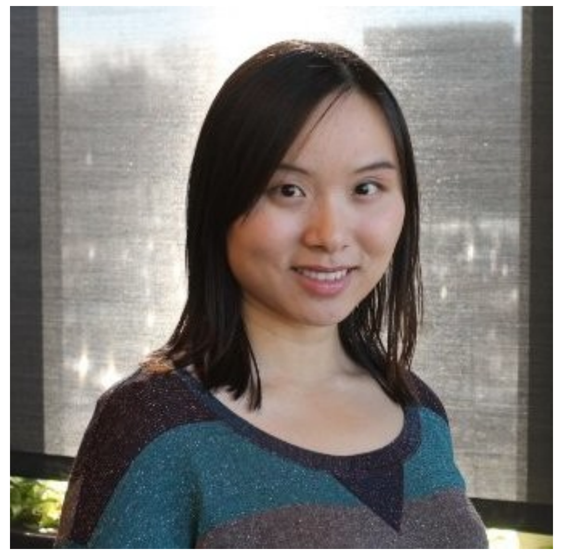
Bad Assumptions about Neural Networks

SignalFire
San Francisco, CA
4 March 2020



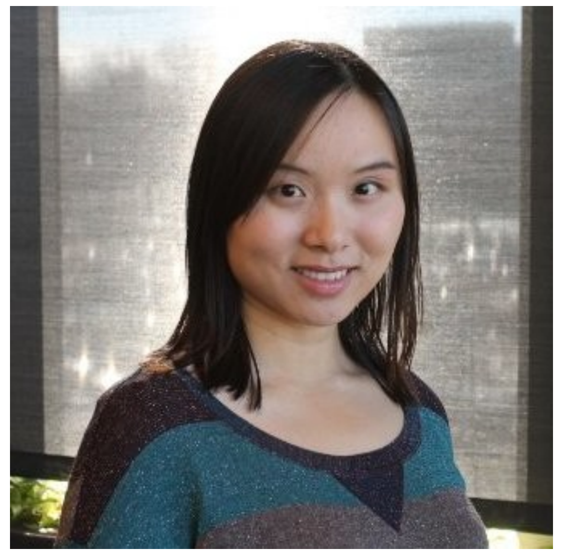
Rosanne Liu

Jason Yosinski

Deep Collective Uber AI

Bad Assumptions about Neural Networks

SignalFire
San Francisco, CA
4 March 2020



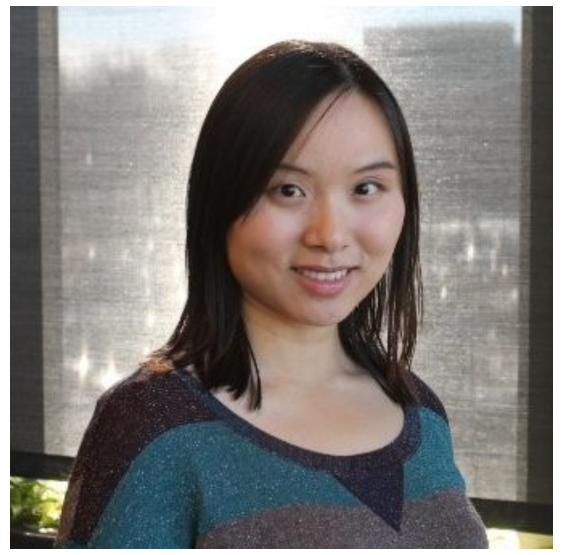
Rosanne Liu

Jason Yosinski

Deep Collective Uber AI

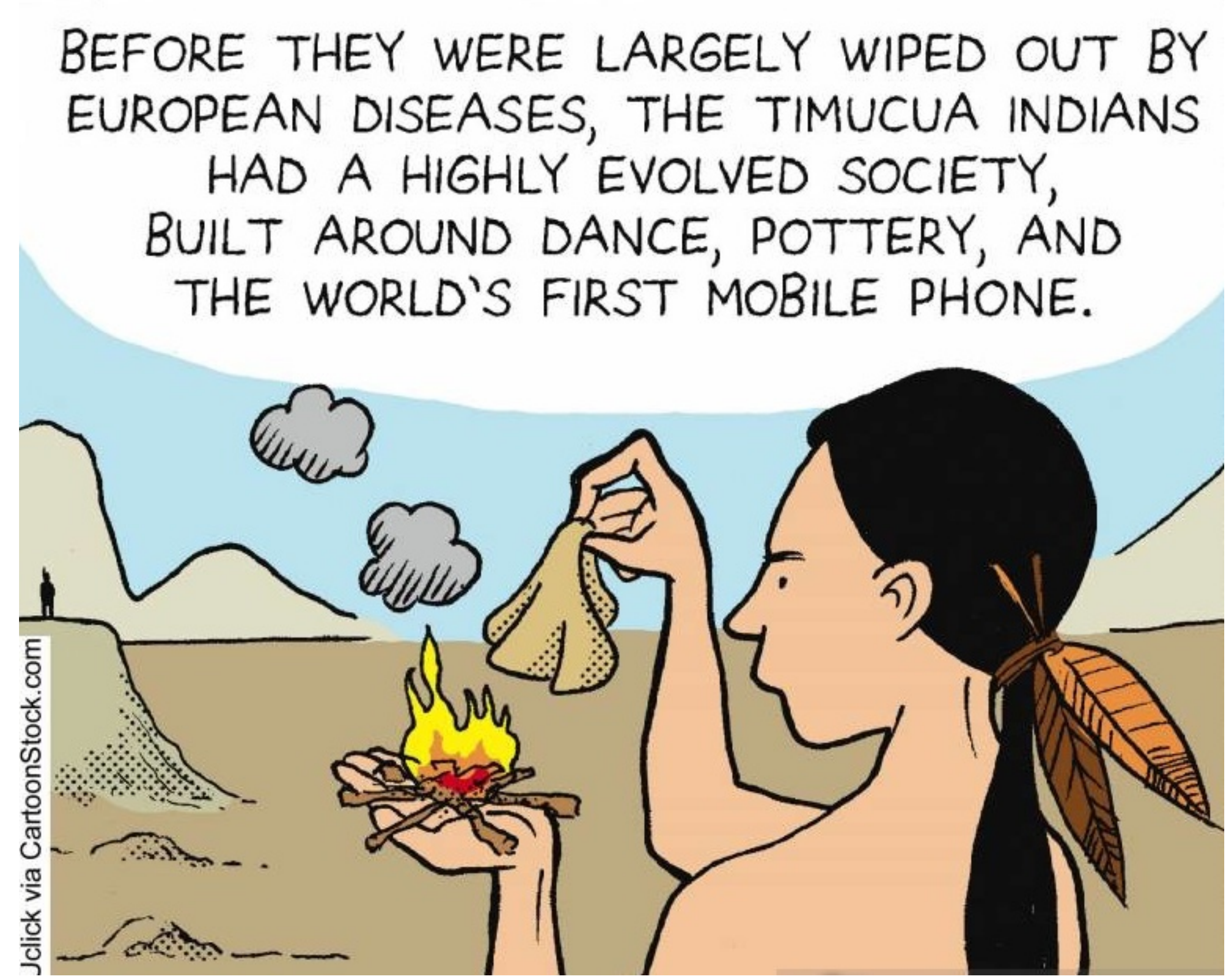
Bad Assumptions about Neural Networks

SignalFire
San Francisco, CA
4 March 2020



Rosanne Liu

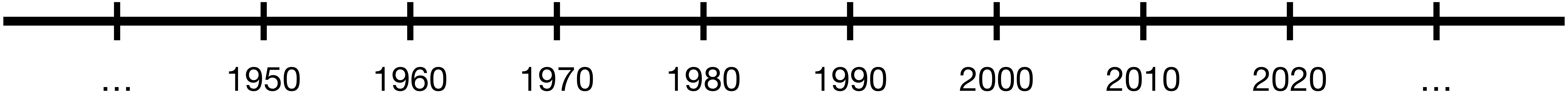
Jason Yosinski


BEFORE THEY WERE LARGELY WIPED OUT BY EUROPEAN DISEASES, THE TIMUCUA INDIANS HAD A HIGHLY EVOLVED SOCIETY, BUILT AROUND DANCE, POTTERY, AND THE WORLD'S FIRST MOBILE PHONE.

