRIORS FOR DATA EFFICIENCY

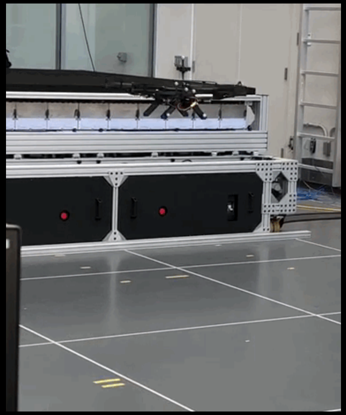
Learning = Data + Priors

How to use structure and domain knowledge to design Priors?

Examples of Priors
- Tensors and graphs
- Symbolic rules
- Physical laws
- Simulations

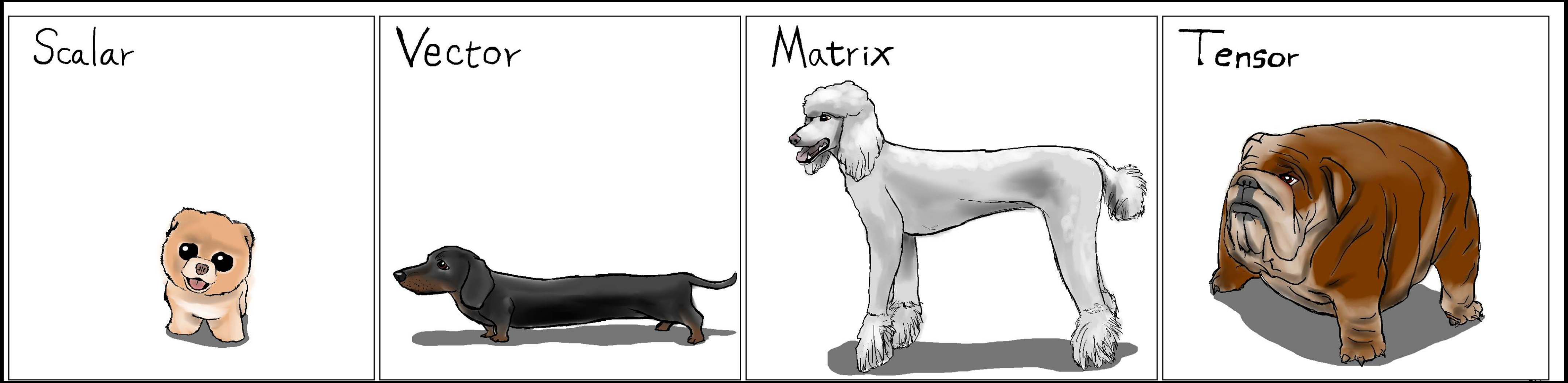


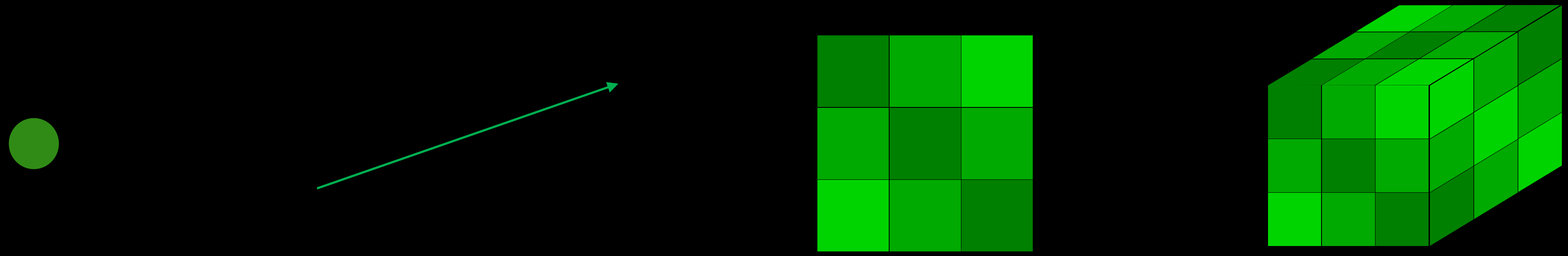




Learning in Many Dimensions

TENSOR : EXTENSION OF MATRIX





TENSORS FOR DATA
ENCODE MULTI-DIMENSIONALITY

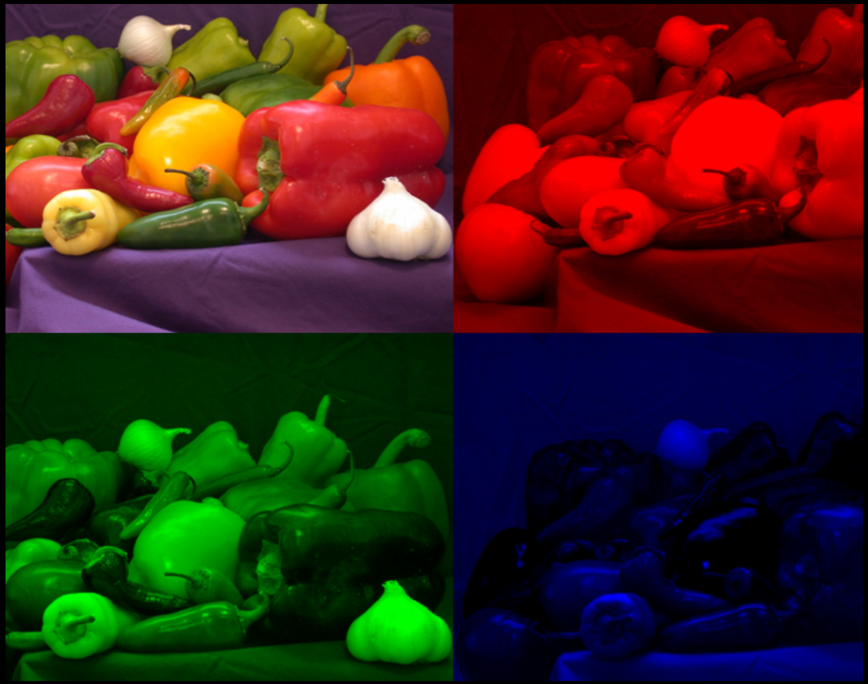


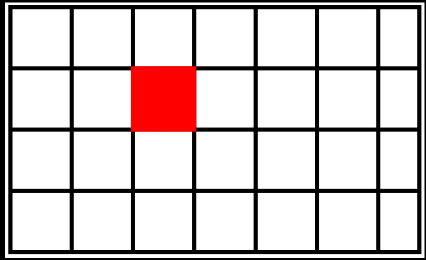
Image: 3 dimensions
Width * Height * Channels

