Shawn Jain

Open AI

December 7, 2020

shawnjain.com

Origins

- Digital Cameras
- Computational Photography
- Wearable Cameras

Research

- Perception and Understanding (MIT)
- Understanding and Behavior (Uber ATG)

Recent Work

- Scale and Speed (MSR)

Origins

Digital Cameras

Computational Photography

Wearable Cameras

Research

Perception and Understanding (MIT)

Understanding and Behavior (Uber ATG)

Recent Work

Scale and Speed (MSR)

Digital Cameras

Digital Cameras

Origins

Digital Cameras

Computational Photography

Wearable Cameras

Research

Perception and Understanding (MIT)

Understanding and Behavior (Uber ATG)

Recent Work

Scale and Speed (MSR)

Origins

Digital Cameras

Computational Photography

Wearable Cameras

Research

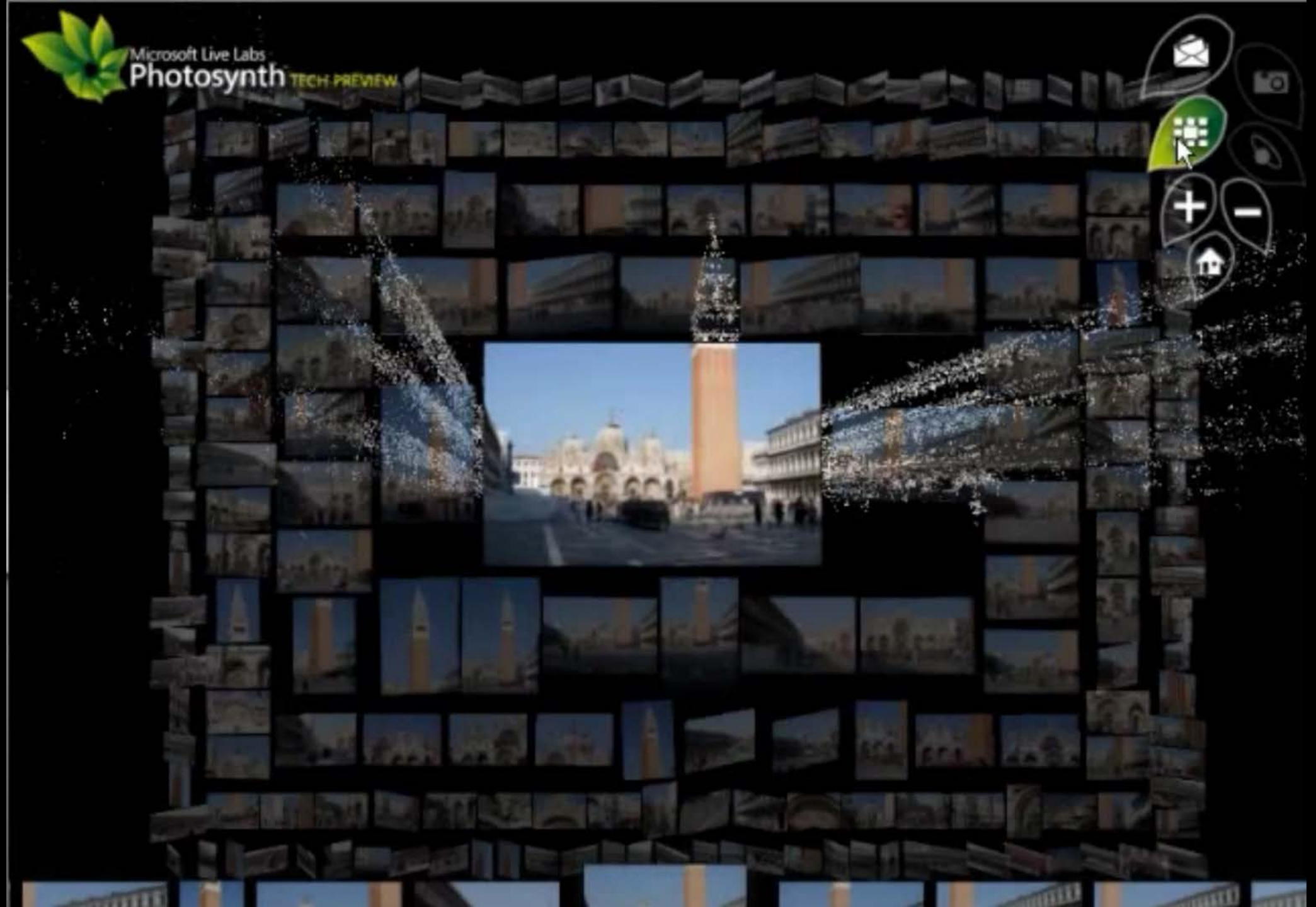
Perception and Understanding (MIT)

Understanding and Behavior (Uber ATG)

Recent Work

Scale and Speed (MSR)

Computational Photography



[1] [2]

Origins

Digital Cameras

Computational Photography

Wearable Cameras

Research

Perception and Understanding (MIT)

Understanding and Behavior (Uber ATG)

Recent Work

Scale and Speed (MSR)

Origins

Digital Cameras

Computational Photography

Wearable Cameras

Research

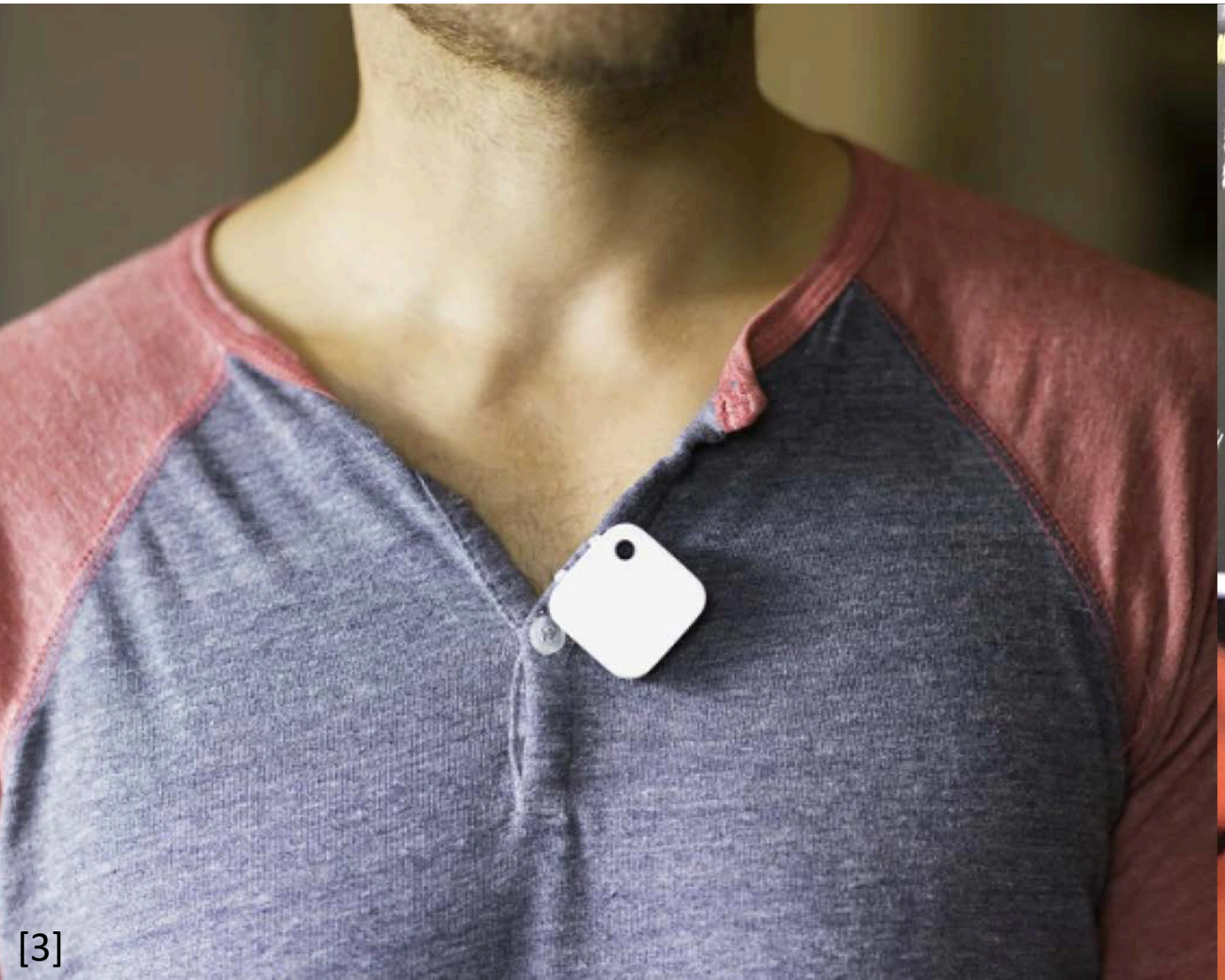
Perception and Understanding (MIT)

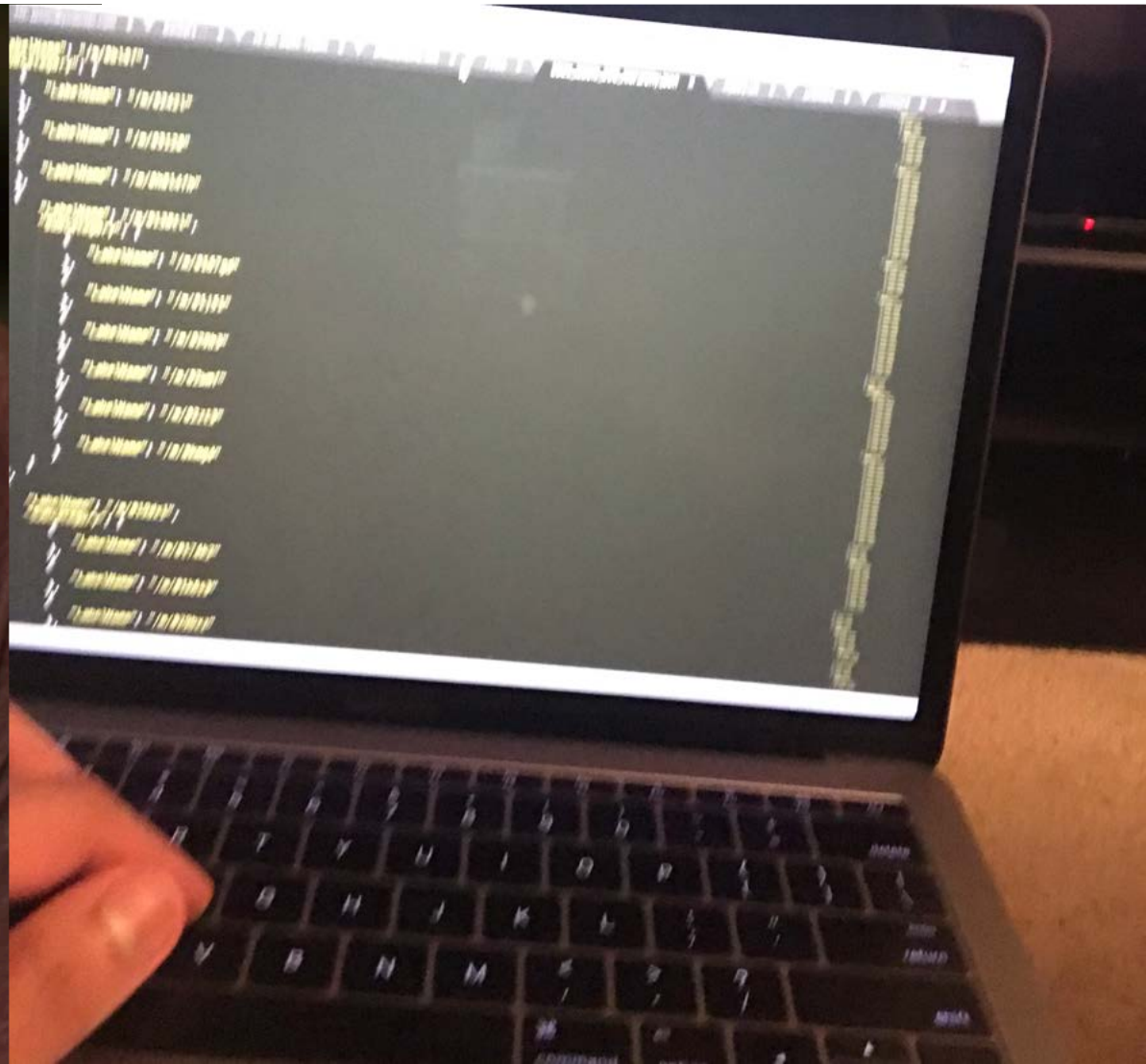
Understanding and Behavior (Uber ATG)

Recent Work

Scale and Speed (MSR)

Wearable Cameras - Lifelogging





[3]

Origins

Digital Cameras

Computational Photography

Wearable Cameras

Research

Perception and Understanding (MIT)

Understanding and Behavior (Uber ATG)

Recent Work

Scale and Speed (MSR)

Origins


Digital Cameras


Computational Photography


Wearable Cameras

Research


Perception and Understanding (MIT)

Understanding and Behavior (Uber ATG)

Recent Work

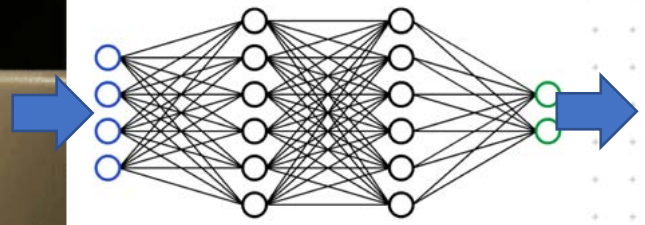
Scale and Speed (MSR)

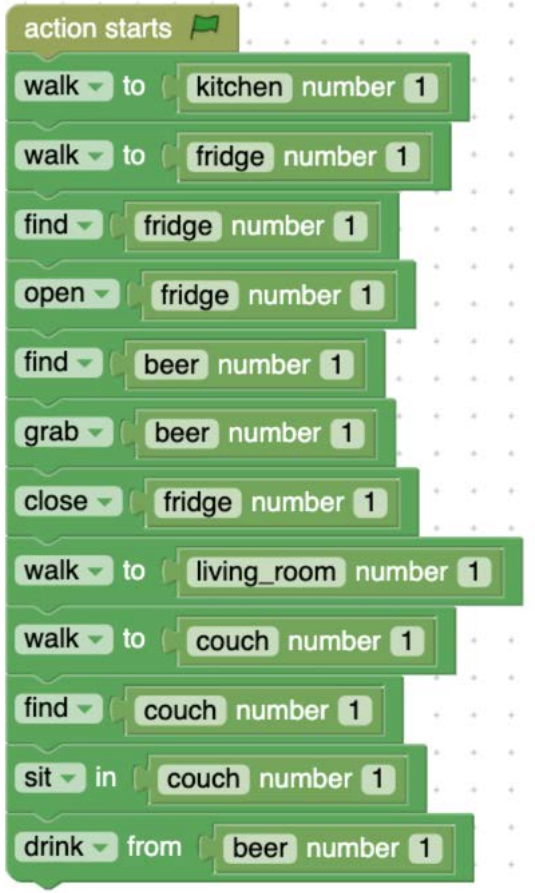
Perception and Understanding

Collaborators: X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, A. Torralba

[5]

Perception and Understanding





[5]