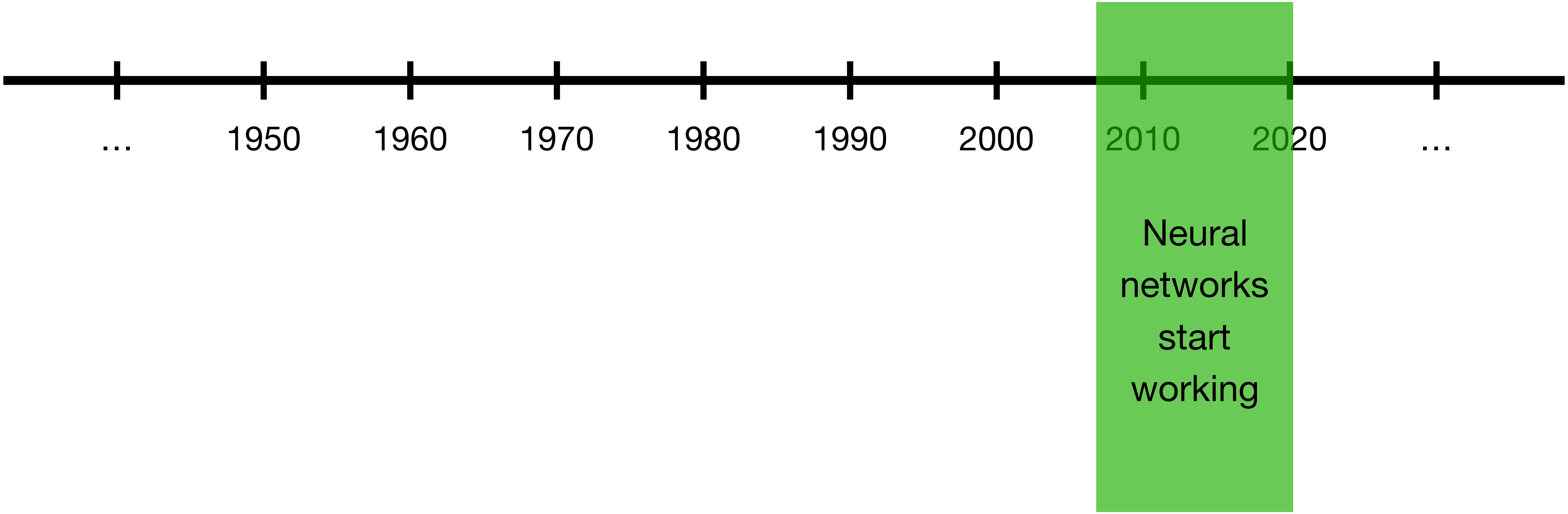
Jclick via CartoonStock.com

Deep Collective Uber AI

Progress in AI

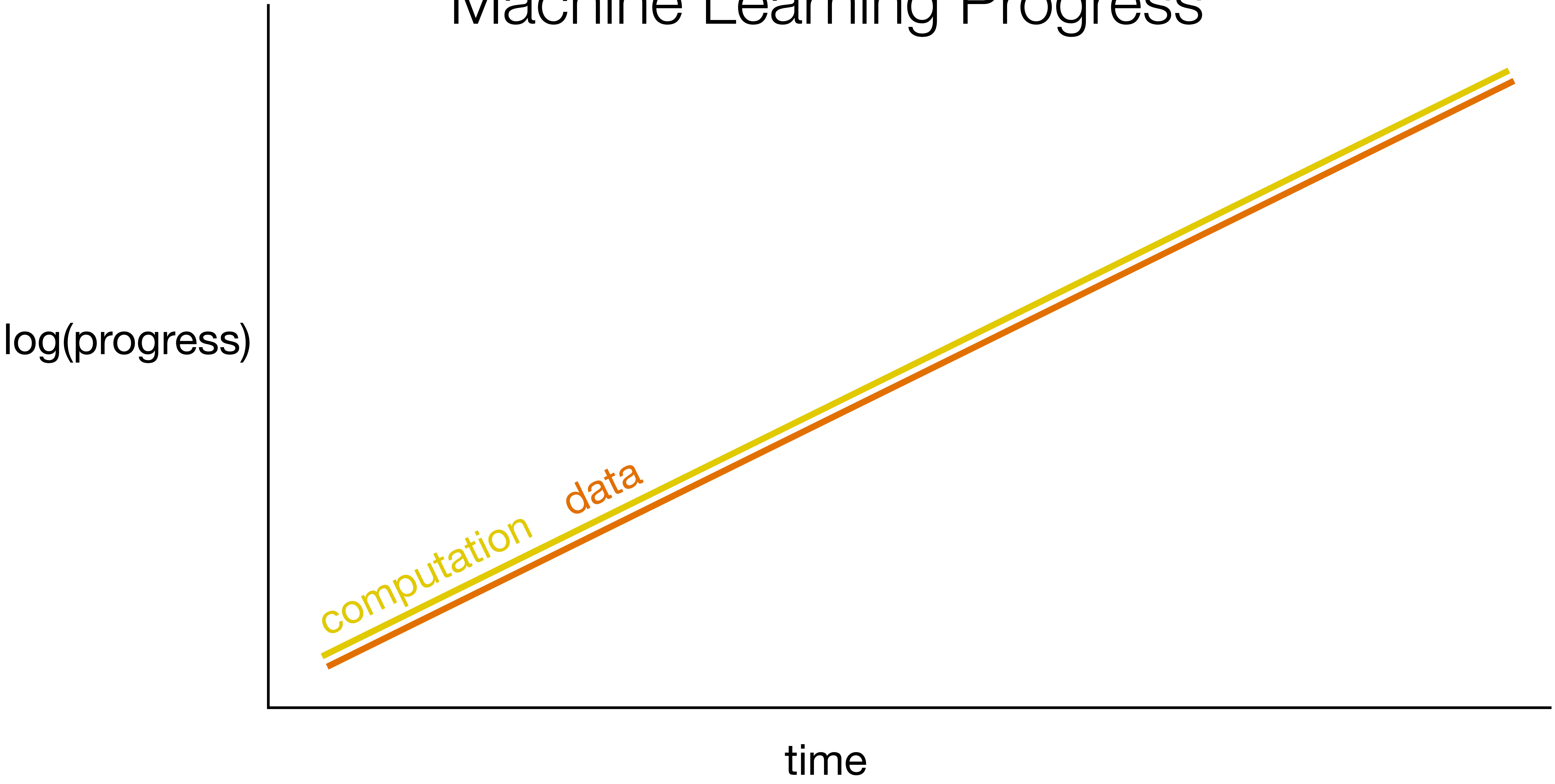


Progress in AI

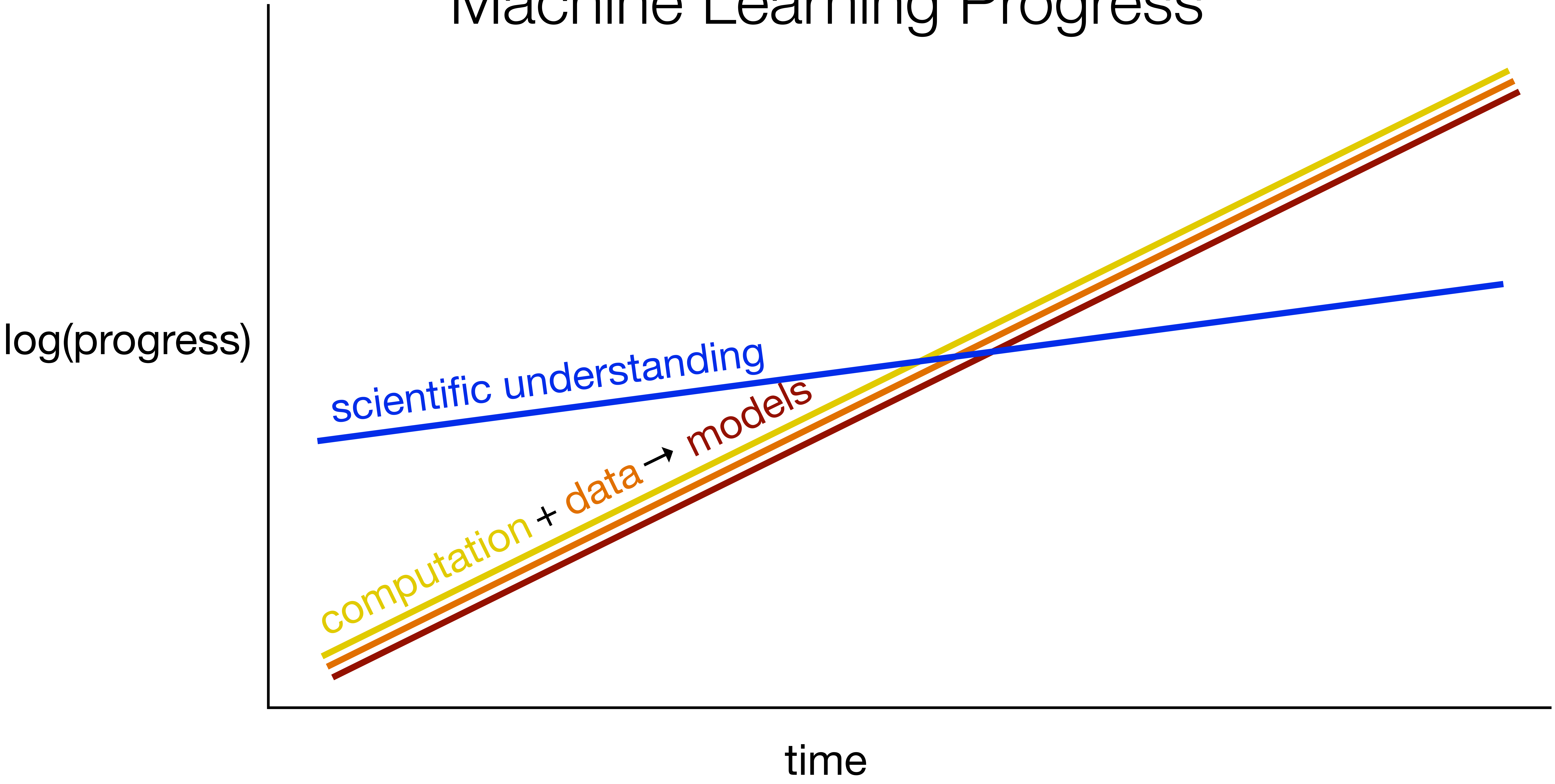


Neural networks start working

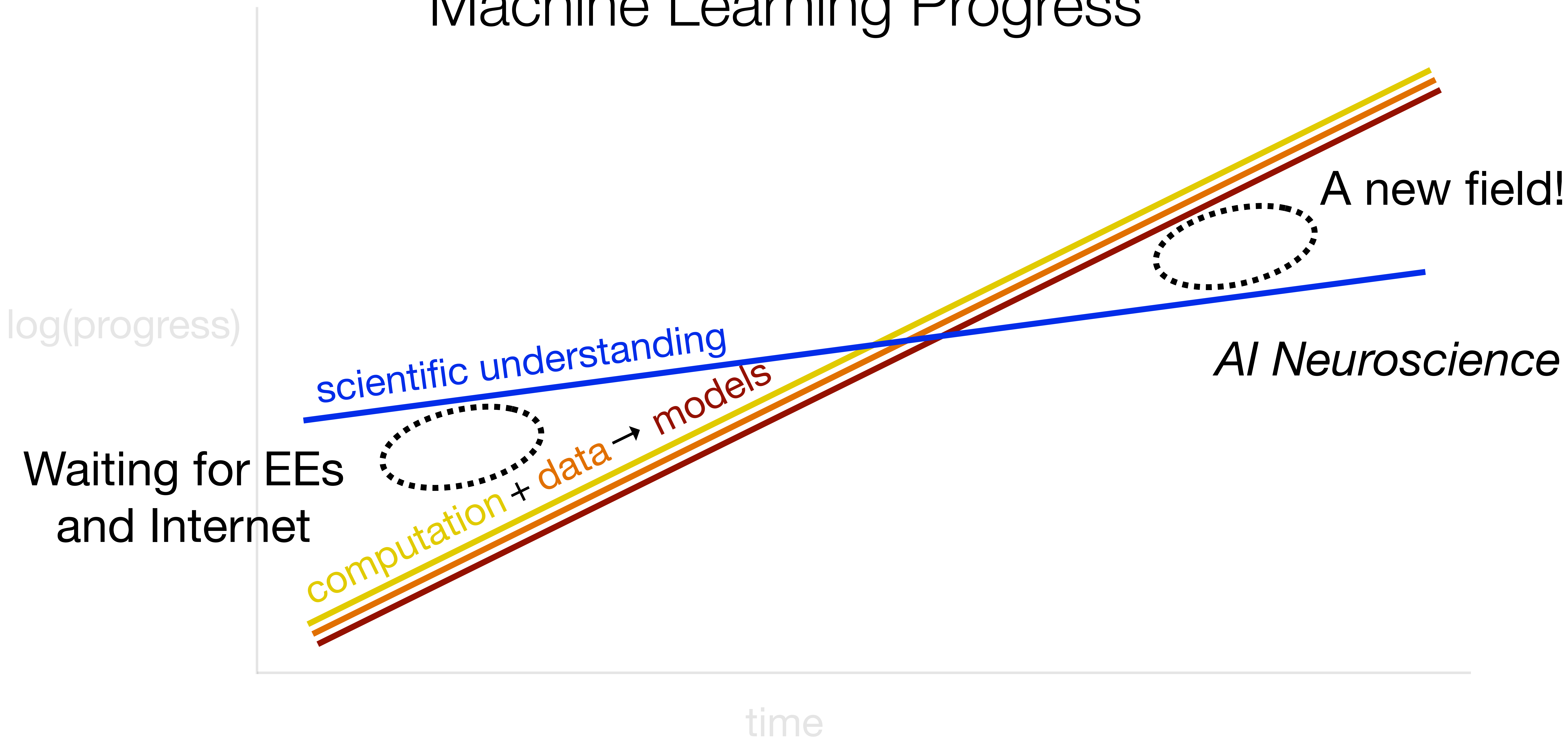
Machine Learning Progress



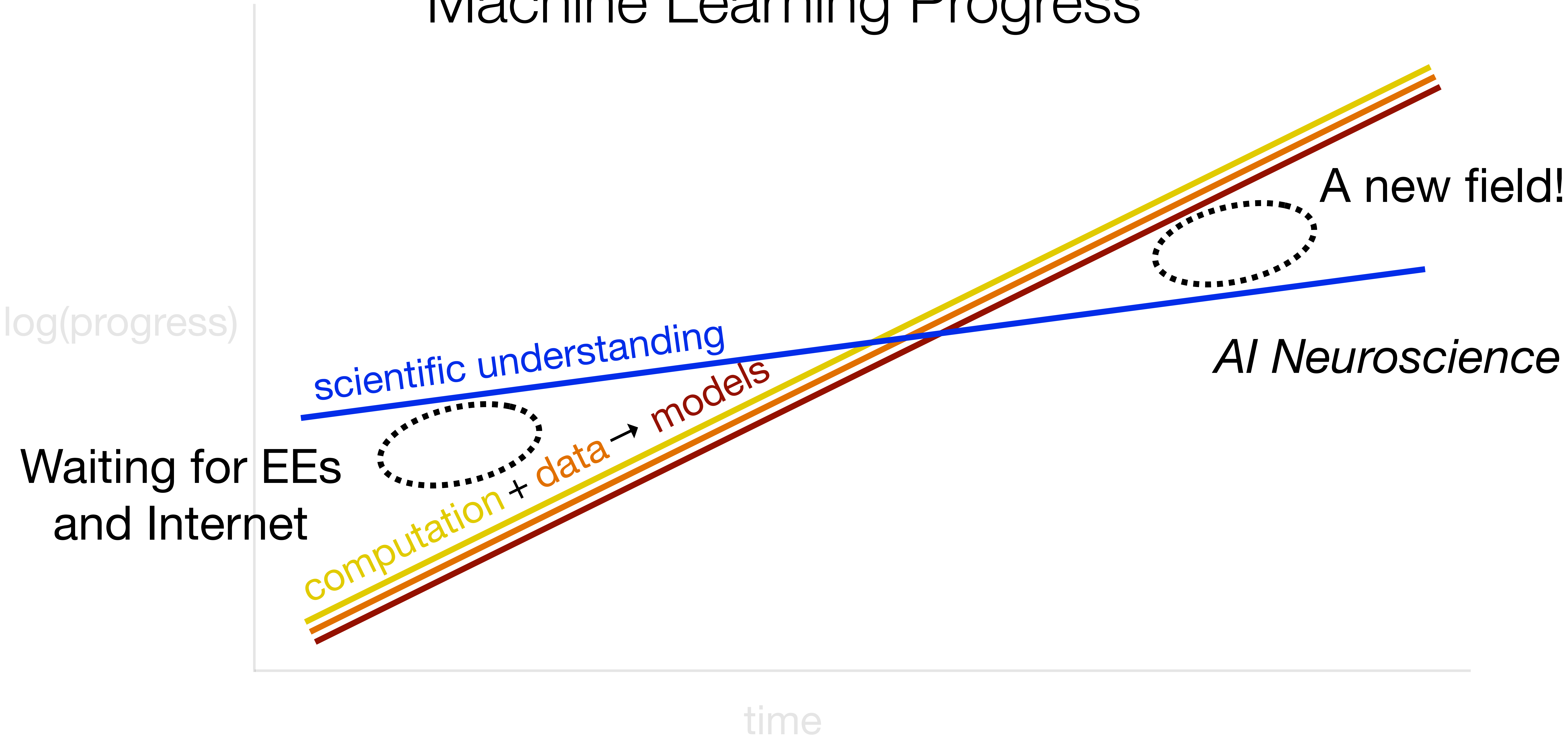
Machine Learning Progress



Machine Learning Progress



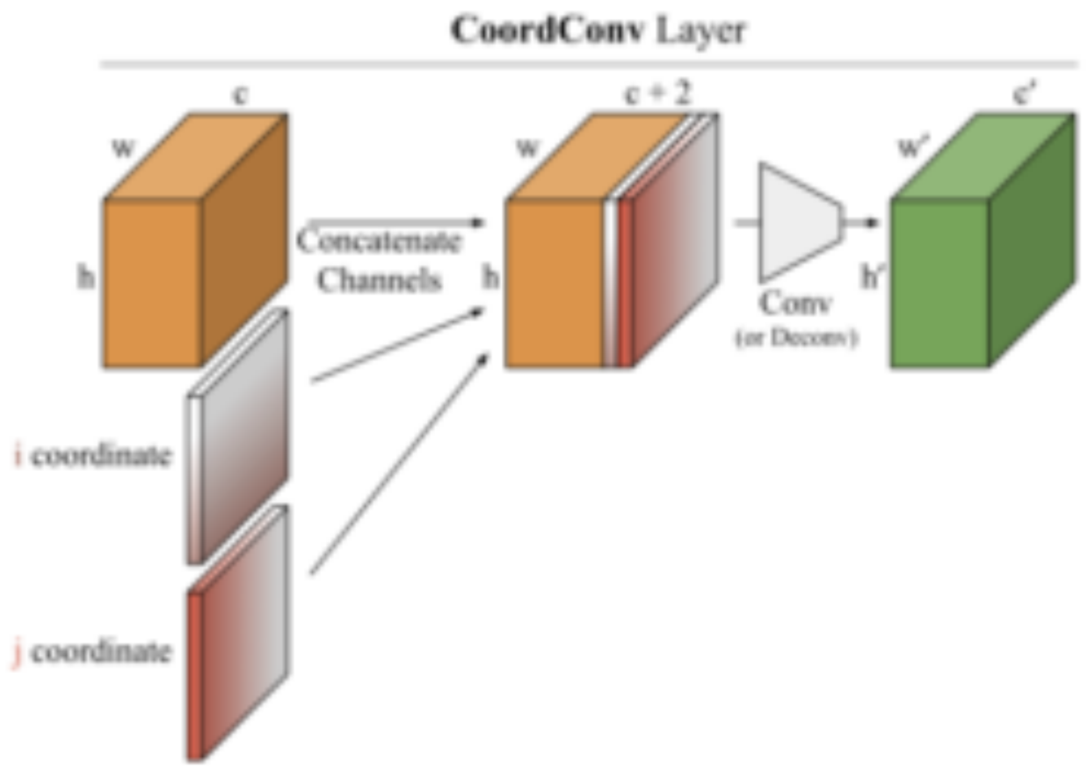
Machine Learning Progress

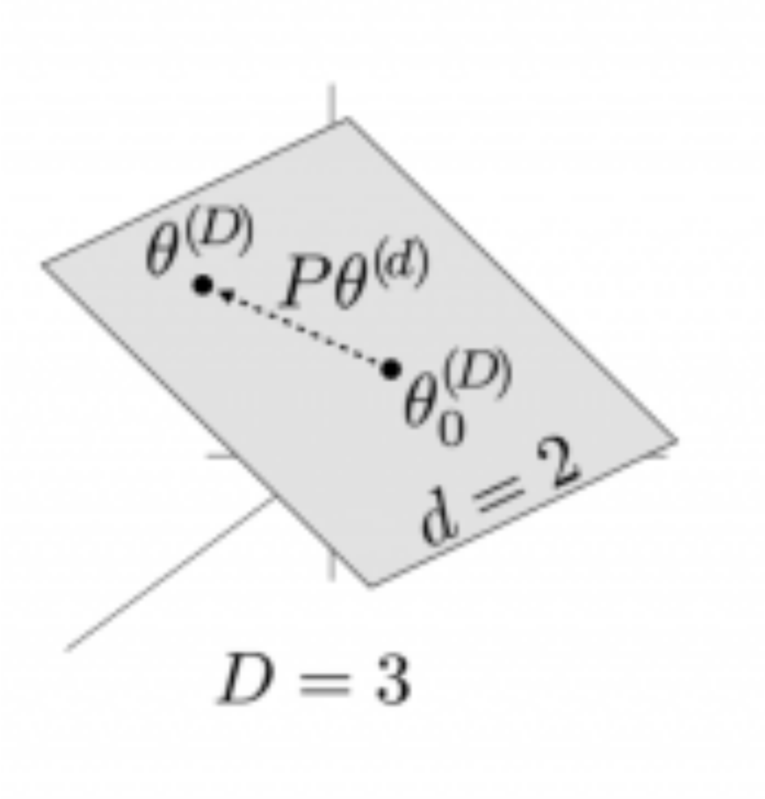


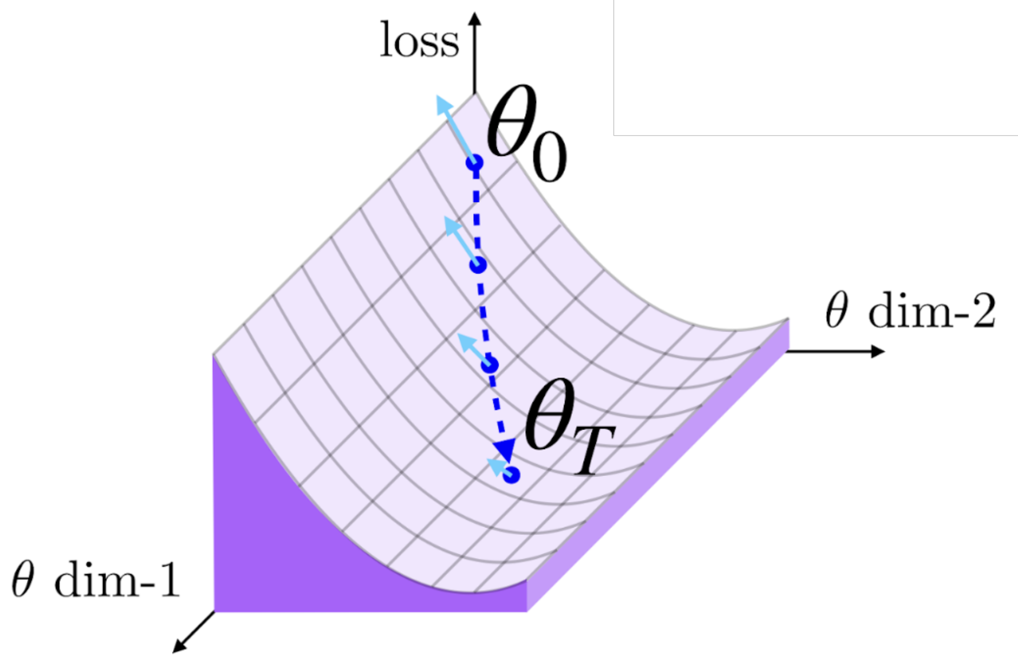
Waiting for EEs and Internet

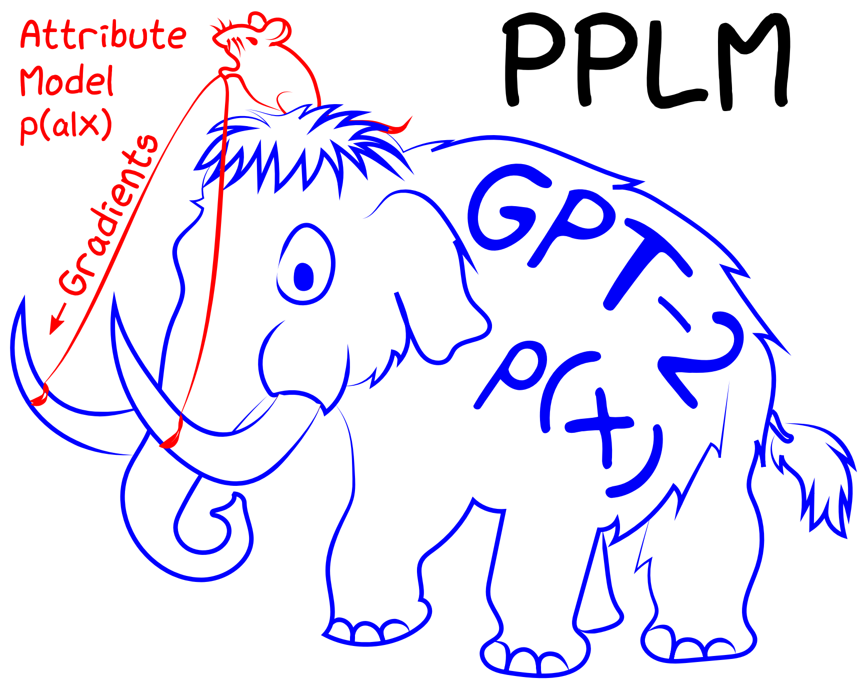
A new field!

AI Neuroscience

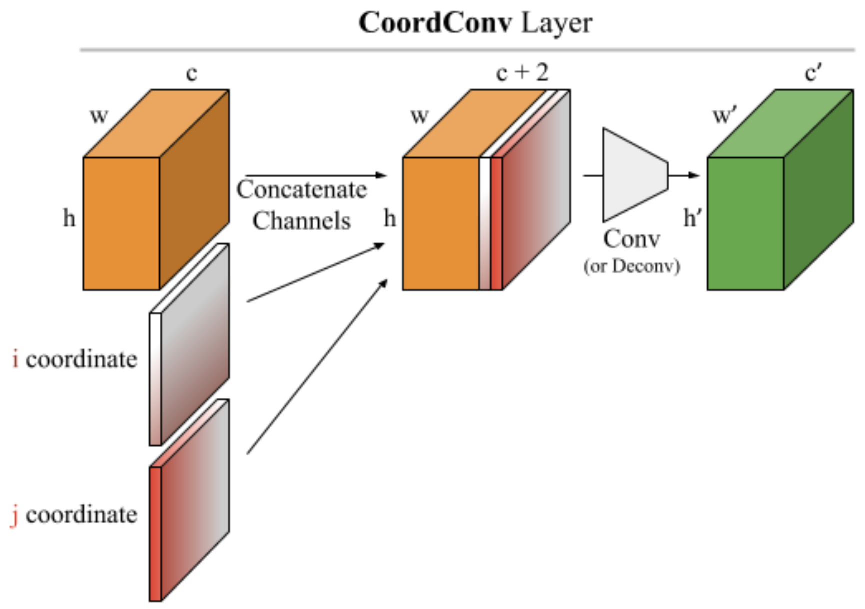

CoordConv Layer
c
w
h
Concatenate Channels
c + 2
w
h
Conv
(or Deconv)
c'
w'
h'
i coordinate
j coordinate


$\theta^{(D)}$
$P\theta^{(d)}$
$\theta_0^{(D)}$
$d = 2$
$D = 3$


loss
θ_0
θ_T
θ dim-2
θ dim-1


PPLM
Attribute
Model
p(a|x)
Gradients
GPT-2
p(x)

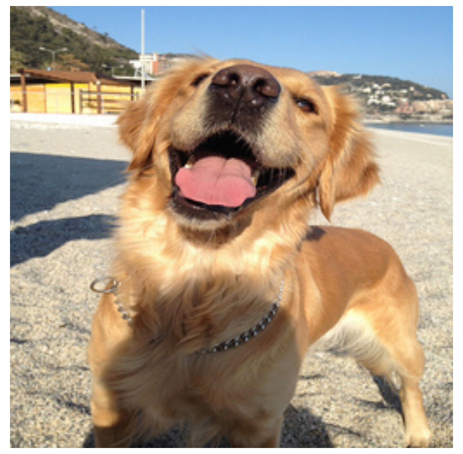
CoordConv



+ Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev

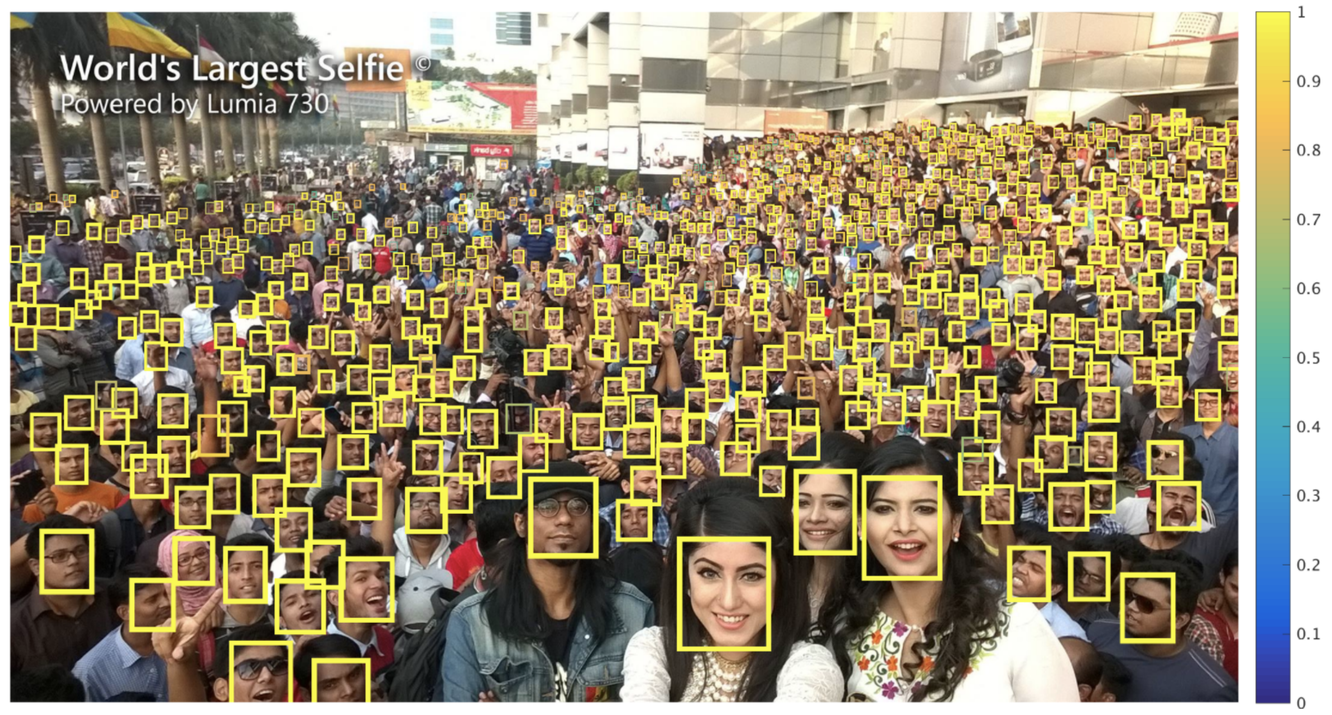
NeurIPS 2018

http://www.rosanneliu.com/publication/coordconv/

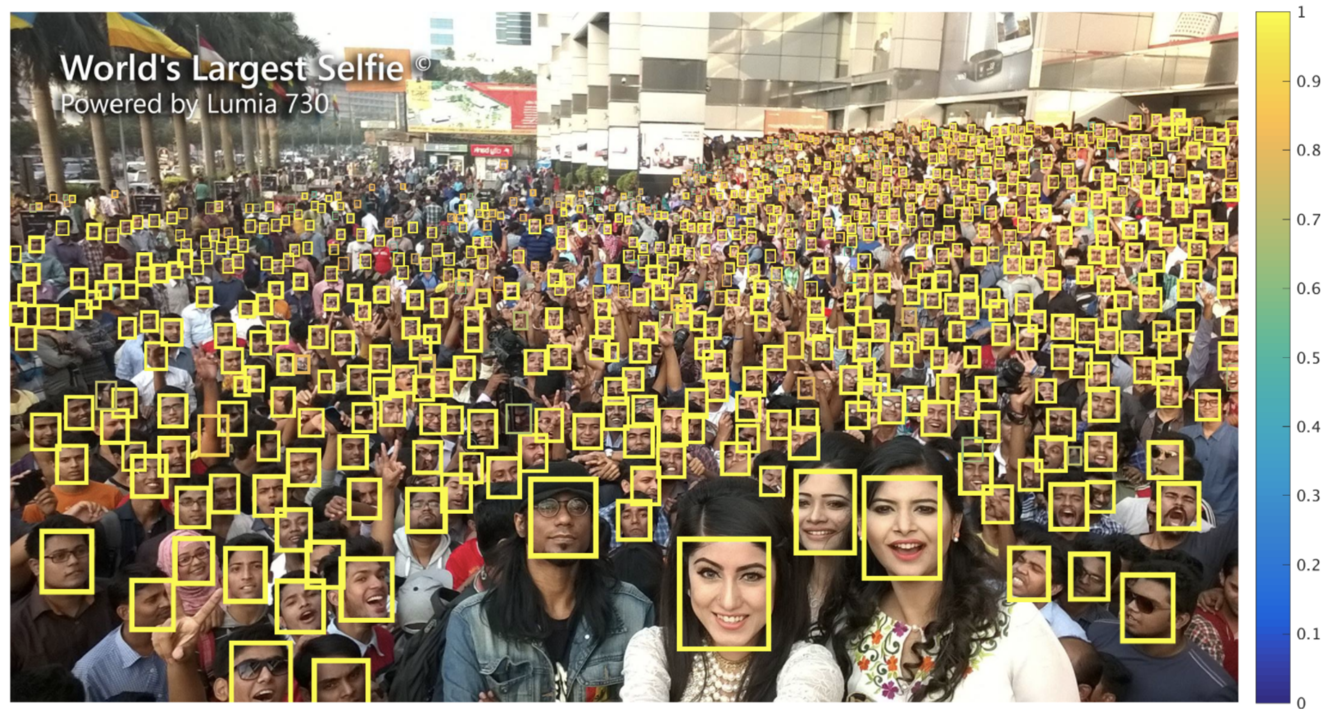


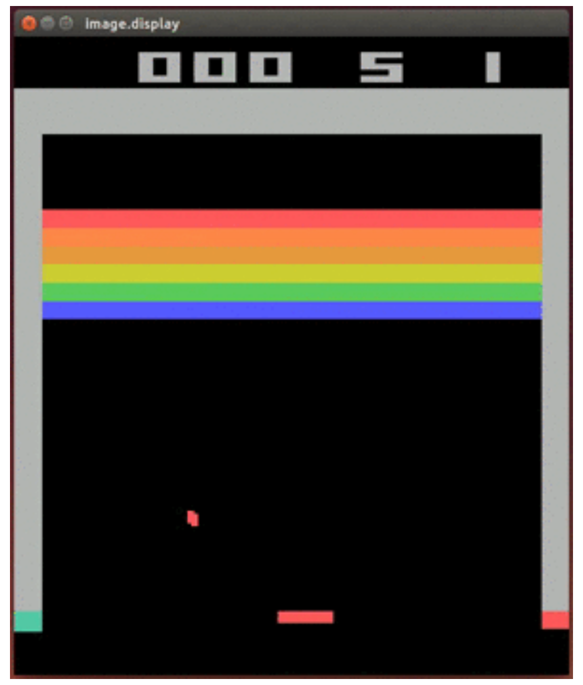
Good boy!
(golden retriever)

Good boy!
(golden retriever)



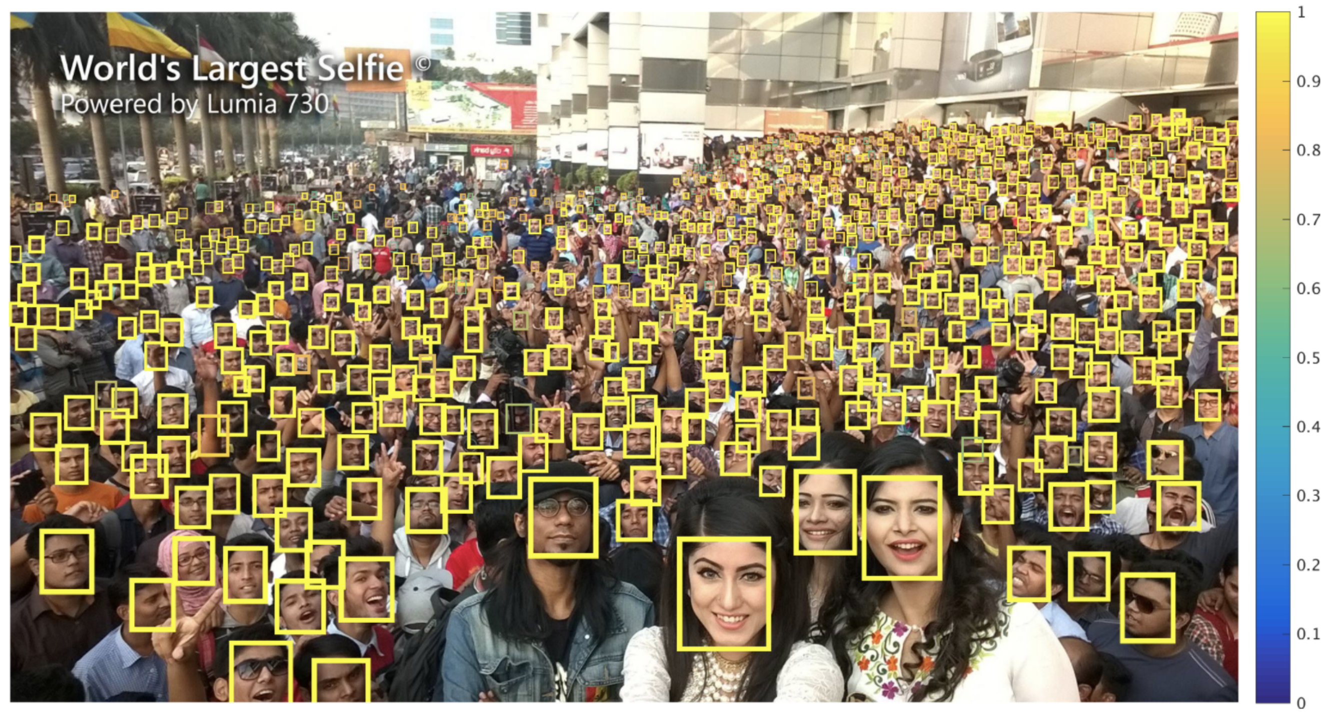
Good boy!
(golden retriever)