Video: 4 dimensions
Width * Height * Channels * Time

TENSORS FOR ML ALGORITHMS
ENCODE HIGHER ORDER MOMENTS

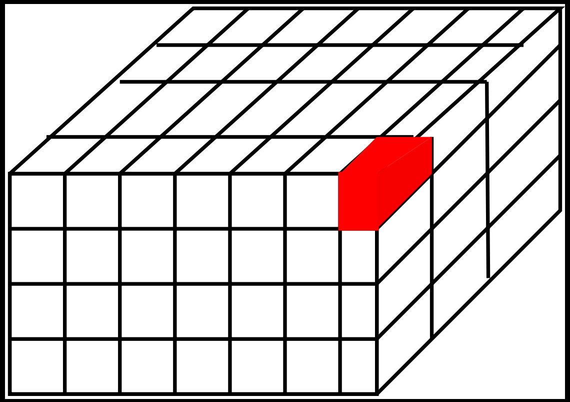
Pairwise correlations

$$E(x \otimes x)_{i,j} = E(x_i x_j)$$



Third order correlations

$$E(x \otimes x \otimes x)_{i,j,k} = E(x_i x_j x_k)$$



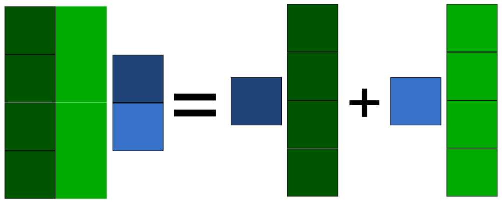
TENSORS FOR COMPUTE

TENSOR CONTRACTION PRIMITIVE

Extends the notion of matrix product

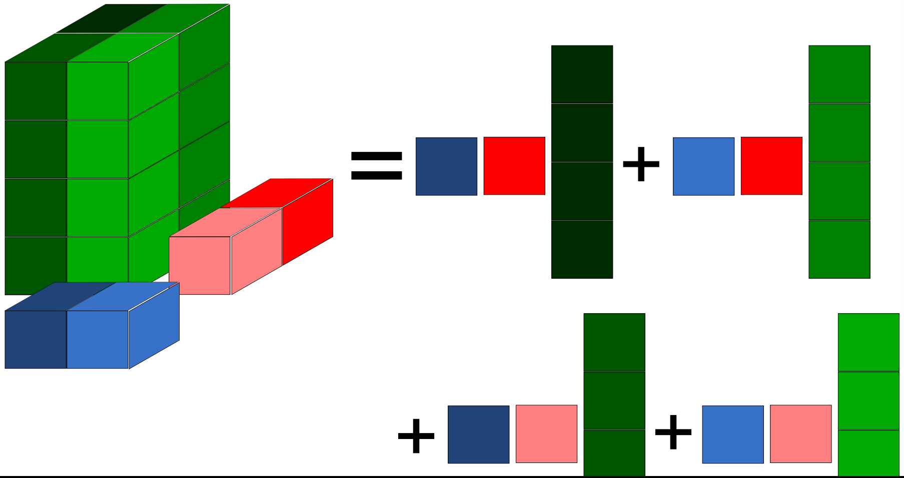
Matrix product

$$Mv = \sum_j v_j M_j$$



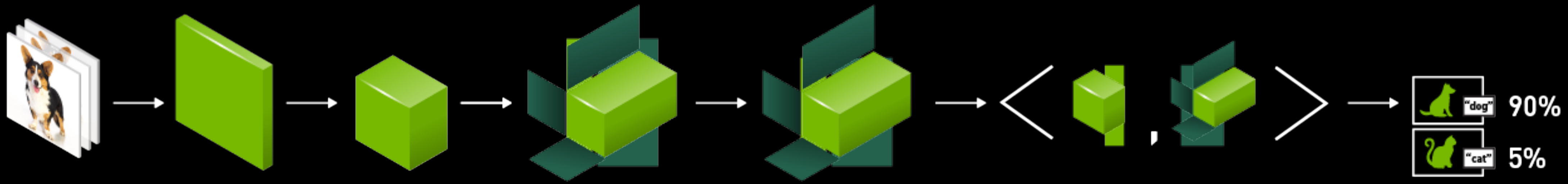
Tensor Contraction

$$T(u, v, \cdot) = \sum_{i,j} u_i v_j T_{i,j,:}$$

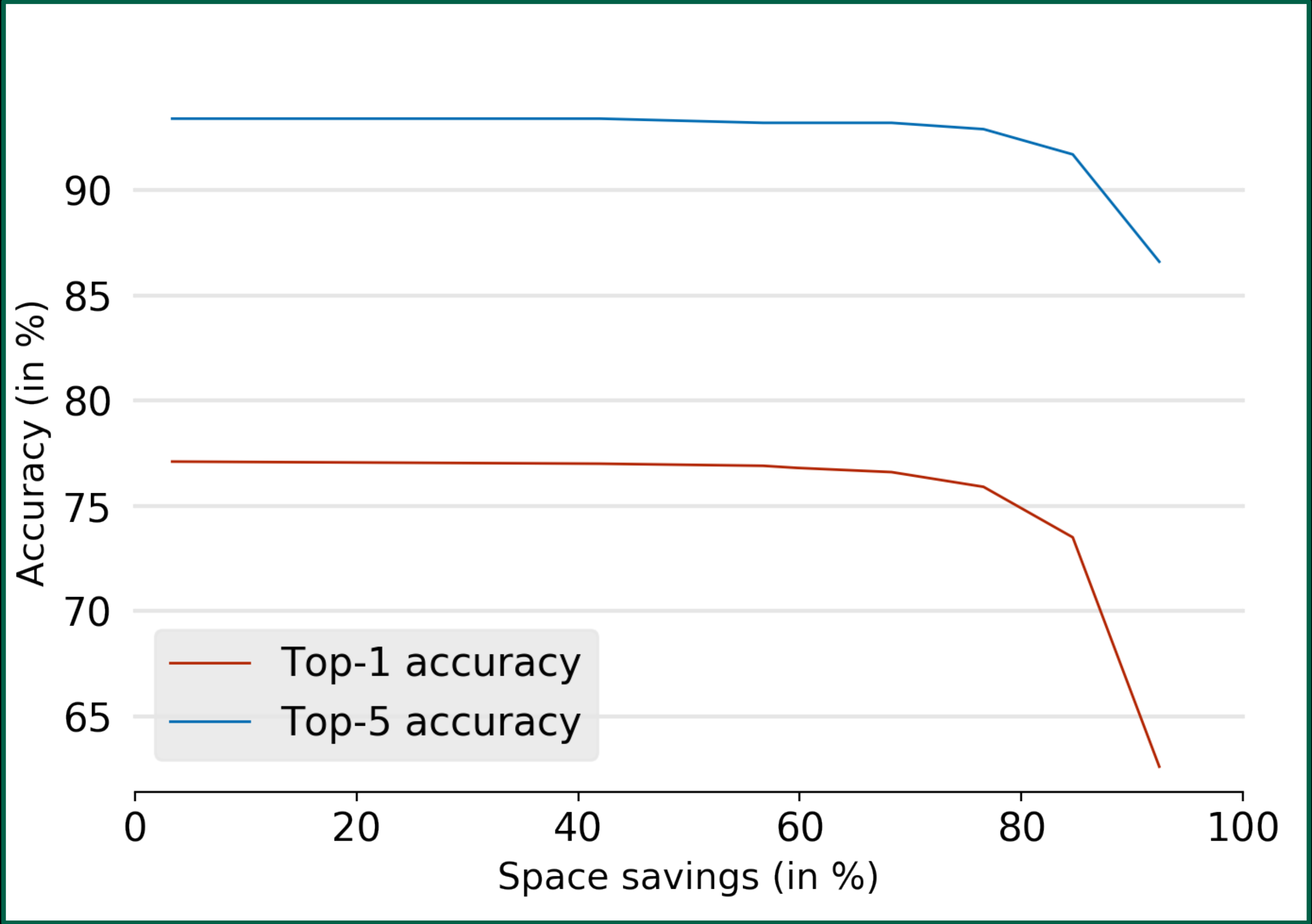


TENSORS FOR MODELS

TENSORS IN NEURAL NETWORKS
STANDARDIZED ON LINEAR ALGEBRA



SPACE SAVING IN DEEP TENSORIZED NETWORKS



Jean Kossaifi

Zachary Lipton

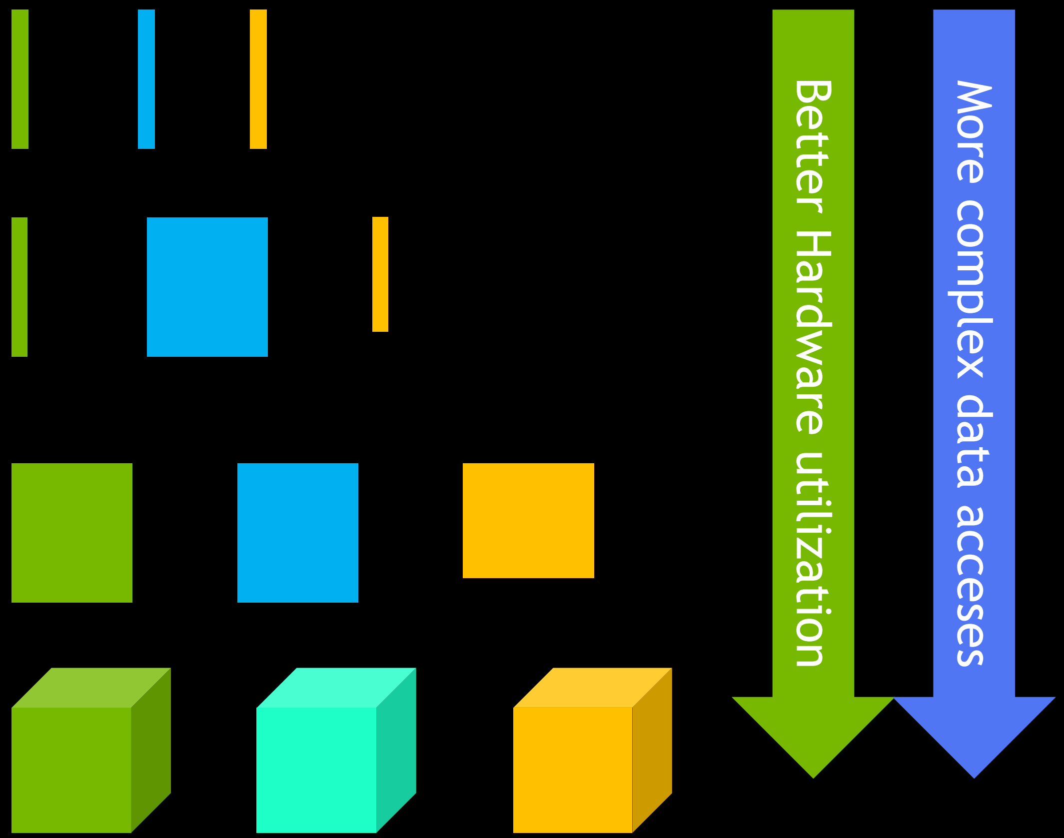
Aran Khanna

Tommaso Furlanello

TENSOR PRIMITIVES?

History & Future

- 1969 – BLAS Level 1: Vector-Vector
- 1972 – BLAS Level 2: Matrix-Vector
- 1980 – BLAS Level 3: Matrix-Matrix
- Now? – BLAS Level 4: Tensor-Tensor



Kim, Jinsung, et al. "Optimizing Tensor Contractions in CCSD (T) for Efficient Execution on GPUs." (2018).

UNSUPERVISED LEARNING TOPIC MODELS THROUGH TENSORS

Topics

- Justice
- Education
- Sports

SECTIONS HOME SEARCH The New York Times

COLLEGE FOOTBALL

At Florida State, Football Clouds Justice

By MIKE McINTIRE and WALT BOGDANICH OCT. 10, 2014

Now, an examination by The New York Times of police and court records, along with interviews with crime witnesses, has found that, far from an aberration, the treatment of the Winston complaint was in keeping with the way the police on numerous occasions have soft-pedaled allegations of wrongdoing by Seminoles football players. From criminal mischief and motor-vehicle theft to domestic violence, arrests have been avoided, investigations have stalled and players have escaped serious consequences.

In a community whose self-image and economic well-being are so tightly bound to the fortunes of the nation's top-ranked college football team, law enforcement officers are finely attuned to a suspect's football connections. Those ties are cited repeatedly in police reports examined by The Times. What's more, dozens of officers work second jobs directing traffic and providing security at home football games, and many express their devotion to the Seminoles on social media.

TMZ, the gossip website, also requested the police report and later asked the school's deputy police chief, Jim L. Russell, if the campus police had interviewed Mr. Winston about the rape report. Mr. Russell responded by saying his officers were not investigating the case, omitting any reference to the city police, even though the campus police knew of their involvement. "Thank you for contacting me regarding this rumor — I am glad I can dispel that one!" Mr. Russell told TMZ in an email. The university said Mr. Russell was unaware of any other police investigation at the time of the inquiry. Soon after, the Tallahassee police belatedly sent their files to the news media and to the prosecutor, William N. Meggs. By then critical evidence had been lost and Mr. Meggs, who criticized the police's handling of the case, declined to

lson after the Seminoles' first game; five

am's second-leading receiver.

On Jan. 10, 2013, a female student at Florida State spotted the man she believed had raped her the previous month. After learning his name, Jameis Winston, she reported him to the Tallahassee police.

In the 21 months since, Florida State officials have said little about how they handled the case, which is no

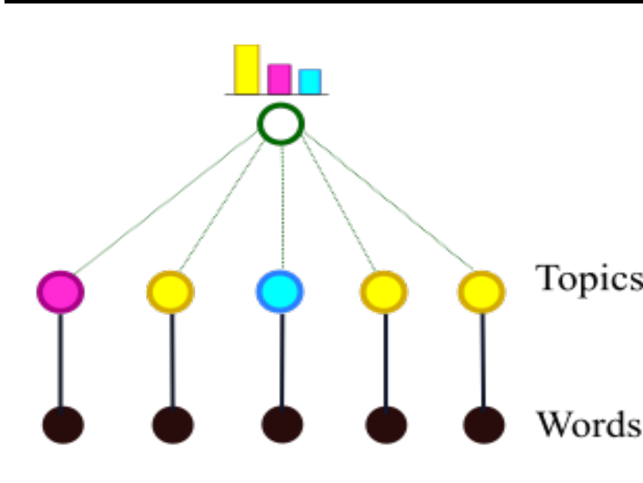
investigated by the federal Depart

Most recently, university officials suspended Mr. Winston for one game after he stood in a public place on campus and, playing off a running Internet gag, shouted a crude reference to a sex act. In a news conference afterward, his coach, Jimbo Fisher, said, "Our hope and belief is Jameis will learn from this and use better judgment and language and decision-making."

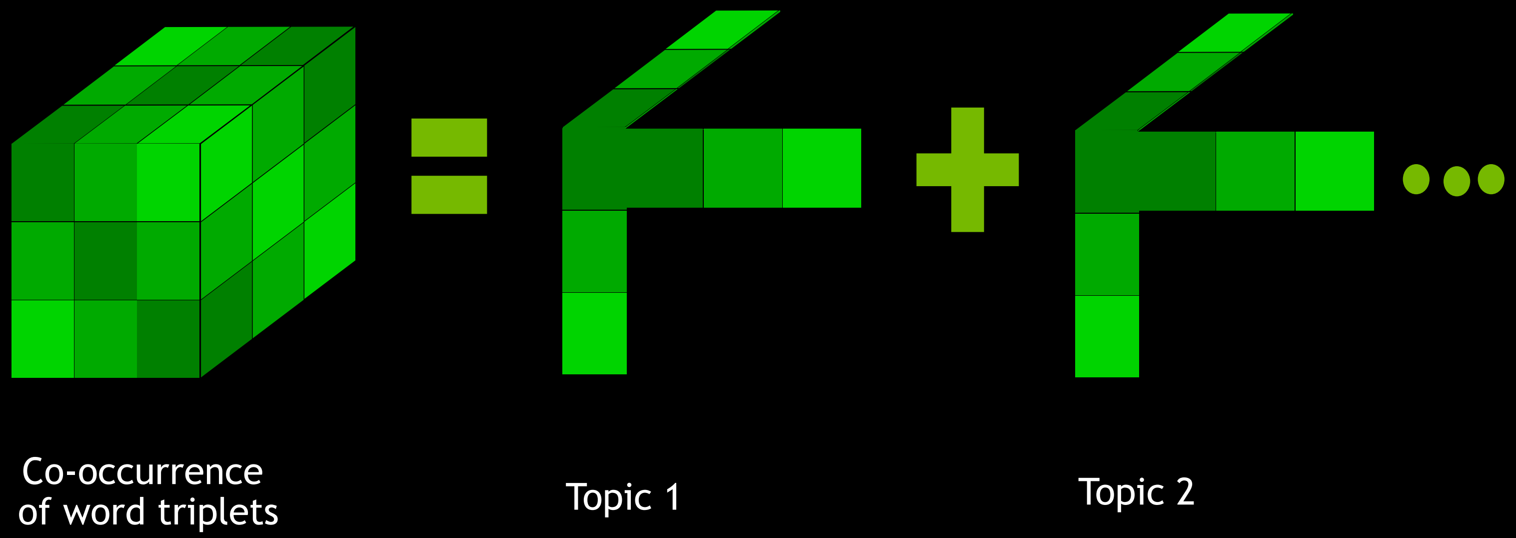
As The Times reported last April, the Tallahassee police also failed to aggressively investigate the rape accusation. It did not become public until November, when a Tampa reporter, Matt Baker, acting on a tip, sought records of the police investigation.

Upon learning of Mr. Baker's inquiry, Florida State, having shown little curiosity about the rape accusation, suddenly took a keen interest in the journalist seeking to report it, according to emails obtained by The Times.

"Can you share any details on the requesting source?" David Perry, the university's police chief, asked the Tallahassee police. Several hours later, Mr.