Collecting Household Activities

1. Name a Household Activity:

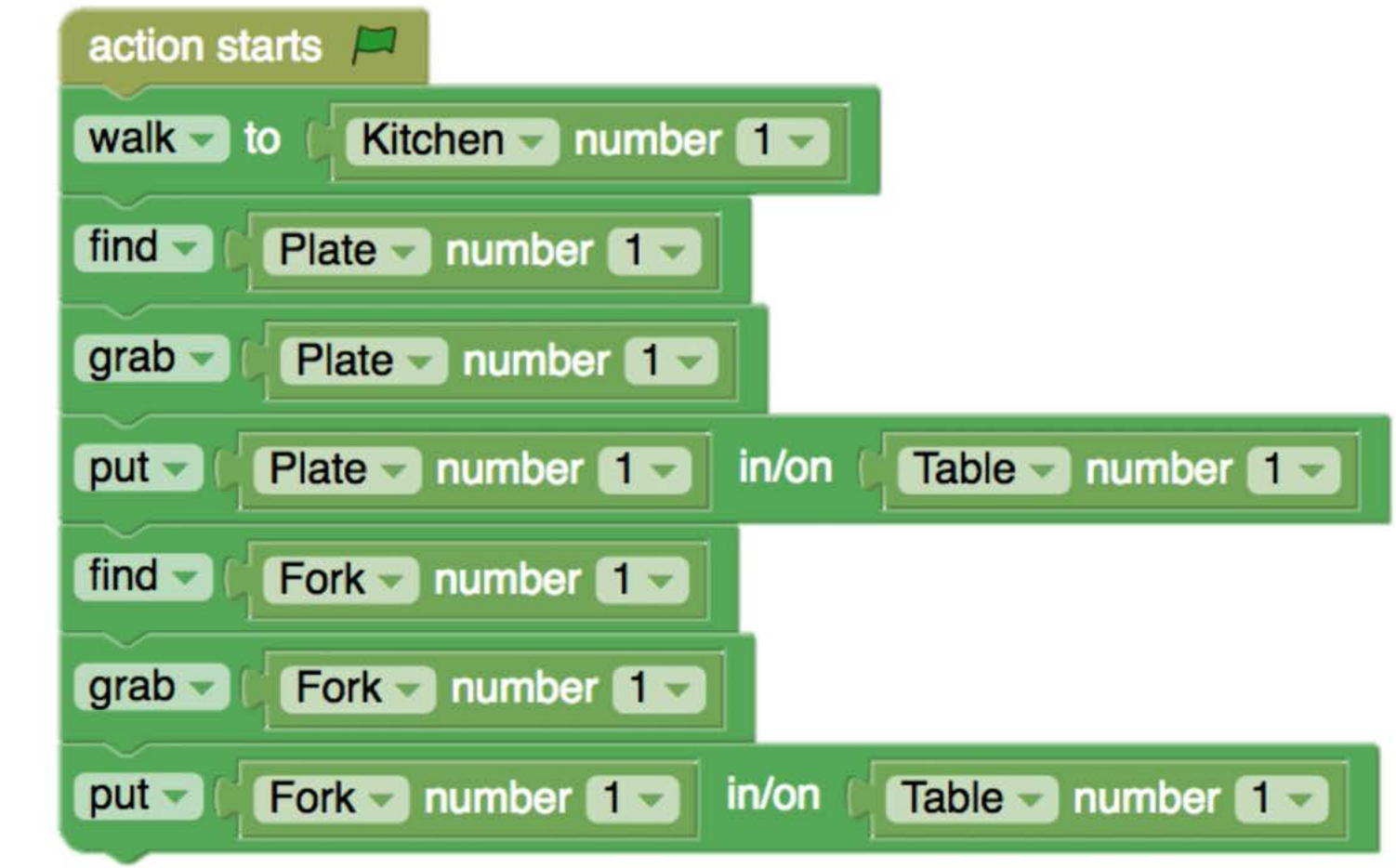
Set the table

2. Give a Description:

Go to the kitchen, find the tableware, and put it onto the table.

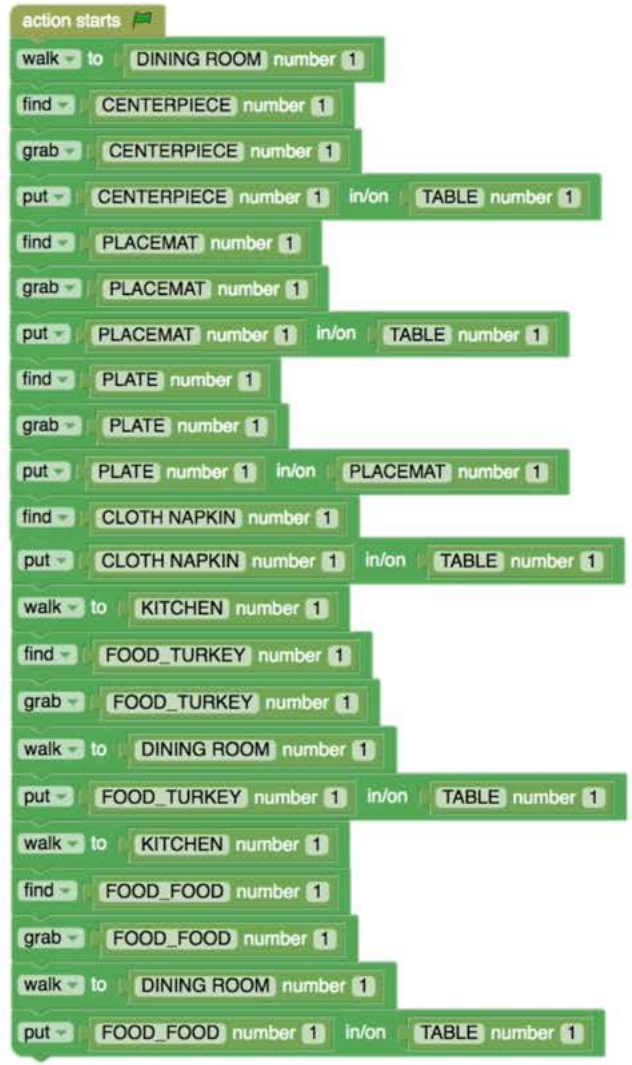
3. Build a program:

Robots wikiHow set up table



thanksgiving

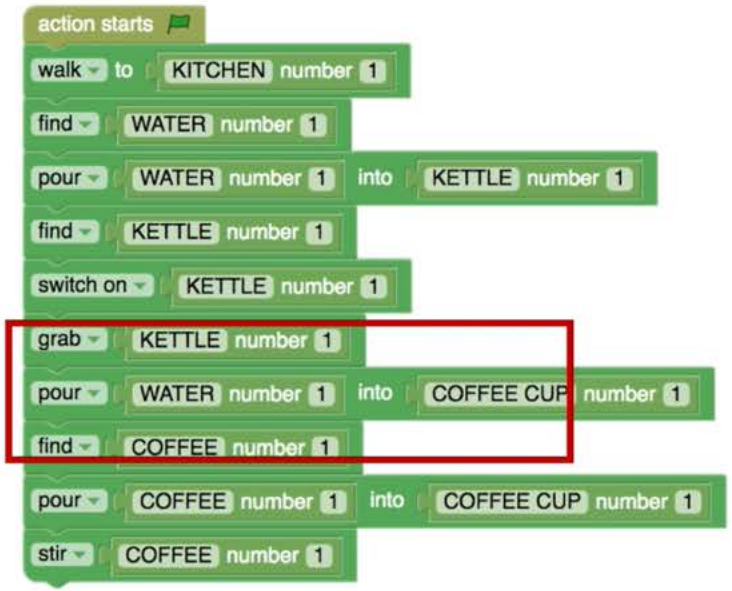
Description: Set out centerpieces, fancy china, etc. Then, bring out turkey and rest of food.



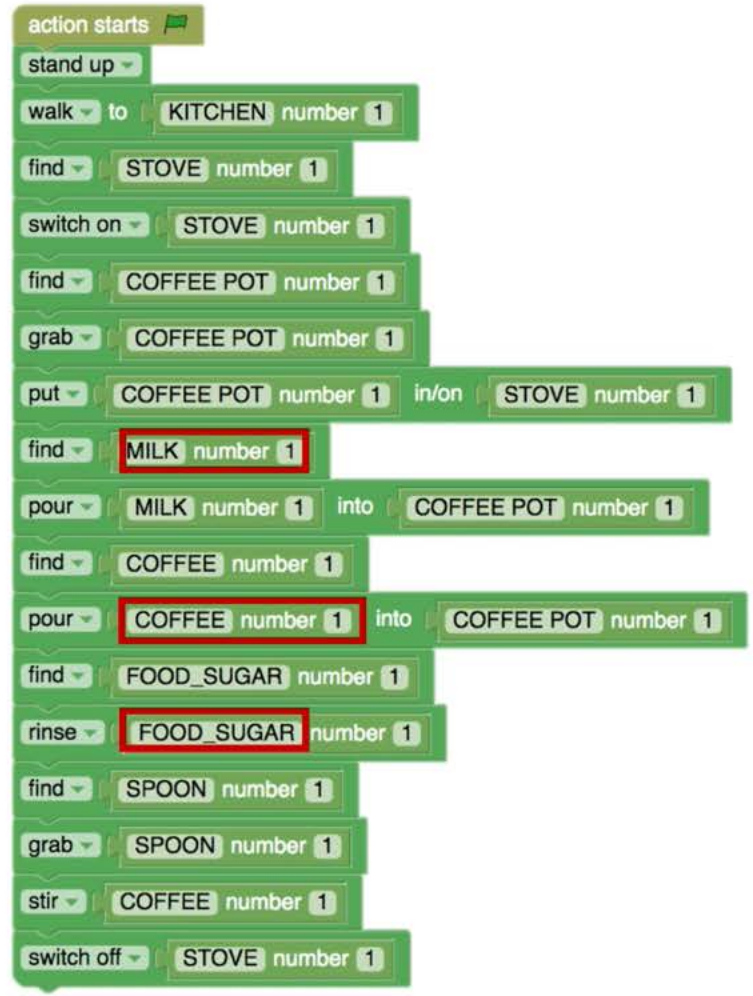
make coffee

Coffee Maker

Kettle



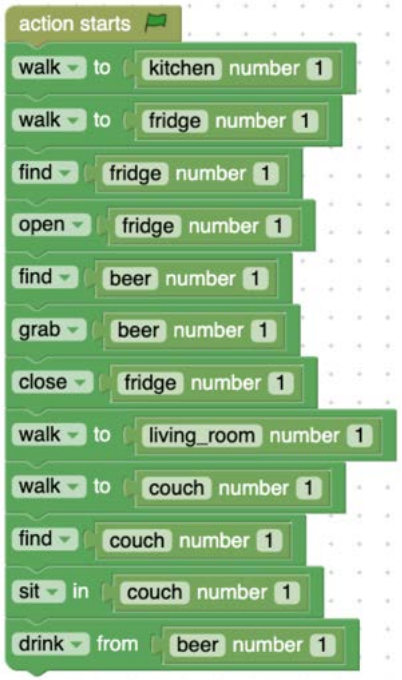
Sugar and Milk



Generating videos from programs

Simulator

Generating videos from programs



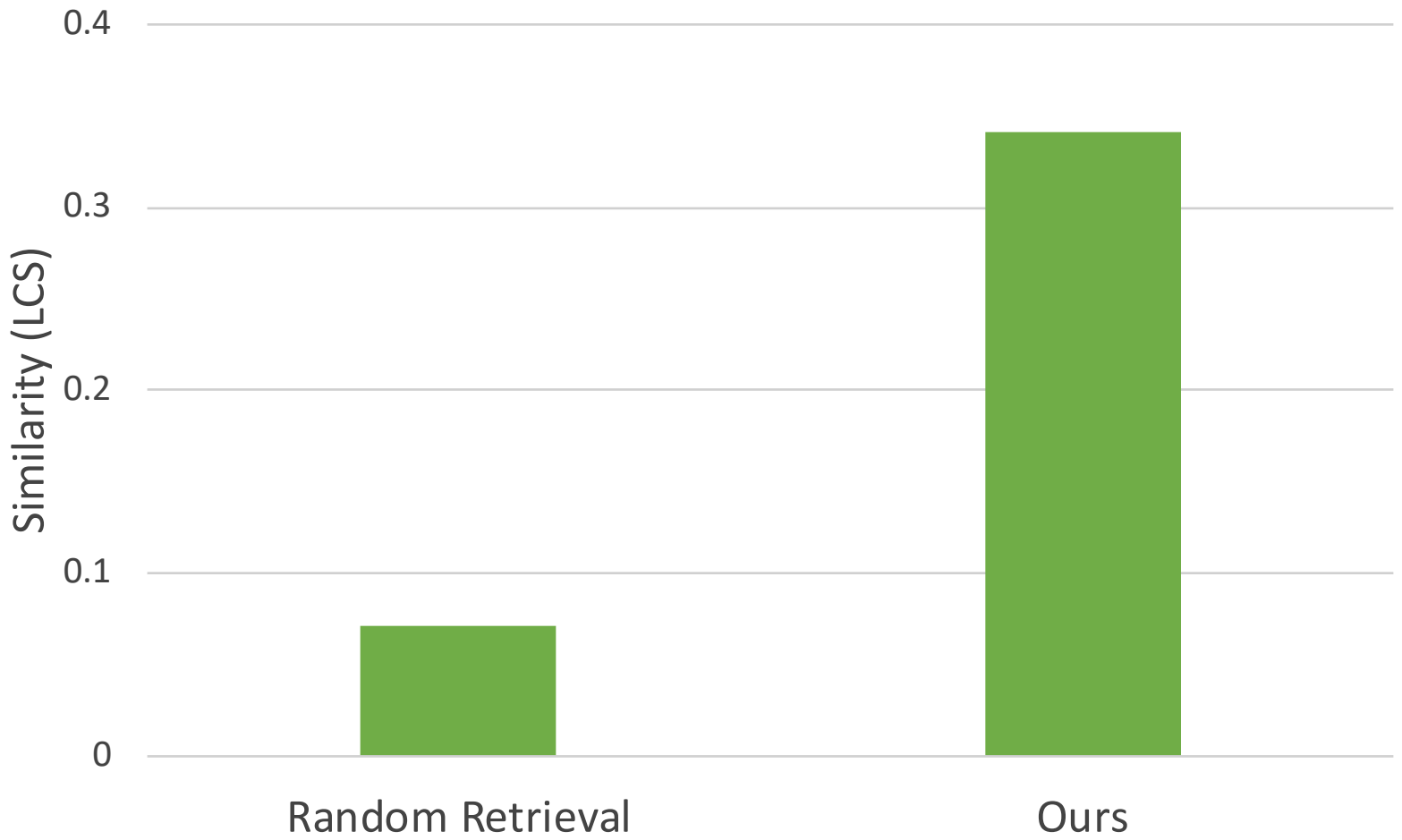
Program

Simulator

Video

[19]

Model Performance



Origins


Digital Cameras


Computational Photography


Wearable Cameras

Research


Perception and Understanding (MIT)

Understanding and Behavior (Uber ATG)

Recent Work

Scale and Speed (MSR)

Origins


Digital Cameras


Computational Photography


Wearable Cameras

Research


Perception and Understanding (MIT)


Understanding and Behavior (Uber ATG)