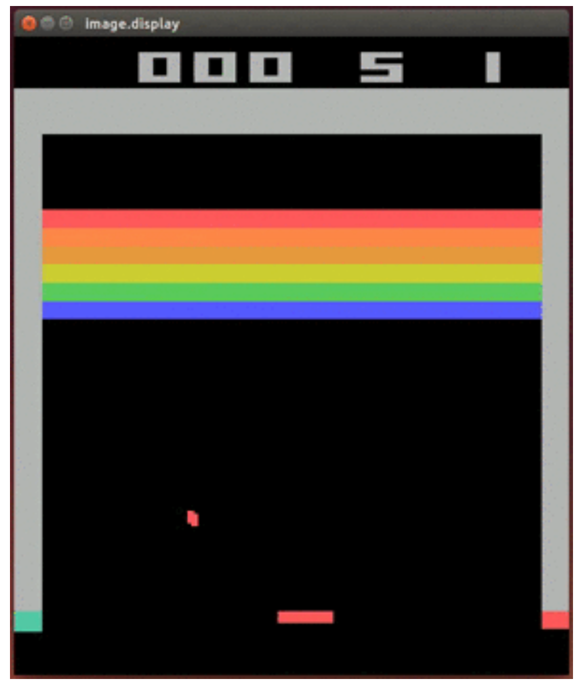




Actions

Good boy!
(golden retriever)





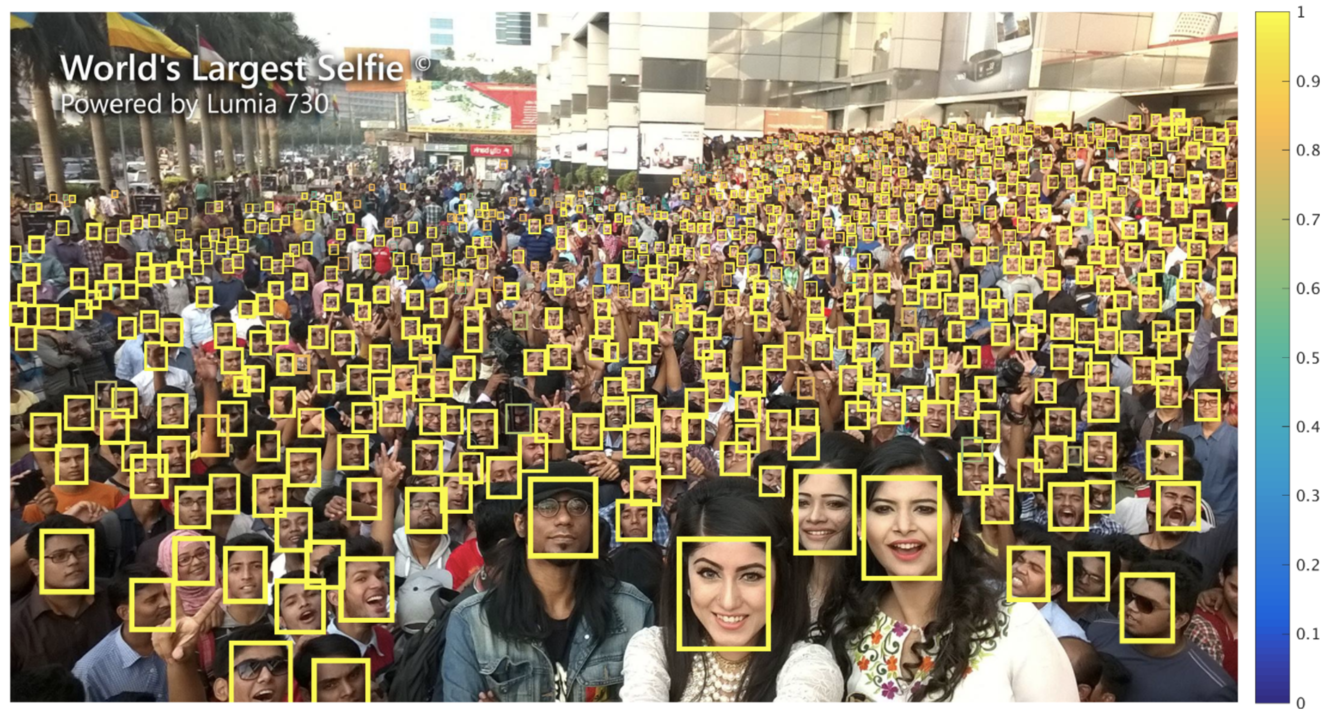
Actions

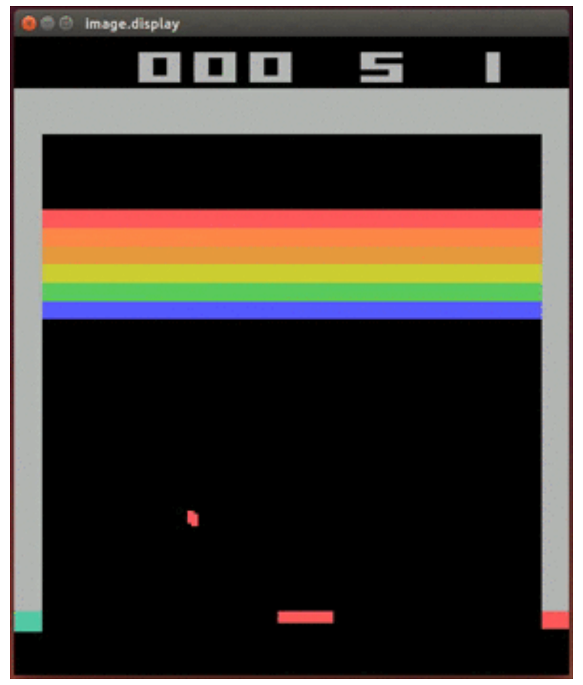
Noise

Convolutional neural networks
(convnet)

Good boy!
(golden retriever)

convnet





convnet

Actions

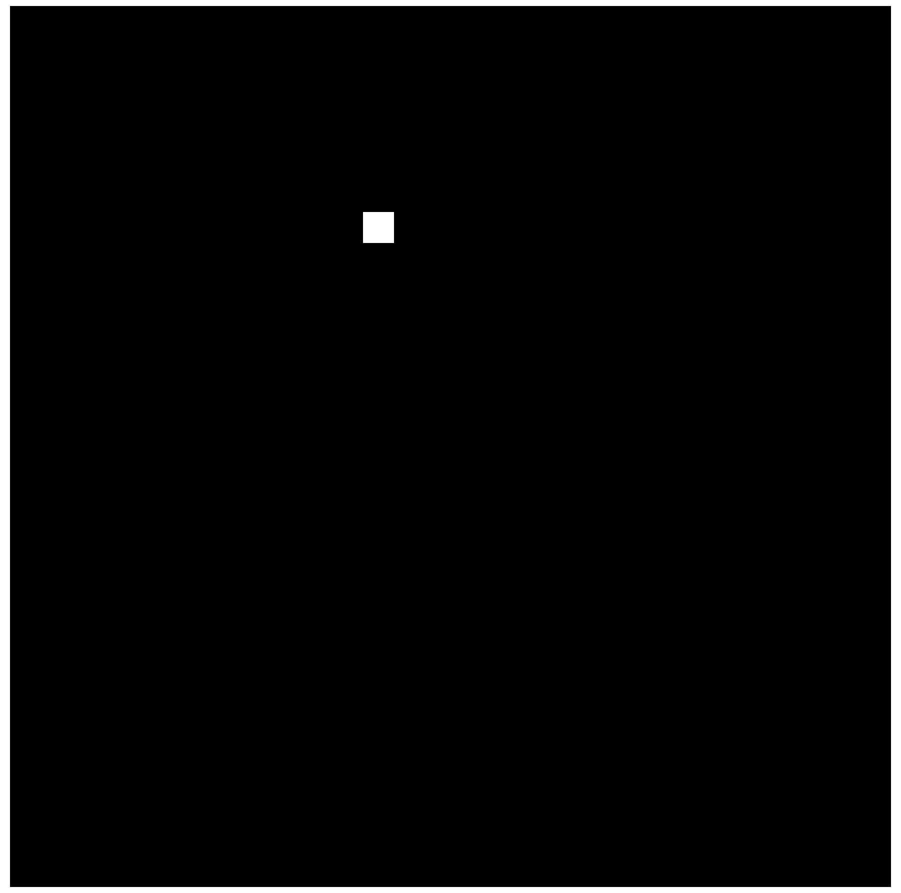
convnet

Noise

[0.23, 1,45, 2.3, 3,03, 1,21, …] $\xrightarrow{\text{convnet}}$

[0.23, 1,45, 2.3, 3,03, 1,21, …] $\xrightarrow{\text{convnet}}$

“(4, 6)” $\xrightarrow{\text{convnet}}$

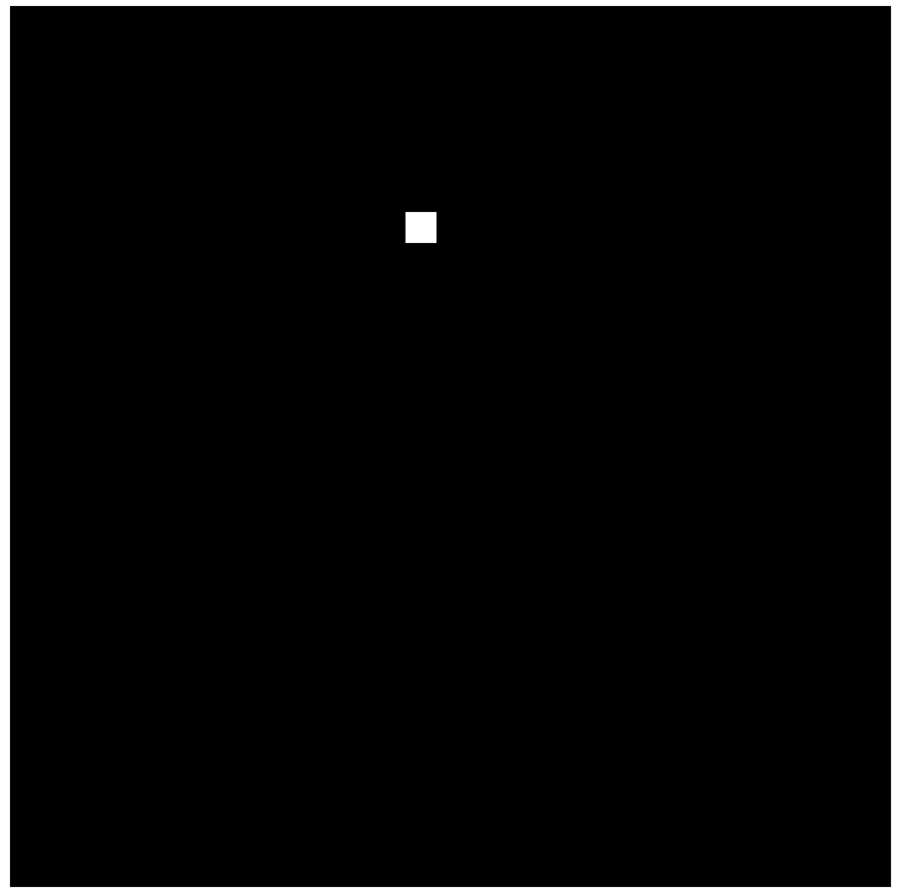


[0.23, 1,45, 2.3, 3,03, 1,21, …]

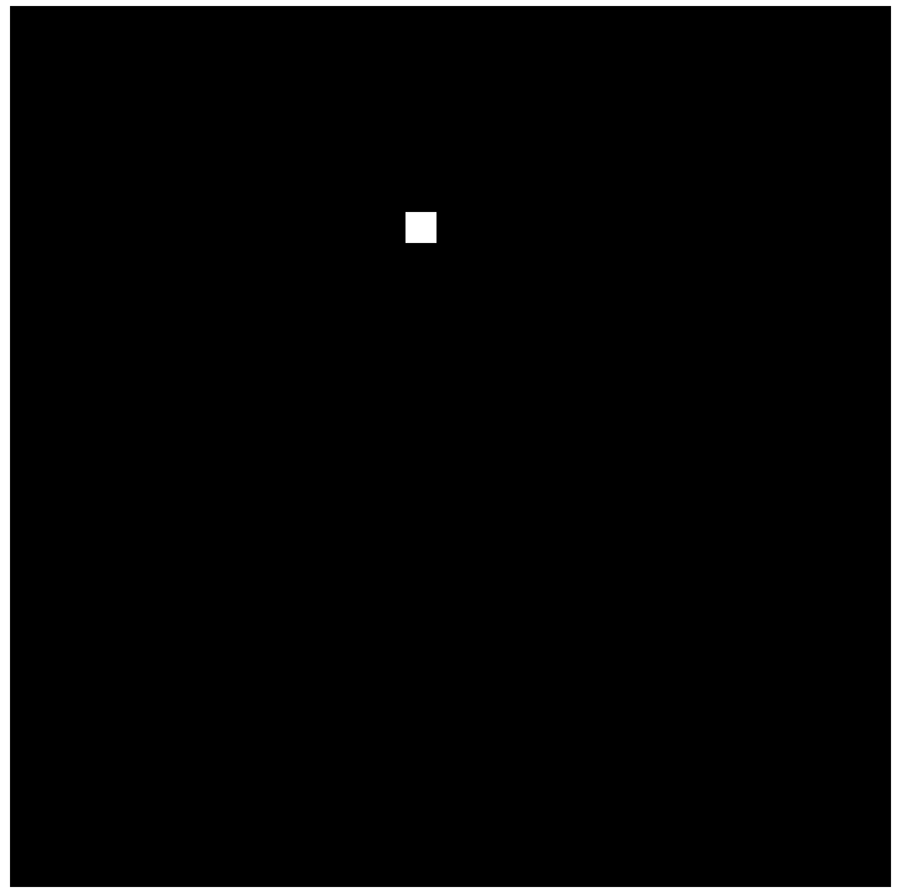
convnet

“(4, 7)”

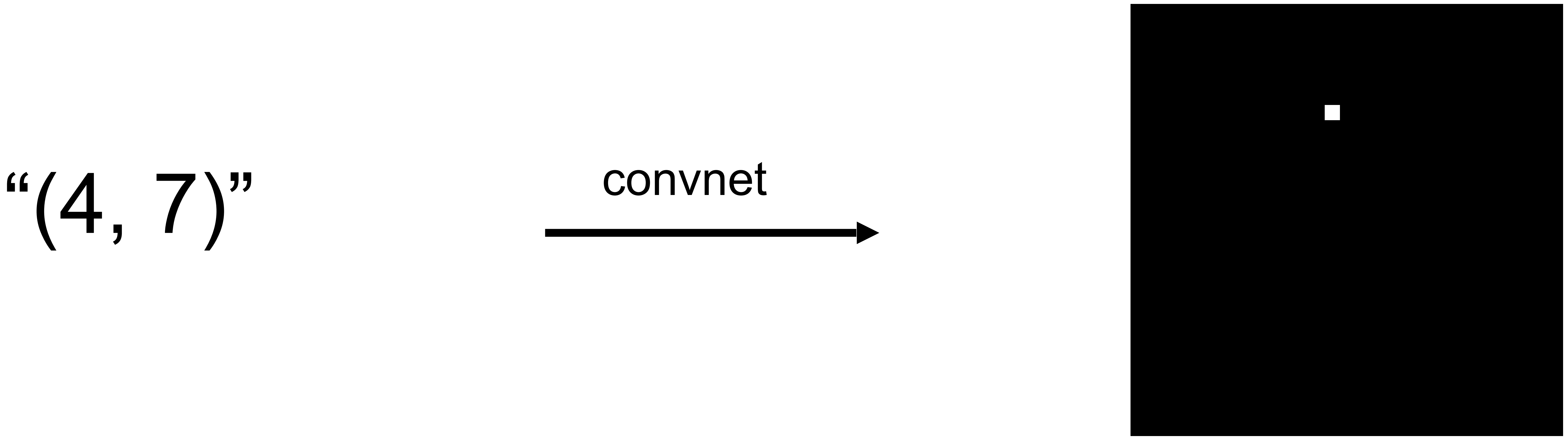
convnet



“(4, 7)” $\xrightarrow{\text{convnet}}$



Coordinate Transform:
Given a *Cartesian* location, highlight *that pixel* on a canvas.

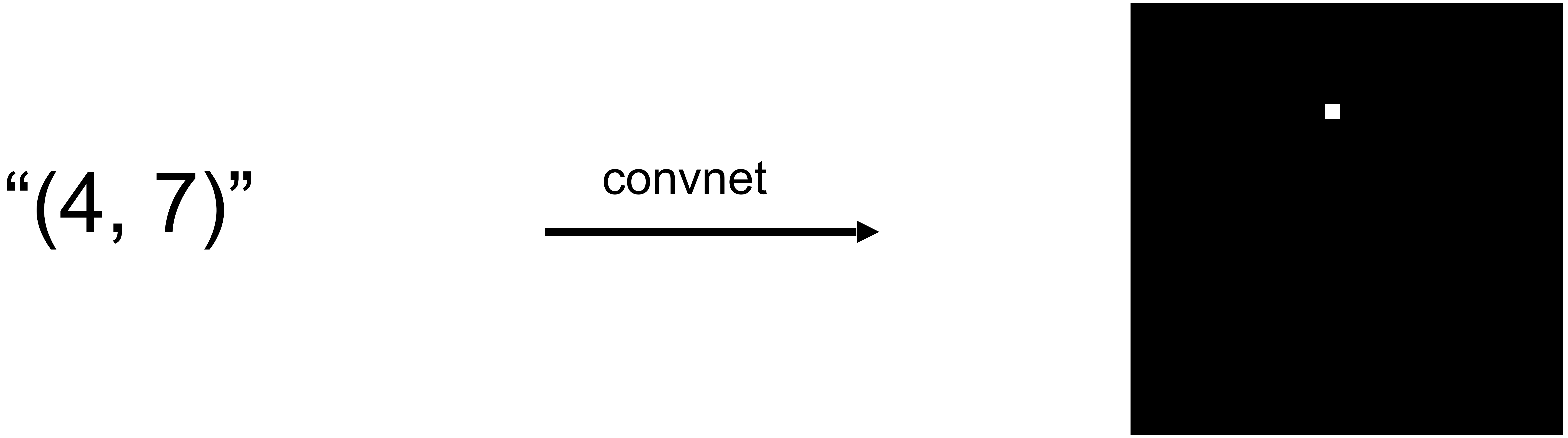


“(4, 7)”

convnet

Convnets fail at this simple pixel task

Coordinate Transform:
Given a *Cartesian* location, highlight *that pixel* on a canvas.

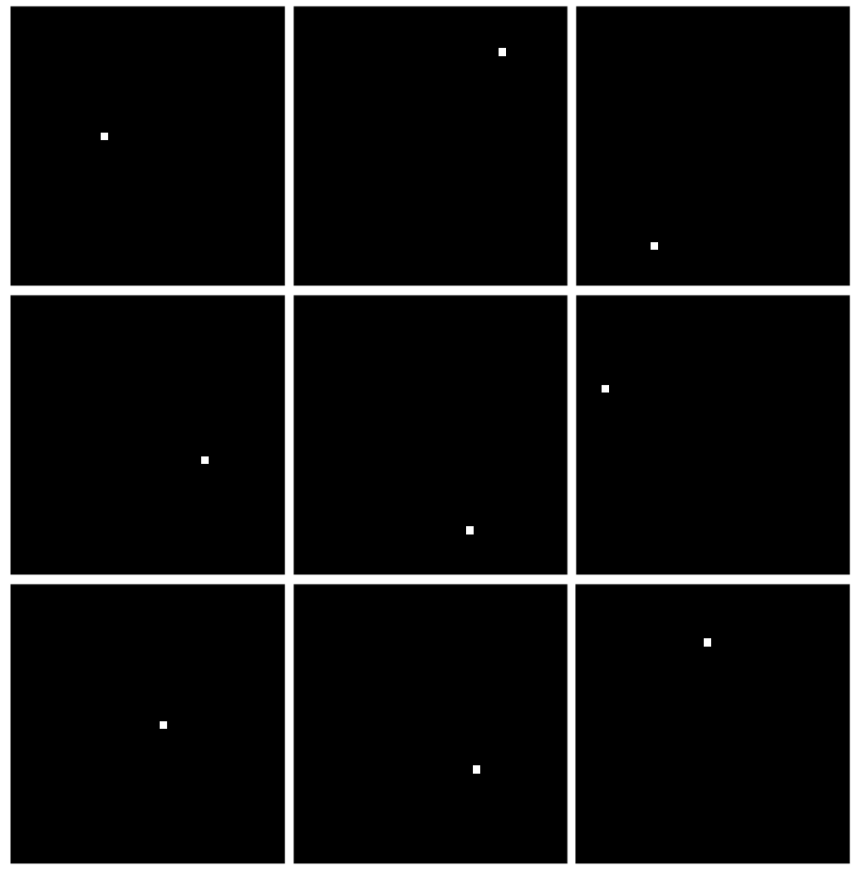


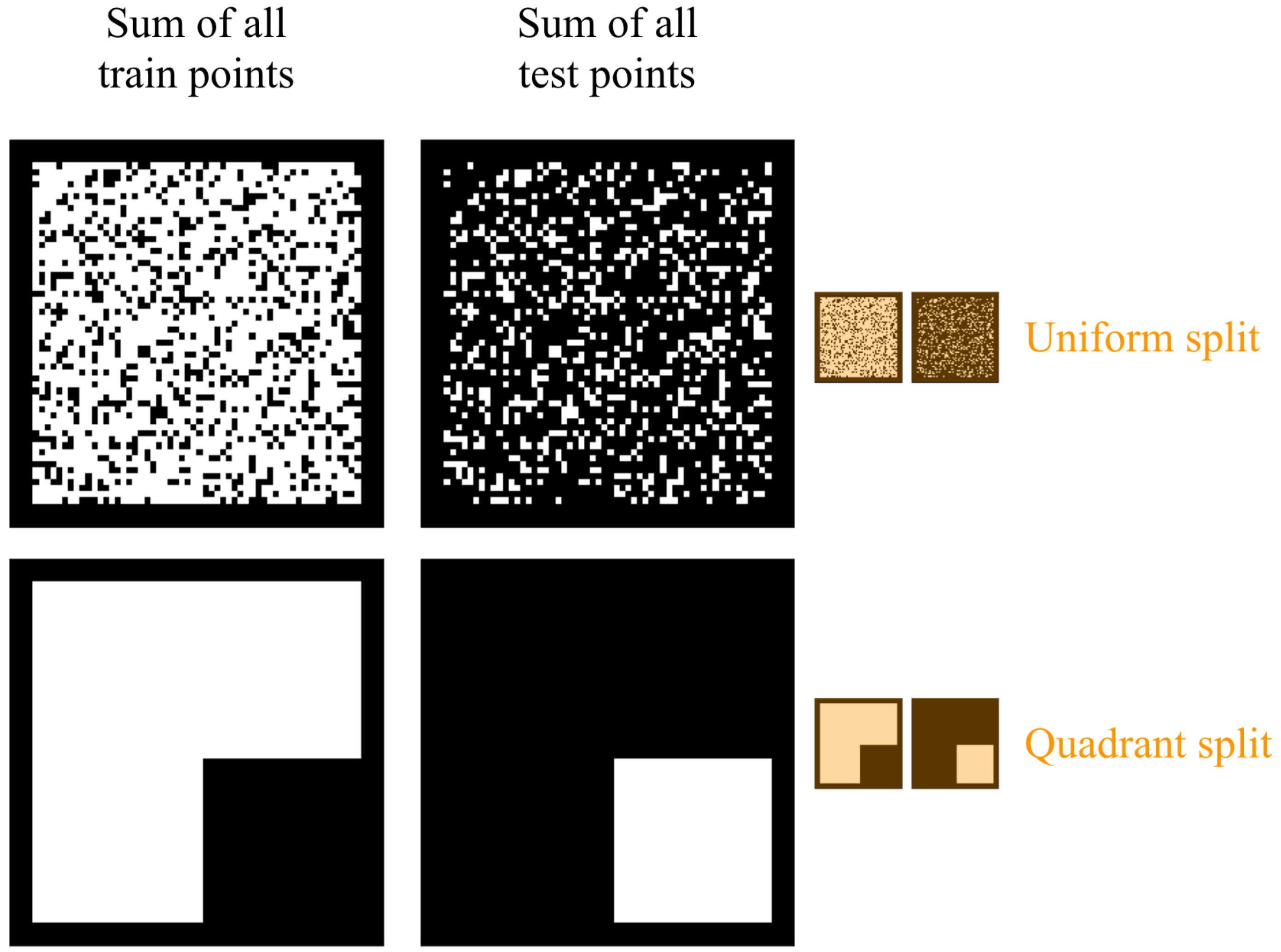
“(4, 7)”