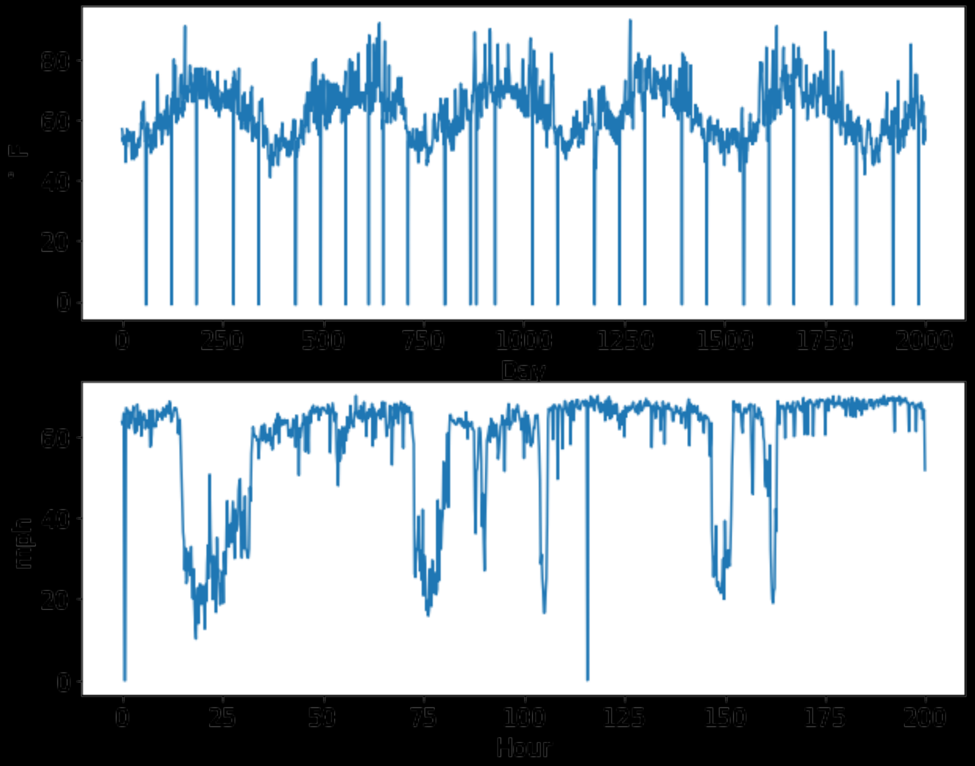
TENSORS FOR MODELING:
TOPIC DETECTION IN TEXT



TENSORS FOR LONG-TERM FORECASTING

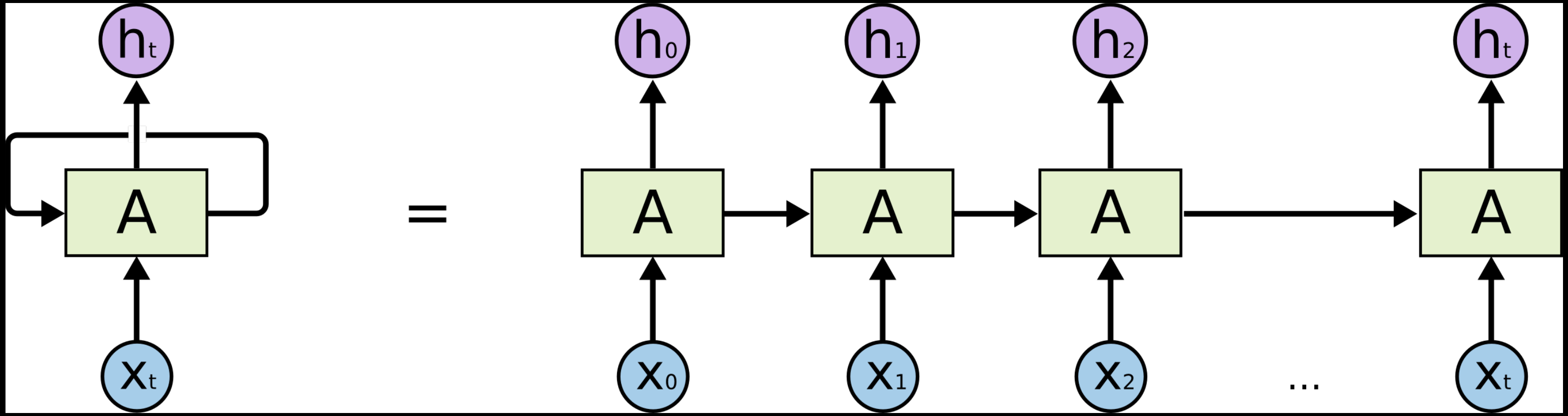
Difficulties in long term forecasting:

- Long-term dependencies
- High-order correlations
- Error propagation



RNN: FIRST-ORDER MARKOV MODELS

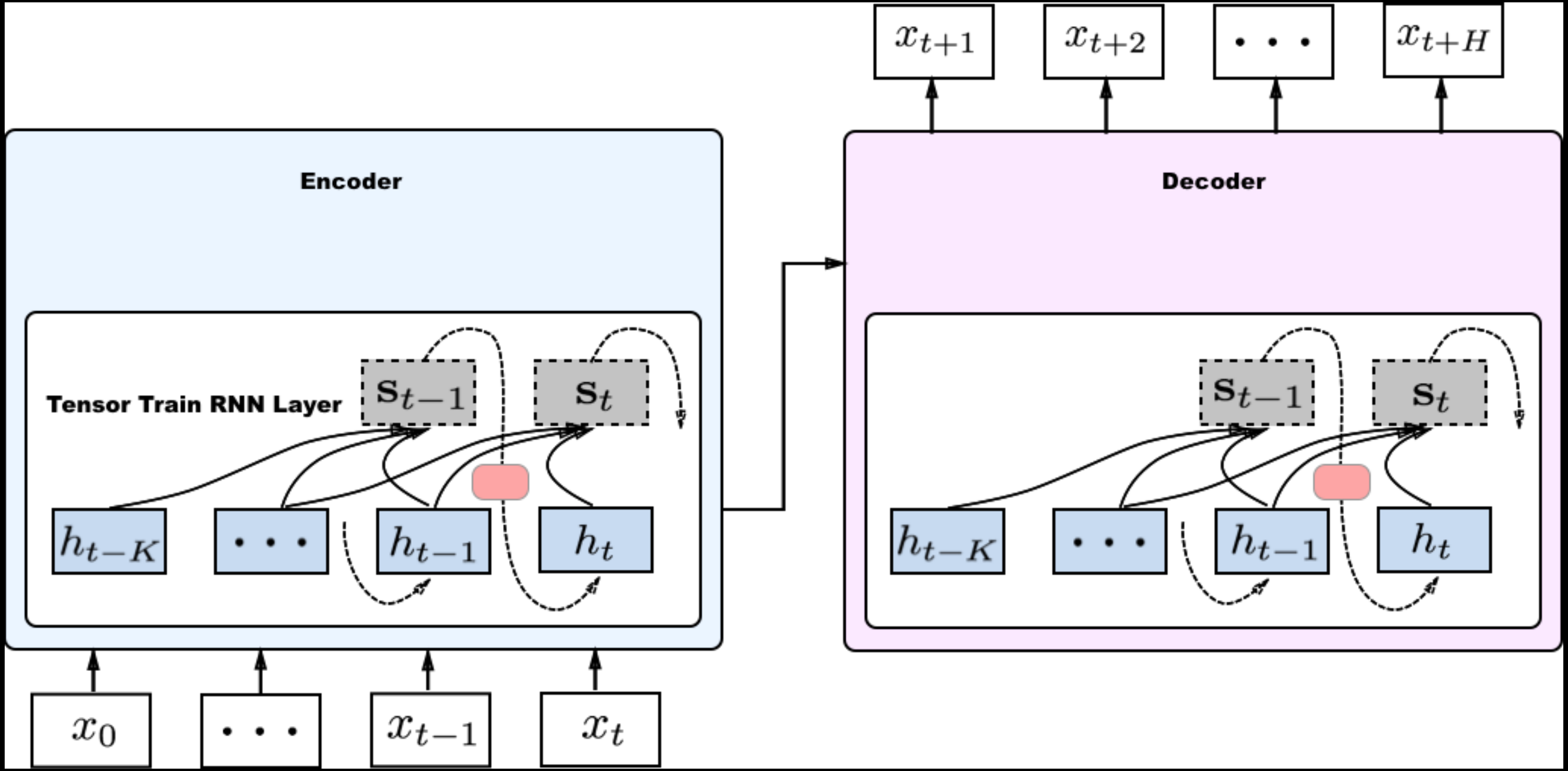
Input state x_t, hidden state h_t, output y_t,