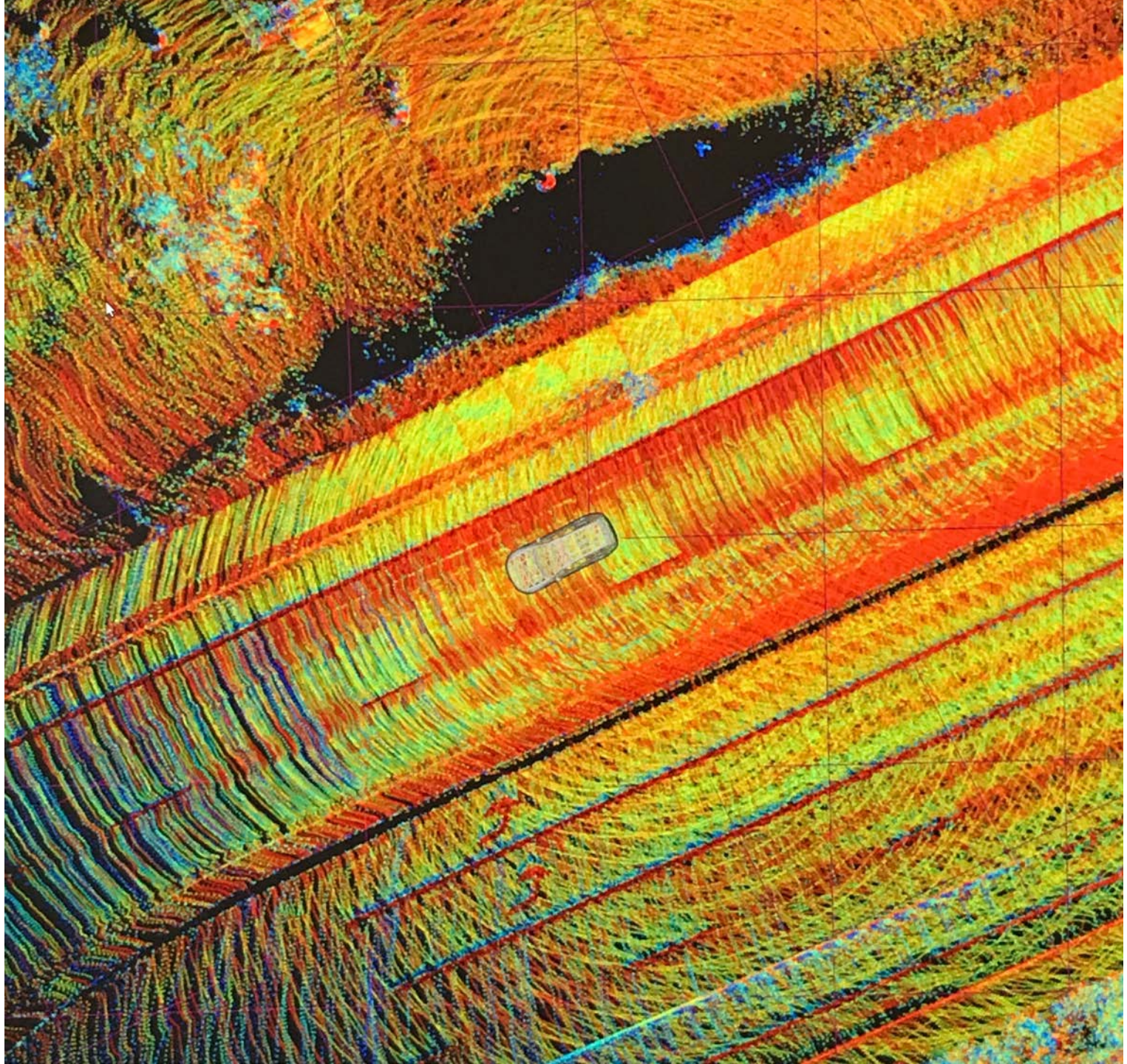
Recent Work

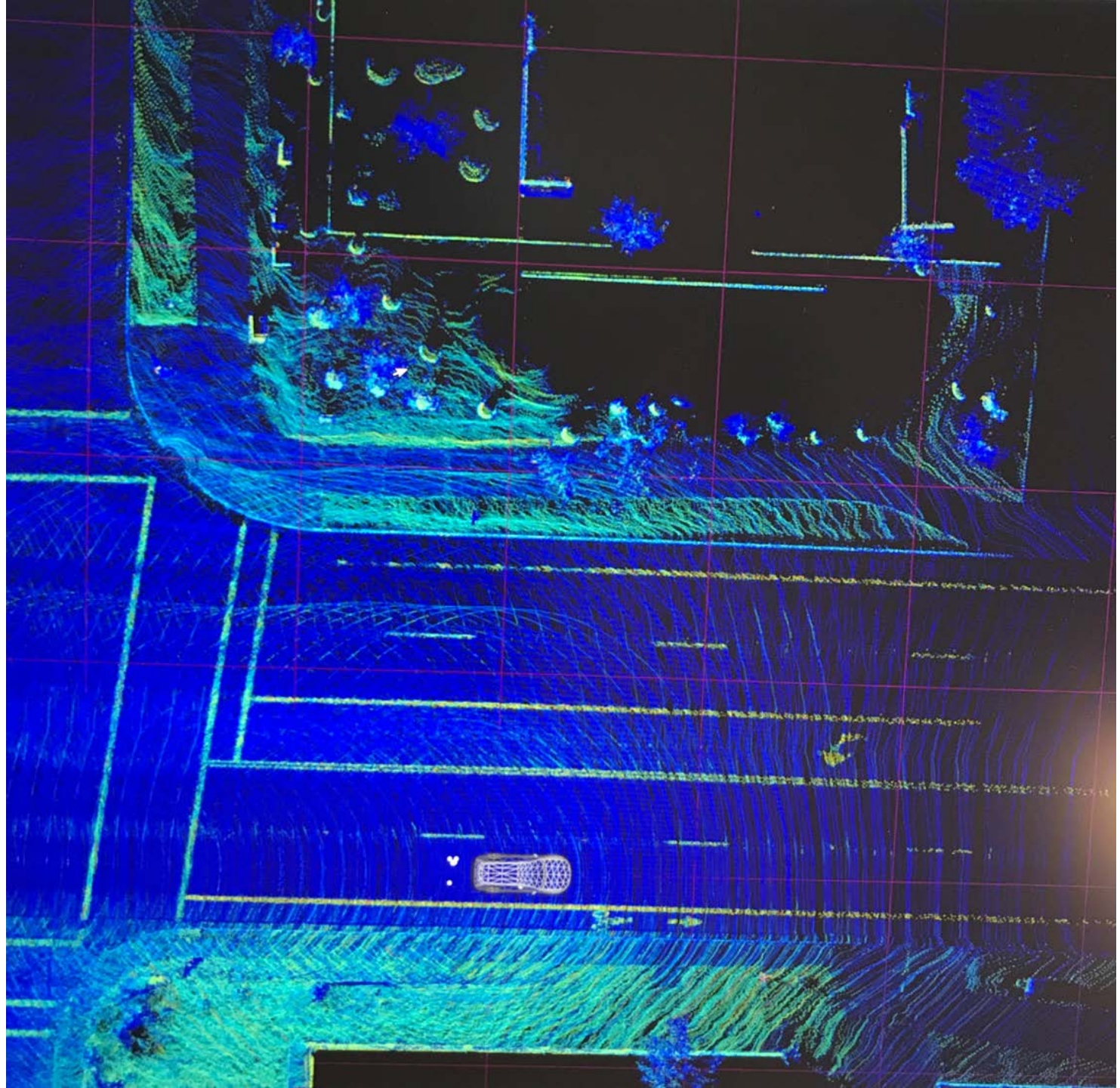
Scale and Speed (MSR)

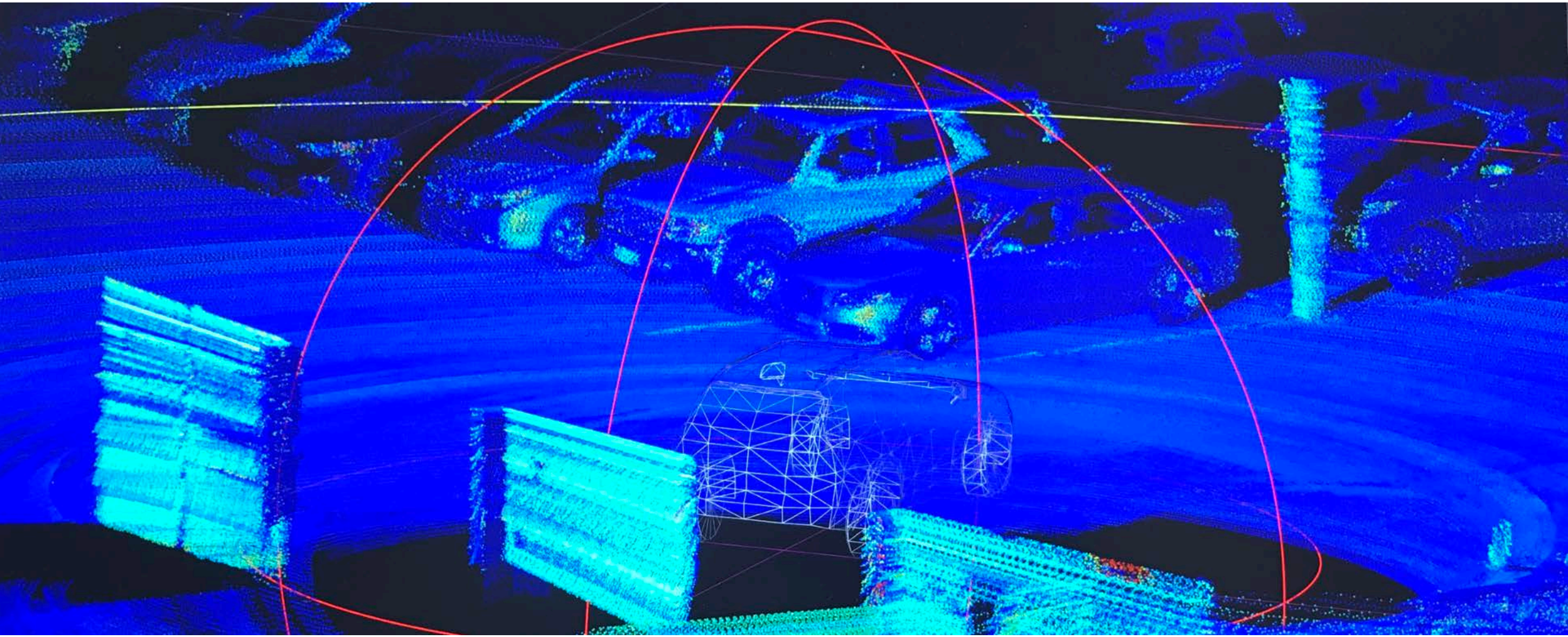
Understanding and Behavior

Uber ATG


Velodyne







Understanding and Behavior

Patent US 10,598,791 B2

Research + Engineering

[6]

Origins


Digital Cameras


Computational Photography


Wearable Cameras

Research


Perception and Understanding (MIT)


Understanding and Behavior (Uber ATG)

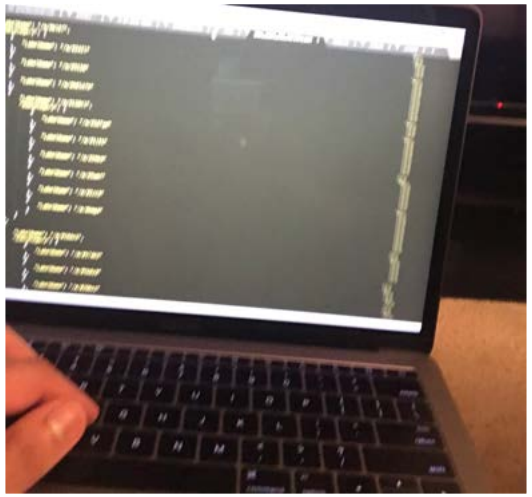
Recent Work

Scale and Speed (MSR)

Scaling Up


VirtualHome
About Docs RobotHow GitHub
VirtualHome
A Multi-Agent Household Simulator

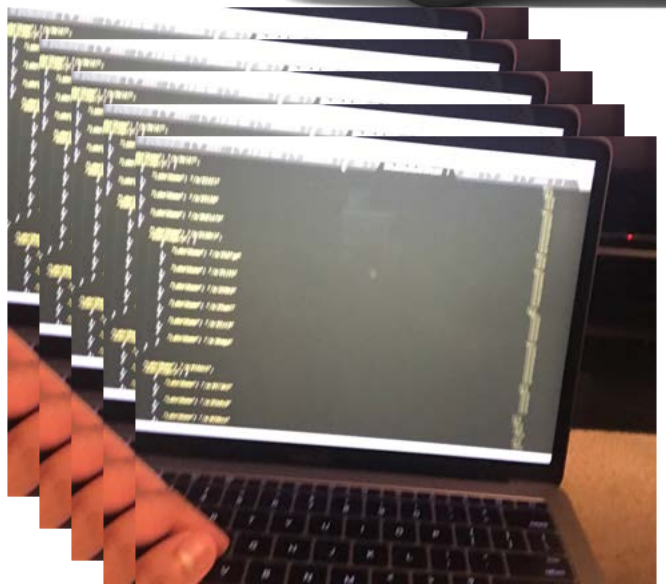

Uber



Scaling Up


VirtualHome
About Docs RobotHow GitHub
VirtualHome
A Multi-Agent Household Simulator


Uber



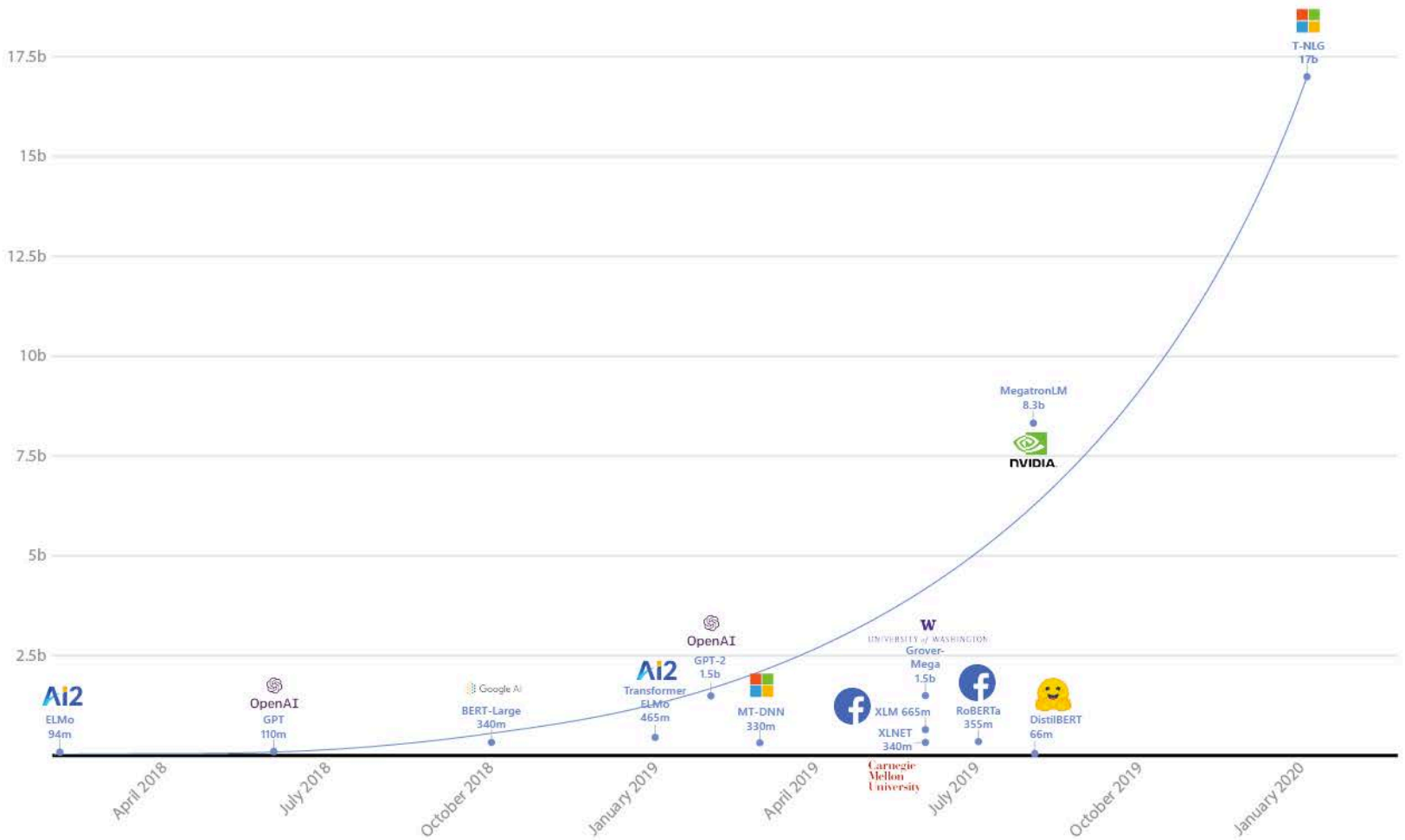
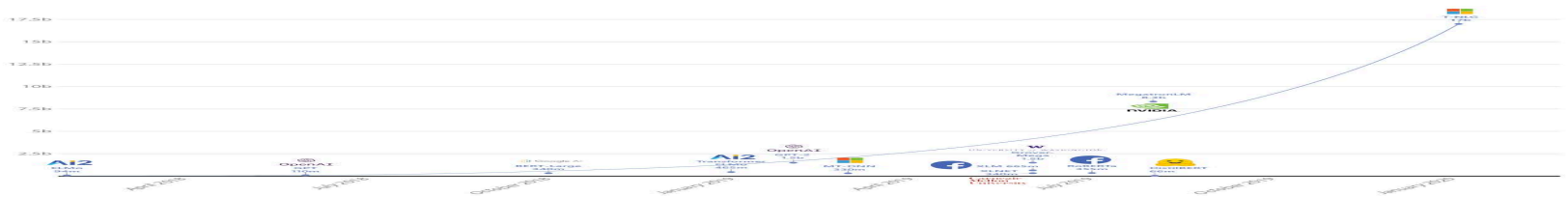