convnet

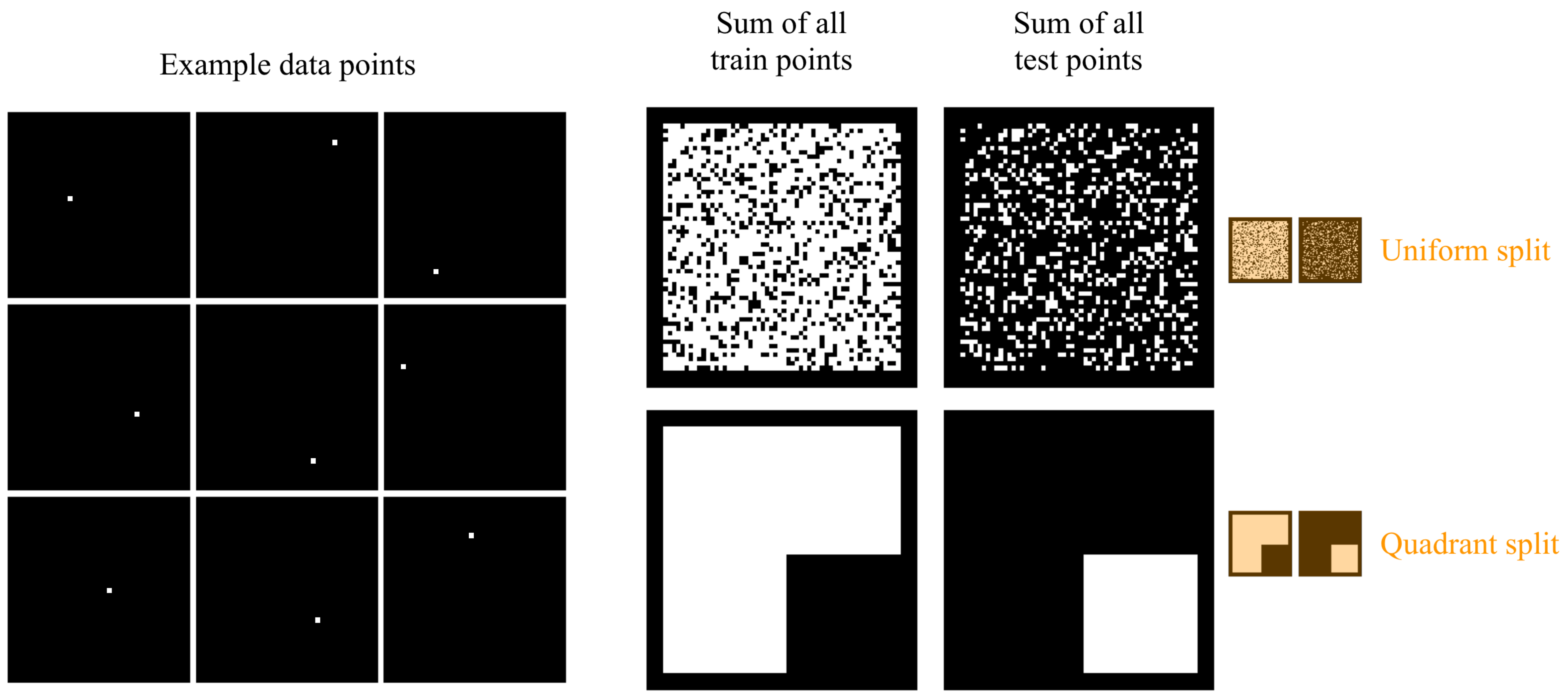
Convnets fail at this simple pixel task

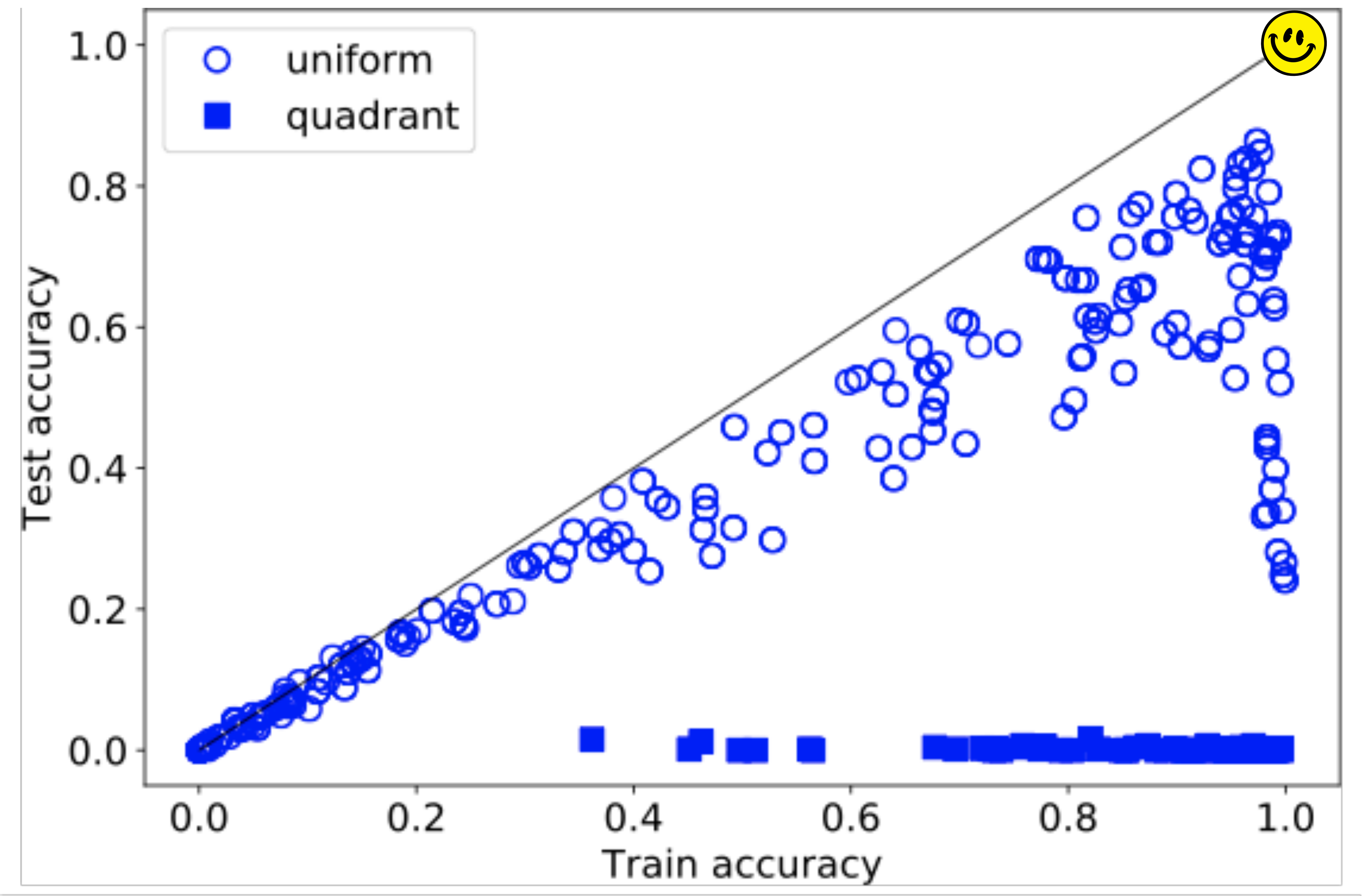
Example data points



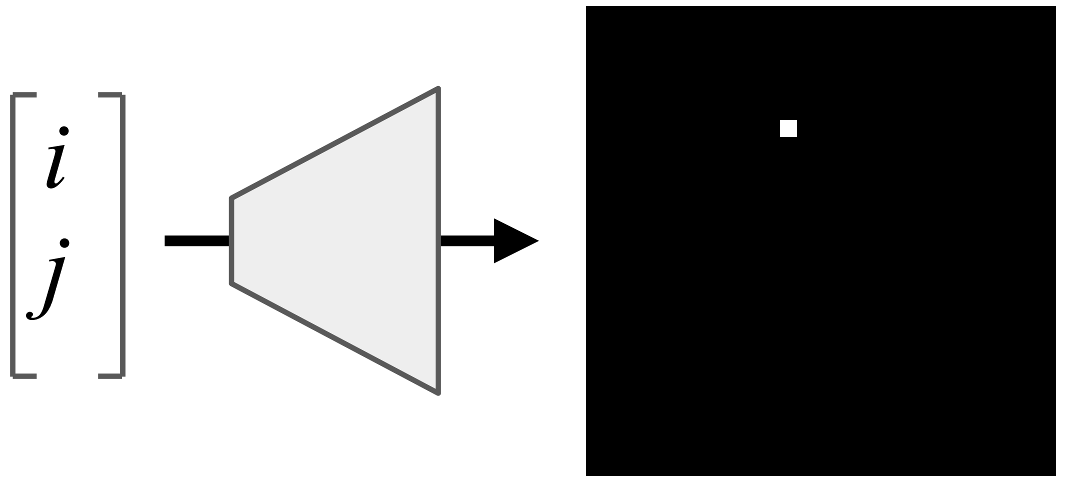


Convnets fail at this simple pixel task





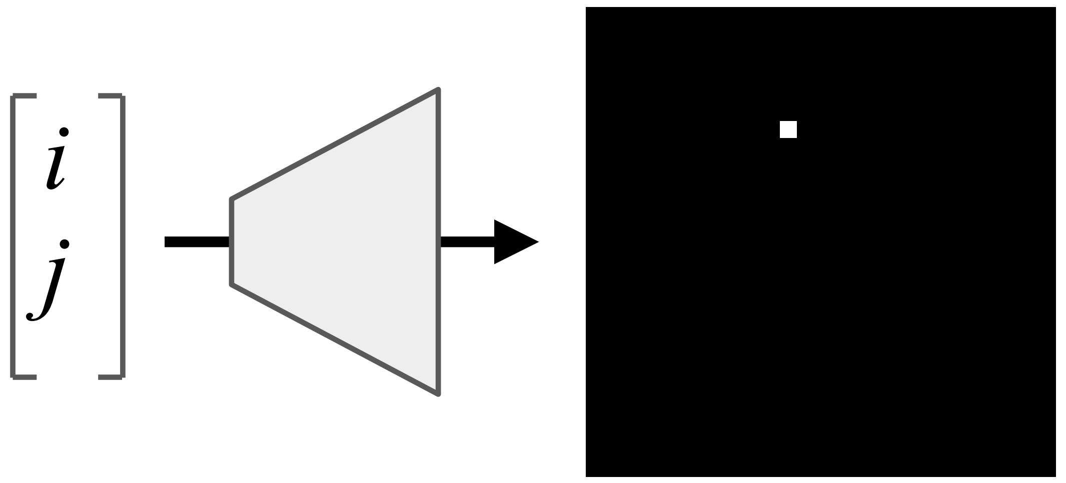
Convnets don’t know how to paint a pixel



Coordinate Transform
Output: per-pixel sigmoid
Loss: supervised cross-entropy

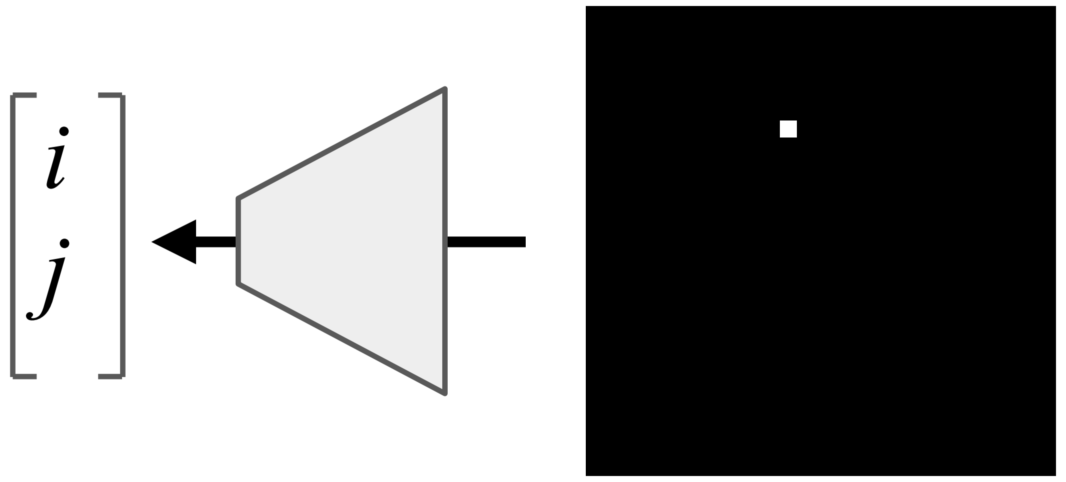
✗ Harder than expected

Convnets don’t know how to paint a pixel, or to locate one



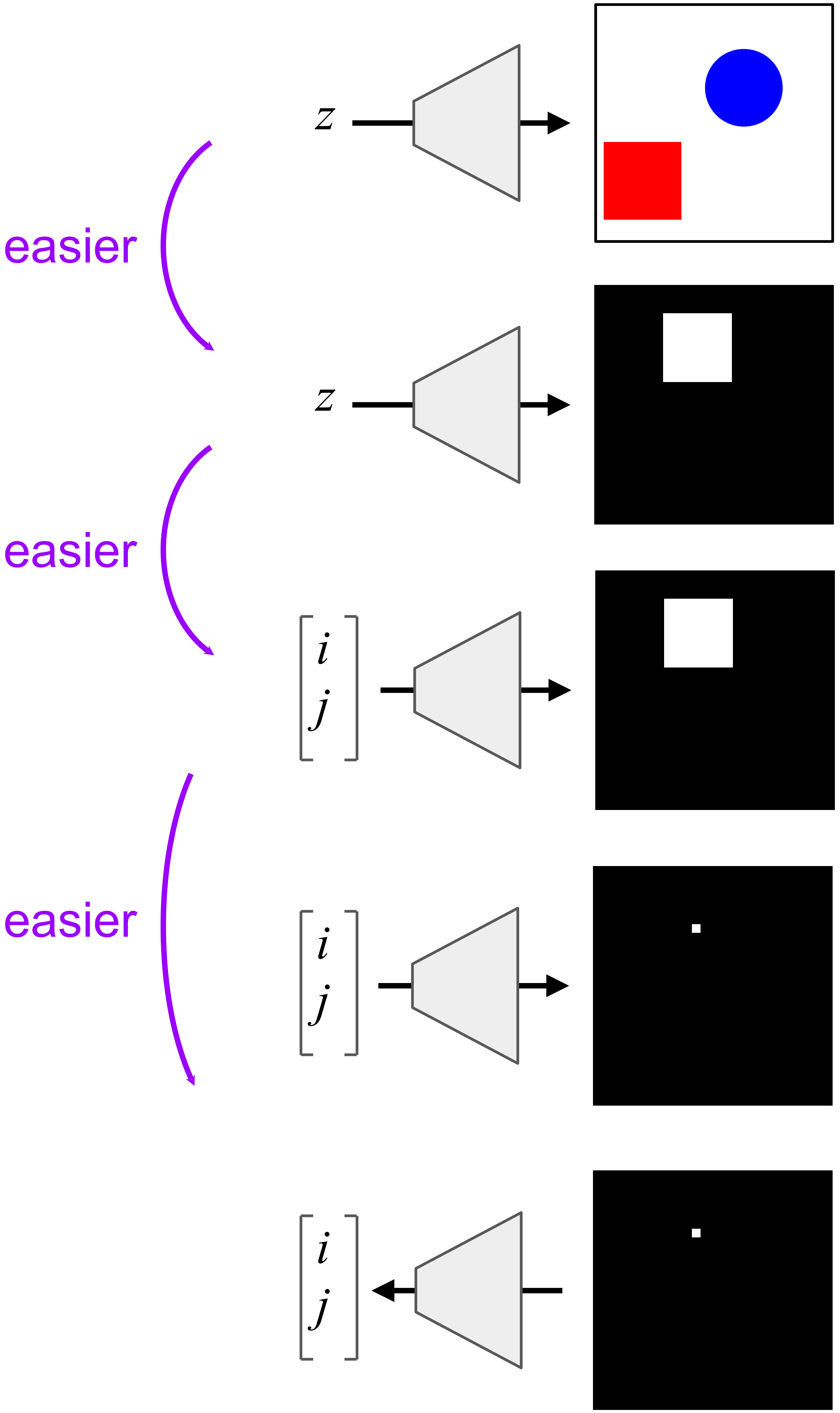
Coordinate Transform
Output: per-pixel sigmoid
Loss: supervised cross-entropy

✗ Harder than expected



Coordinate Transform
Output: linear
Loss: supervised mse

✗ Harder than expected



Output: per-pixel, per-channel sigmoid
Loss: learned GAN discriminator

X Harder than expected

Output: per-pixel sigmoid
Loss: learned GAN discriminator

X Harder than expected

Output: per-pixel sigmoid
Loss: supervised cross-entropy

X Harder than expected

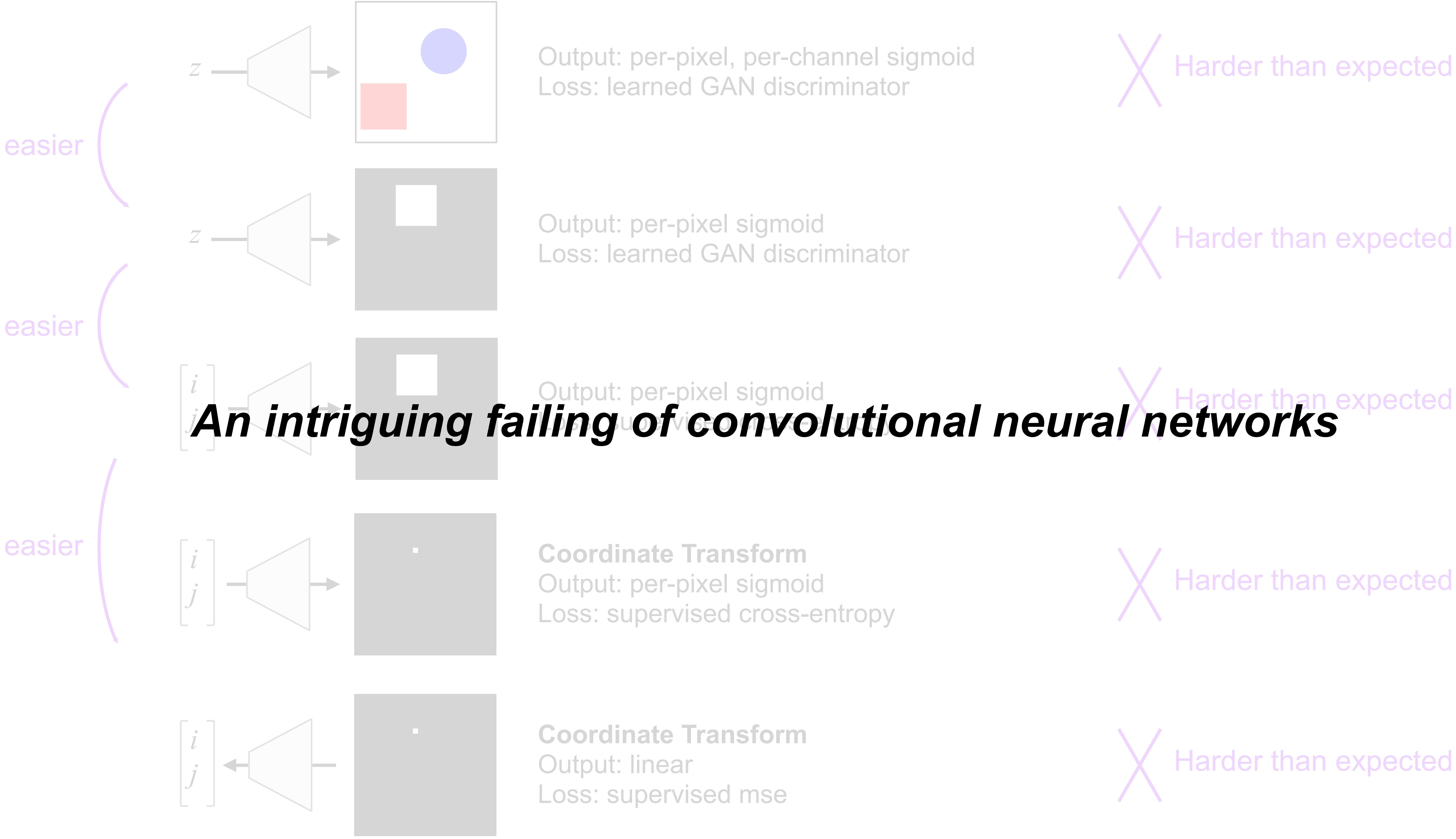
Coordinate Transform
Output: per-pixel sigmoid
Loss: supervised cross-entropy