TENSOR-TRAIN RNNS AND LSTMS

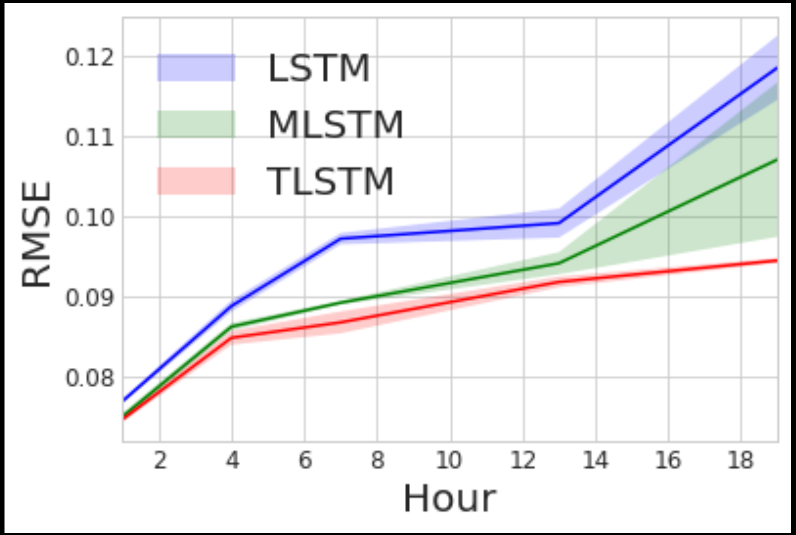
Seq2seq architecture

TT-LSTM cells

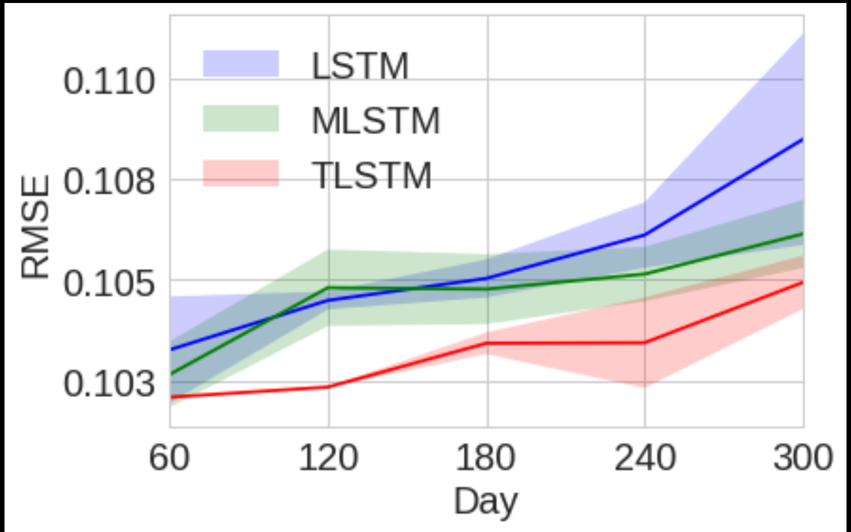


TENSOR LSTM FOR LONG-TERM FORECASTING

Traffic dataset



Climate dataset



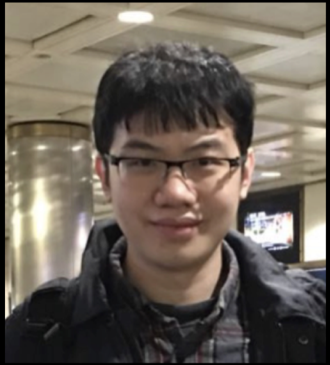
Rose Yu

Stephan Zhang

Yisong Yue

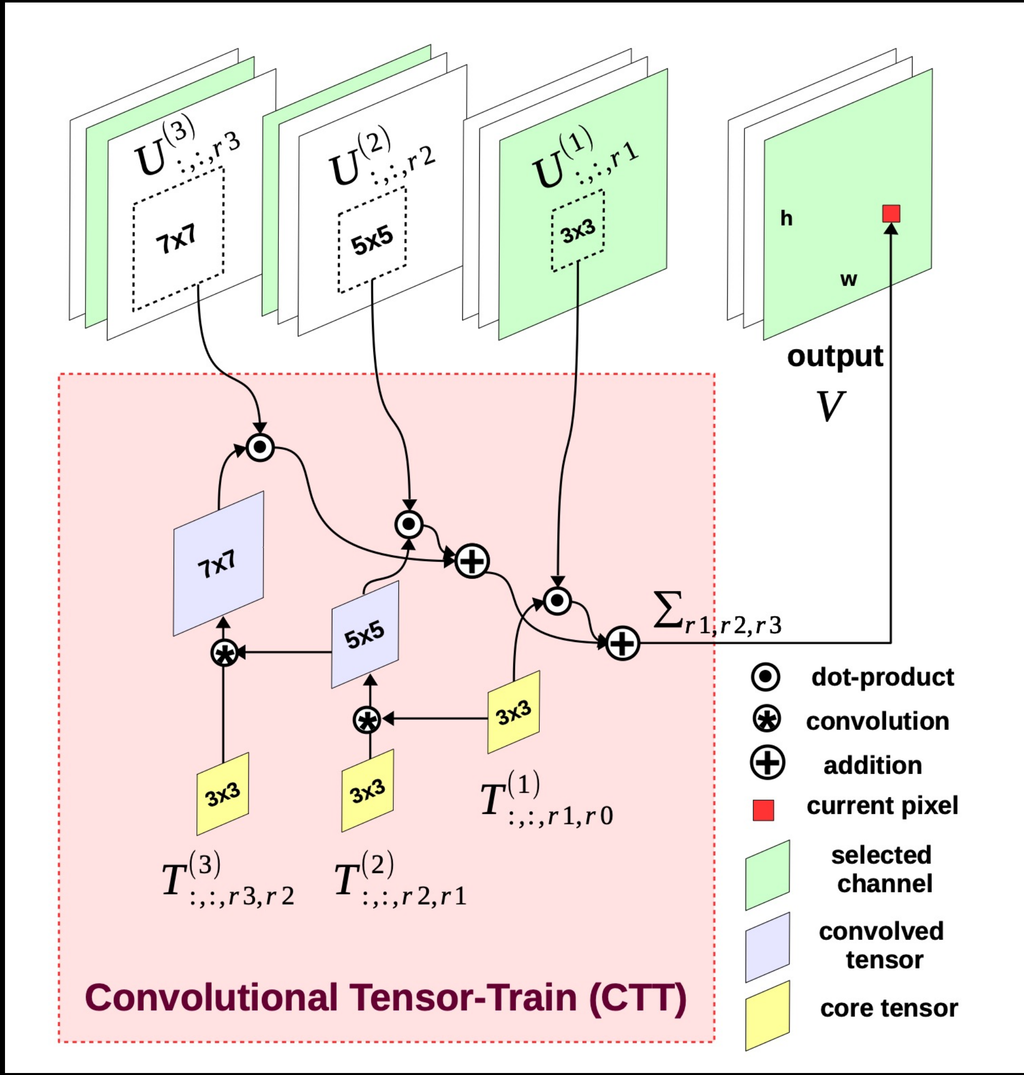
LONG-TERM VIDEO PREDICTION WITH CONVOLUTIONAL TENSOR-TRAIN LSTM

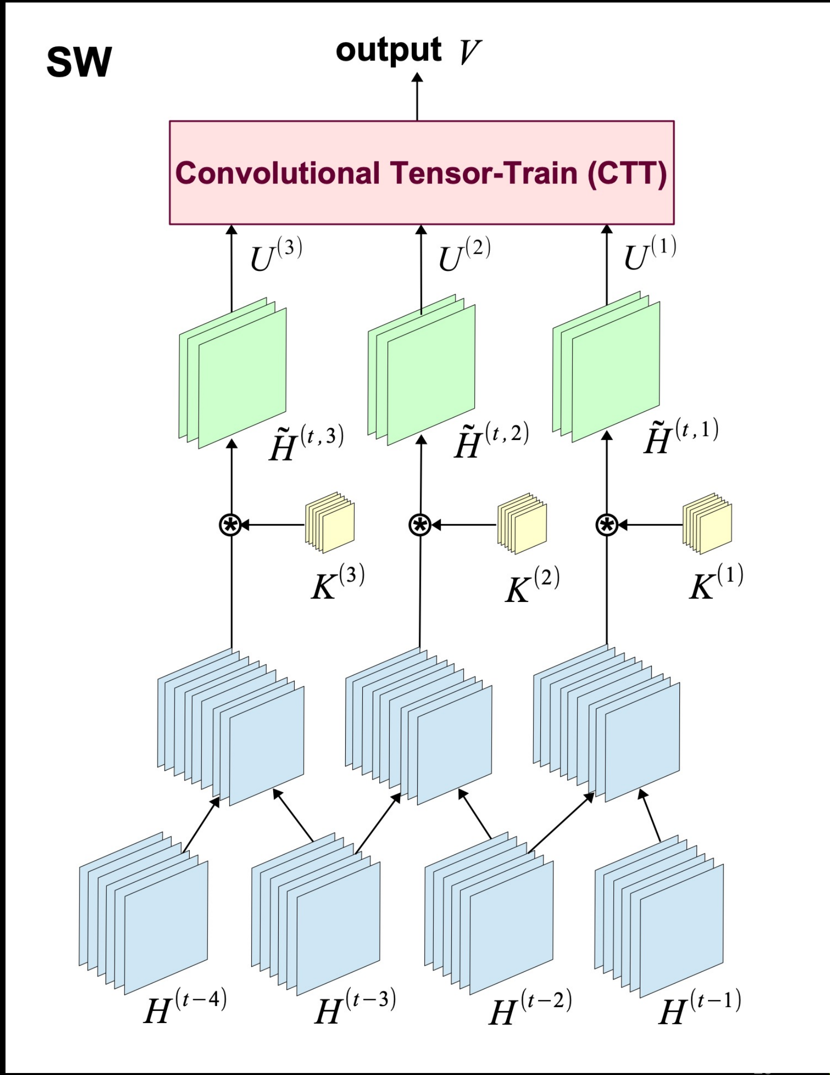
Jiahao Su, Wonmin Byeon, Furong Huang, Jan Kautz, Anima Anandkumar