Figure from [17]

OpenAI
GPT3
175B



Adapted from [17]

Origins


Digital Cameras


Computational Photography


Wearable Cameras

Research


Perception and Understanding (MIT)


Understanding and Behavior (Uber ATG)

Recent Work

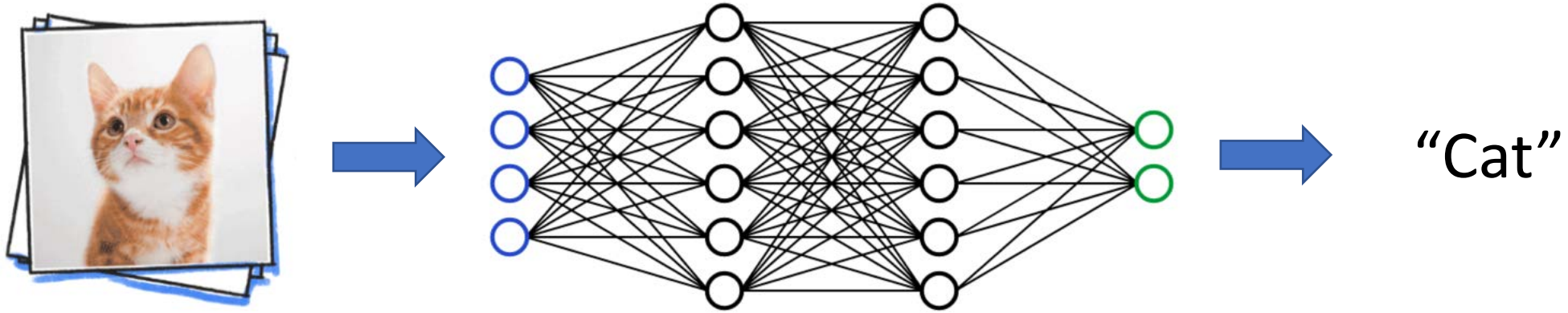

Scale and Speed (MSR)

Scaling Up > Speed Up

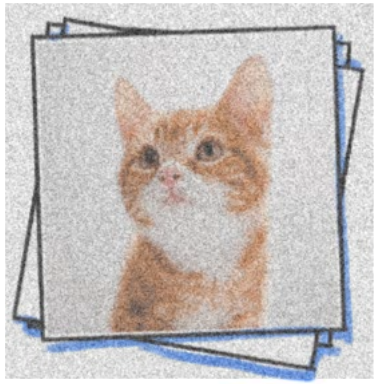
GPT3

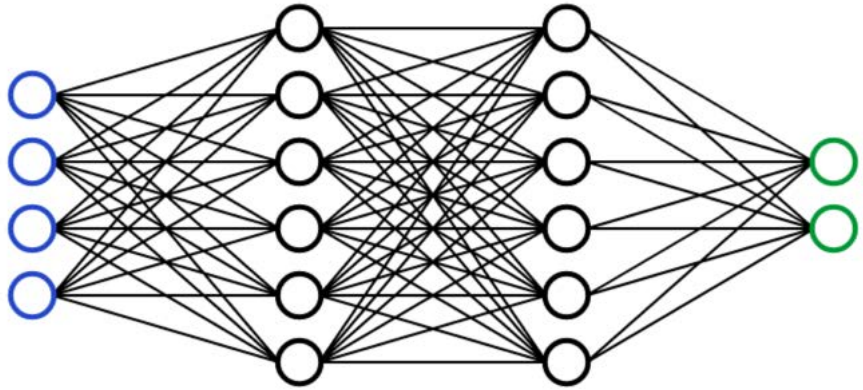
A general purpose method to speed up training and inference?

Collaborator: Greg Yang, Microsoft Research


"Cat"

+ Input Noise

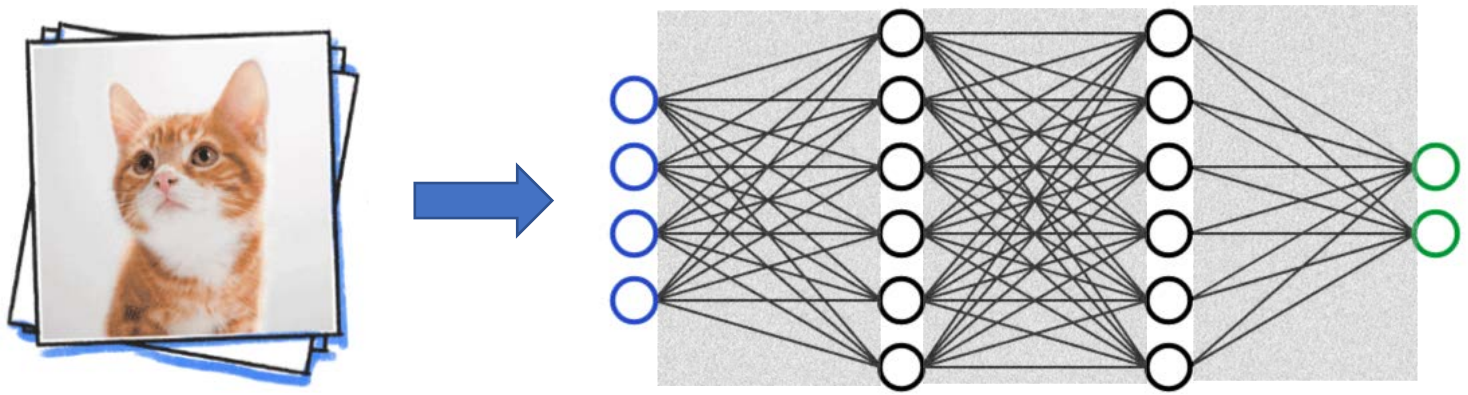




“Cat”

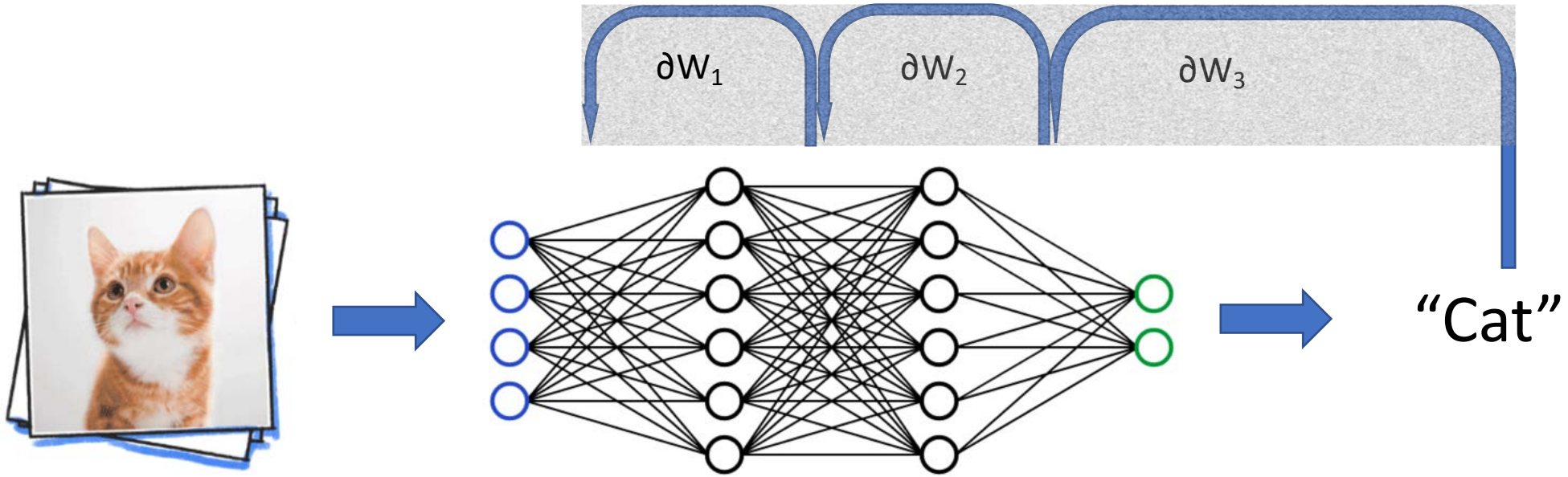
[7] [8] [9]

+ Weight Noise



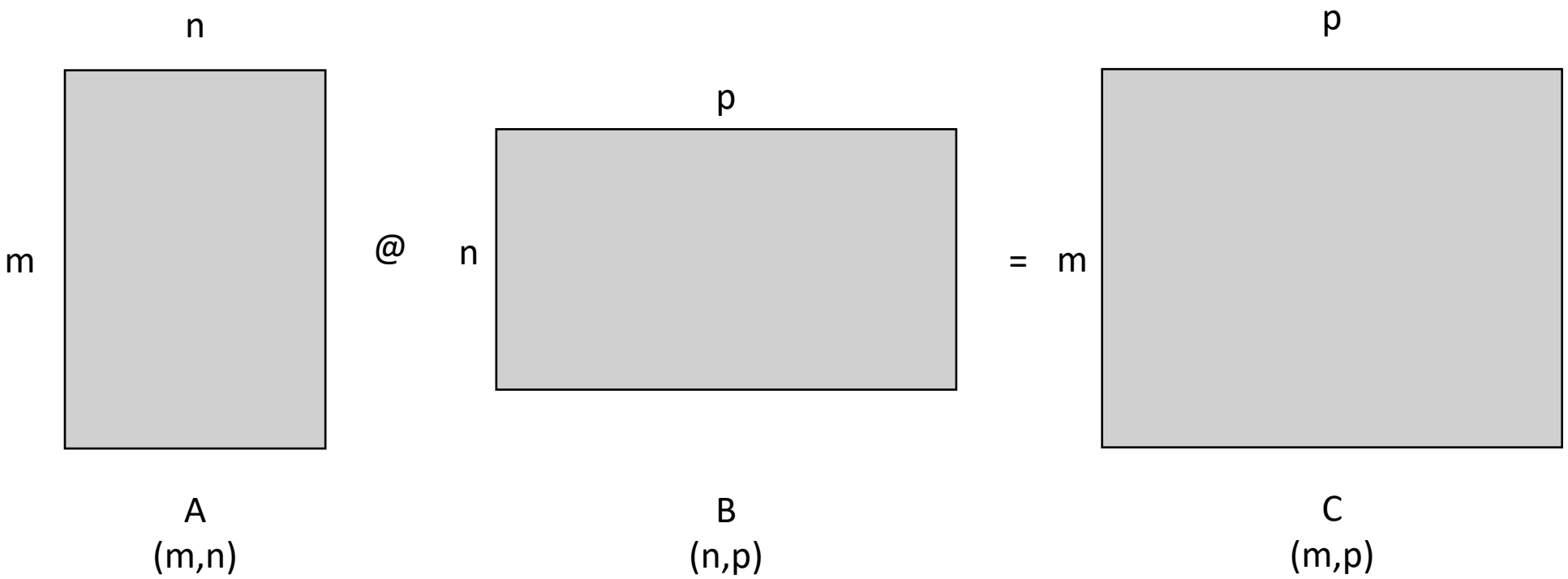
[7] [8] [9]

+ Gradient Noise

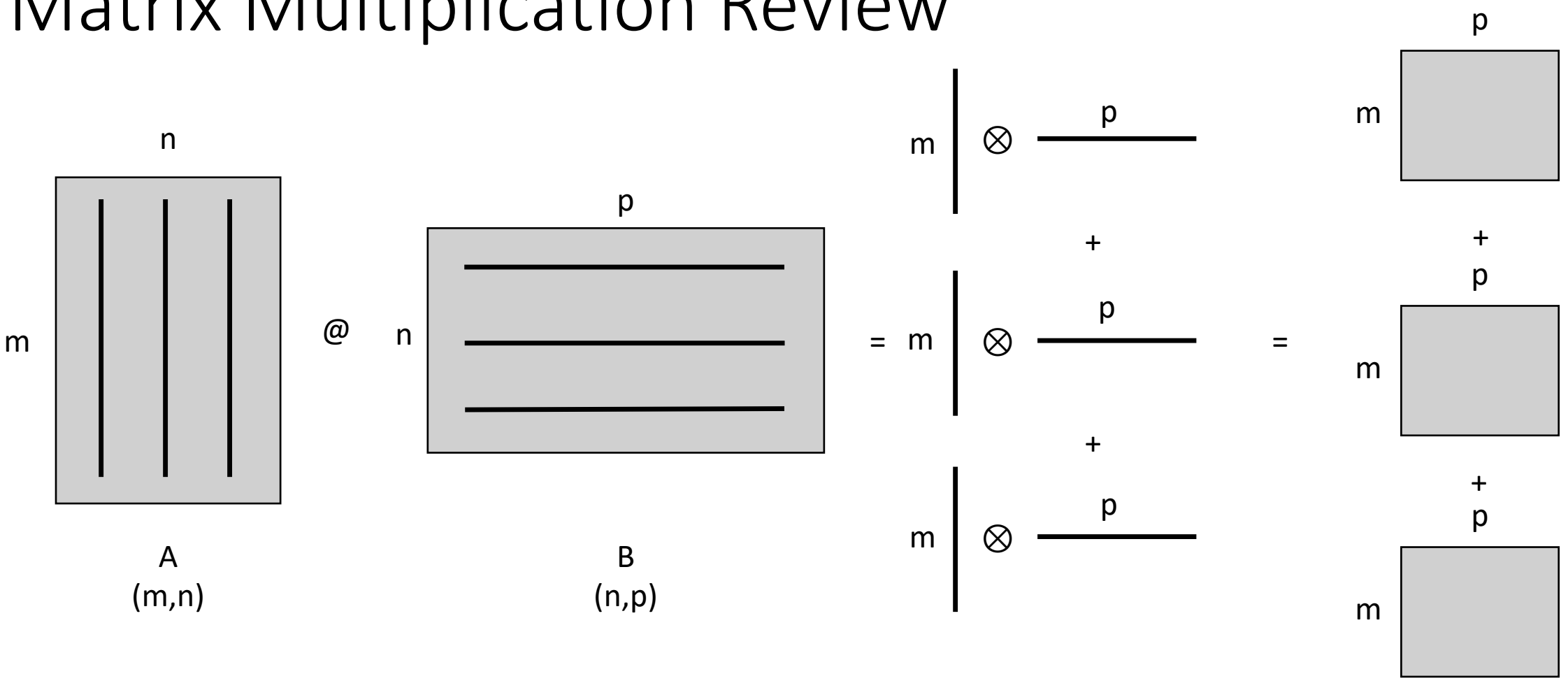


[7] [8] [9]