X Harder than expected

Coordinate Transform
Output: linear
Loss: supervised mse

X Harder than expected

An intriguing failing of convolutional neural networks



Output: per-pixel, per-channel sigmoid
Loss: learned GAN discriminator

Harder than expected

Output: per-pixel sigmoid
Loss: learned GAN discriminator

Harder than expected

Output: per-pixel sigmoid
Loss: supervised cross-entropy

Harder than expected

Coordinate Transform
Output: per-pixel sigmoid
Loss: supervised cross-entropy

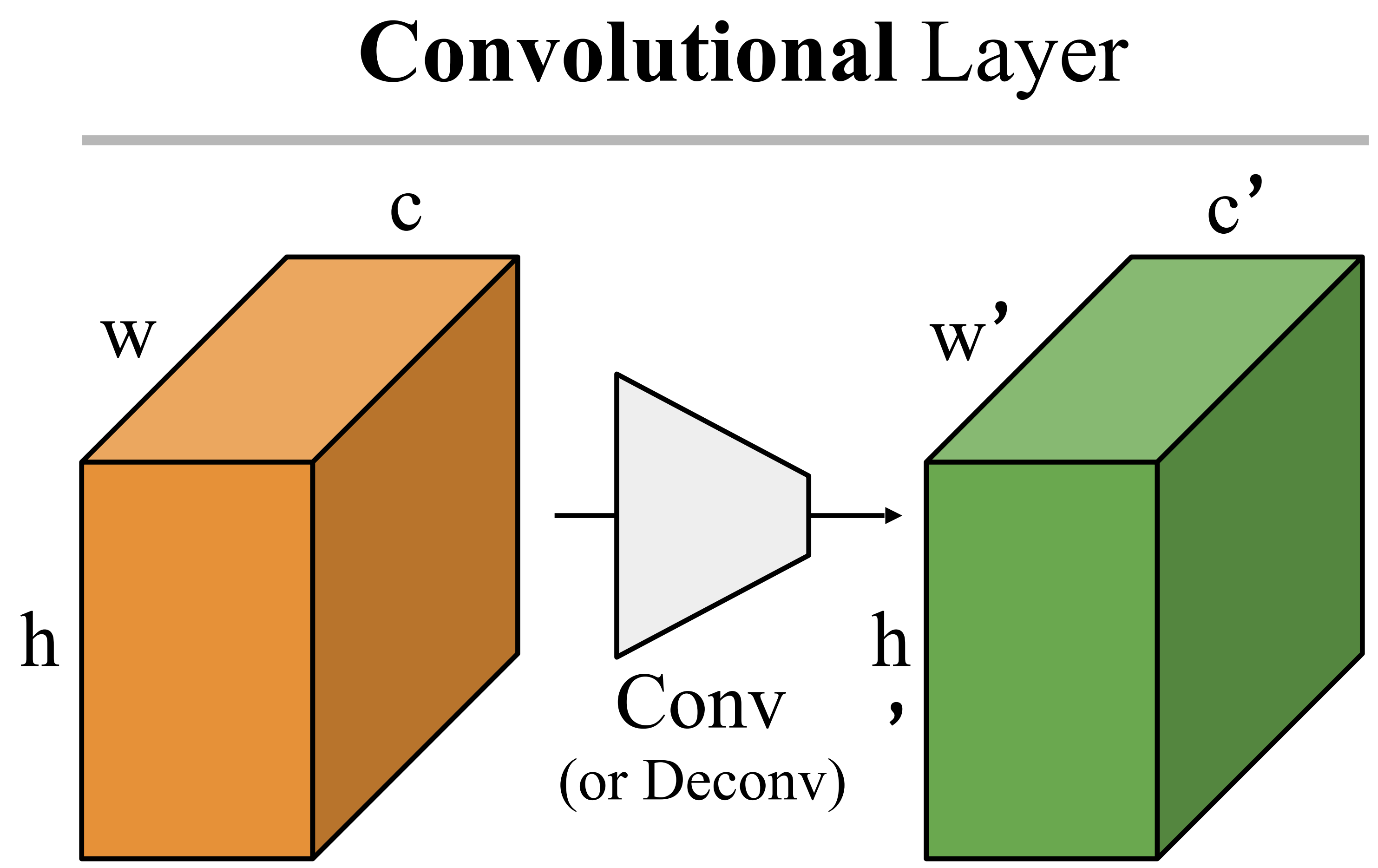
Harder than expected

Coordinate Transform
Output: linear
Loss: supervised mse

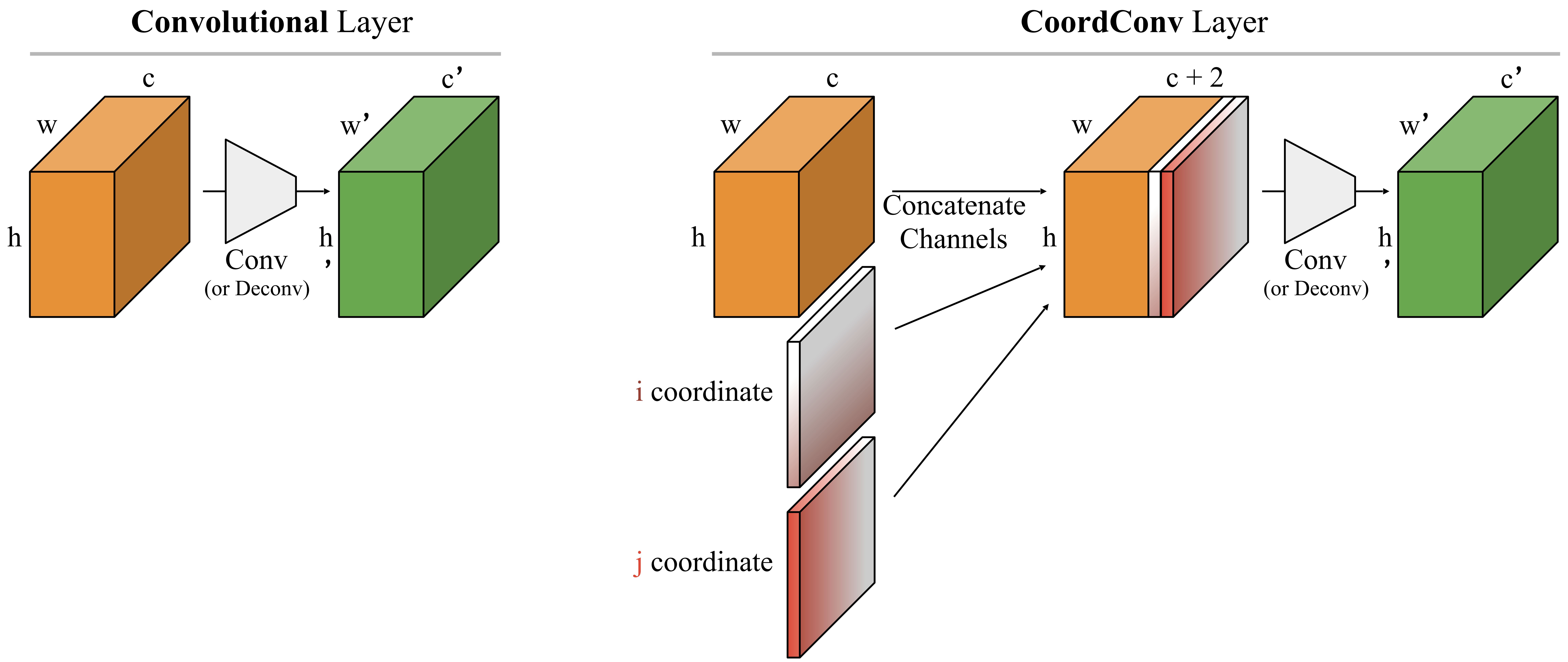
Harder than expected

The *CoordConv* solution

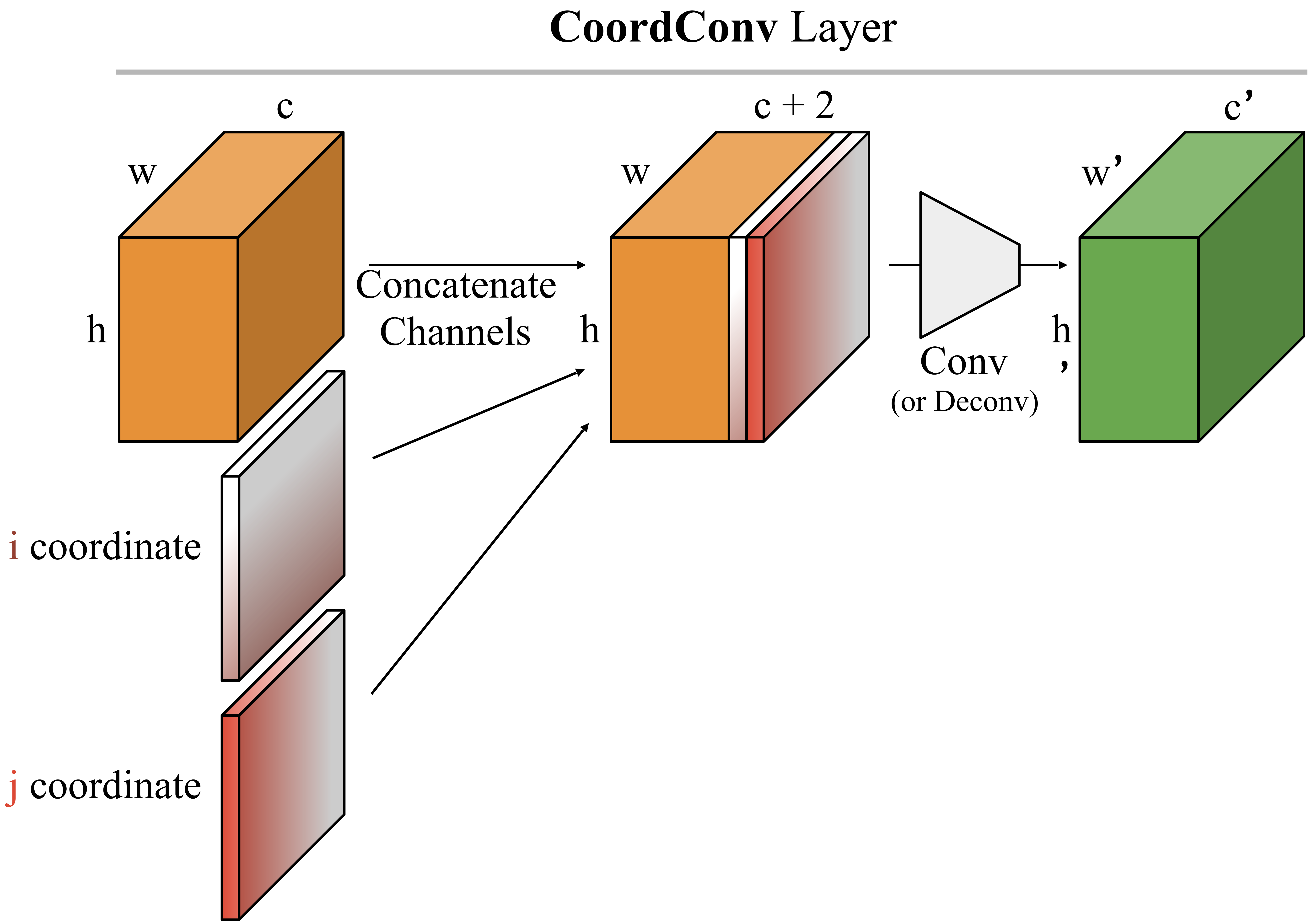
The *CoordConv* solution



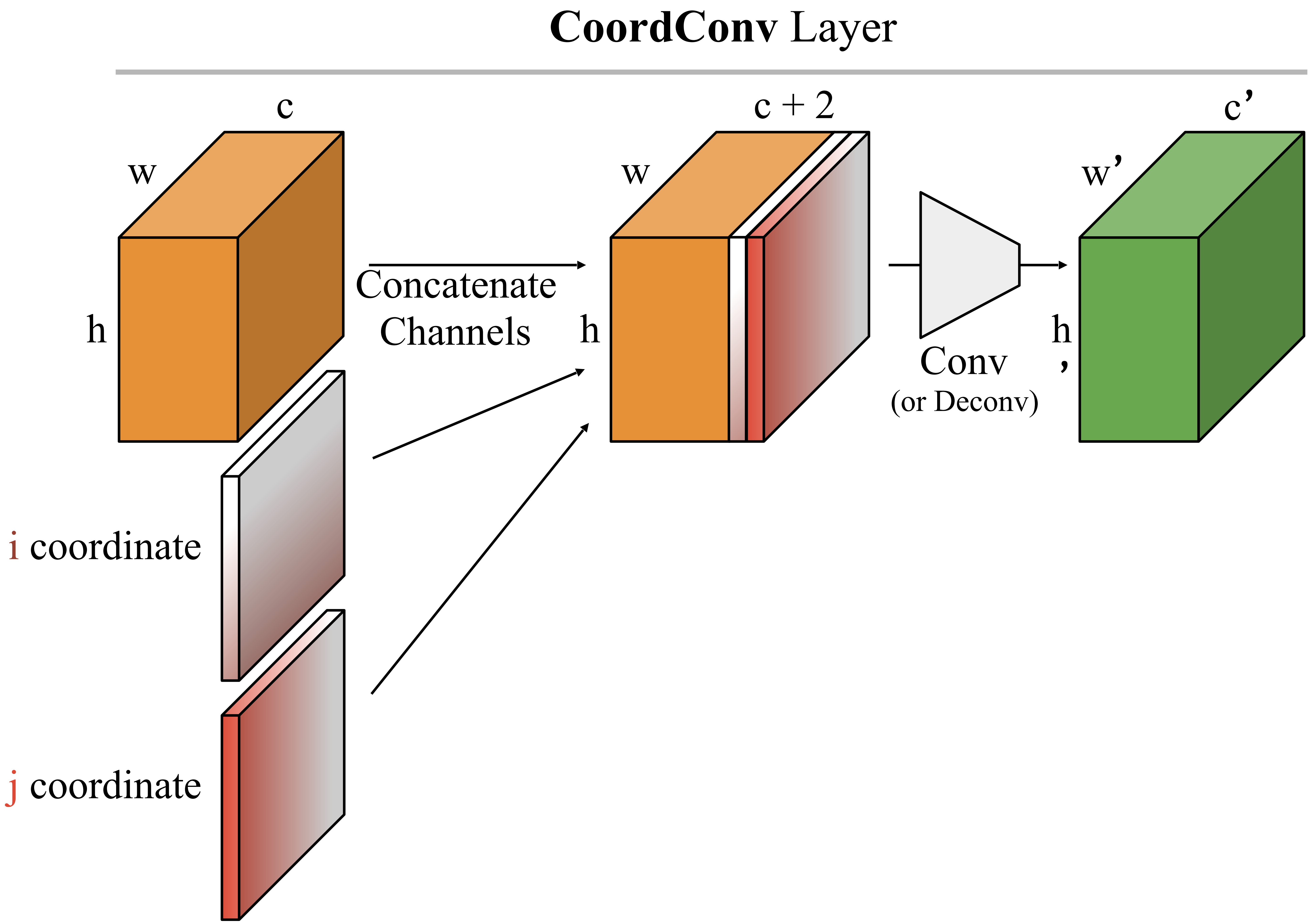
The *CoordConv* solution



The *CoordConv* solution



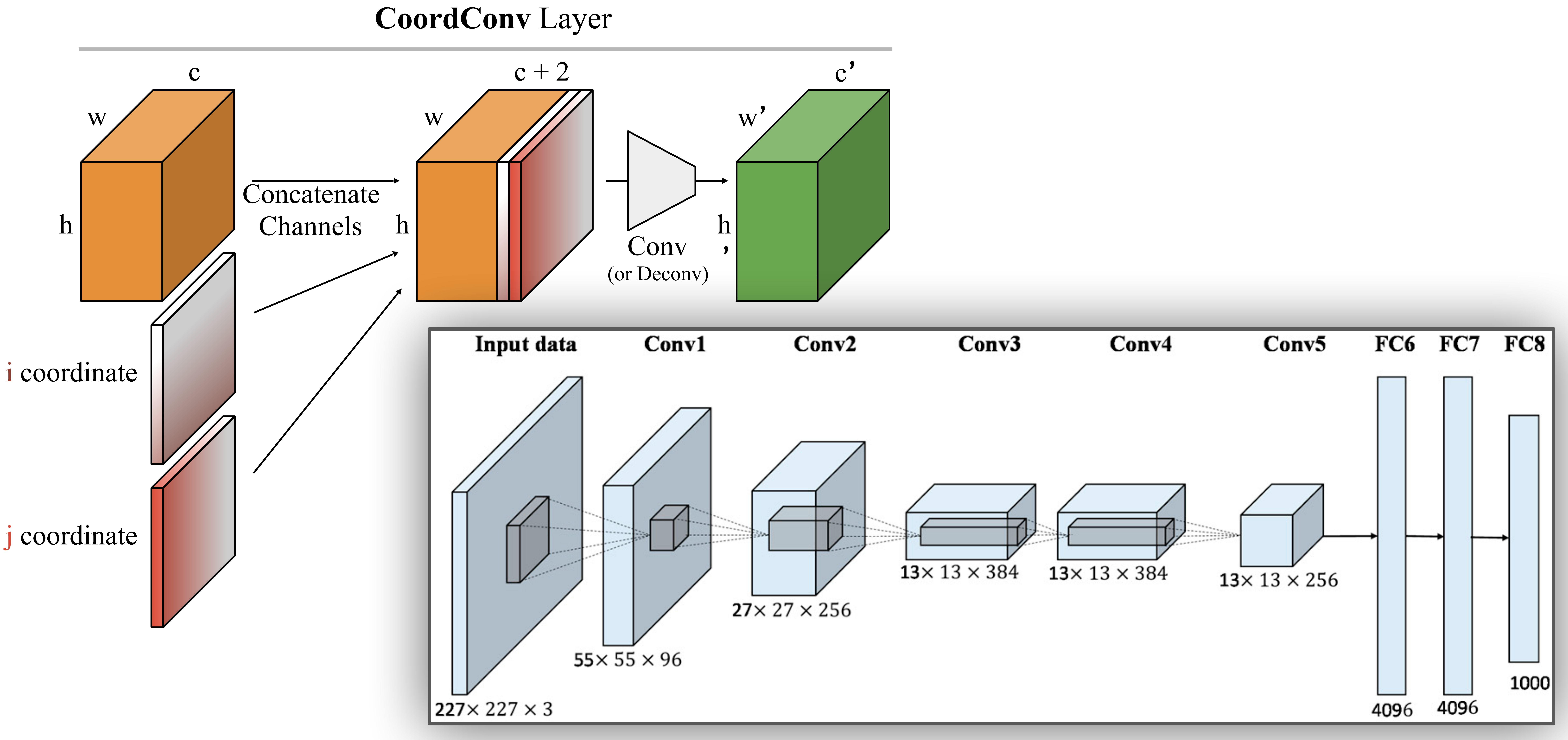
The *CoordConv* solution



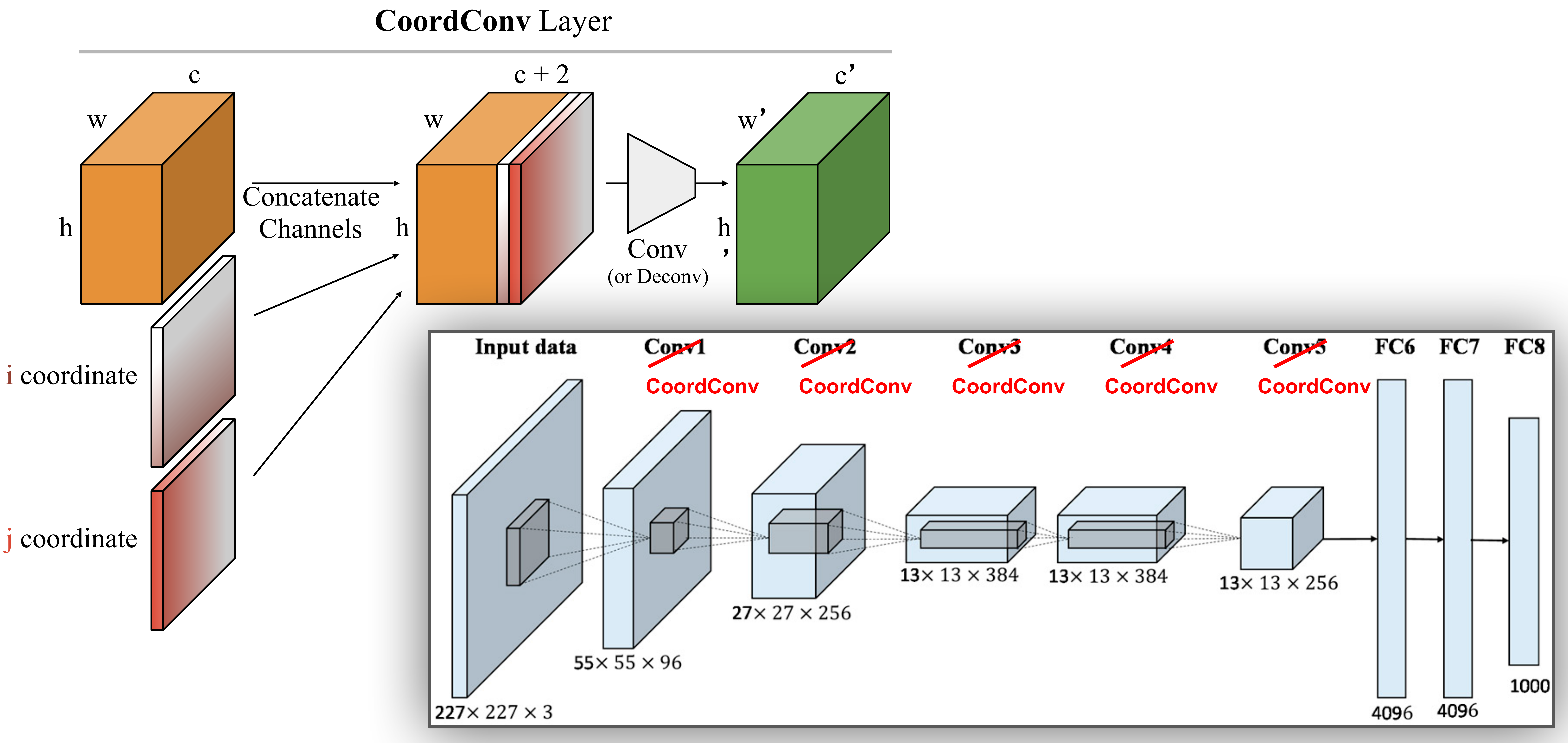
Great things about convolution:

- Few parameters
  (keep this)
- Fast computation on GPU
  (keep this)
- Translation equivariance
  (optionally learned or discarded, as needed)