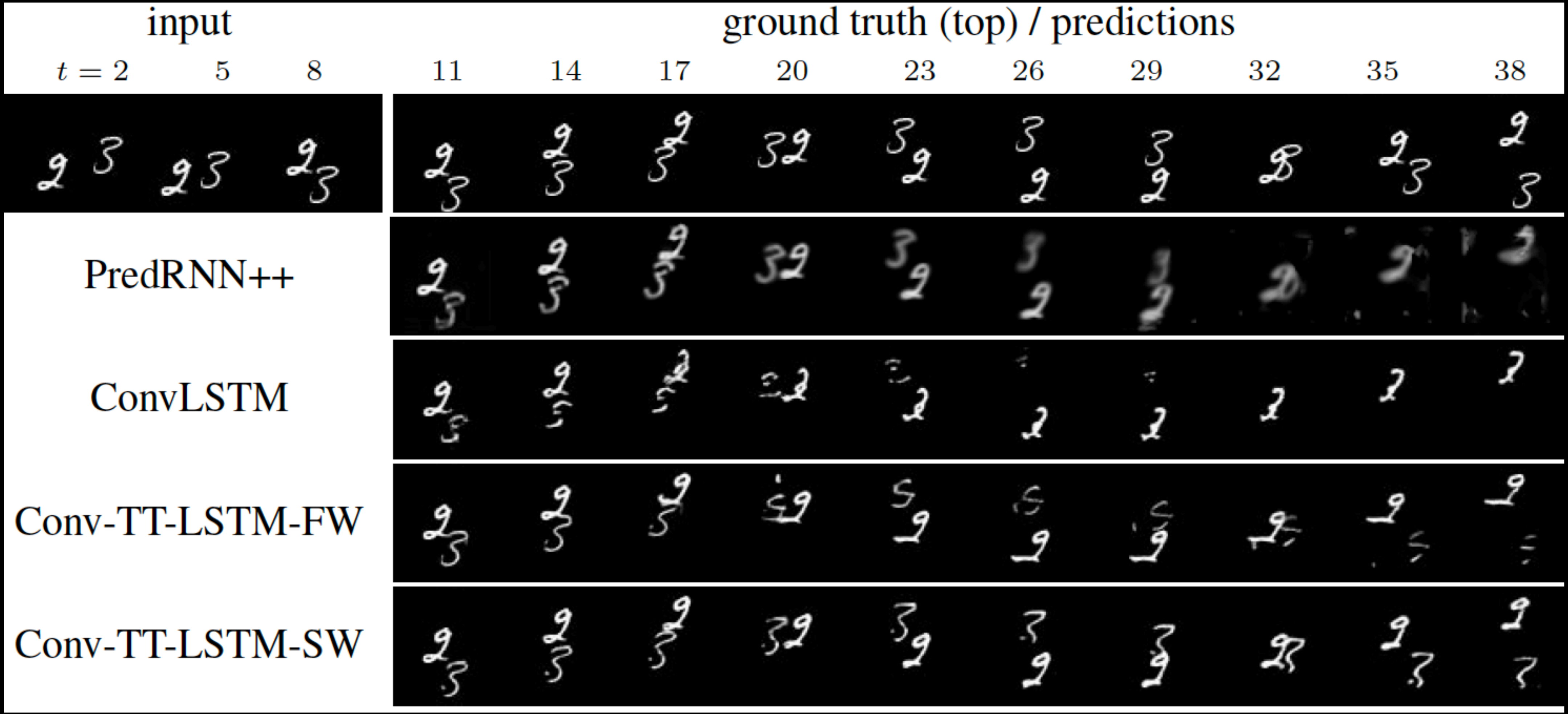


CONVOLUTIONAL TENSOR-TRAIN LSTM

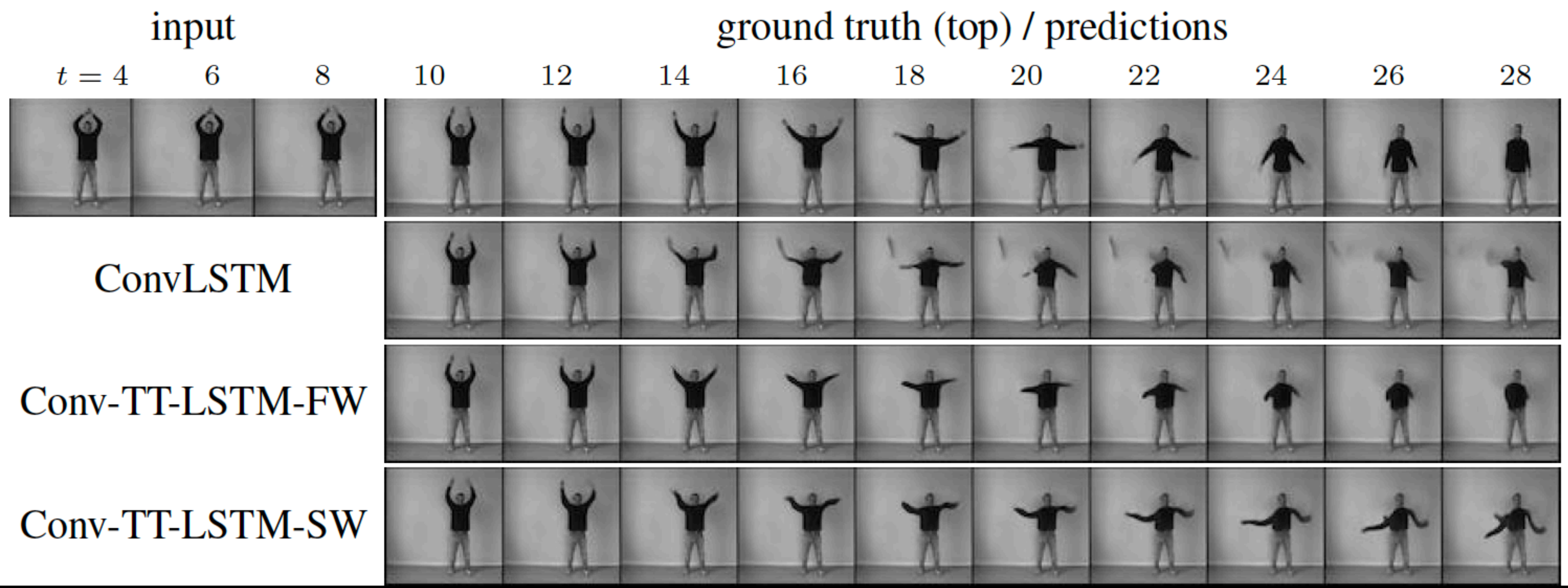




PREDICTION RESULTS



PREDICTION RESULTS



PREDICTION RESULTS

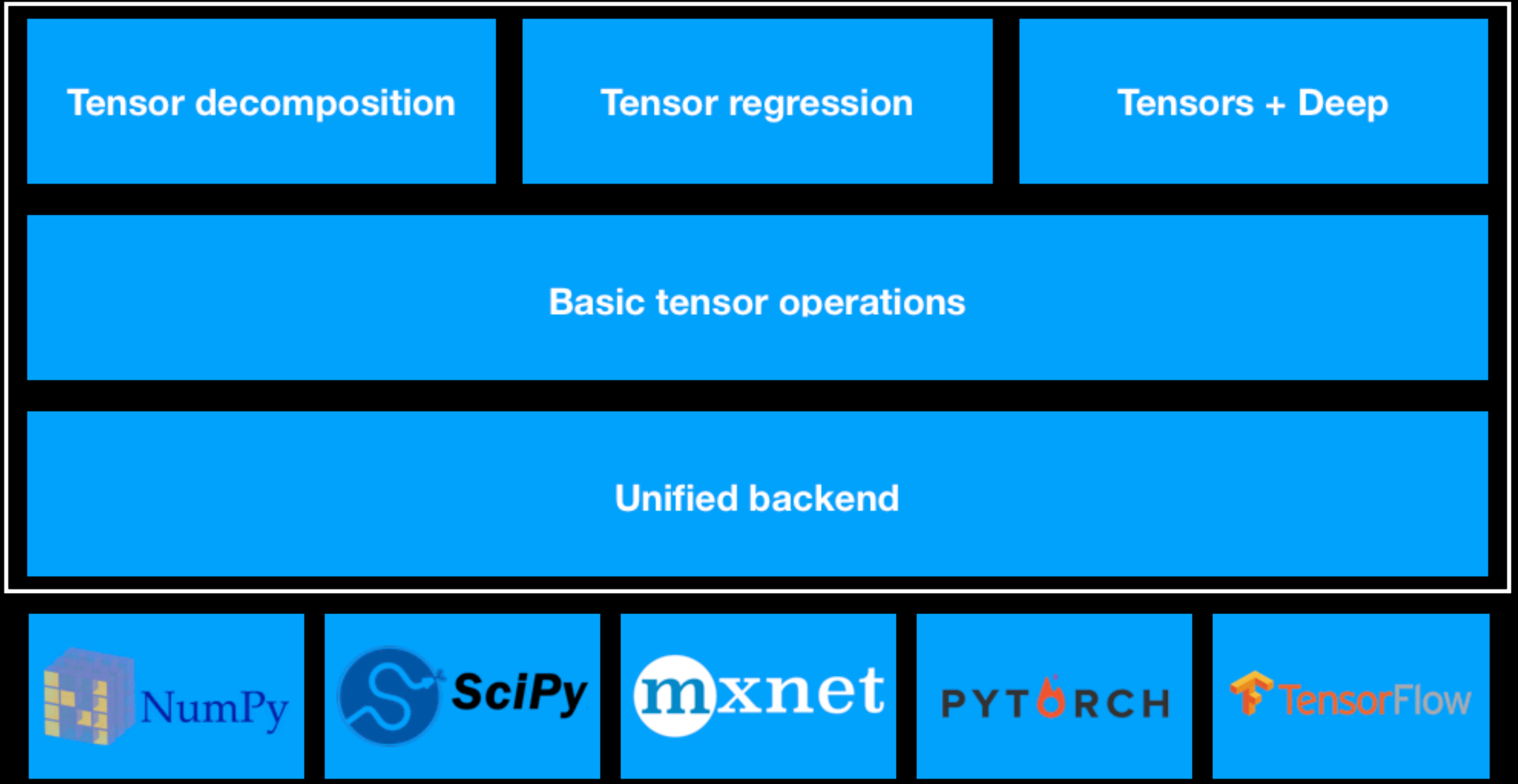
Moving MNIST

Method	(10 -> 30) MSE($\times 10^{-3}$)	SSIM	# parameters
Baseline ConvLSTM (4-layers model)	37.19	0.791	11.48M
Conv-TT-LSTM-FW (4-layers model)	**31.46**	**0.819**	**5.65M**
Baseline ConvLSTM ($\mathcal{L}_1$ loss only)	33.96	0.805	3.97M
Conv-TT-LSTM-FW ($\mathcal{L}_1$ loss only)	**30.27**	**0.827**	**2.65M**
Baseline ConvLSTM (teacher forcing)	36.95	0.802	3.97M
Conv-TT-LSTM-FW (teacher forcing)	**34.84**	**0.807**	**2.65M**
Baseline ConvLSTM (our strategy)	33.08	0.806	3.97M
Conv-TT-LSTM-FW (our strategy)	**28.88**	**0.831**	**2.65M**

KTH

Method	(10 -> 20) PSNR	(10 -> 20) SSIM	(10 -> 40) PSNR	(10 -> 40) SSIM	# Parameters
ConvLSTM (Xingjian et al., 2015)	23.58	0.712	22.85	0.639	7.58M
PredRNN++ (Wang et al., 2018a)	28.46	0.865	25.21	0.741	15.05M
E3D-LSTM (Wang et al., 2018b)	29.31	0.879	**27.24**	0.810	≈15M[3]
ConvLSTM-12 (baseline)	27.16	0.871	25.32	0.840	3.97M
Conv-TT-LSTM-FW (ours)	27.38	0.874	25.60	0.845	2.65M
Conv-TT-LSTM-SW (ours)	27.51	0.875	25.78	**0.846**	2.69M

TENSORLY: HIGH-LEVEL API FOR TENSOR ALGEBRA



- Python programming
- User-friendly API
- Multiple backends: flexible + scalable
- Example notebooks

Jean Kossaifi

TENSORLY WITH PYTORCH BACKEND

```
import tensorly as tl
from tensorly.random import tucker_tensor

tl.set_backend('pytorch')
core, factors = tucker_tensor((5, 5, 5),
                              rank=(3, 3, 3))
core = Variable(core, requires_grad=True)
factors = [Variable(f, requires_grad=True) for f in factors]

optimiser = torch.optim.Adam([core]+factors, lr=lr)

for i in range(1, n_iter):
    optimiser.zero_grad()
    rec = tucker_to_tensor(core, factors)
    loss = (rec - tensor).pow(2).sum()
    for f in factors:
        loss = loss + 0.01*f.pow(2).sum()

    loss.backward()
    optimiser.step()
```

Set Pytorch backend

Tucker Tensor form

Attach gradients

Set optimizer

Blending Data Driven Learning with Symbolic Reasoning

AGE-OLD DEBATE IN AI

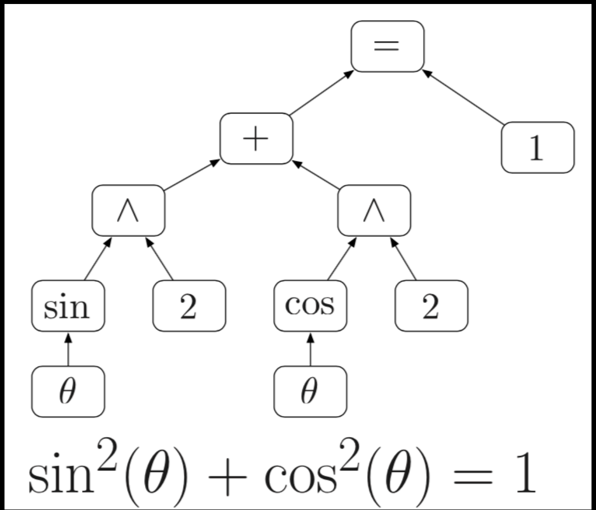
Symbols vs. Representations

Symbolic reasoning:

- Humans have impressive ability at symbolic reasoning
- Compositional: can combine different concepts

Representation learning:

- Data driven: Do not need to know the base concepts
- Black box and not compositional

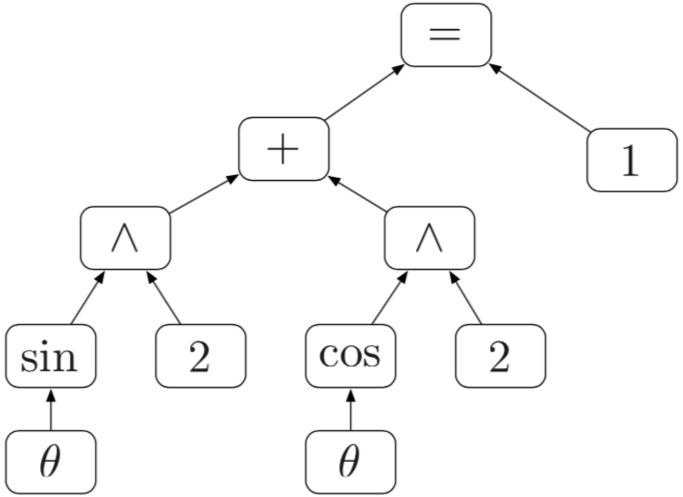


Forough Arabshahi

Sameer Singh

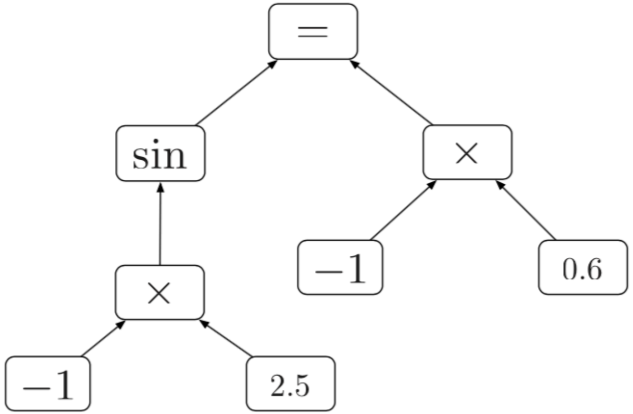
Combining Symbolic Expressions & Black-box Function Evaluations in Neural Programs, ICLR 2018

EXPLOITING HIERARCHICAL REPRESENTATIONS



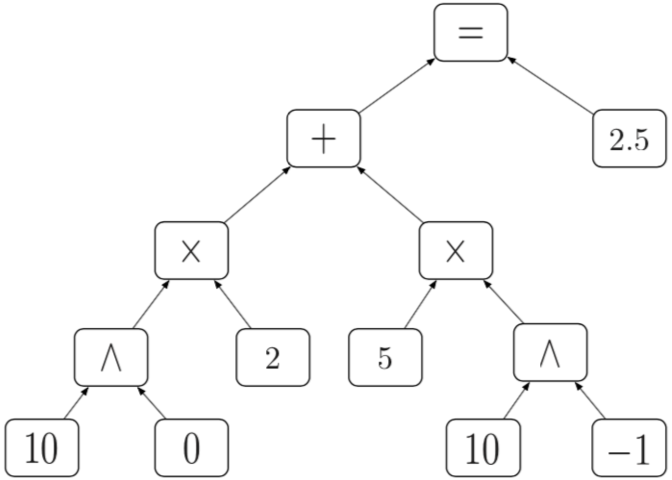
$$\sin^2(\theta) + \cos^2(\theta) = 1$$

Symbolic expression



$$\sin(-2.5) = -0.6$$

Function Evaluation Data Point



decimal tree for 2.5

Number Encoding Data Point

EQUATION VERIFICATION

	Majority Class	Sympy	LSTM : sym	TreeLSTM : sym	TreeLSTM:sym+num
Generalization	50.24%	81.74%	81.71%	95.18%	97.20%
Extrapolation	44.78%	71.93%	76.40%	93.27%	96.17%

TAKE-AWAYS

Vastly Improved numerical evaluation: 90% over function-fitting baseline.

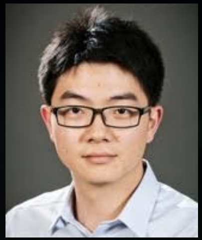
Generalization to verifying symbolic equations of higher depth

LSTM: Symbolic	TreeLSTM: Symbolic	TreeLSTM: symbolic + numeric
76.40 %	93.27 %	**96.17 %**

Combining symbolic + numerical data helps in better generalization for both tasks: symbolic and numerical evaluation.