Matrix Multiplication Review

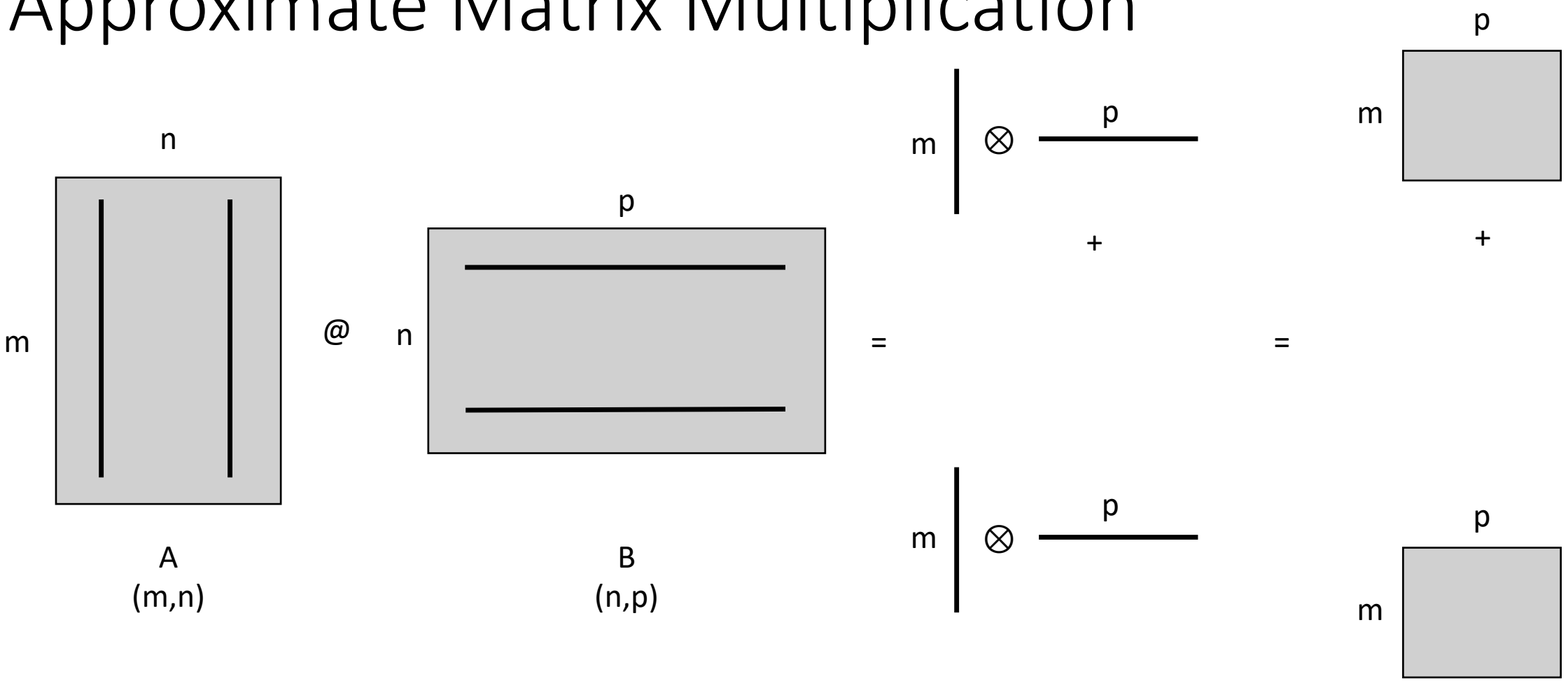


Matrix Multiplication Review



Runtime:
O(mnp)

Approximate Matrix Multiplication



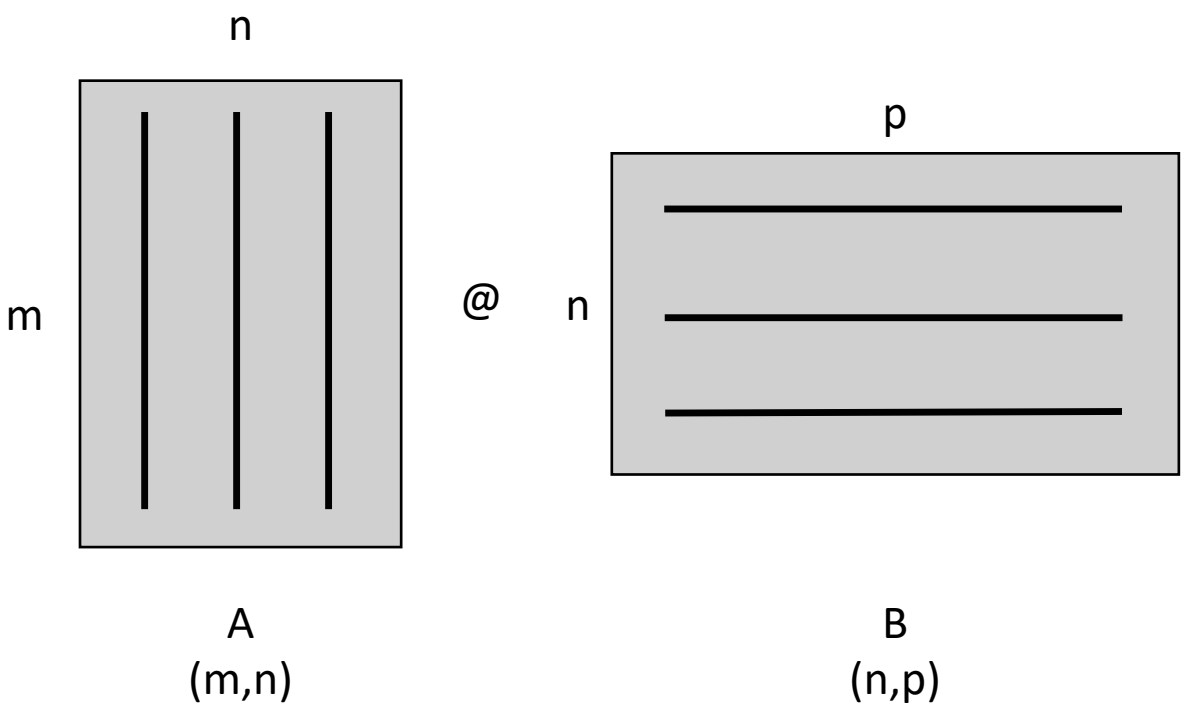
Select k << n
Column-Row Pairs

k total
outer products

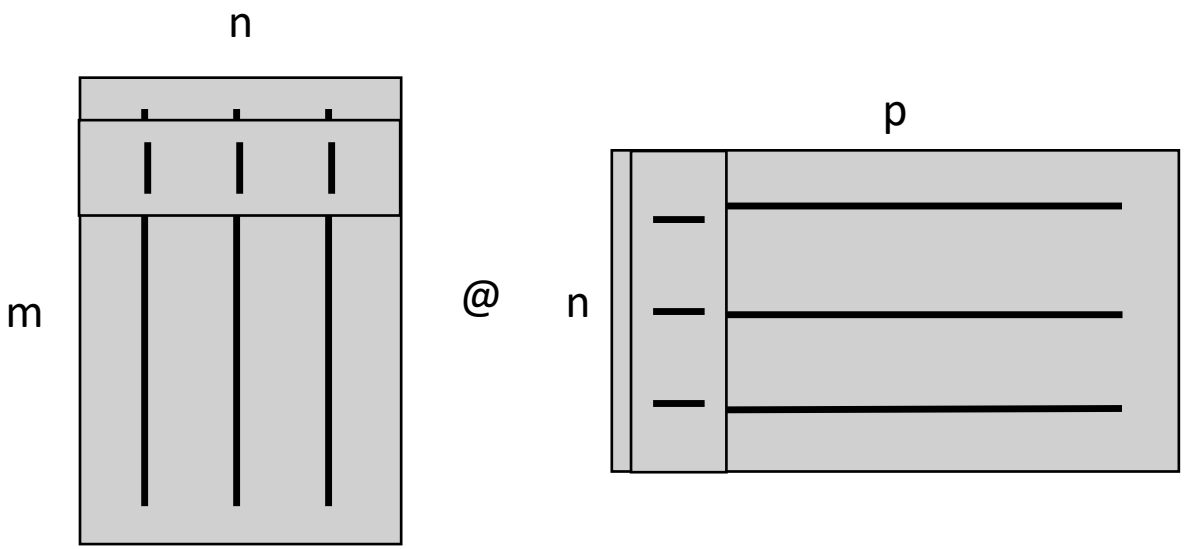
Runtime:
O(mkp)

[10] [11] [12] Adelmen et. al.'s idea

How to select Column-Row pairs?



How to select Column-Row pairs?



Pick pairs with the largest norm products!
Assumption: bigger is more important

- Norm-Proportional Sampling (NPS)
- Deterministic Top-K (det_top_k)
- Single Norm Top-K (single_norm)

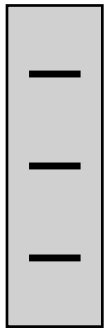
ColNorm(A)
(n,)

x

RowNorm(B)
(n,)

=

Weight Vector
(n,)



Top_K(Weight Vector)
(k,)

Backward with Column-Row Sampling

Save on dy and dW with the same sampling mask!