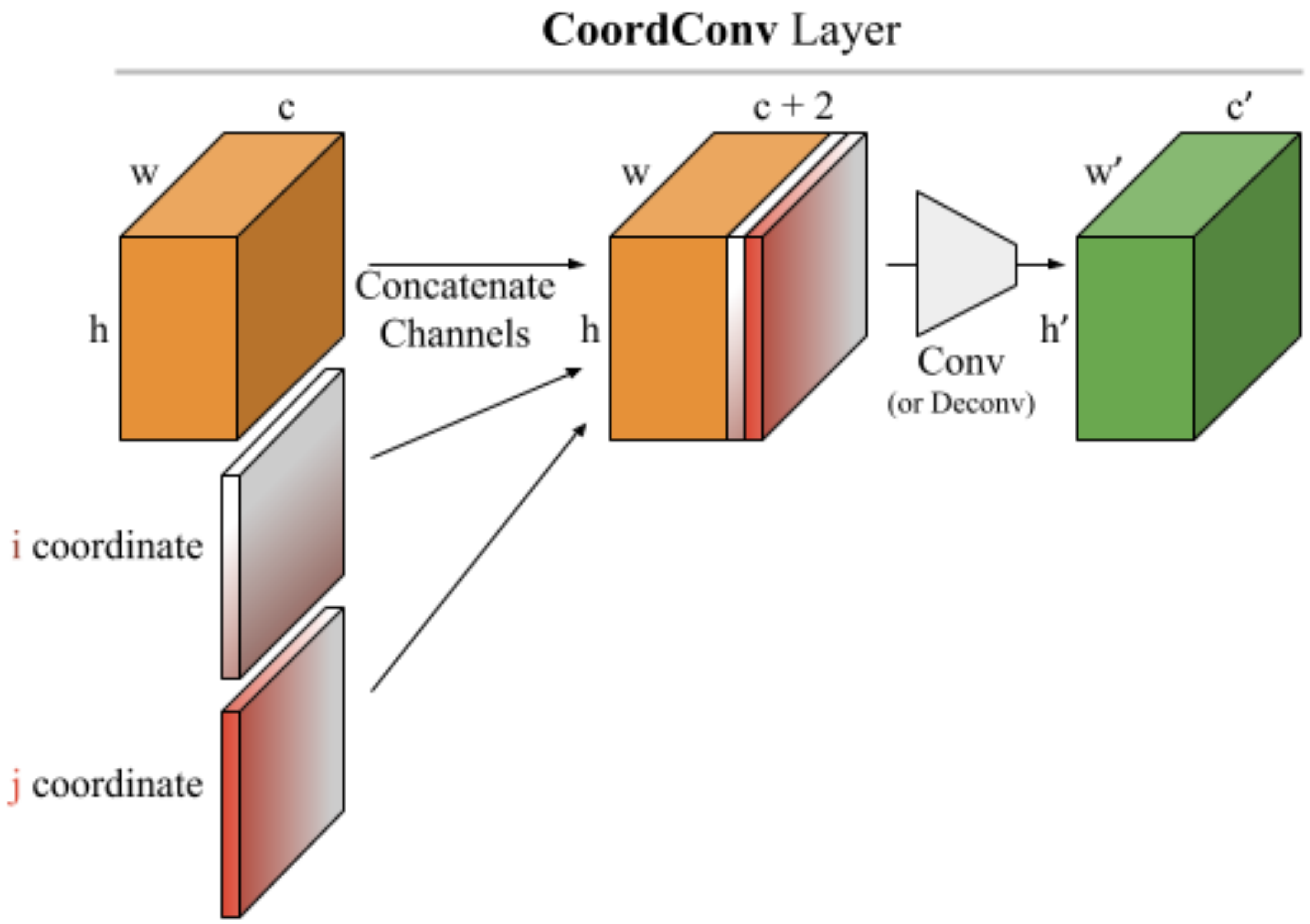
The *CoordConv* solution



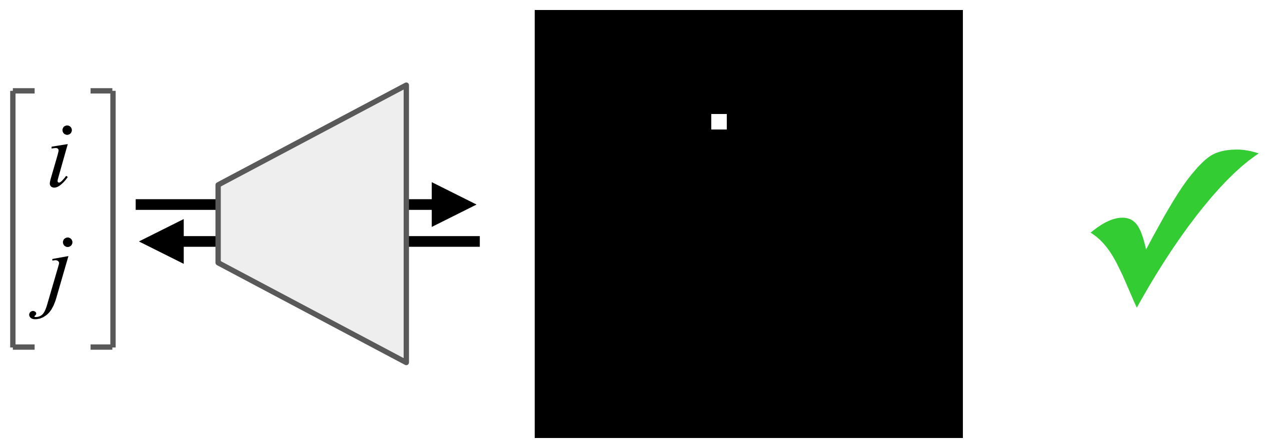
The *CoordConv* solution

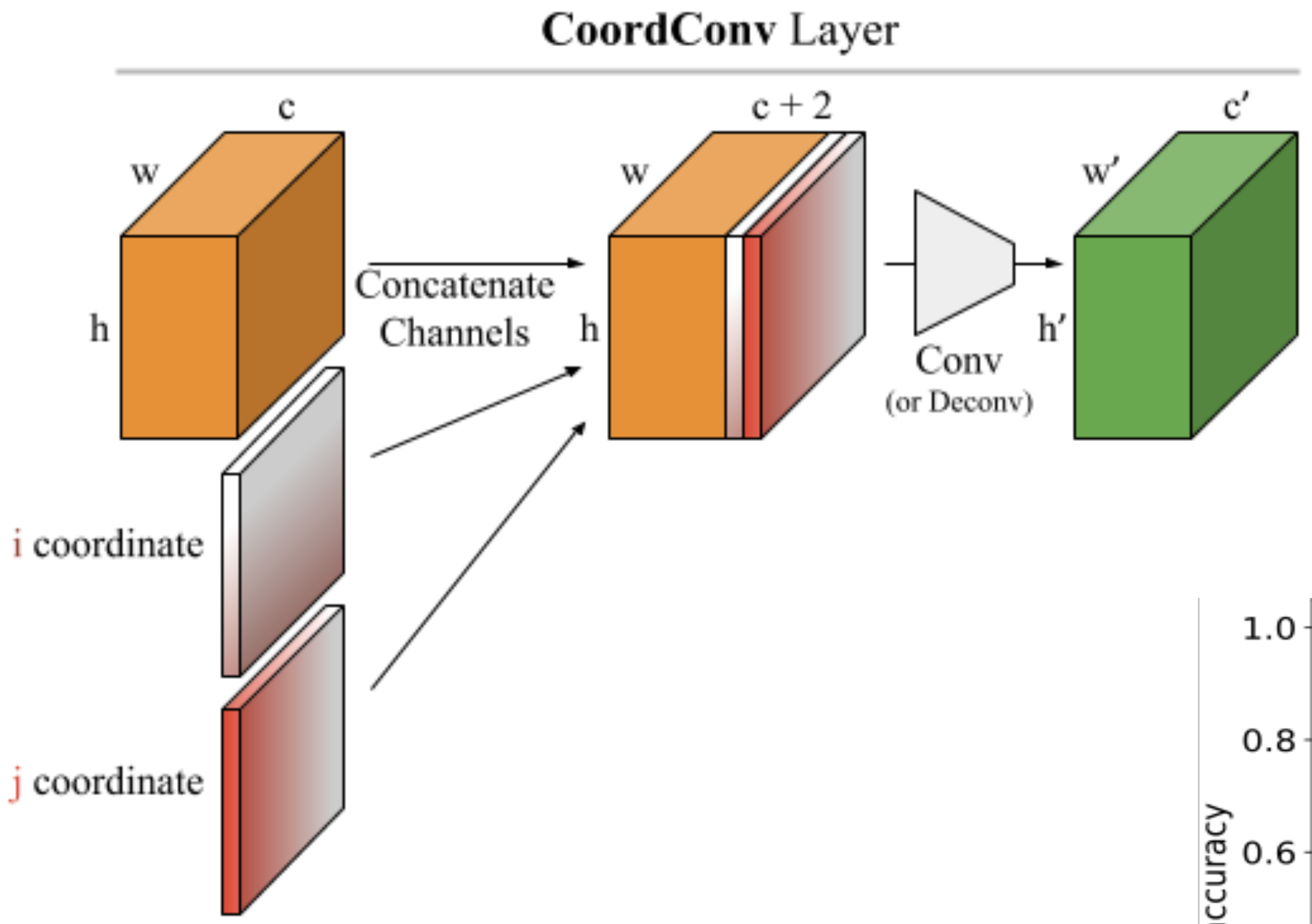


The *CoordConv* solution

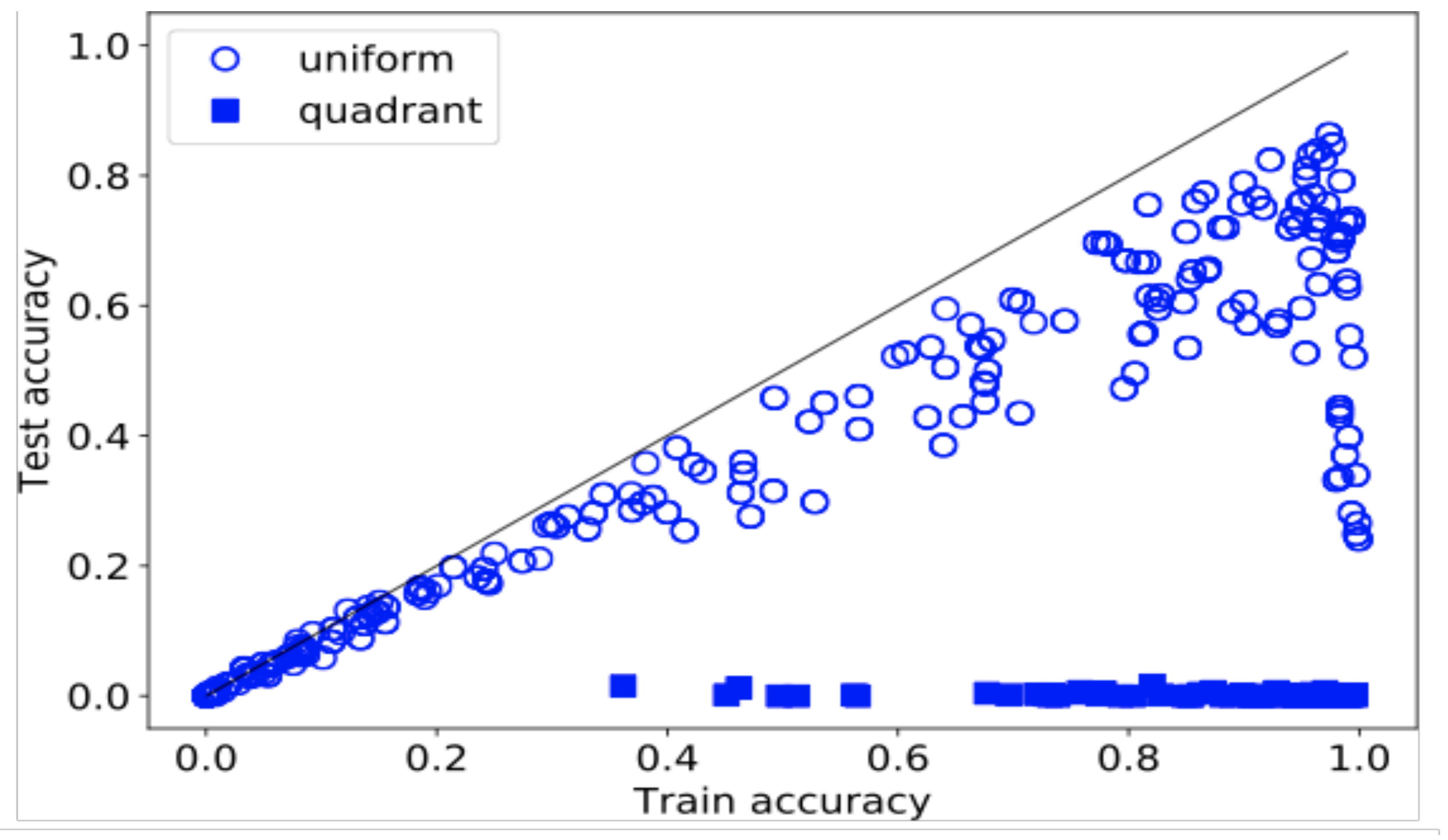


The *CoordConv* solution

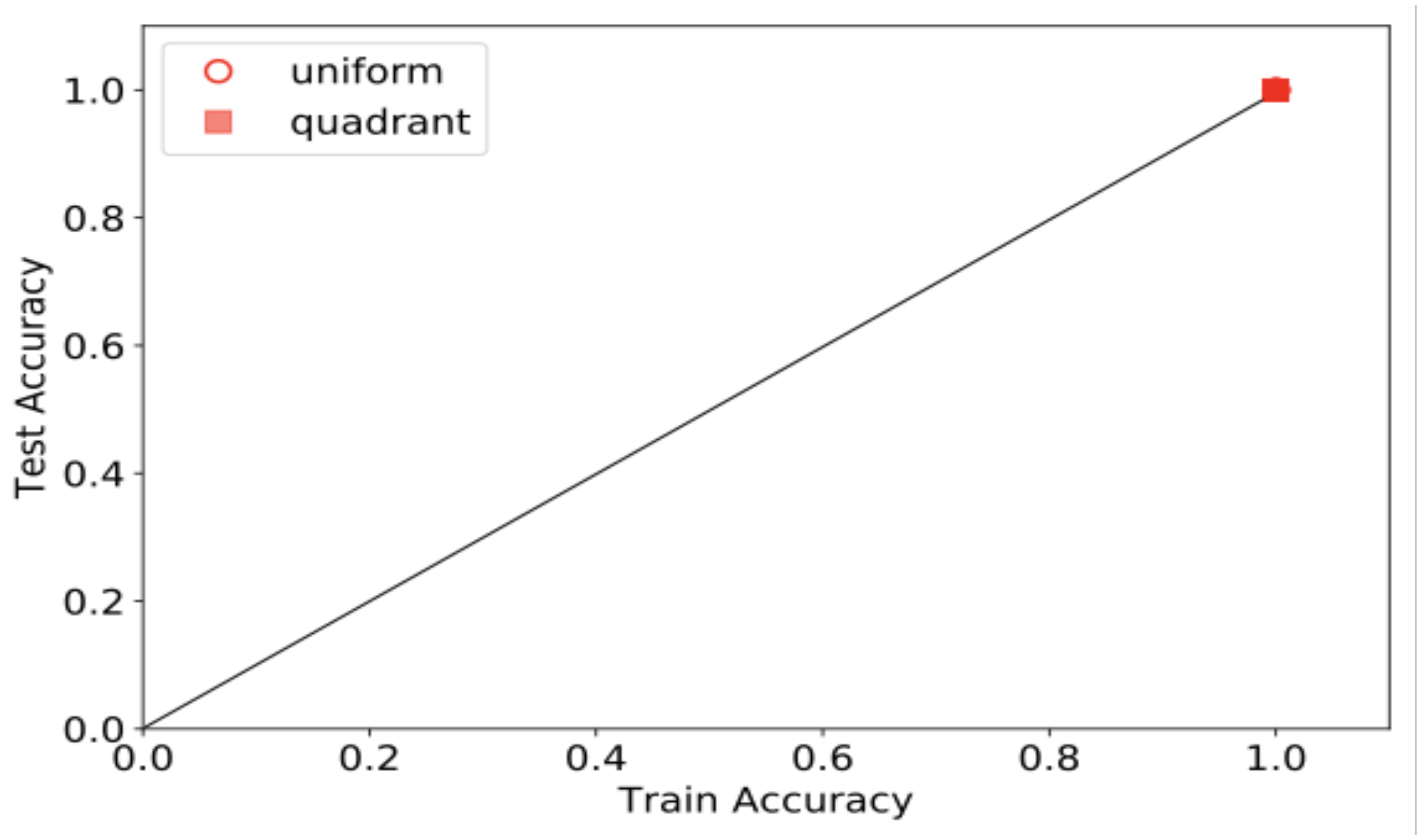




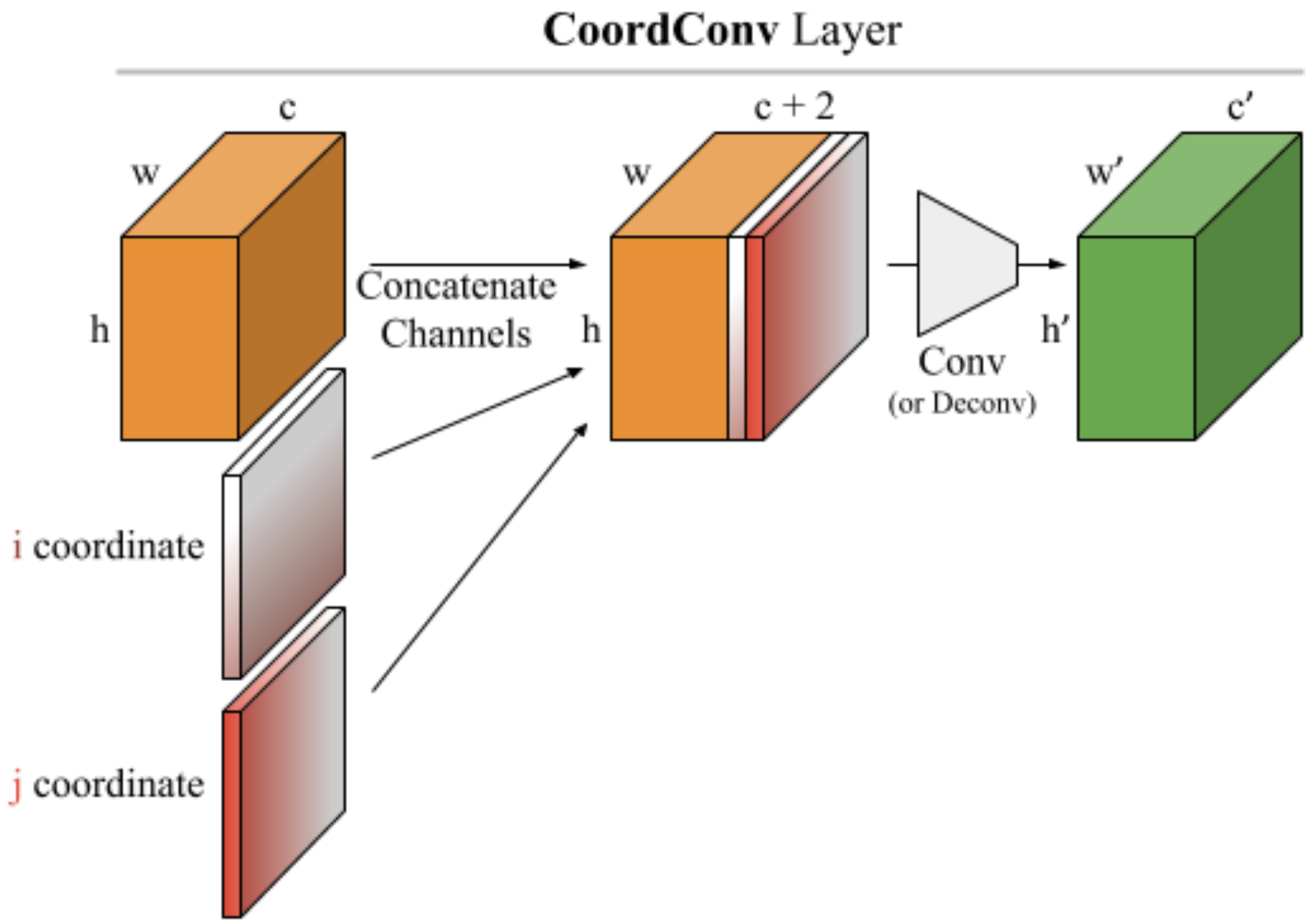
Conv

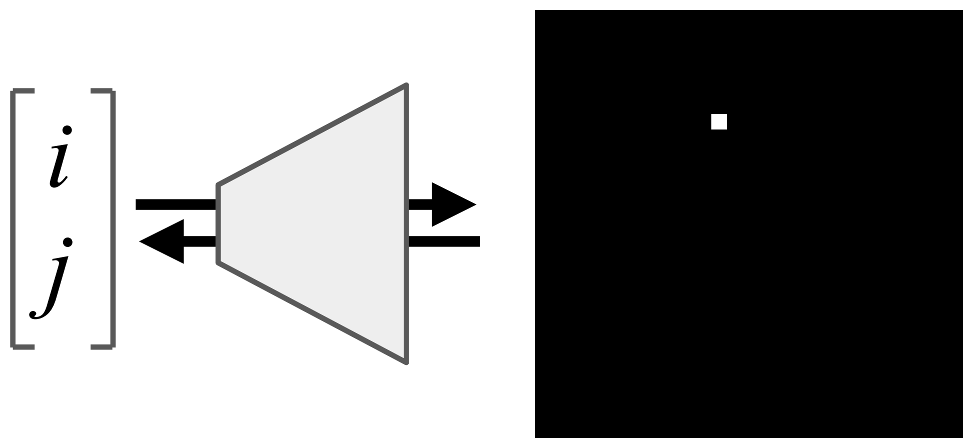


CoordConv

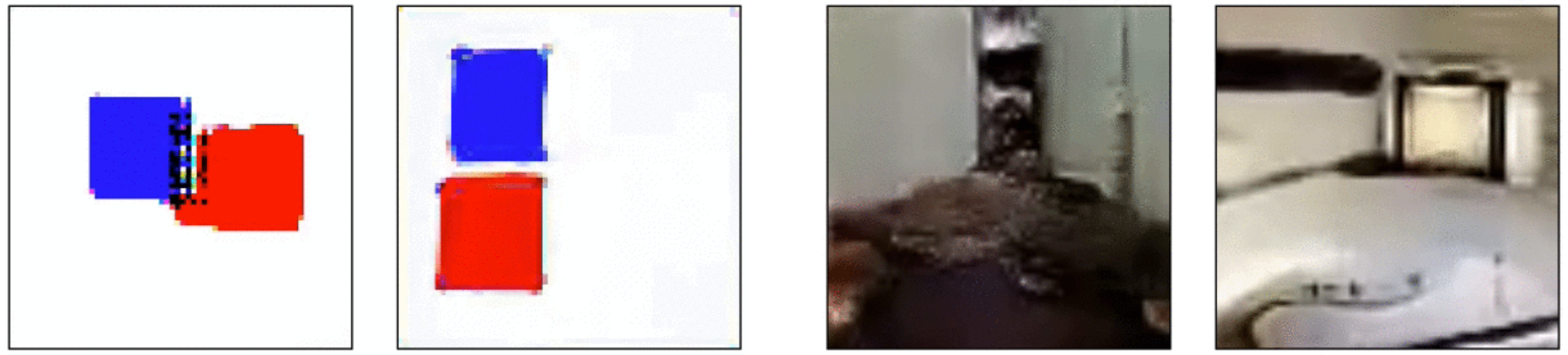


The *CoordConv* solution



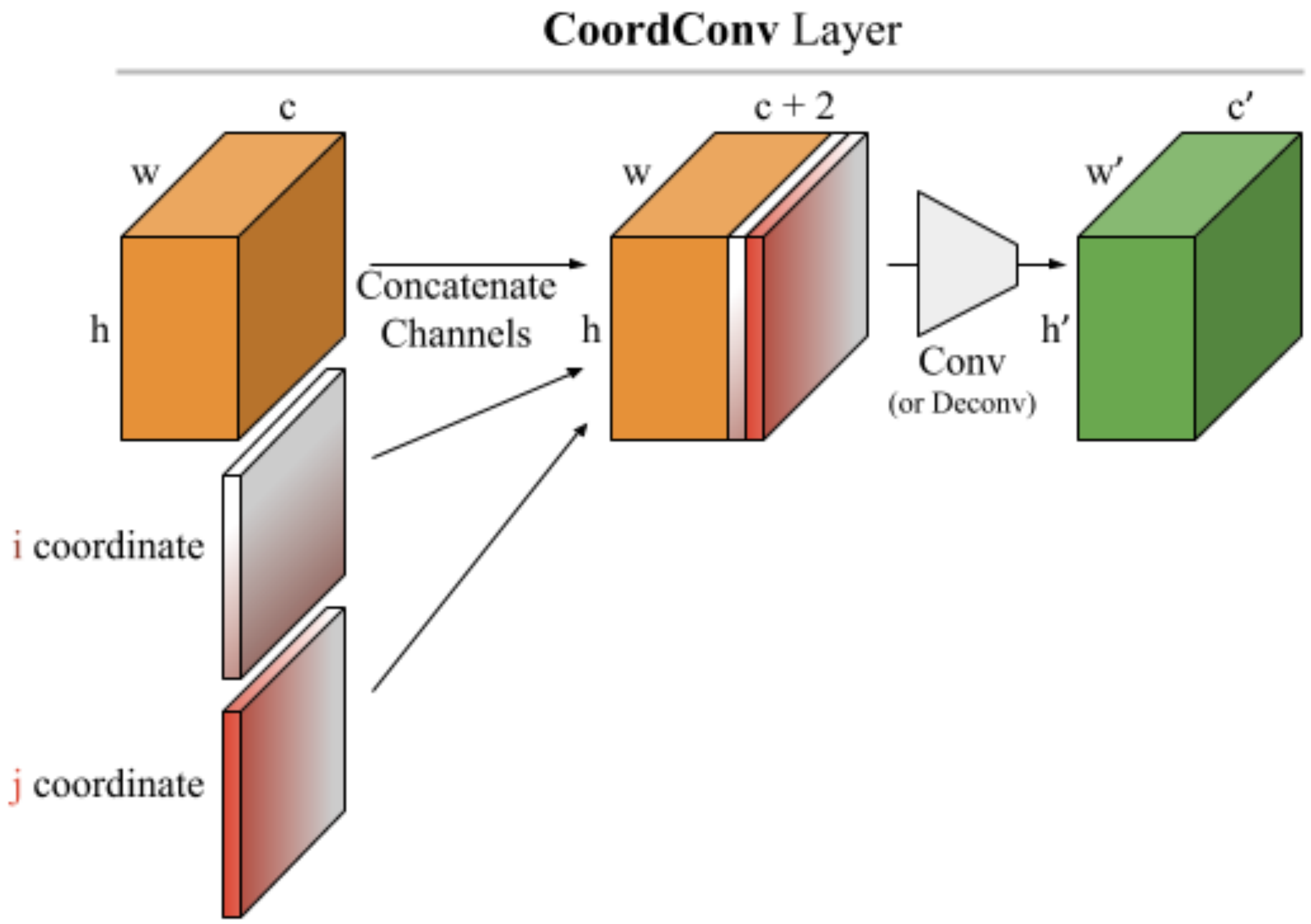


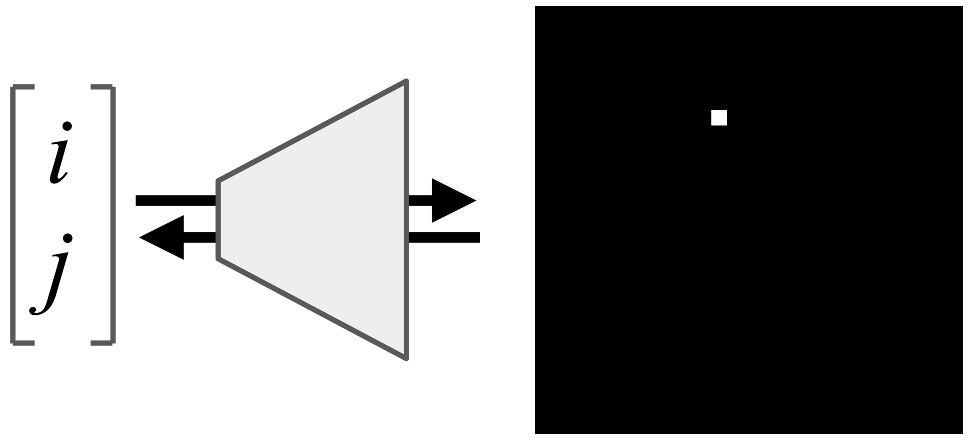
✓



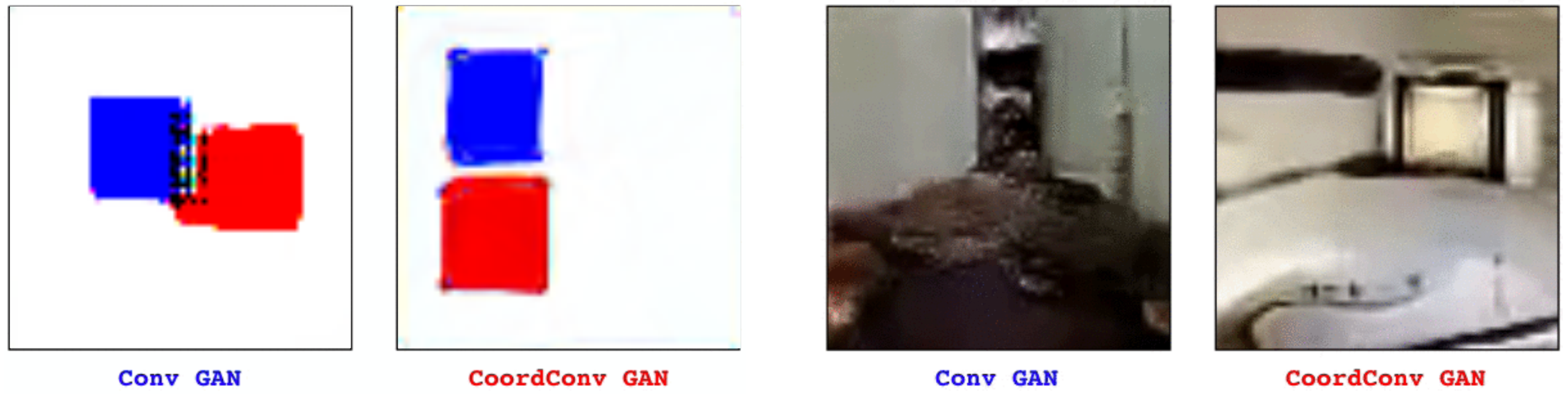
✓

The *CoordConv* solution