Learning in Control Systems

Guanya Shi



Xichen Shi

Michael O'Connell

Rose Yu

Kamyar Azizzadenesheli

Soon-Jo Chung

Yisong Yue

Neural Lander: Stable Drone Landing Control using Learned Dynamics, ICRA 2019

LEARNING RESIDUAL DYNAMICS FOR DRONE LANDING

f = nominal dynamics
$\tilde{f}$ = learned dynamics

New State

Current Action (aka control input)

$$s_{t+1} = f(s_t, a_t) + \tilde{f}(s_t, a_t) + \epsilon$$

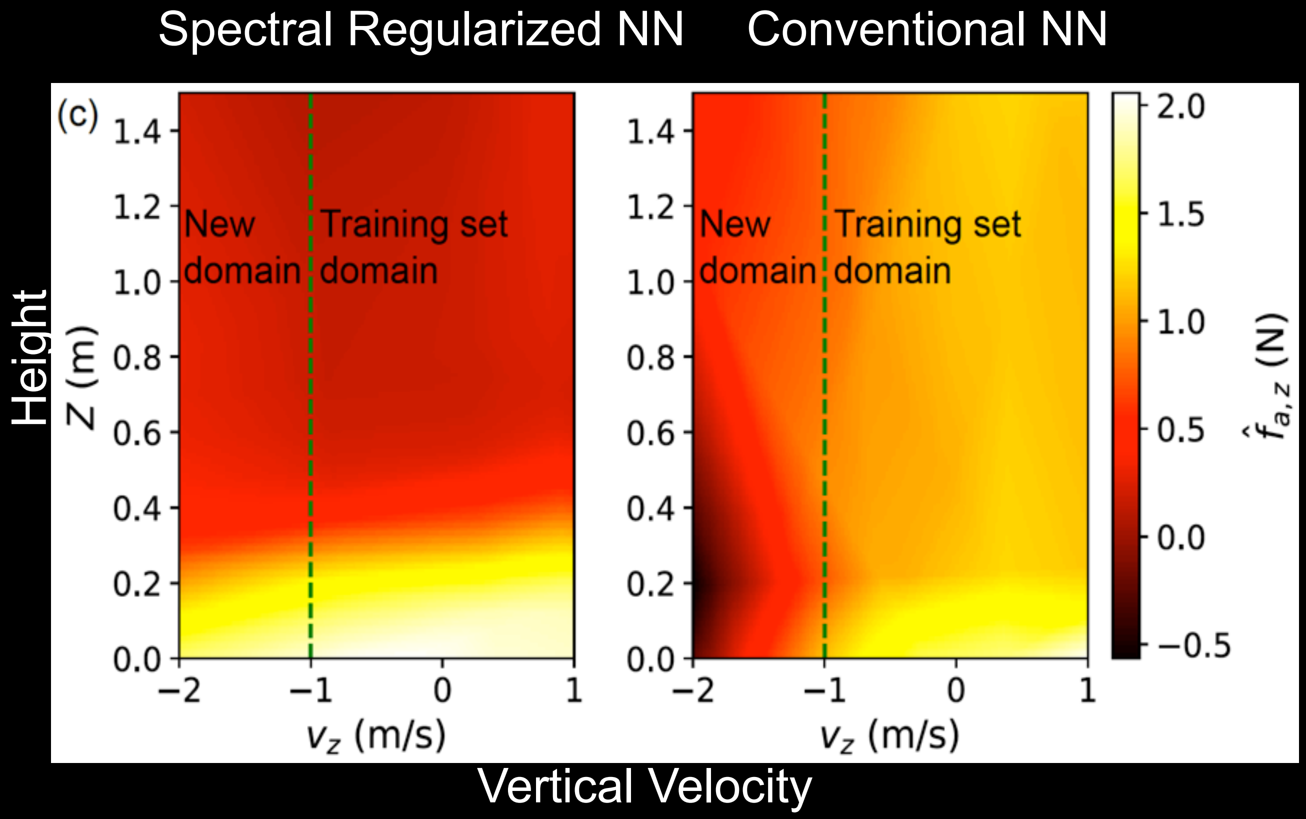
Current State

Unmodeled Disturbance

Use existing control methods to generate actions

- Provably robust (even using deep learning)
- Requires $\tilde{f}$ Lipschitz & bounded error

GENERALIZATION PERFORMANCE ON DRONE



Spectral Normalization:
Ensures $\widetilde{F}$ is Lipshitz
[Bartlett NeurIPS 2017]
[Miyato et al., ICLR 2018]

Spectral-Normalized
4-Layer Feed-Forward

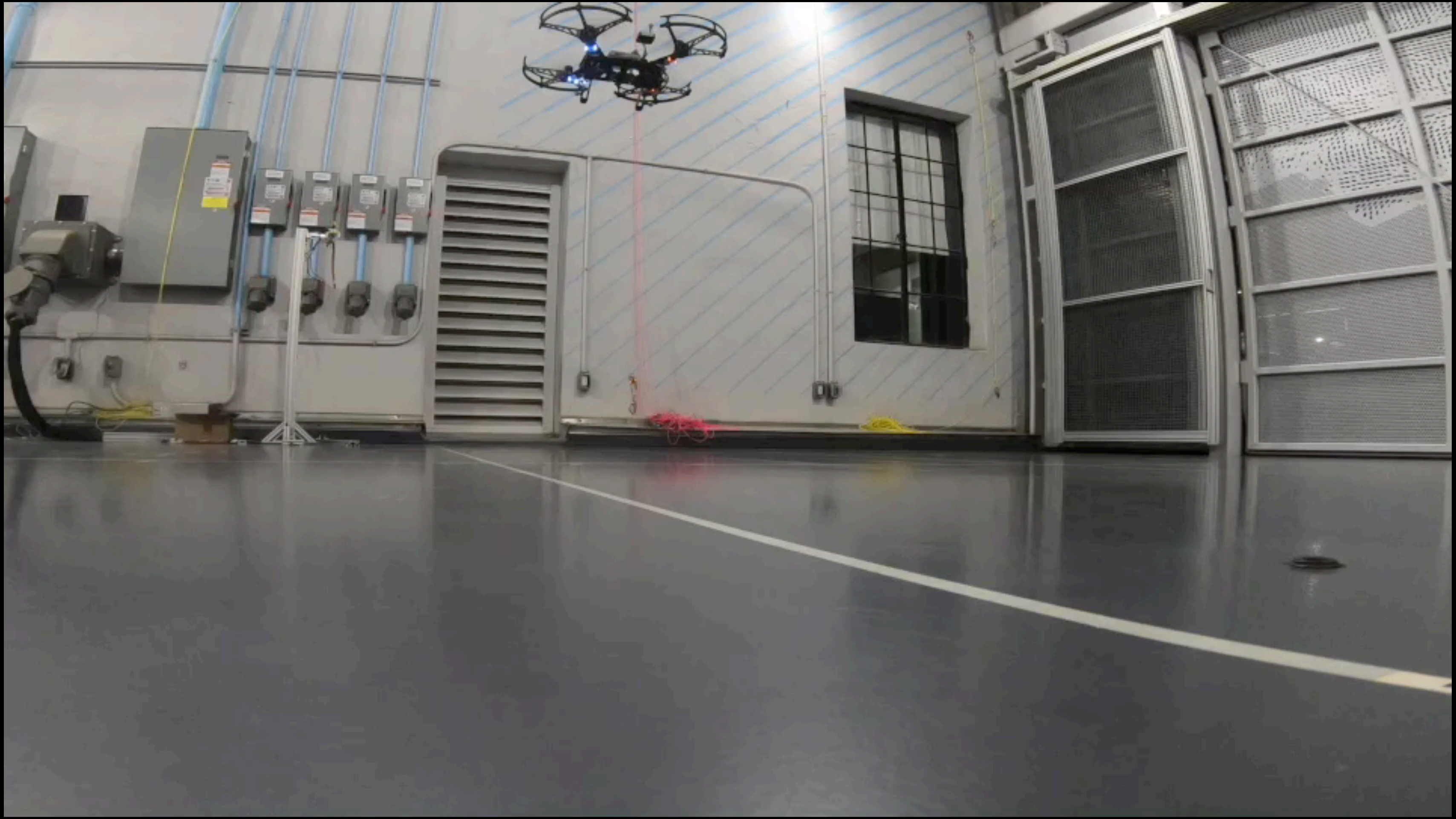
Ongoing Research:
Safe Exploration

CAST @ CALTECH
LEARNING TO LAND

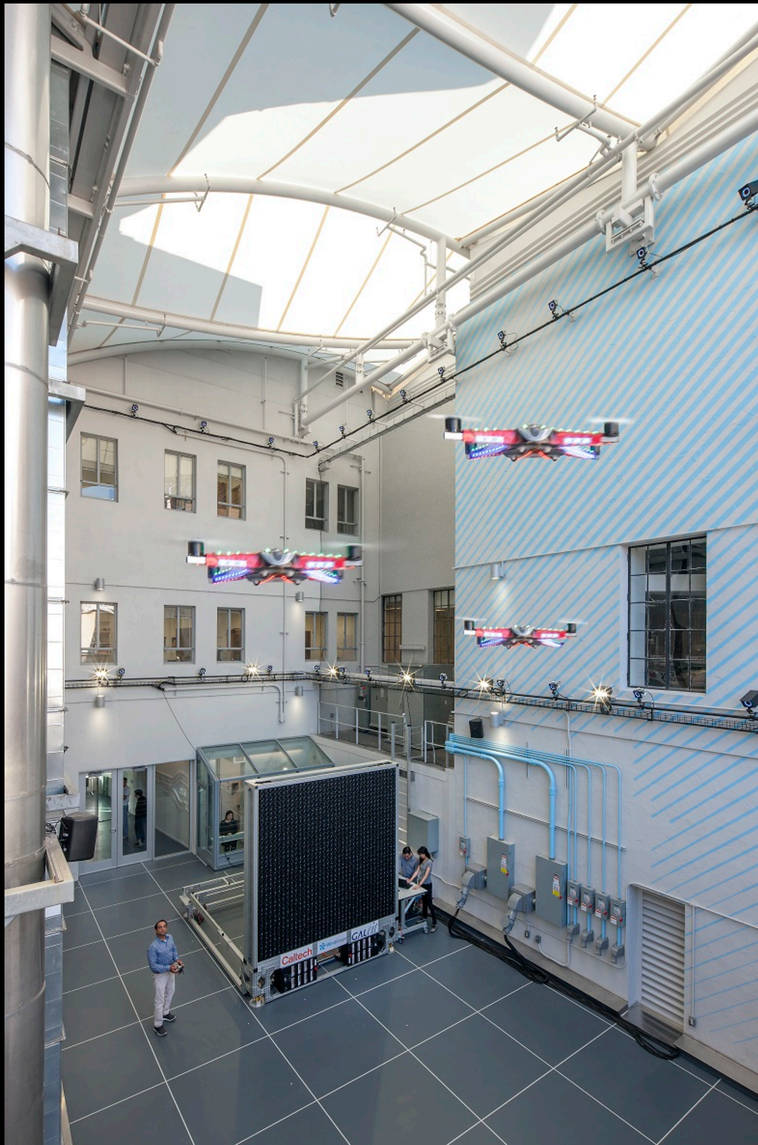
3D Landing Performance

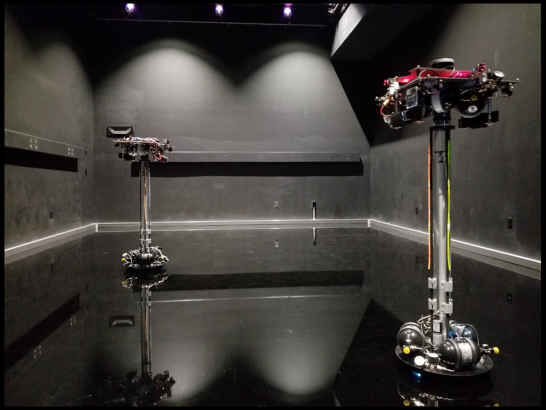
TESTING TRAJECTORY TRACKING

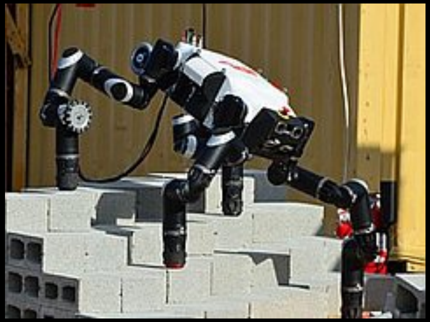
Move around a circle super close to the ground



AUTONOMOUS DYNAMIC ROBOTS





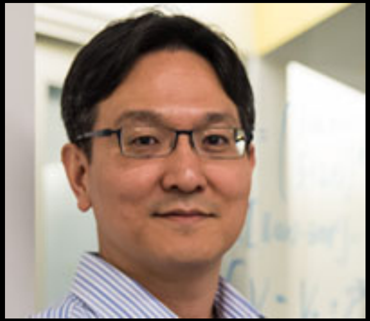


http://cast.caltech.edu

Postdoc Openings!