Normal Forward

$y = x @ W$

Ours Forward

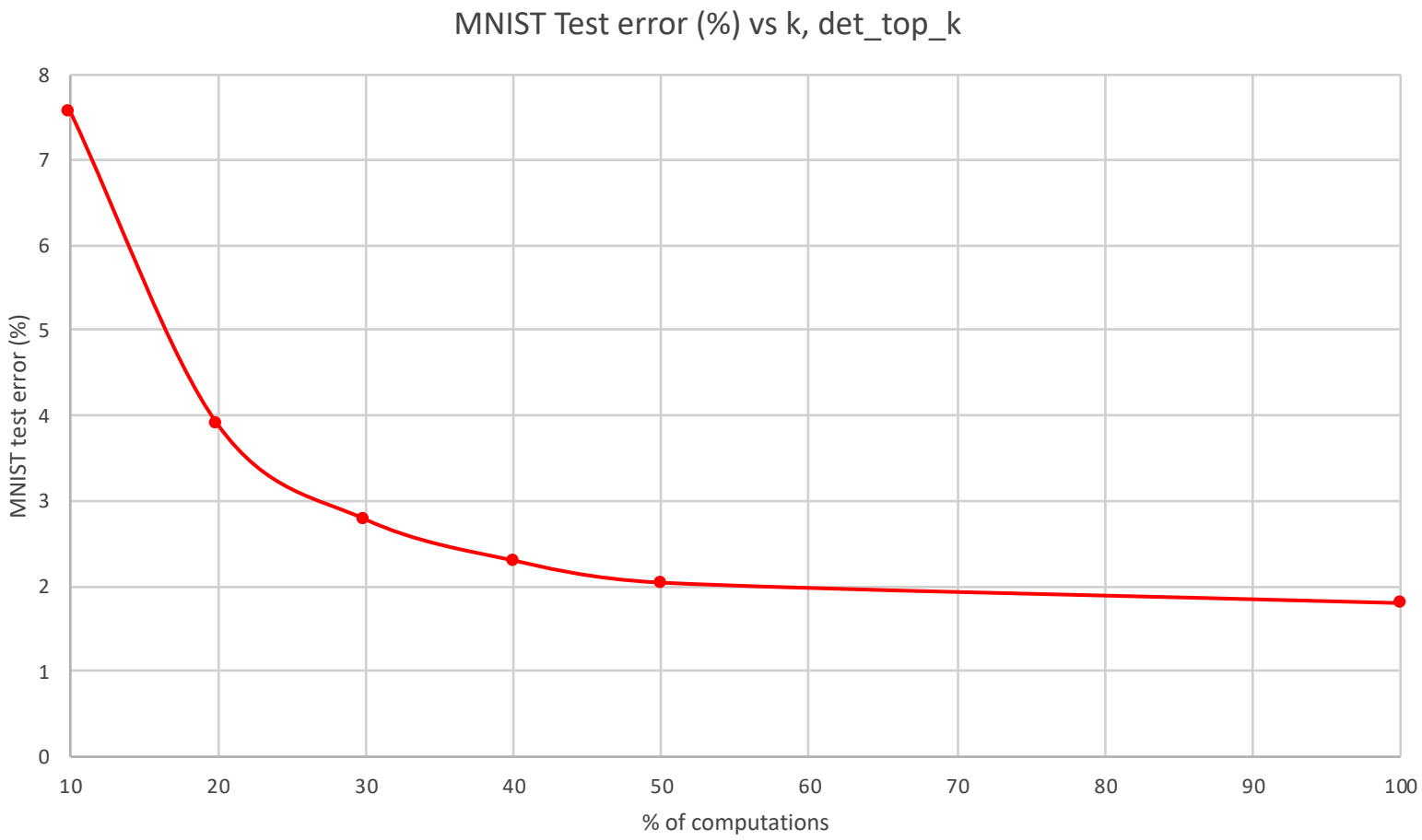
$y = x[mask, :] @ W[:, mask]$

Ours Backward

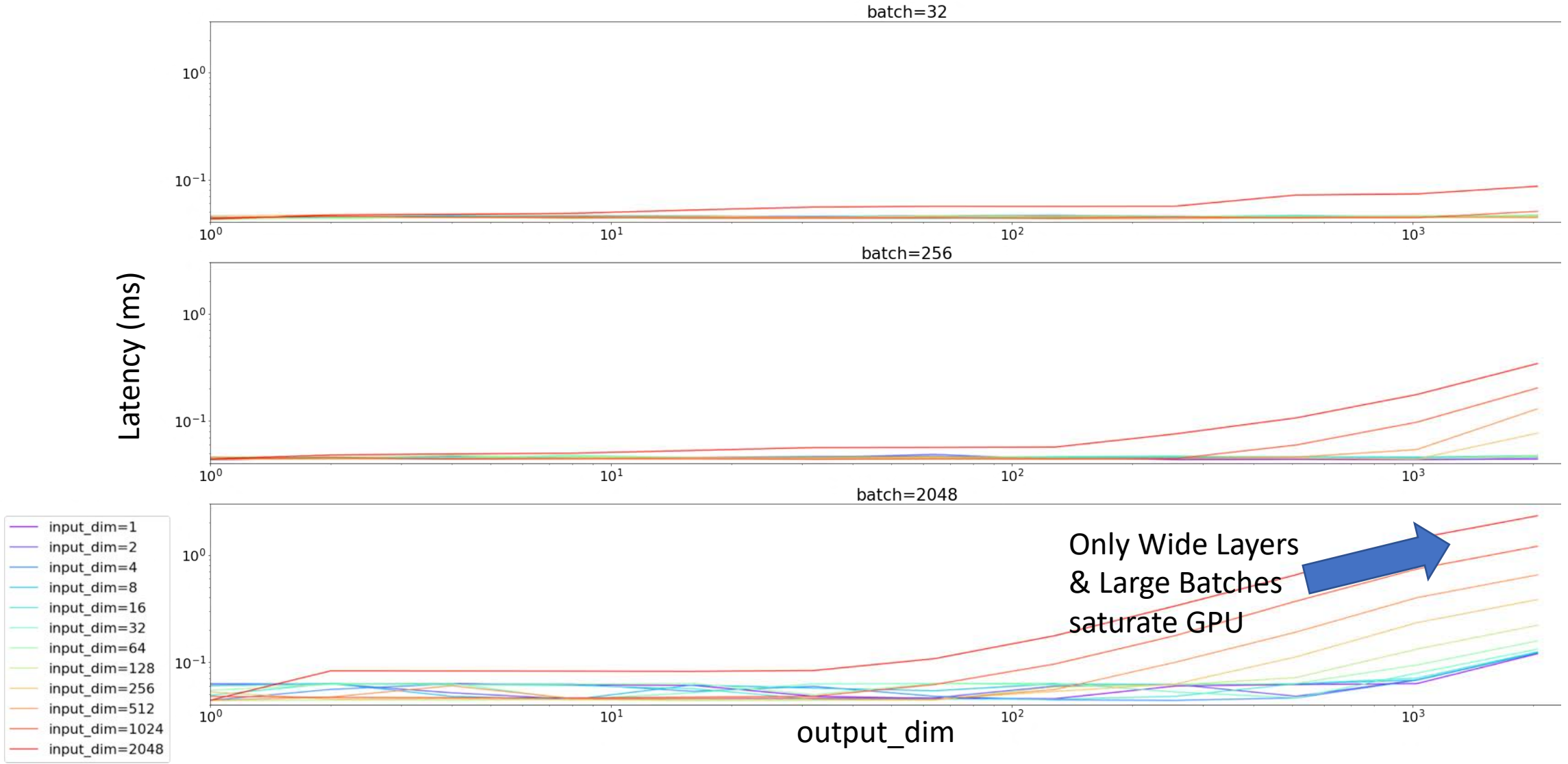
$dx = W[:, mask]' @ dy$

$dW = dy @ x[mask, :]$

Model Performance - MNIST



Regime of speedup



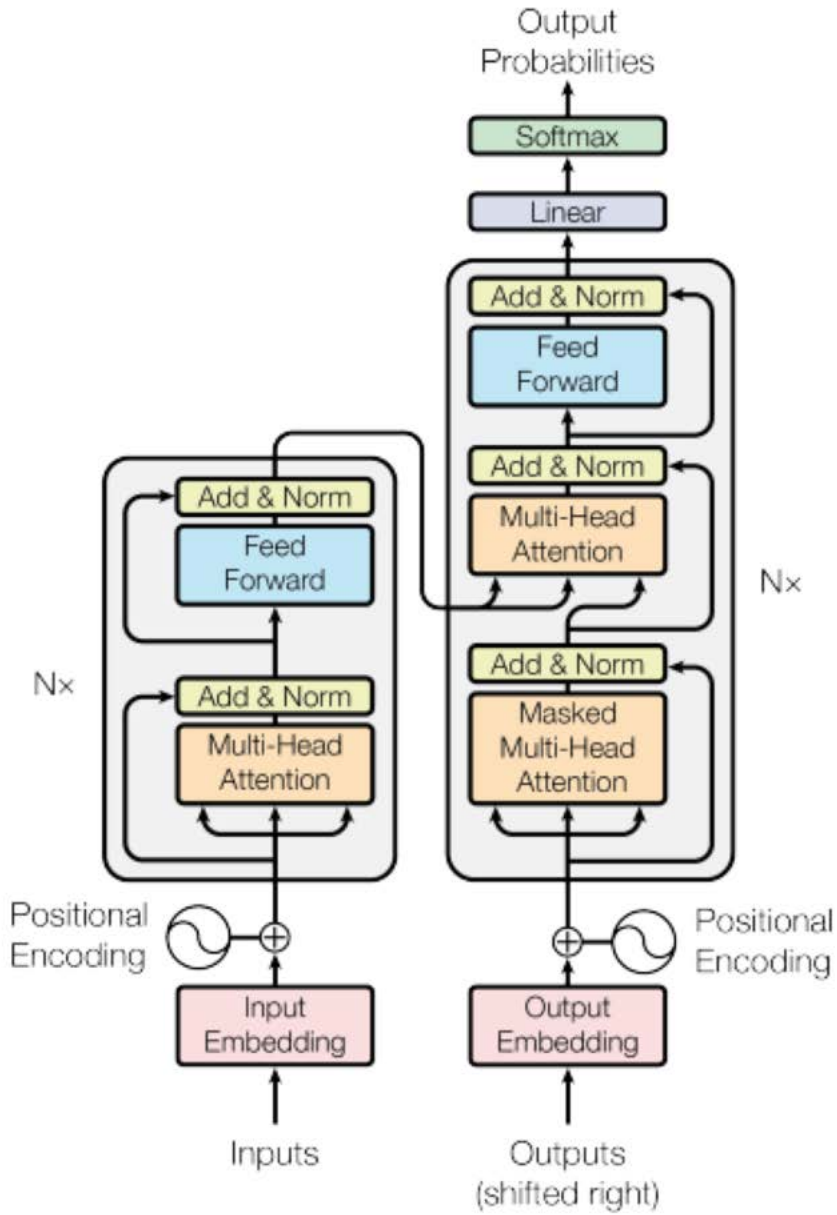
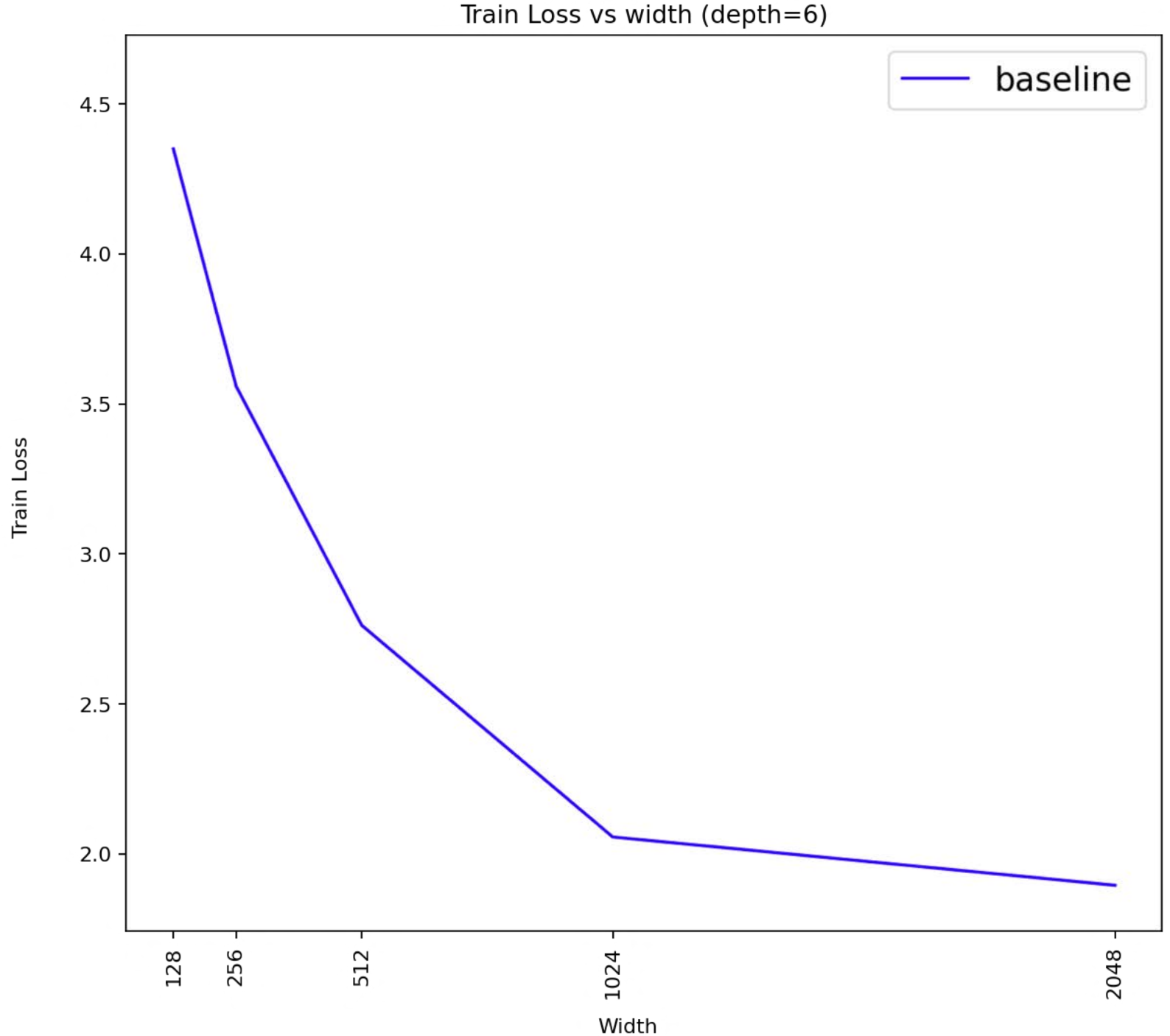


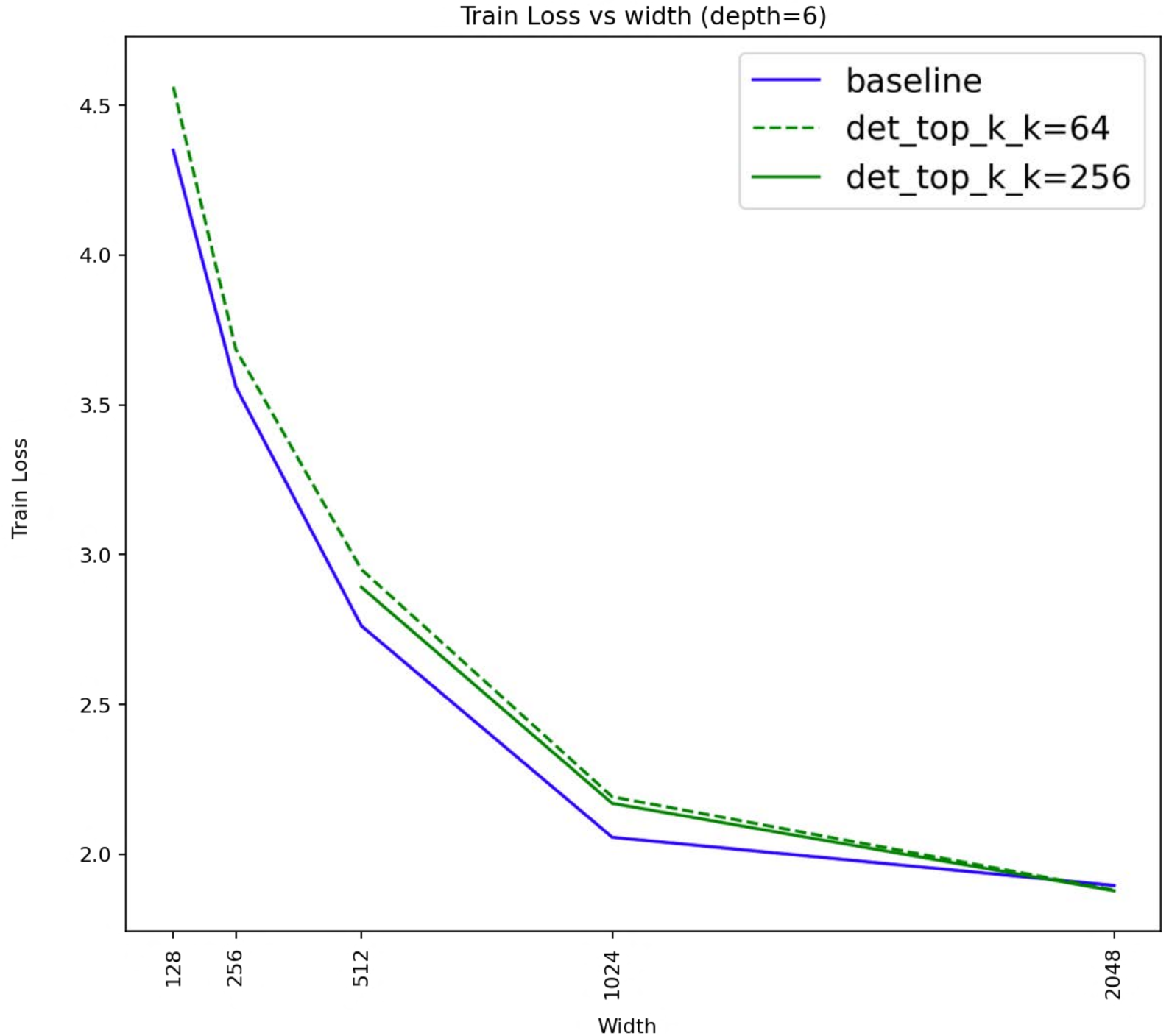
Figure 1: The Transformer - model architecture.

[18]

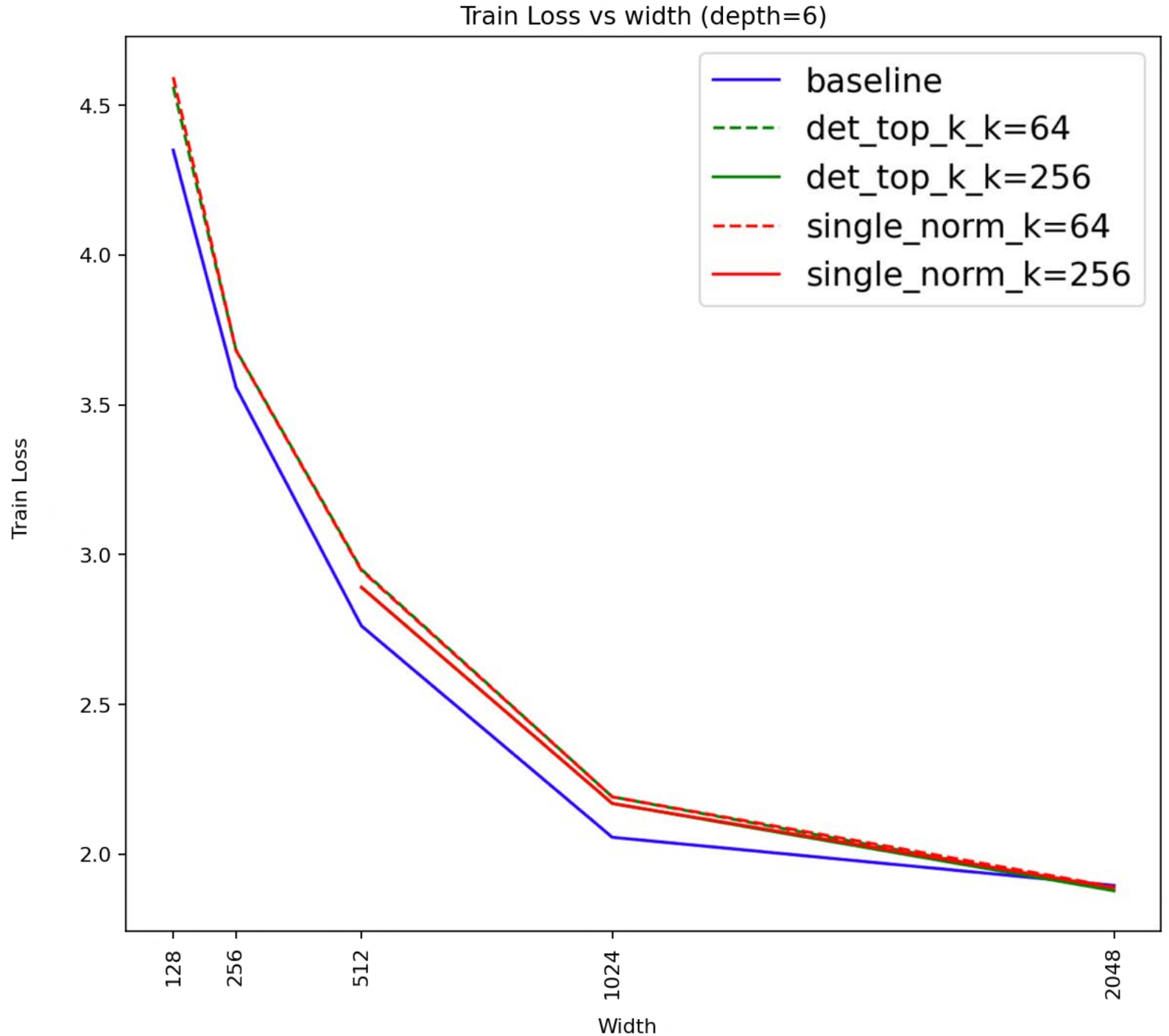
(Vaswani et. al. 2017)