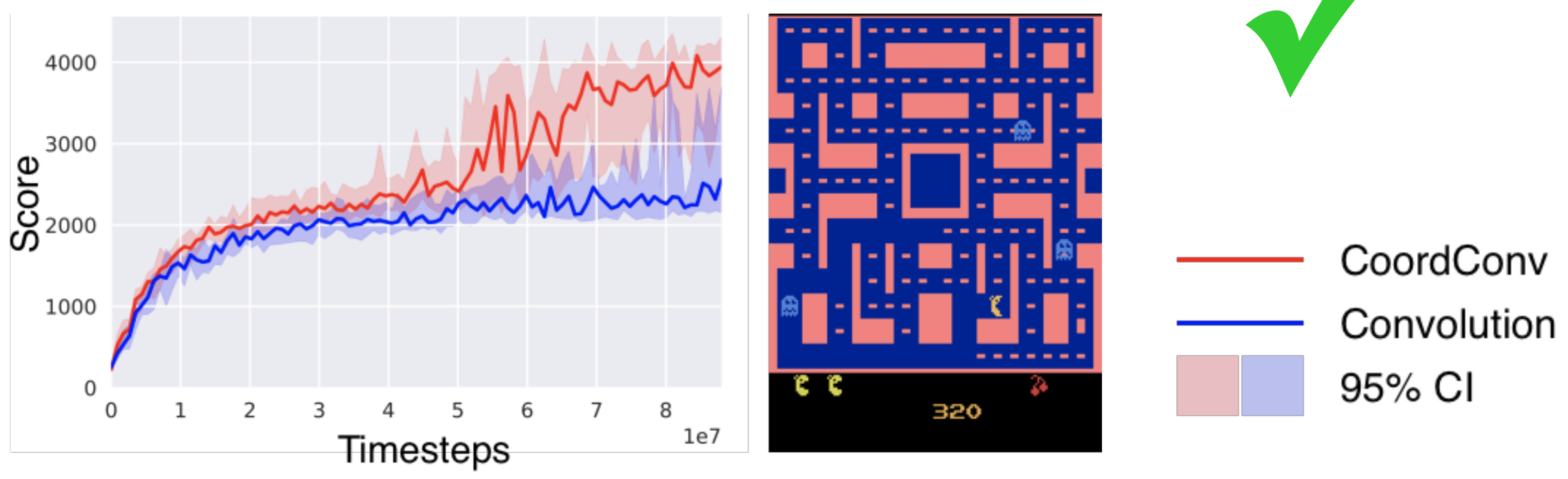




✓

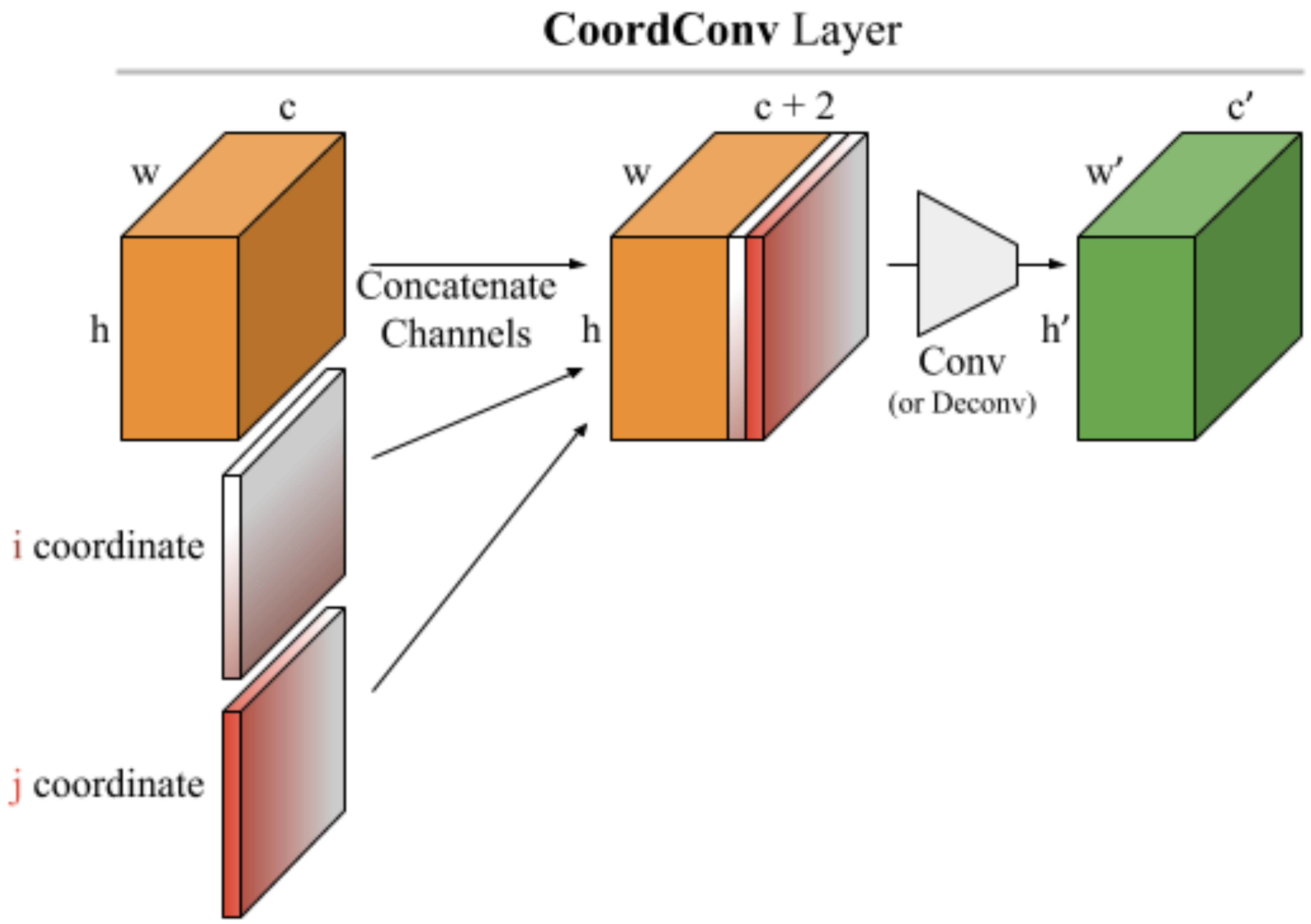


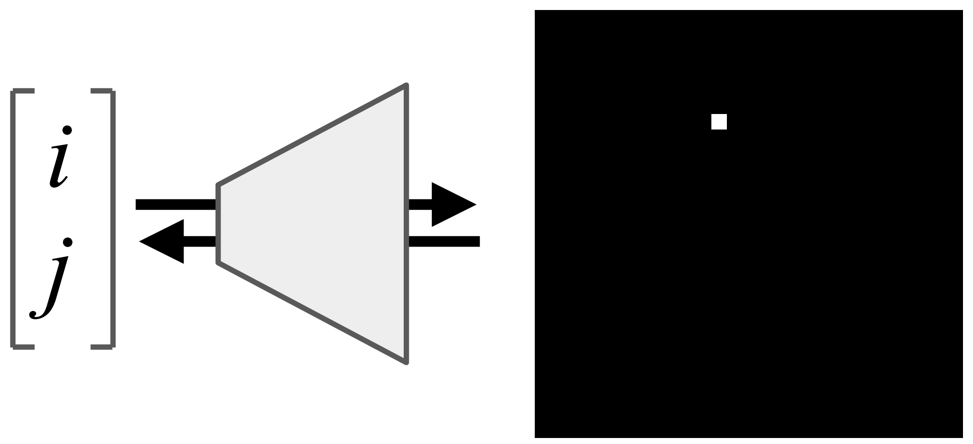
✓



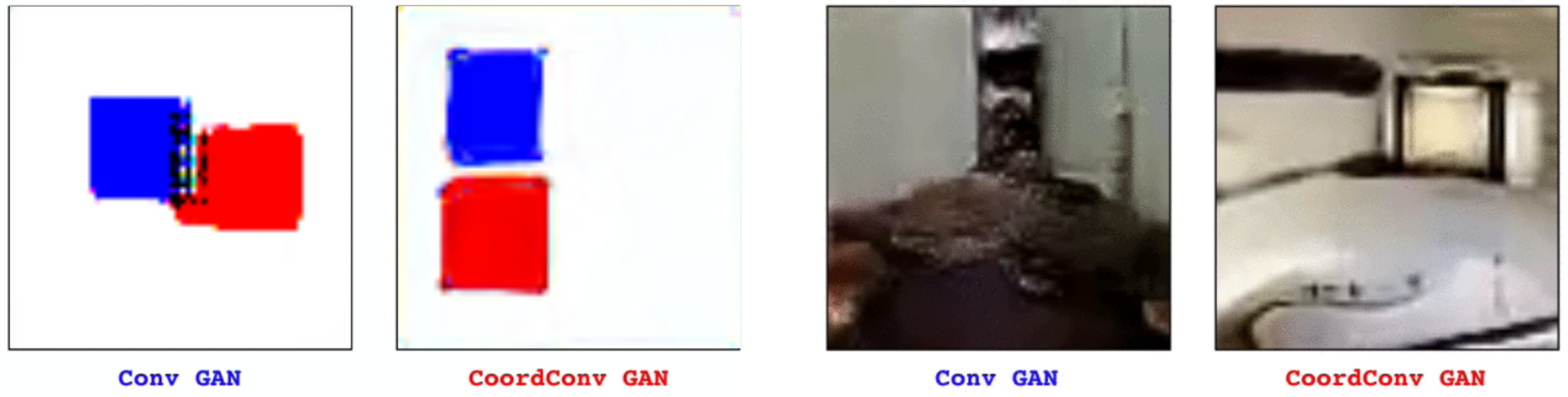
✓

The *CoordConv* solution



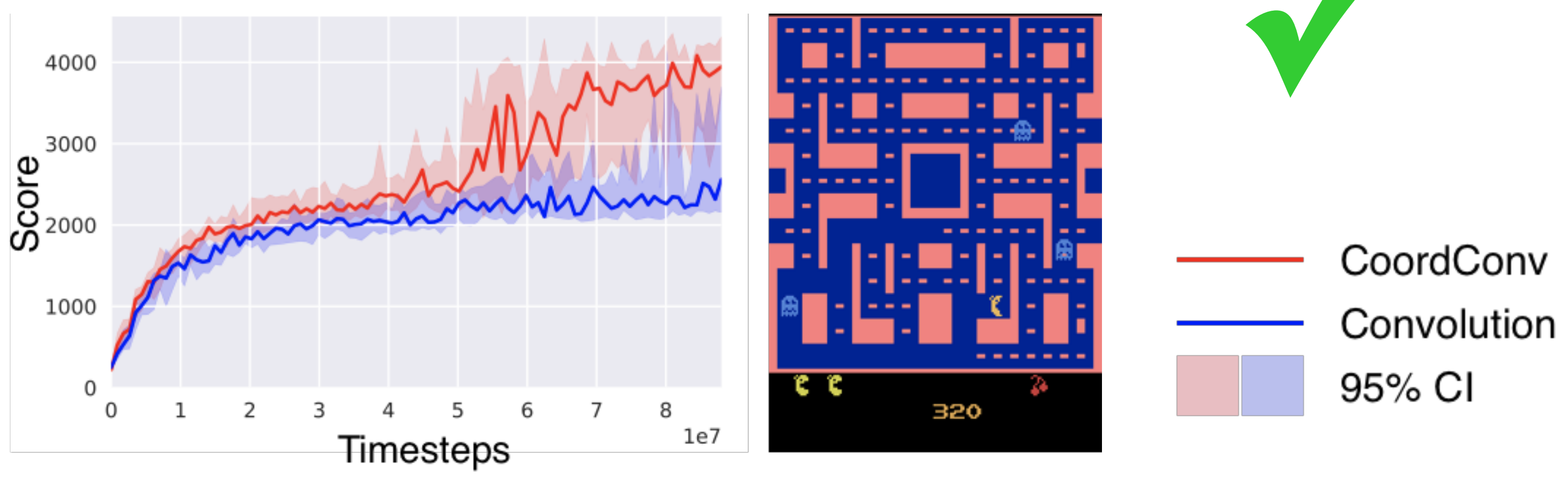


✓



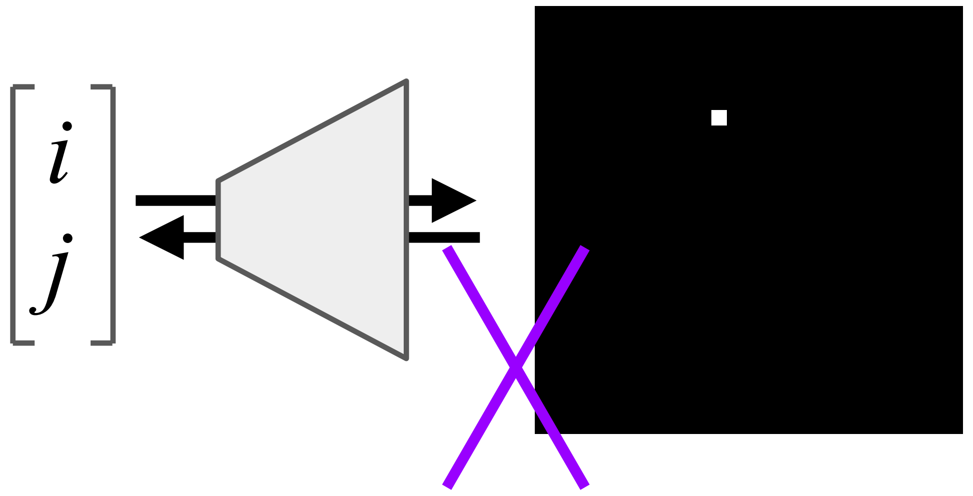
✓

✓

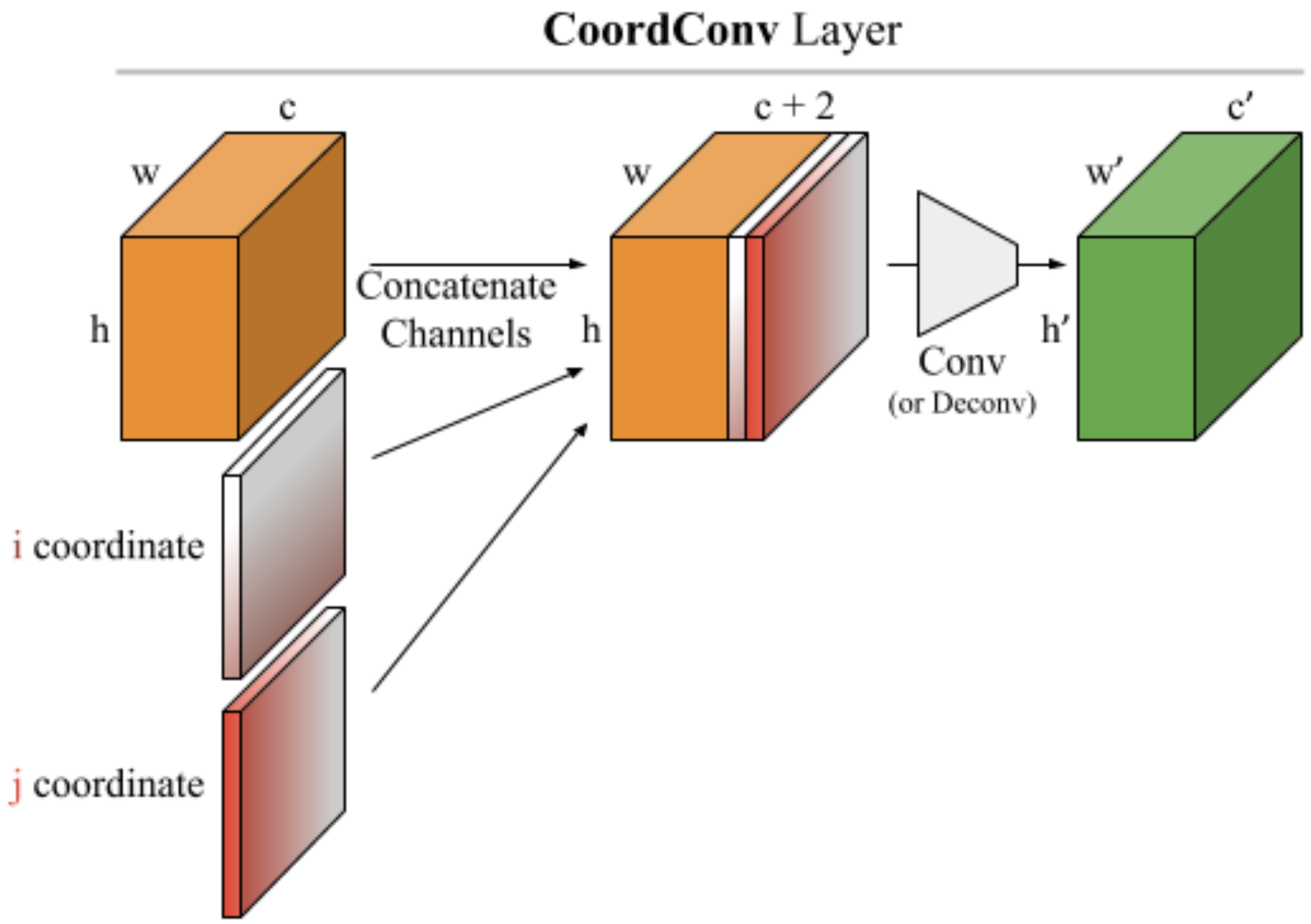


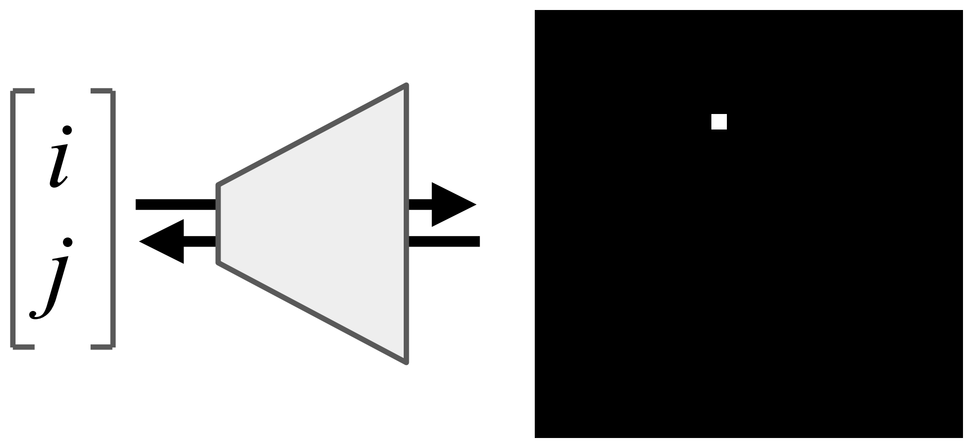
Good boy!
(golden retriever) ✗

In Summary



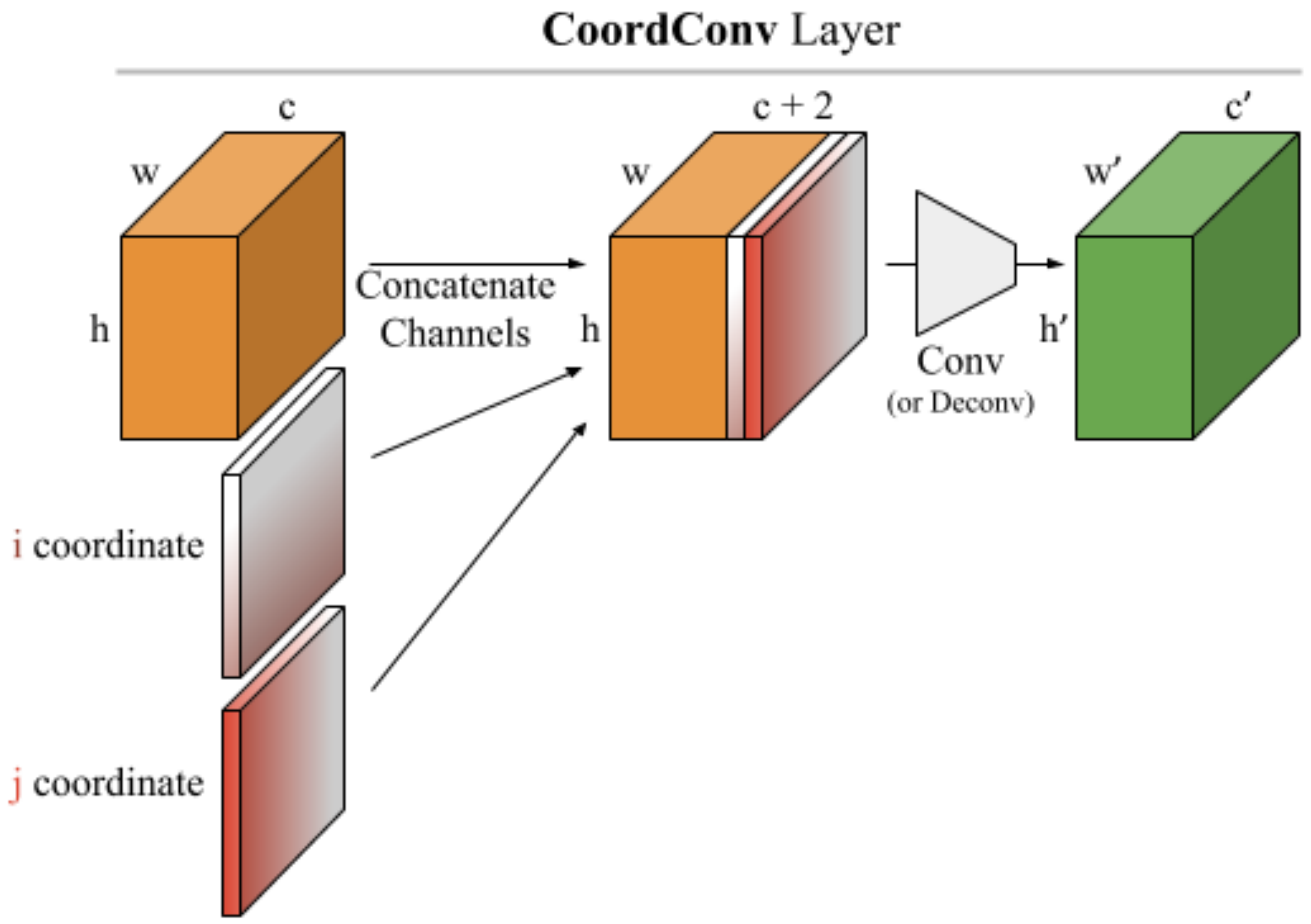
In Summary

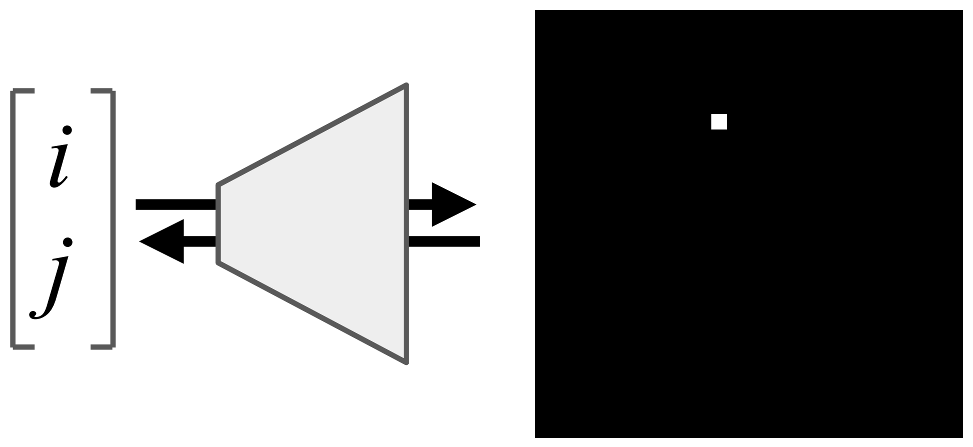

CoordConv Layer
c
w
h
Concatenate Channels
i coordinate
j coordinate
c + 2
w
h
Conv
(or Deconv)
c′
w′
h′