(applications considered starting January)

Mory Gharib



Soon-Jo Chung

Aaron Ames

Anima Anandkumar

Yisong Yue

Joel Burdick

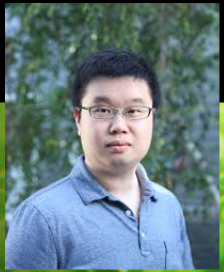
Katie Bouman

Pietro Perona

DETECTING VISUAL HARDNESS

Beidi Chen, Weiyang Liu, Animesh Garg, Zhiding Yu, Anshumali Shrivastava, Jan Kautz, Anima Anandkumar

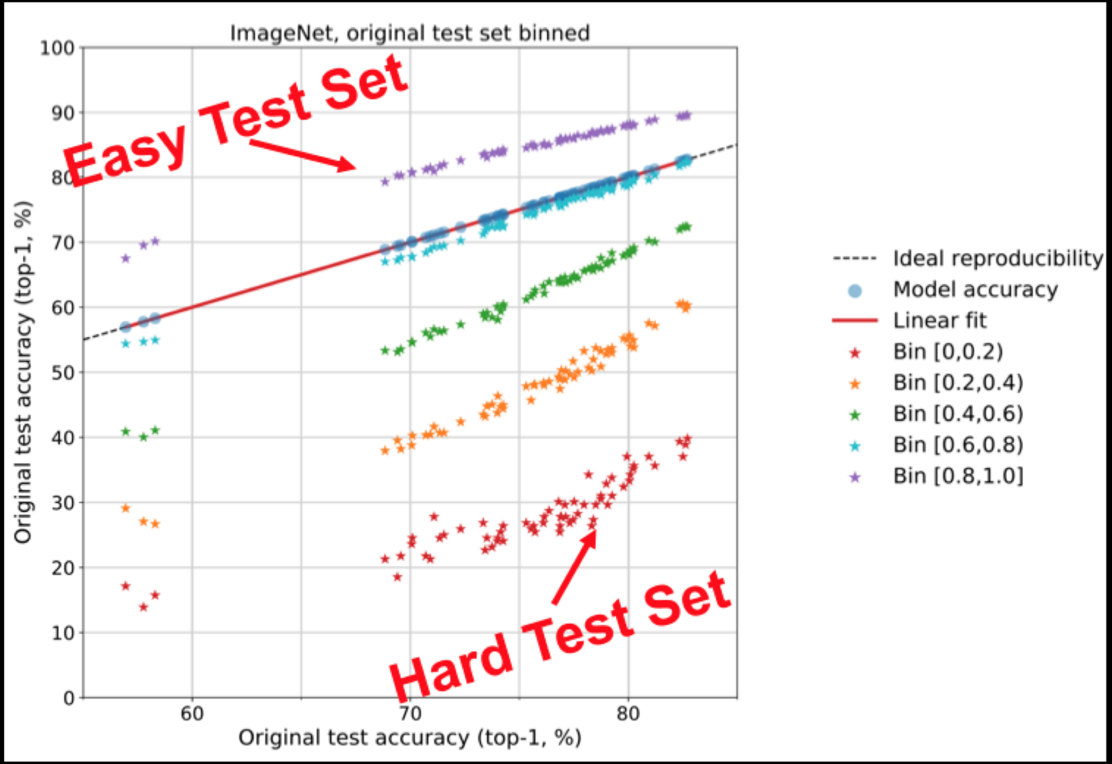



RECORDING HUMAN SELECTION FREQUENCIES

Selection frequency is a measure of human visual hardness + annotator bias

Human Labelling Interface



Recht et al. “Do ImageNet Classifiers Generalize to ImageNet?” ICML 2019

MODEL CONFIDENCE > 0.9 HUMAN SELECTION FREQUENCY < 0.5

MODEL CONFIDENCE < 0.5 HUMAN SELECTION FREQUENCY > 0.9

LOSS FUNCTION OF CNNS IN VISUAL RECOGNITION

- Softmax cross-entropy loss - one of the most popular loss functions in CNN

$$L = \frac{1}{N}\sum_i L_i = \frac{1}{N}\sum_i -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_j}}\right)$$

where,

$$L_i = -\log\left(\frac{e^{\boxed{\|\boldsymbol{W}_{y_i}\|\|\boldsymbol{x}_i\|}\boxed{\cos(\theta_{y_i})}}}{\sum_j e^{\|\boldsymbol{W}_j\|\|\boldsymbol{x}_i\|\cos(\theta_j)}}\right)$$

magnitude

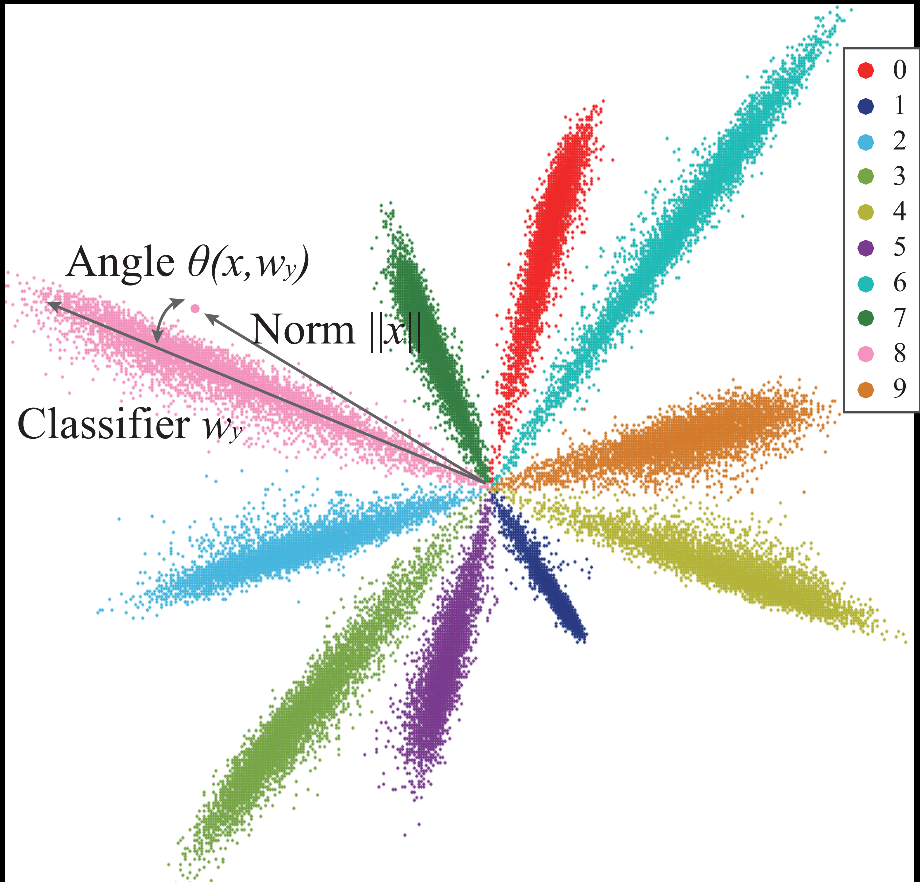
angle between feature and classifier

Model Confidence

NVIDIA.

2D FEATURE EMBEDDING ON MNIST

- Deeply learned features are naturally decoupled with angle and norm.
- The angle reflects the semantic difference.



BRIDGING THE GAP BETWEEN HUMAN VISUAL HARDNESS AND MODEL PREDICTIONS -- ANGULAR VISUAL HARDNESS

- Definition of angular visual hardness (AVH):

  Given a sample x with label y:

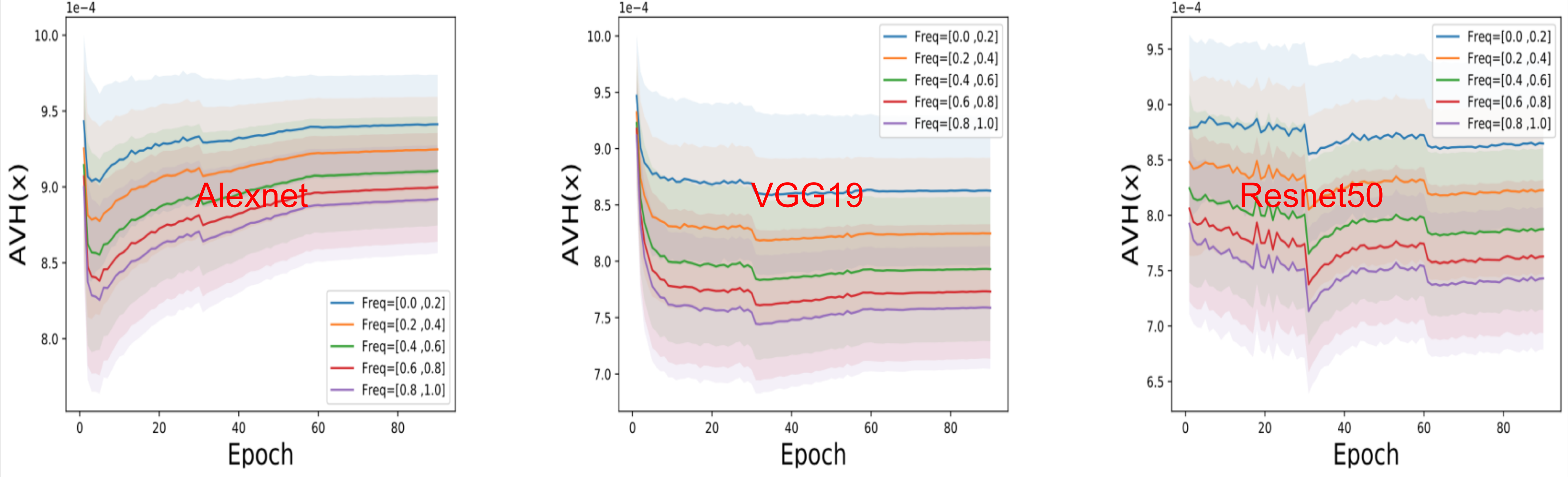
$$\boldsymbol{AVH}(x) = \frac{\mathcal{A}(x, w_y)}{\sum_{i=1}^{C} \mathcal{A}(x, w_i)}$$

  where,

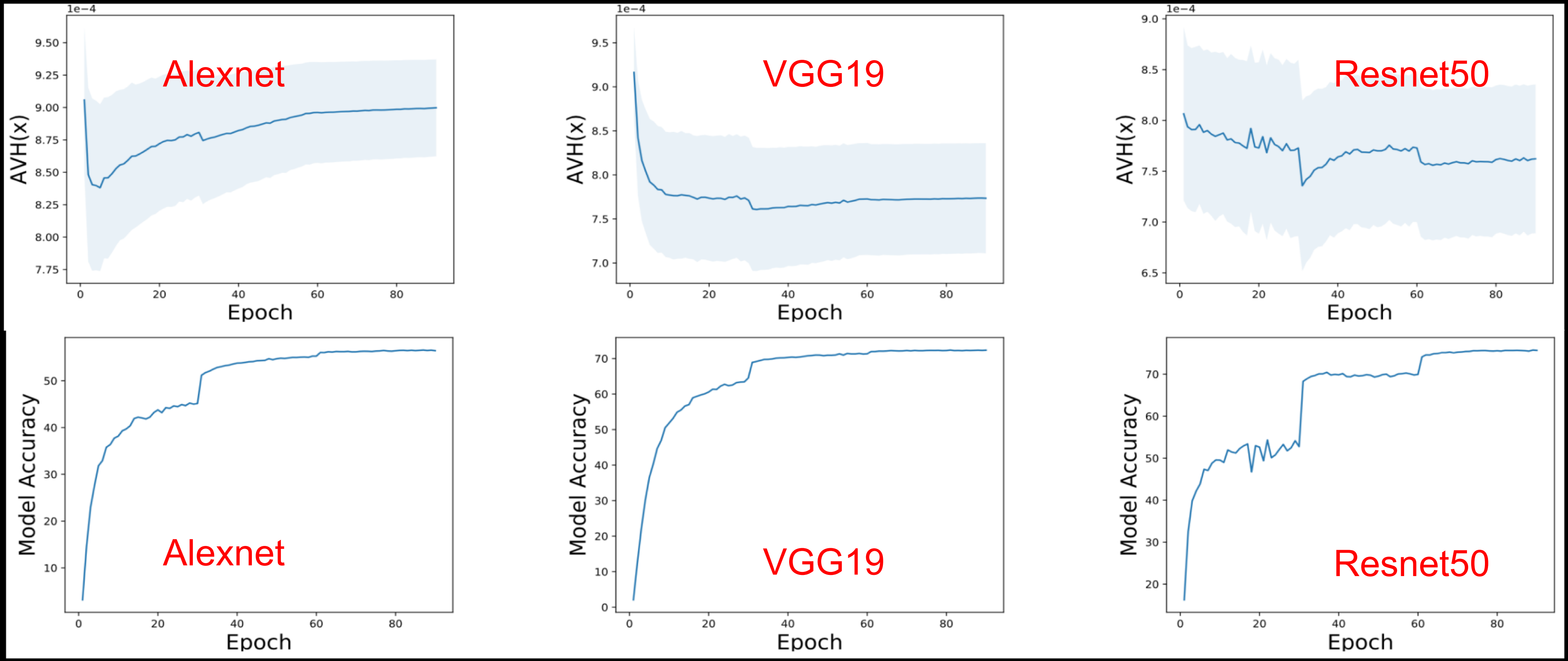
$$\mathcal{A}(\boldsymbol{u}, \boldsymbol{v}) = \arccos\left(\frac{\langle \boldsymbol{u}, \boldsymbol{v} \rangle}{\|\boldsymbol{u}\| \|\boldsymbol{v}\|}\right)$$

w_i is the classifier for the i-th class.