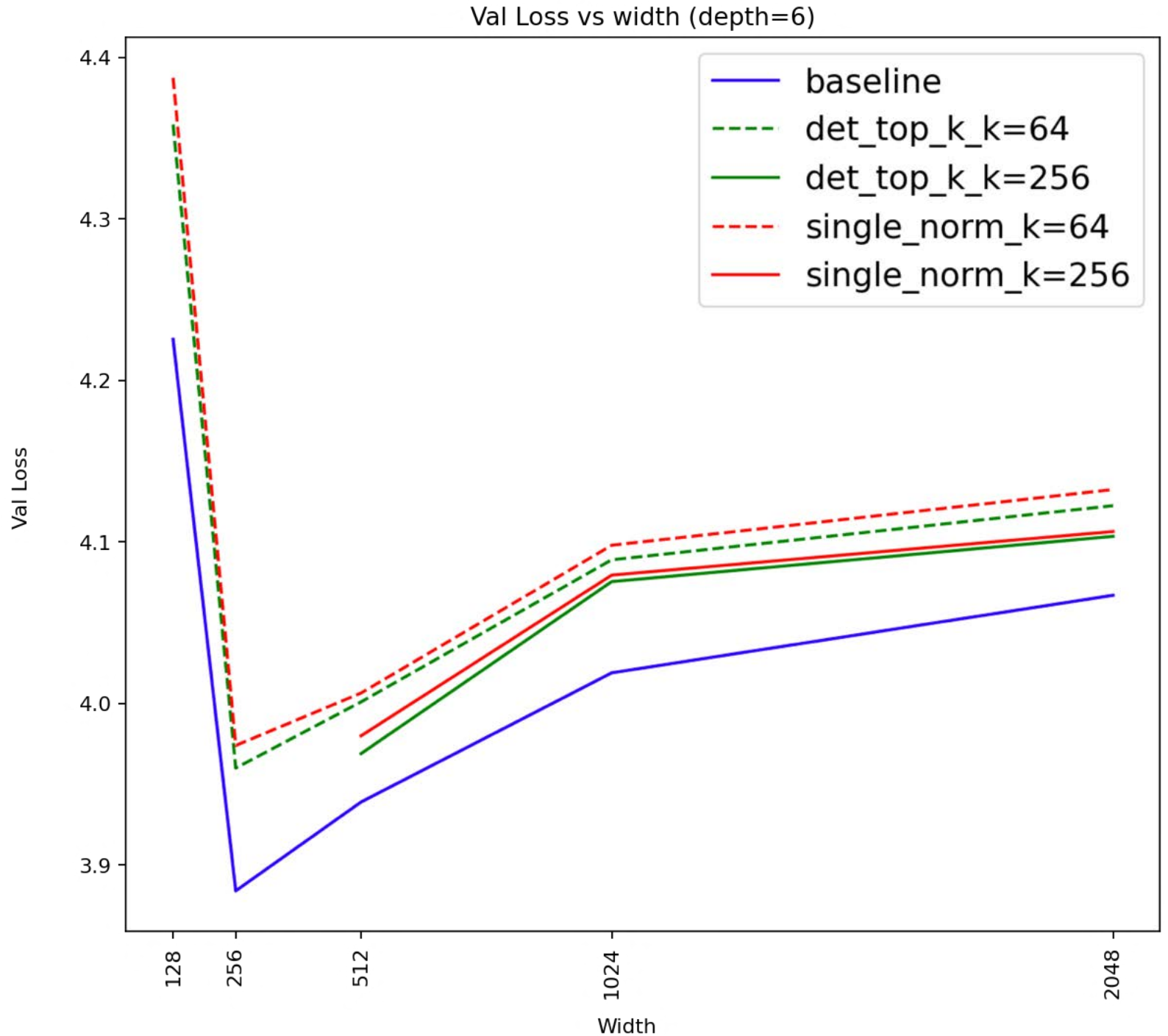
IWSLT DE-EN
Transformer model
Trained from scratch
(no pretraining)



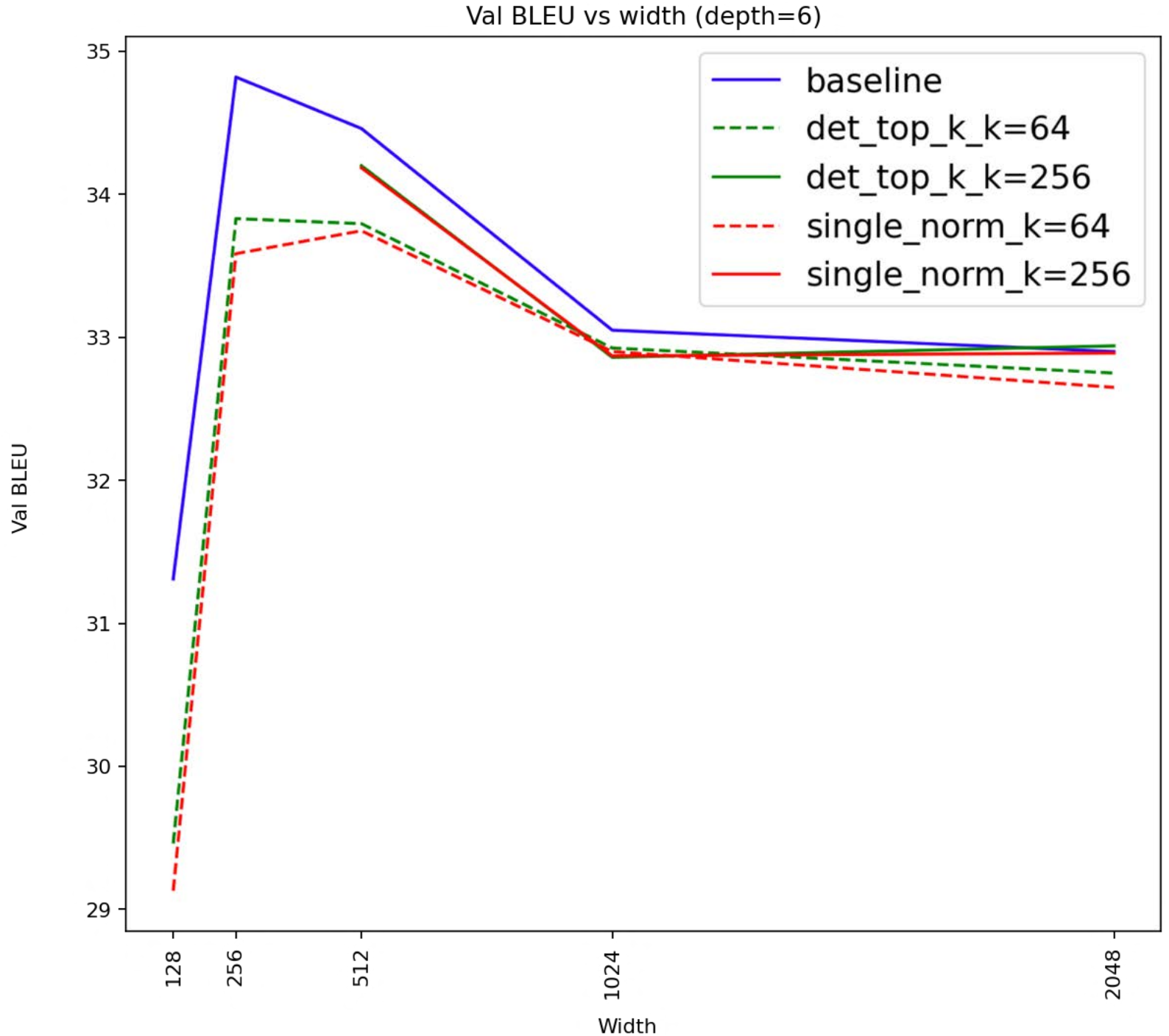
IWSLT DE-EN
Transformer model
Trained from scratch
(no pretraining)



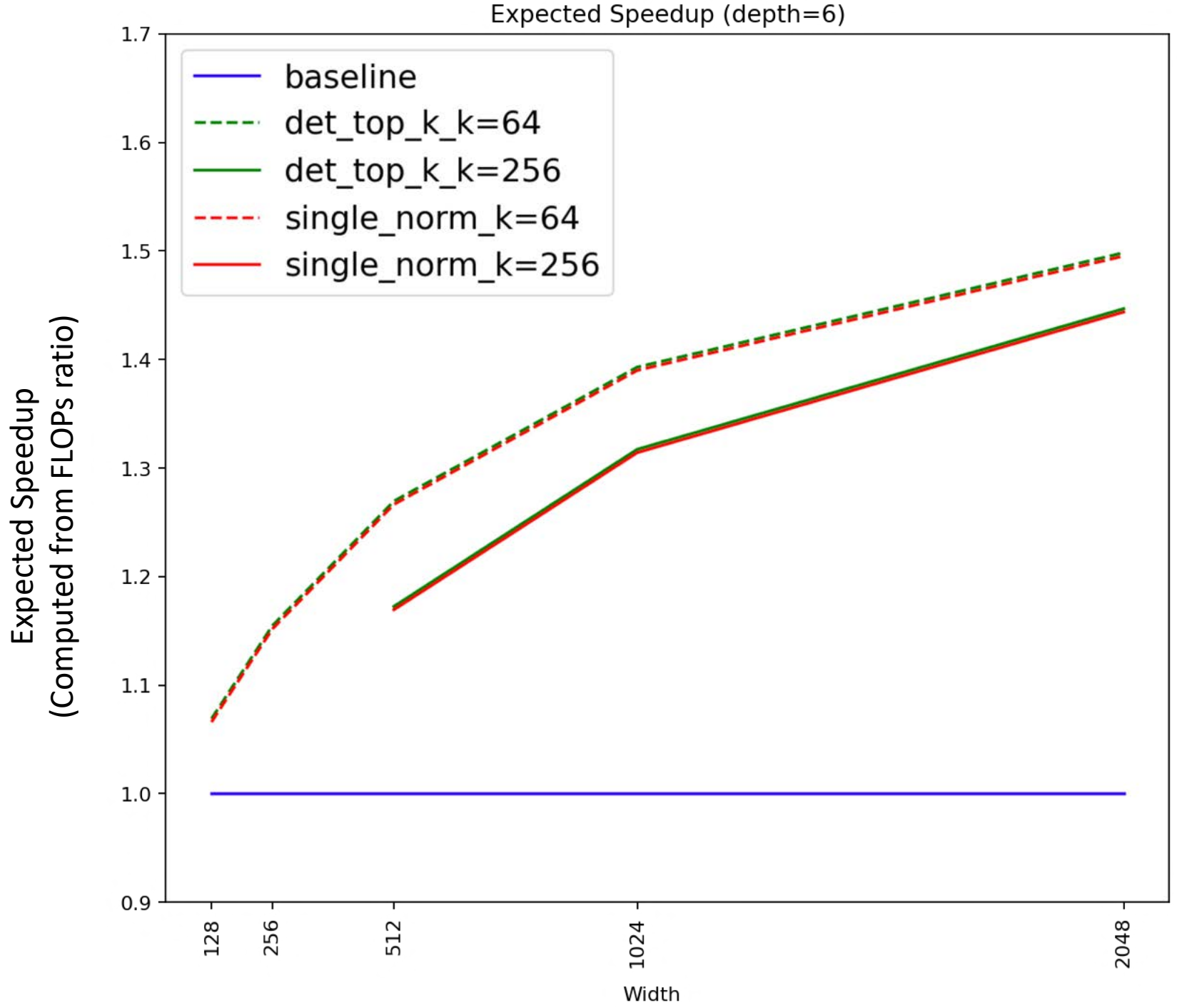
IWSLT DE-EN
Transformer model
Trained from scratch
(no pretraining)

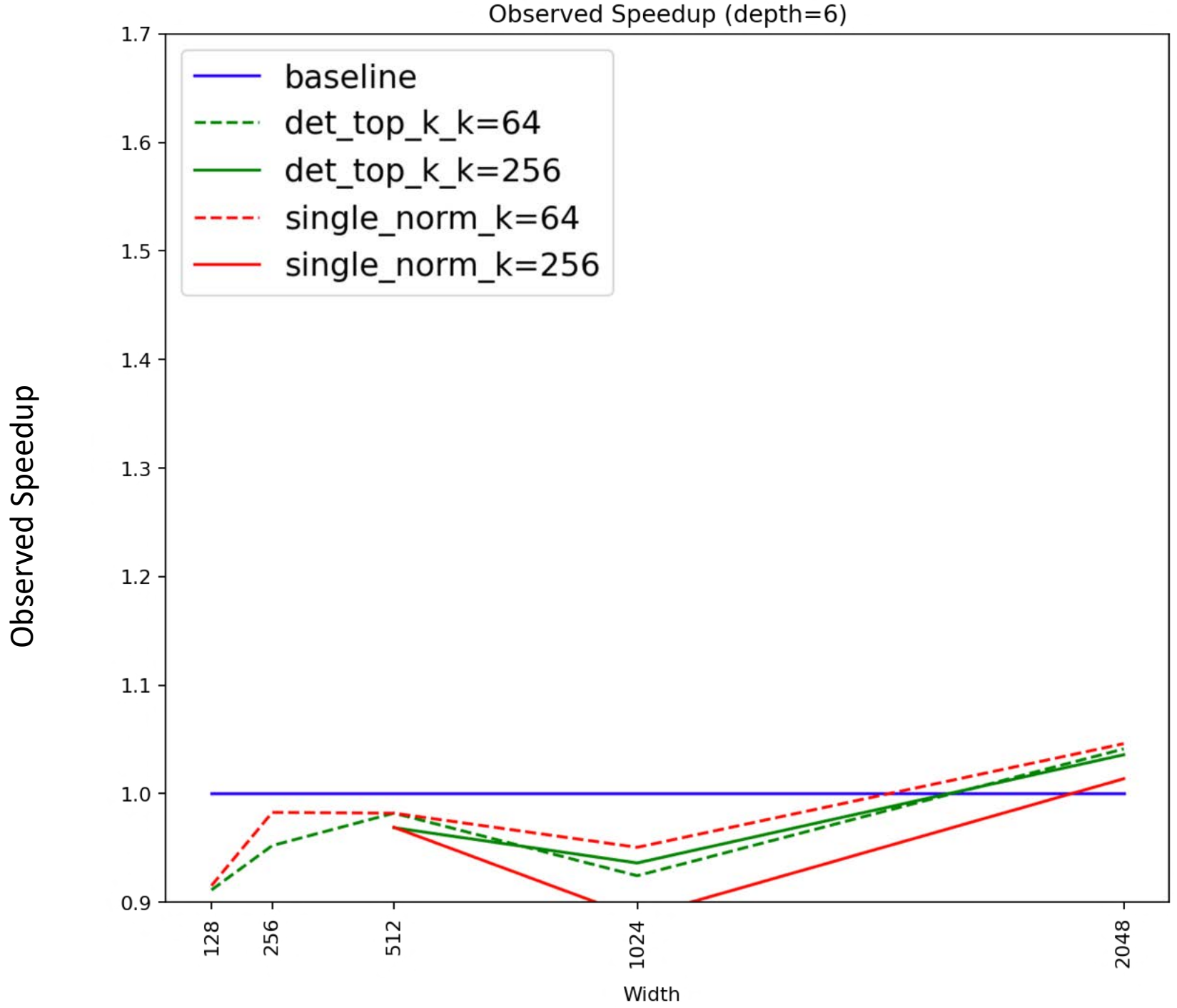


IWSLT DE-EN
Transformer model
Trained from scratch
(no pretraining)



IWSLT DE-EN
Transformer model
Trained from scratch
(no pretraining)


Expected Speedup (depth=6)
baseline
det_top_k_k=64
det_top_k_k=256
single_norm_k=64
single_norm_k=256
Expected Speedup
(Computed from FLOPs ratio)
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
128
256
512
1024
2048
Width


Observed Speedup (depth=6)
baseline
det_top_k_k=64
det_top_k_k=256
single_norm_k=64
single_norm_k=256
Observed Speedup
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
128
256
512
1024
2048
Width

Why no speedup?