AVH IS A UNIVERSAL SCORE OF HARDNESS

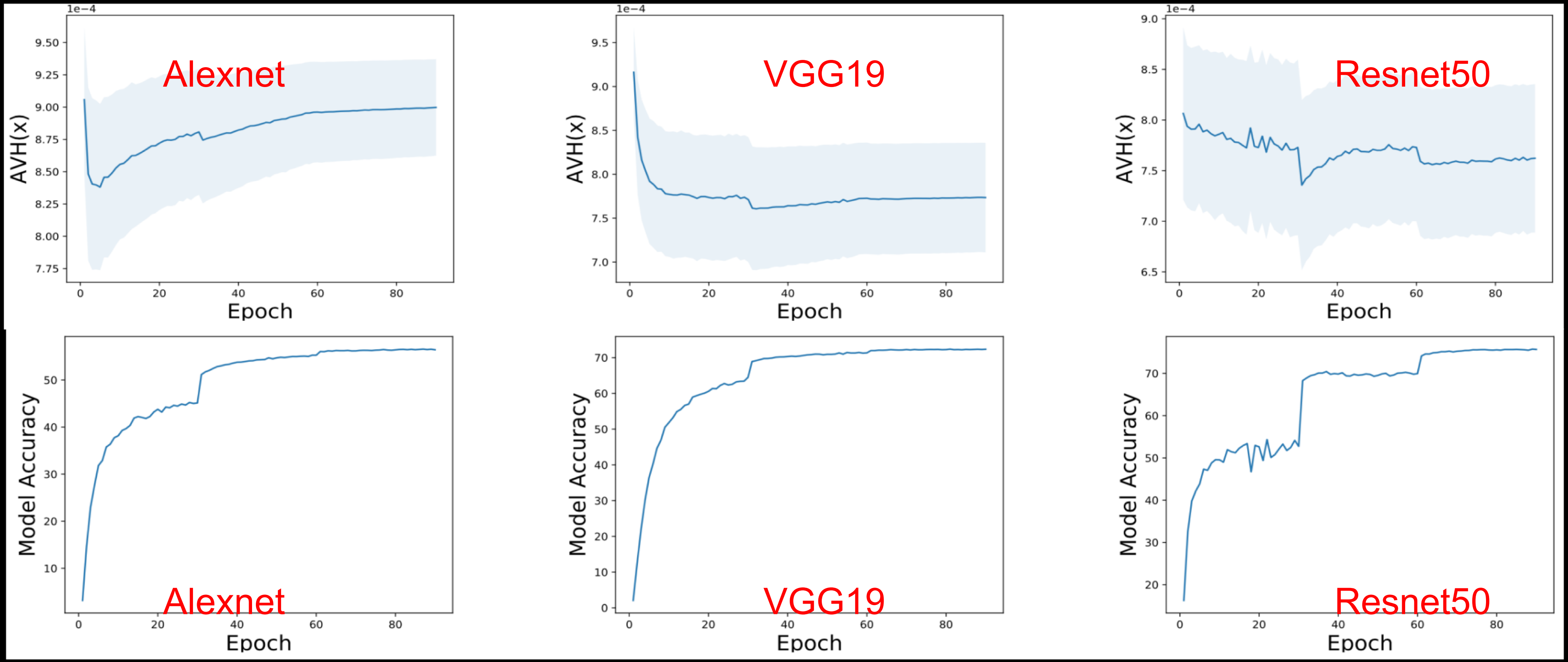


AVH IS AN INDICATOR OF MODEL'S GENERALIZATION ABILITY

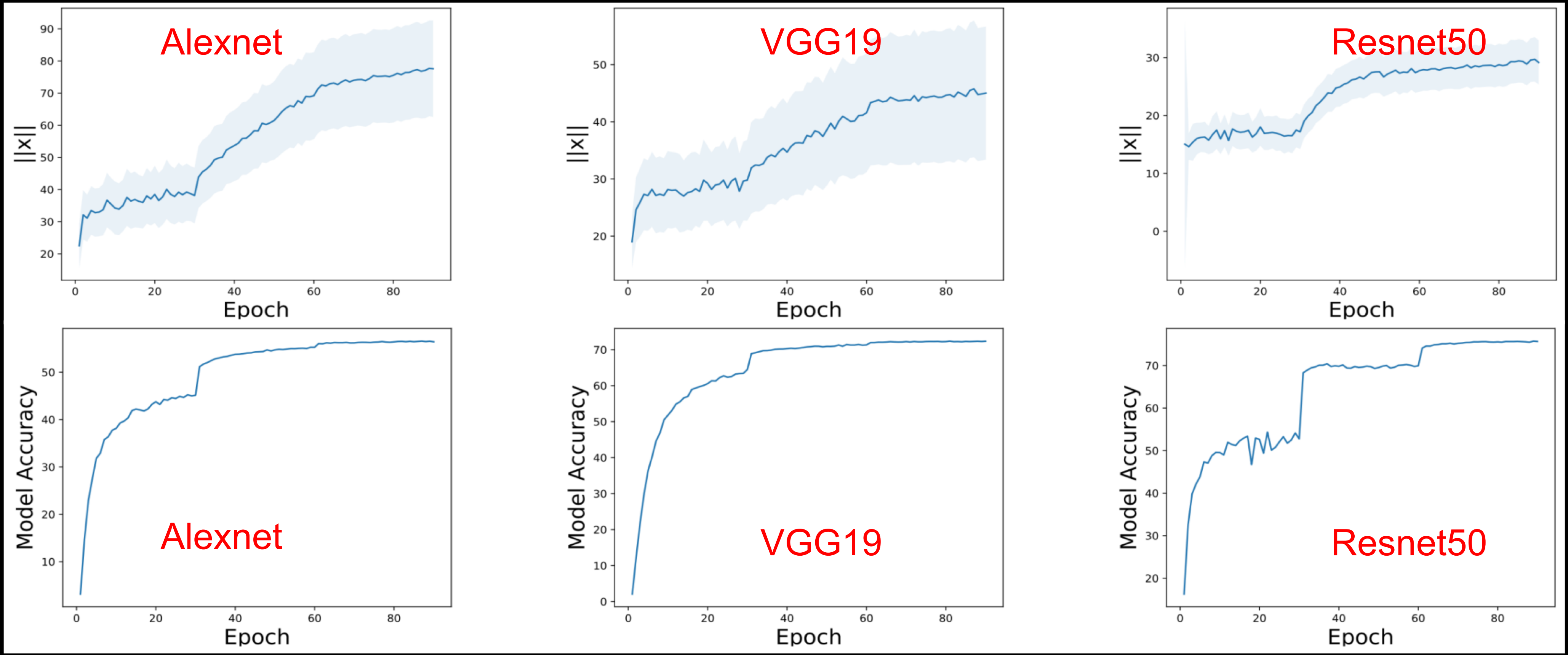


NVIDIA.

AVH HITS A PLATEAU VERY EARLY EVEN WHEN THE ACCURACY OR LOSS IS STILL IMPROVING



THE NORM OF FEATURE EMBEDDING KEEPS INCREASING DURING TRAINING



NVIDIA.

SOME CONJECTURES ABOUT TRAINING DYNAMICS

- Phase 1: Softmax cross-entropy loss first optimize angles among different classes while norm fluctuates and increases very slowly
- Phase 2: Angles become more stable and change slowly while norm increases rapidly
- Easy examples: Angles are matched well for correct classification
- Hard examples: Angles plateau and loss can only be improved by increasing norm

APPLICATION TO SELF-TRAINING

RESULTS ON VIS-DA 17

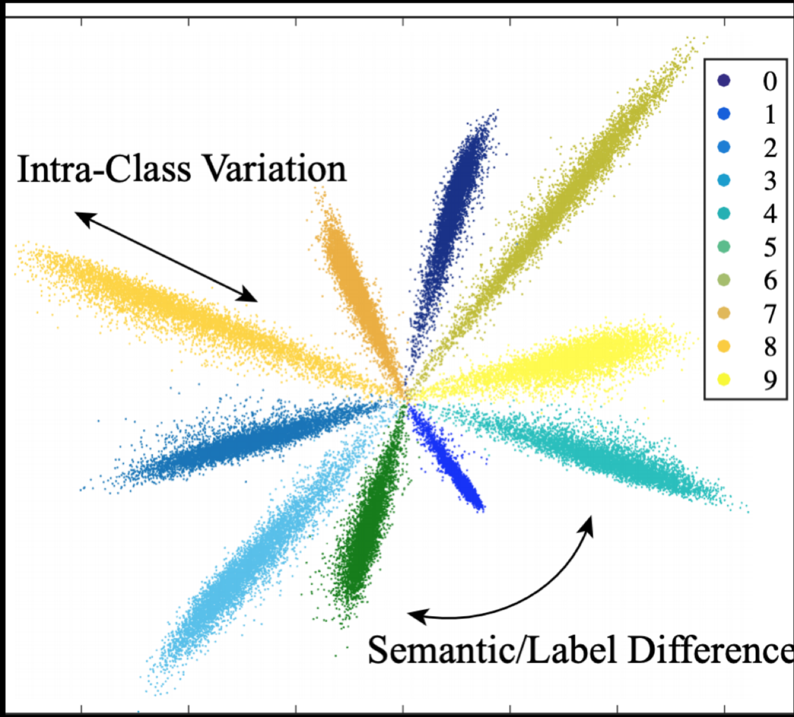
- Self-training sensitive to misclassified pseudo-labels
- Need good measure of hard examples

Method	Aero	Bike	Bus	Car	Horse	Knife	Motor	Person	Plant	Skateboard	Train	Truck	Mean
Source [51]	55.1	53.3	61.9	59.1	80.6	17.9	79.7	31.2	81.0	26.5	73.5	8.5	52.4
MMD [42]	87.1	63.0	76.5	42.0	90.3	42.9	85.9	53.1	49.7	36.3	**85.8**	20.7	61.1
DANN [16]	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
ENT [19]	80.3	75.5	75.8	48.3	77.9	27.3	69.7	40.2	46.5	46.6	79.3	16.0	57.0
MCD [50]	87.0	60.9	**83.7**	64.0	88.9	79.6	84.7	76.9	88.6	40.3	83.0	25.8	71.9
ADR [51]	87.8	79.5	**83.7**	65.3	**92.3**	61.8	**88.9**	73.2	87.8	60.0	85.5	32.3	74.8
Source [65]	68.7	36.7	61.3	**70.4**	67.9	5.9	82.6	25.5	75.6	29.4	83.8	10.9	51.6
CBST [65]	87.2	78.8	56.5	55.4	85.1	79.2	83.8	77.7	82.8	**88.8**	69.0	**72.0**	76.4
CRST [65]	88.0	79.2	61.0	60.0	87.5	81.4	86.3	78.8	85.6	86.6	73.9	68.8	78.1
Proposed	**93.3**	**80.2**	78.9	60.9	88.4	**89.7**	**88.9**	**79.6**	**89.5**	86.8	81.5	60.0	**81.5**

TAKE-AWAYS

- Angular distance (normalized) is a robust measure of human selection frequency, related to visual ambiguity.

- Application to self-training gives SOTA results



CONCLUSION

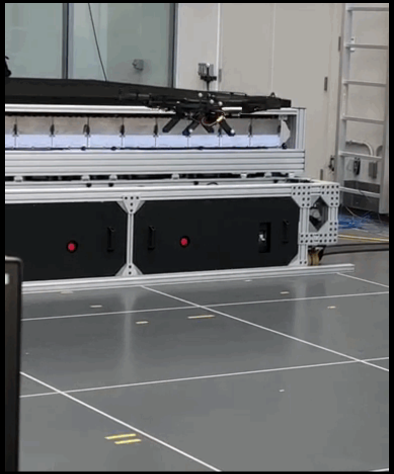
End-to-end learning from scratch is impossible in most settings

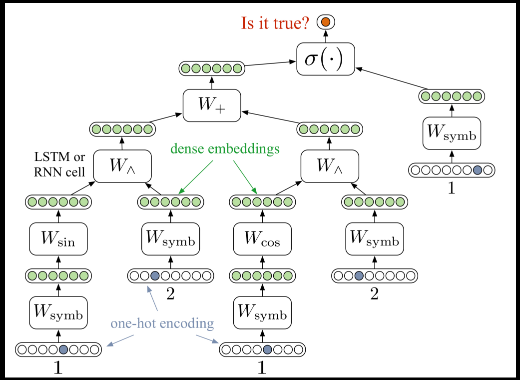
Blend DL w/ prior knowledge => improve data efficiency, generalization, model size

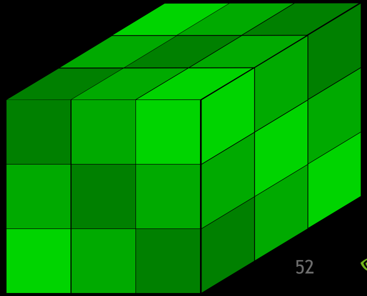
Obtain side guarantees like stability + safety in control

Outstanding challenge (application dependent):

what is right blend of prior knowledge vs data?







Thank you