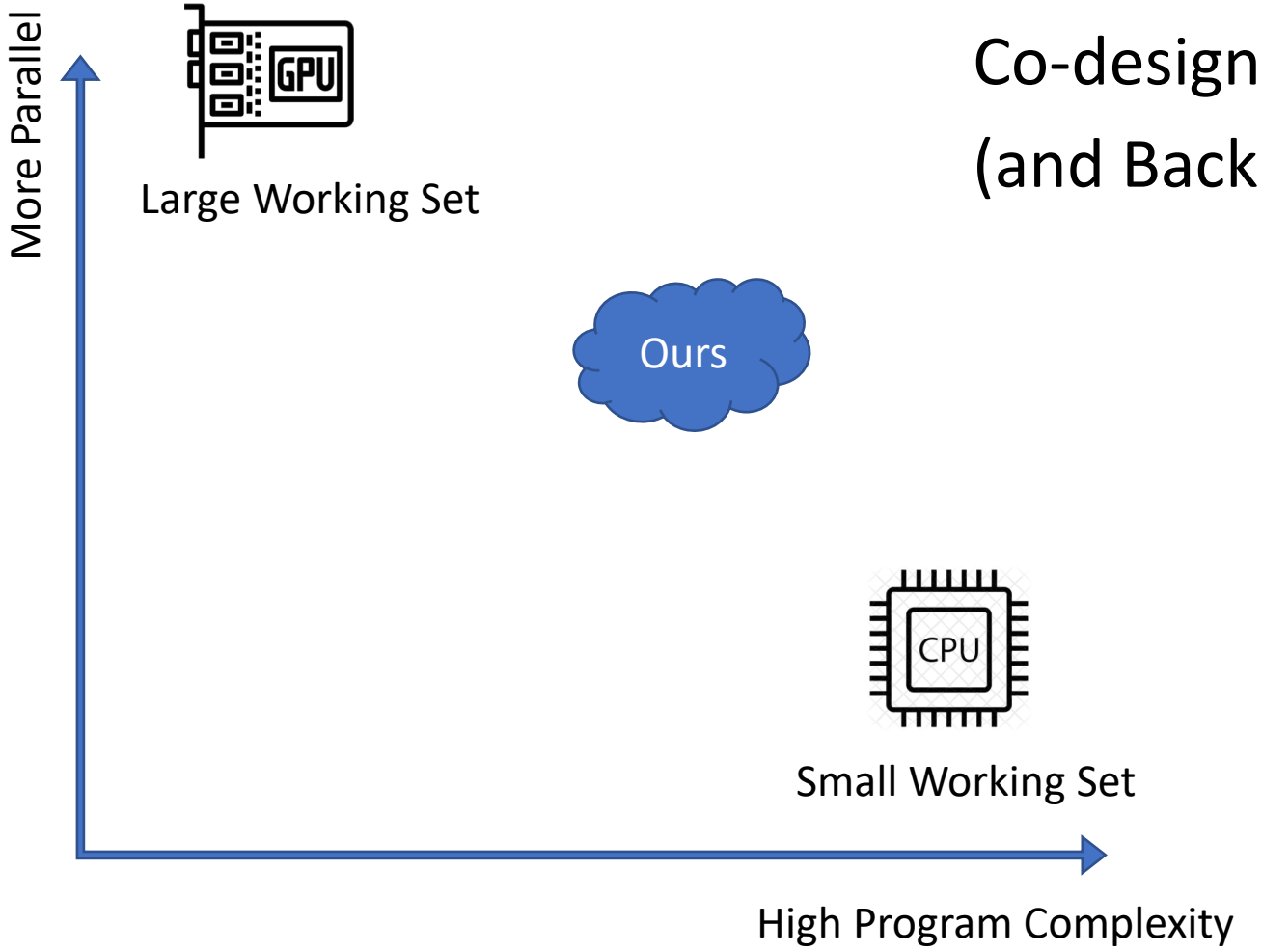
- Theoretical Time != Wall Time
- Discrepancy: memory access
- Memory access 10x slower than compute
- Cache Thrashing

Algo Win != Impl Win

Custom CUDA Kernels

```
$ nvcc -o crs ./crs.cu
```

Hardware and Algorithms



Co-design in the future

(and Backprop may be limiting!)

Extensions

- Reuse Sampling Mask
- Memory friendly sampling mask
- Collaborate with a computer architect

W ide

A pproximate

C olumn-Row

S ampling

Origins

 Digital Cameras

 Computational Photography

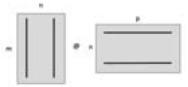
 Wearable Cameras

Research

 Perception and Understanding (MIT)

 Understanding and Behavior (Uber ATG)

Recent Work

 **Scale and Speed (MSR)**

Origins

Digital Cameras

Computational Photography

Wearable Cameras

Research

Perception and Understanding

Understanding and Behavior

Recent Work

Scale and Speed

Origins

Digital Cameras

Computational Photography

Wearable Cameras

Research

Perception and Understanding

Understanding and Behavior

Recent Work

Scale and Speed

Origins

Digital Cameras

Computational Photography

Wearable Cameras

Research

Perception and Understanding

Understanding and Behavior

Recent Work

Scale and Speed

+ Open-Endedness?

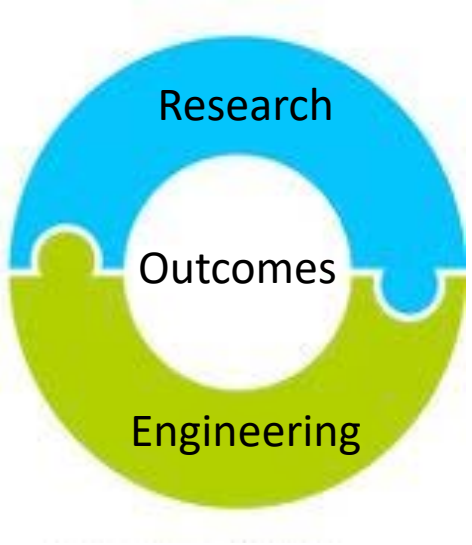
Connections to Open-Endedness

- Backprop Algorithm Relaxed
- Feedback Alignment
- Deep Neuro-Evolution
- Open-Ended Environments

[13] [14] [15]

Why Me?

- Contribute to the full research cycle
  - Motivated from Math/Theory
  - → Impl → Experiments → Eng. Challenges/Solutions → Takeaways
- Side Projects in Open-Endedness
  - Combining GD and NE
  - Regularization – Diversity connection
- Beyond the leaderboard
- Contributions in both Research and Engineering



References

[1] N. Snavely, S. M. Seitz, and R. Szeliski, “Photo tourism: Exploring photo collections in 3d,” ACM Trans. Graph., vol. 25, p. 835846, July 2006 http://phototour.cs.washington.edu/Photo_Tourism.pdf

[2] PhotoSynth Demo. https://www.youtube.com/watch?v=BHzw98qt5Lo

[3] Narrative Clip. https://newatlas.com/narrative-clip-2/35422/

[4] Mistry, Pranav, Pattie Maes, and Liyan Chang. “WUW - wear Ur world.” Proceedings of the 27th international conference extended abstracts on Human factors in computing systems - CHI EA '09. Boston, MA, USA, 2009. 4111. https://www.pranavmistry.com/archived/projects/sixthsense/

[5] X. Puig, K. Ra, M. Boben, J. Li, T. Wang, S. Fidler, and A. Torralba, “VirtualHome: Simulating Household Activities via Programs,” arXiv e-prints, p. arXiv:1806.07011, June 2018. https://arxiv.org/pdf/1806.07011.pdf

[6] Lidar Intensity Calibration Patent https://patents.google.com/patent/US10598791B2/en

[7] H. Salman, G. Yang, J. Li, P. Zhang, H. Zhang, I. Razenshteyn, and S. Bubeck, “Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers,” arXiv e-prints, p. arXiv:1906.04584, June 2019. https://arxiv.org/abs/1906.04584

[8] J. M. Cohen, E. Rosenfeld, and J. Zico Kolter, “Certified Adversarial Robustness via Randomized Smoothing,” arXiv e-prints, p. arXiv:1902.02918, Feb. 2019. https://arxiv.org/abs/1902.02918

[9] A. Neelakantan, L. Vilnis, Q. V. Le, I. Sutskever, L. Kaiser, K. Kurach, and J. Martens, “Adding Gradient Noise Improves Learning for Very Deep Networks,” arXiv e-prints, p. arXiv:1511.06807, Nov. 2015. https://arxiv.org/pdf/1511.06807.pdf

[10] M. Adelman and M. Silberstein, “Faster Neural Network Training with Approximate Tensor Operations,” arXiv e-prints, p. arXiv:1805.08079, May 2018. https://arxiv.org/pdf/1805.08079.pdf

[11] X. Sun, X. Ren, S. Ma, and H. Wang, “meProp: Sparsified Back Propagation for Accelerated Deep Learning with Reduced Overfitting,” arXiv e-prints, p. arXiv:1706.06197, June 2017. https://arxiv.org/abs/1706.06197

[12] X. Sun, X. Ren, S. Ma, B. Wei, W. Li, J. Xu, H. Wang, and Y. Zhang, “Training Simplification and Model Simplification for Deep Learning: A Minimal Effort Back Propagation Method,” arXiv e-prints, p. arXiv:1711.06528, Nov. 2017. https://arxiv.org/pdf/1711.06528.pdf

[13] F. Petroski Such, V. Madhavan, E. Conti, J. Lehman, K. O. Stanley, and J. Clune, “Deep Neuroevolution: Genetic Algorithms Are a Competitive Alternative for Training Deep Neural Networks for Reinforcement Learning,” arXiv e-prints, p. arXiv:1712.06567, Dec. 2017. https://arxiv.org/pdf/1712.06567.pdf

[14] J. Launay, I. Poli, F. Boniface, and F. Krzakala, “Direct Feedback Alignment Scales to Modern Deep Learning Tasks and Architectures,” arXiv e-prints, p. arXiv:2006.12878, June 2020. https://arxiv.org/pdf/2006.12878.pdf

[15] Shawn Jain’s DFA Scales https://github.com/darkmatter08/dfa-scales-to-modern-deep-learning/commit/8dcdbd1e3116dda44b92c1faae6e0725c8fd806e

[16] A. Nøkland, “Direct Feedback Alignment Provides Learning in Deep Neural Networks,” arXiv e-prints, p. arXiv:1609.01596, Sept. 2016.https://arxiv.org/pdf/1609.01596.pdf

[17] Microsoft Turing Blog Post https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention Is All You Need. CoRR abs/1706.03762 (2017). https://arxiv.org/abs/1706.03762

[19] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. CoRR abs/1502.03044 (2015). https://arxiv.org/abs/1502.03044

Shawn Jain

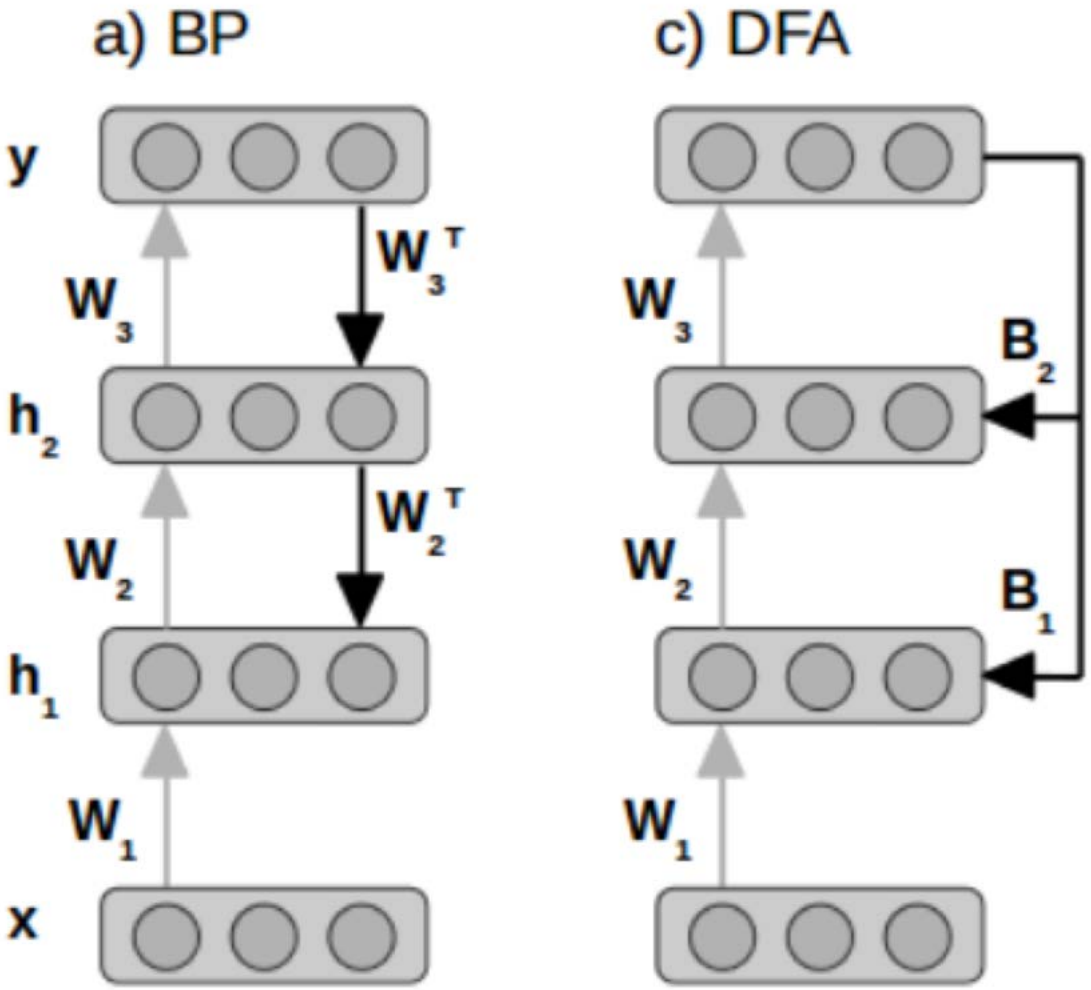
Open AI

December 7, 2020

shawnjain.com

Backup

Feedback Alignment



[16]

Adapted from Figure 1 (Nøkland 2016)

Direct Feedback Alignment vs BP