i
j

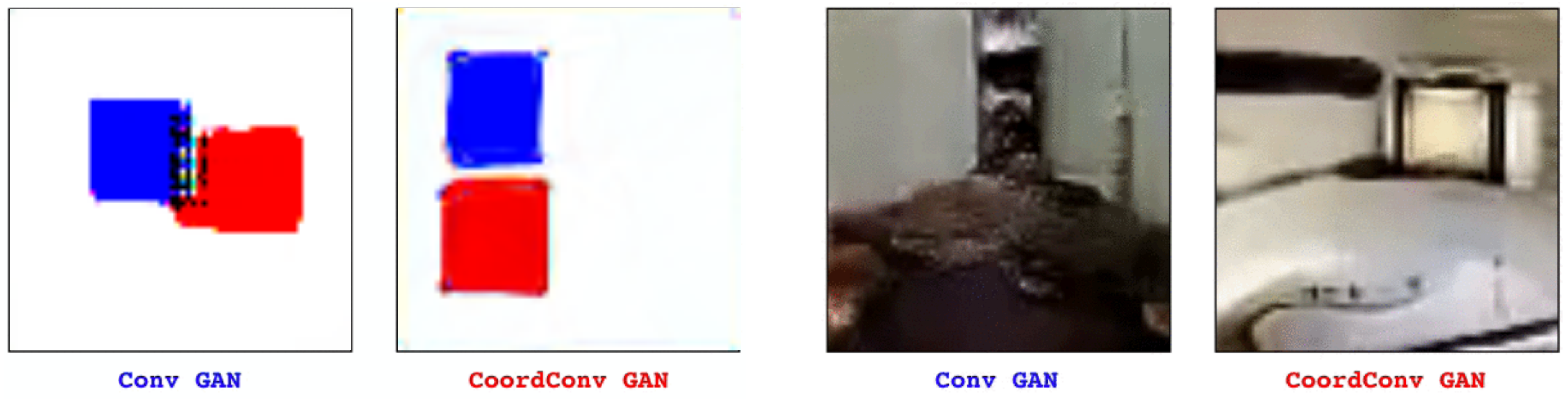
✓

In Summary

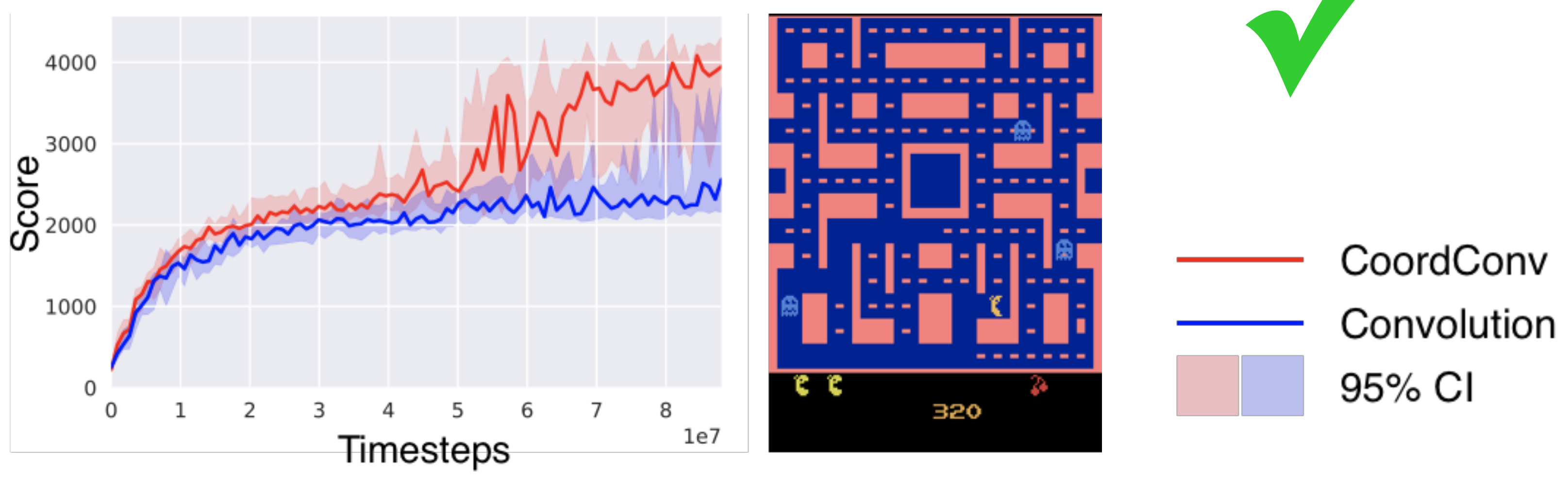




✓



✓

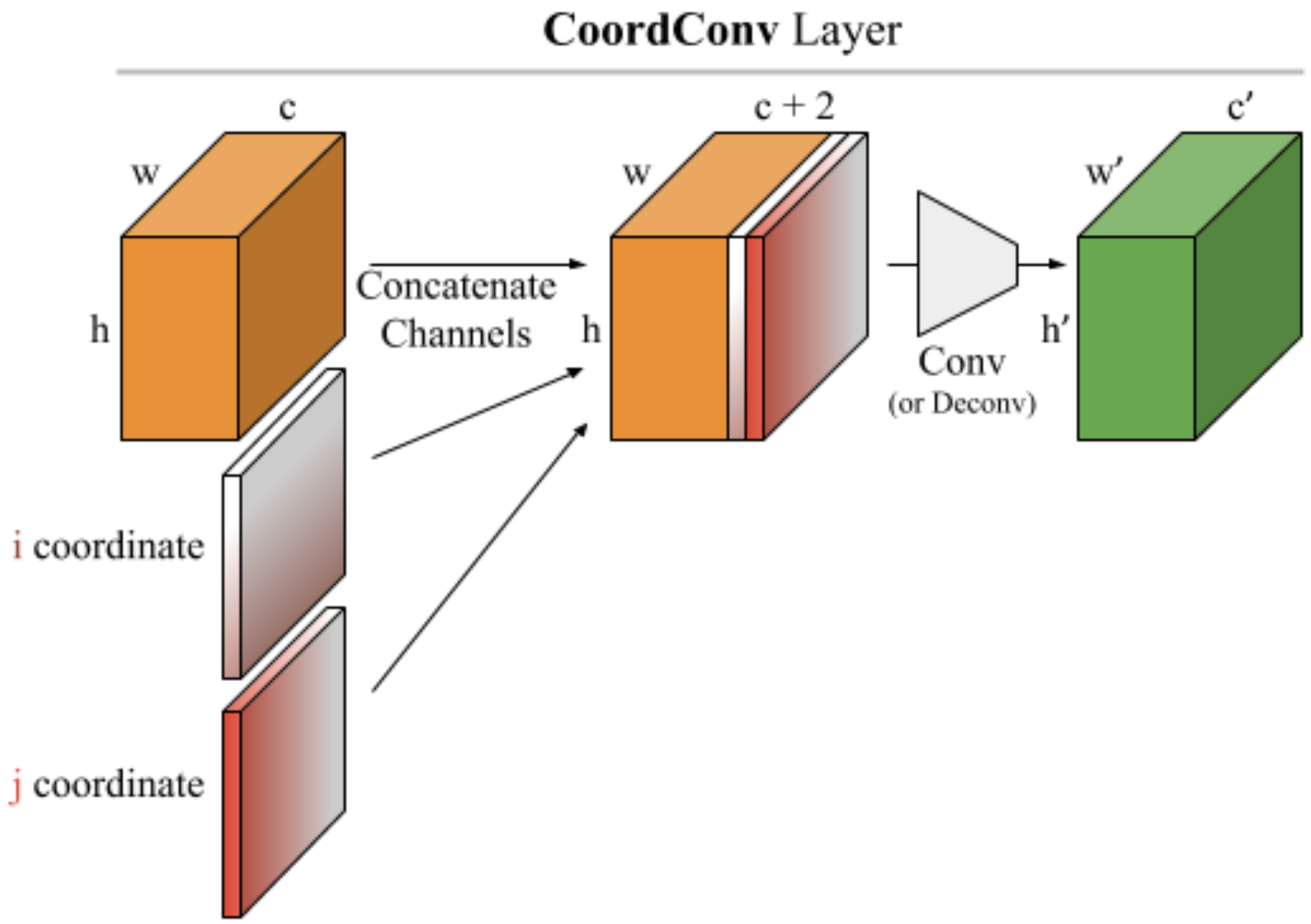


✓

→

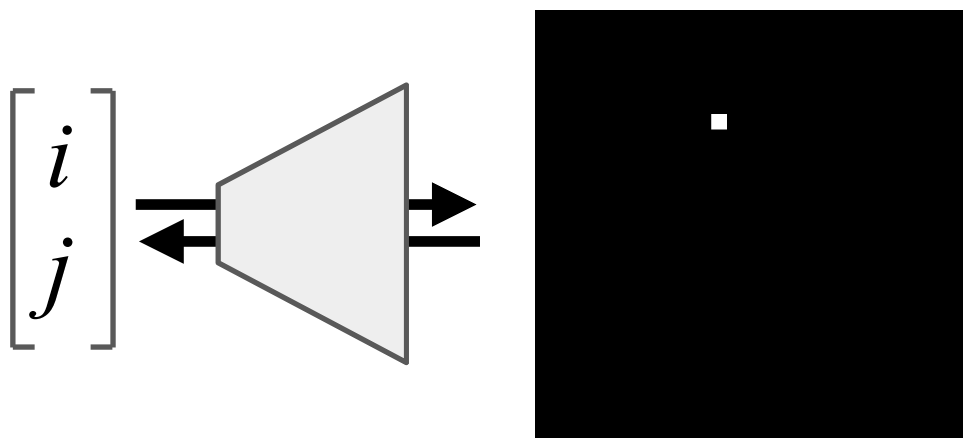
Good boy!
(golden retriever) ✗

In Summary

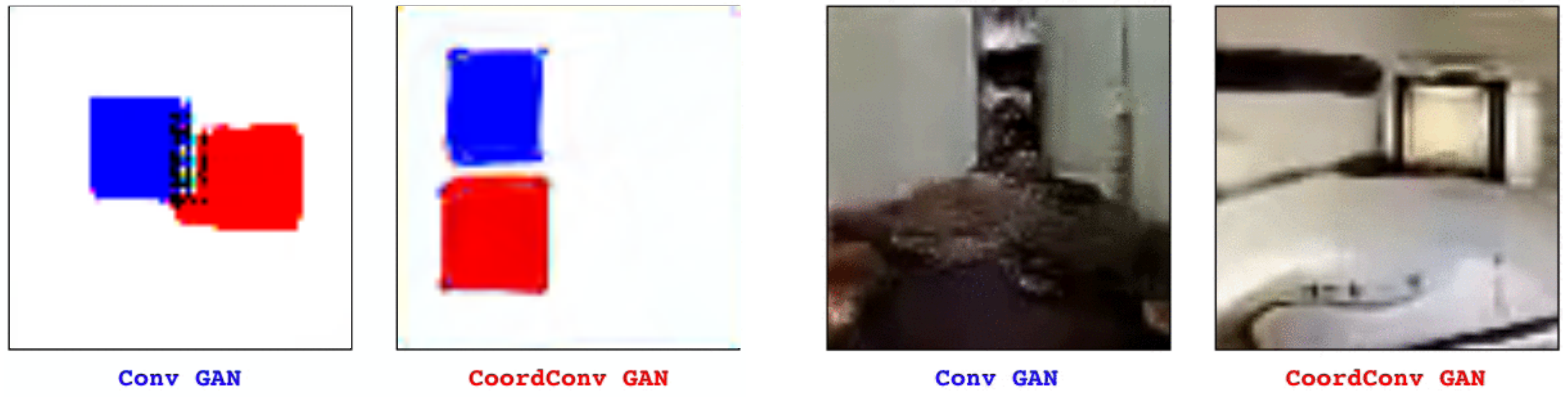


An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution.

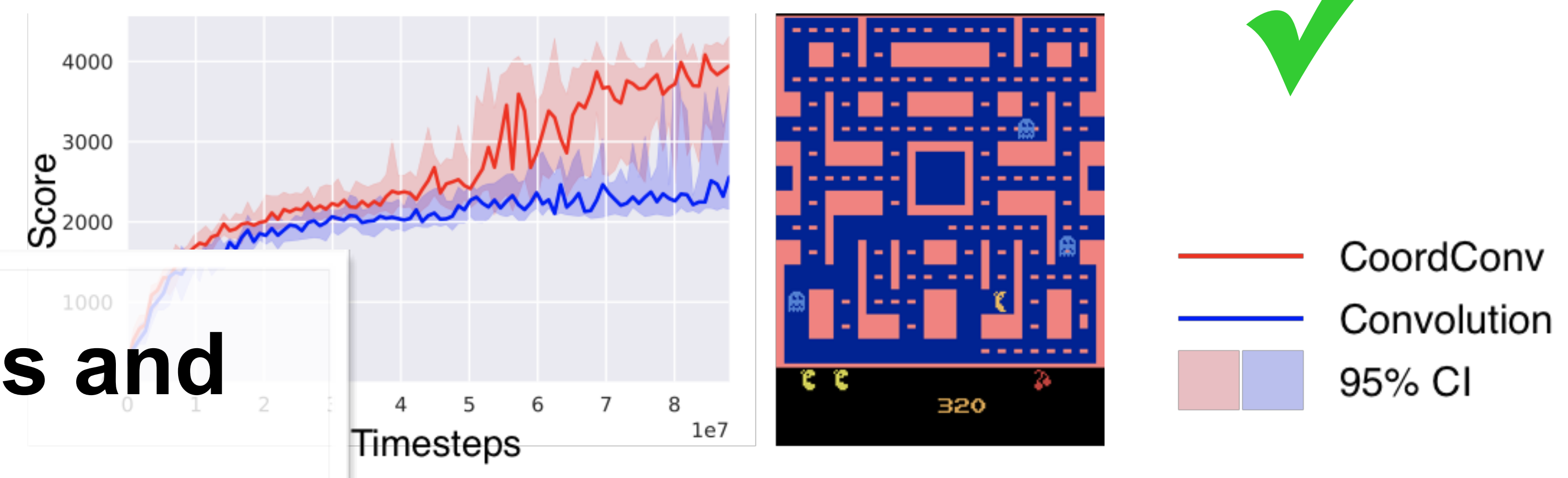
R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, J. Yosinski, *NeurIPS 2018*.



✓



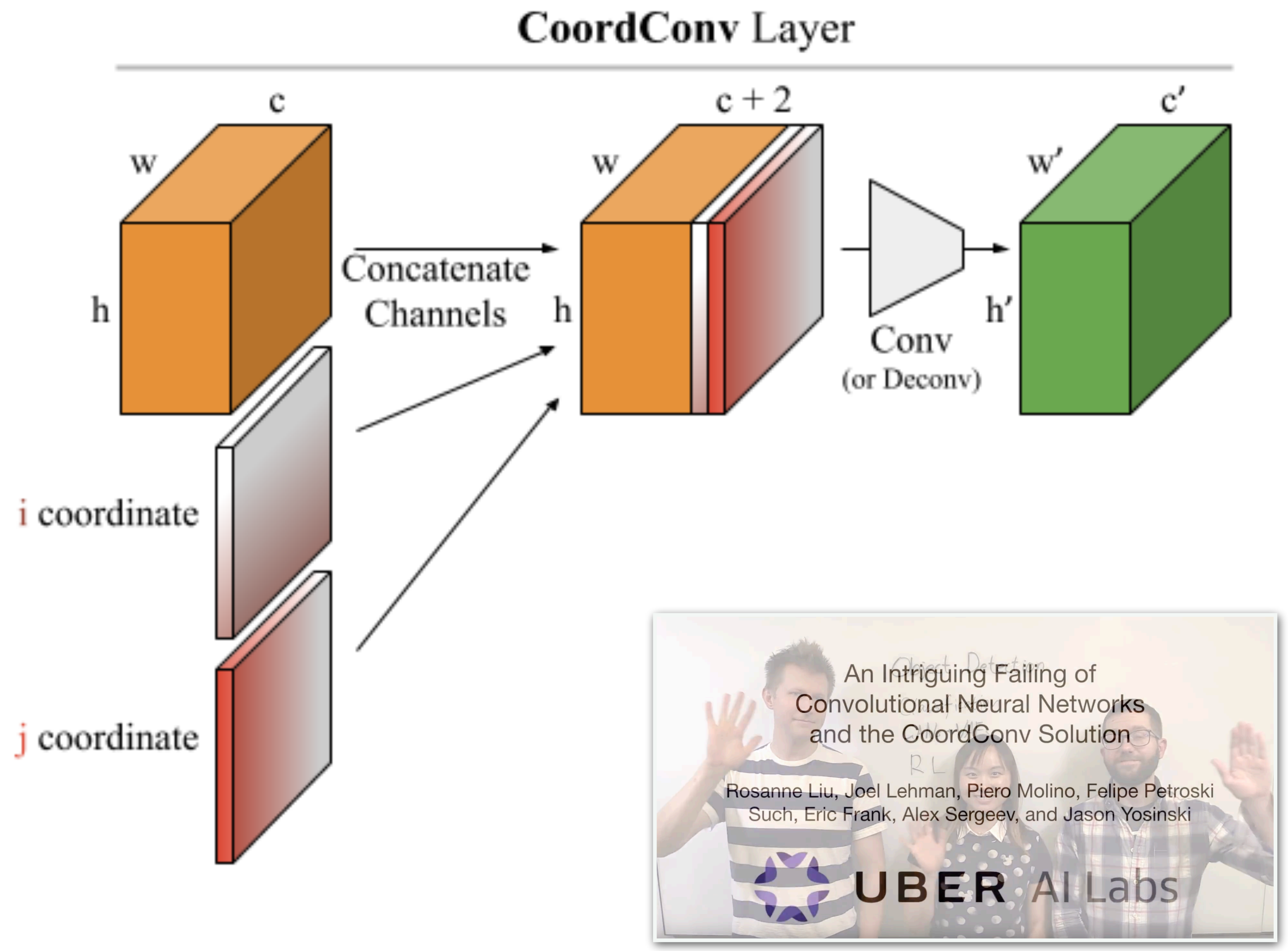
✓



✓

→ Good boy!
(golden retriever) ✗

In Summary



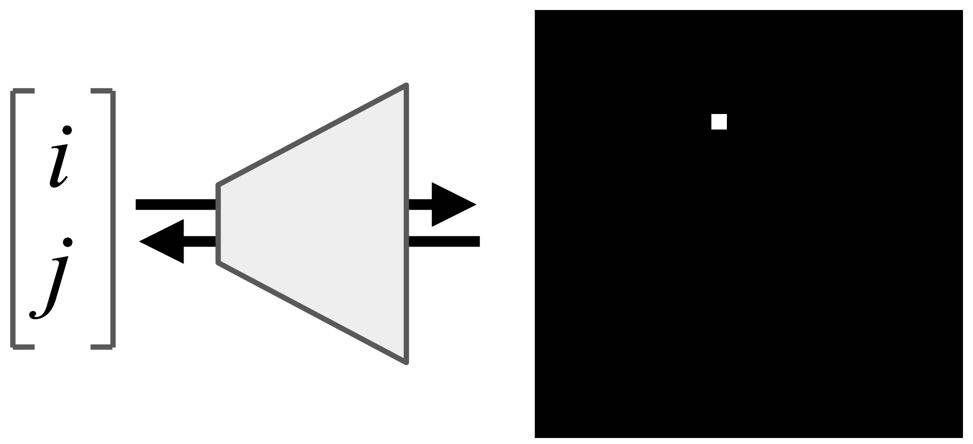
An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution

Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski

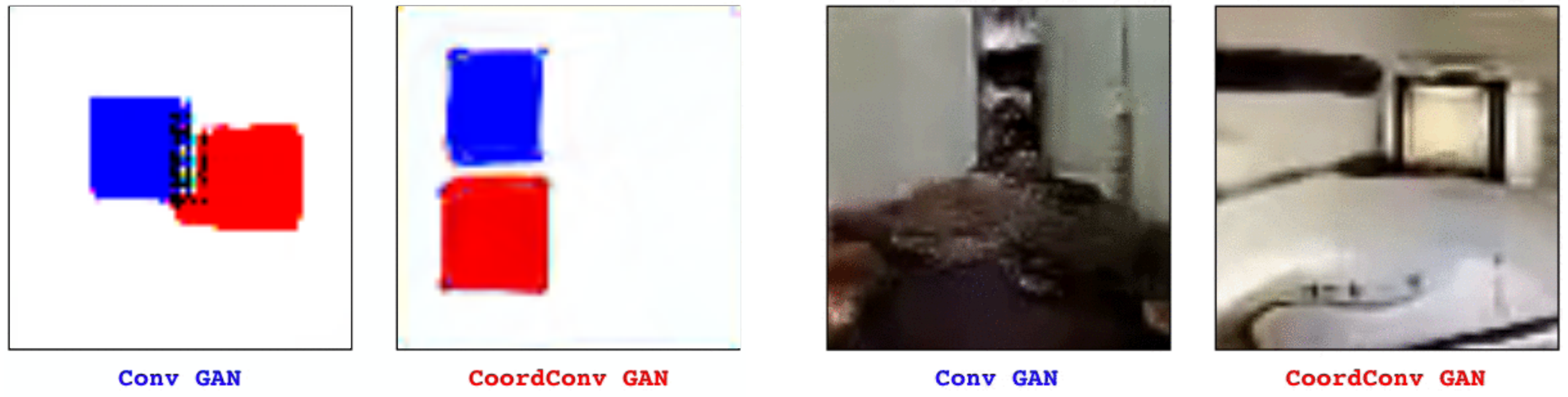
UBER AI Labs

An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution.

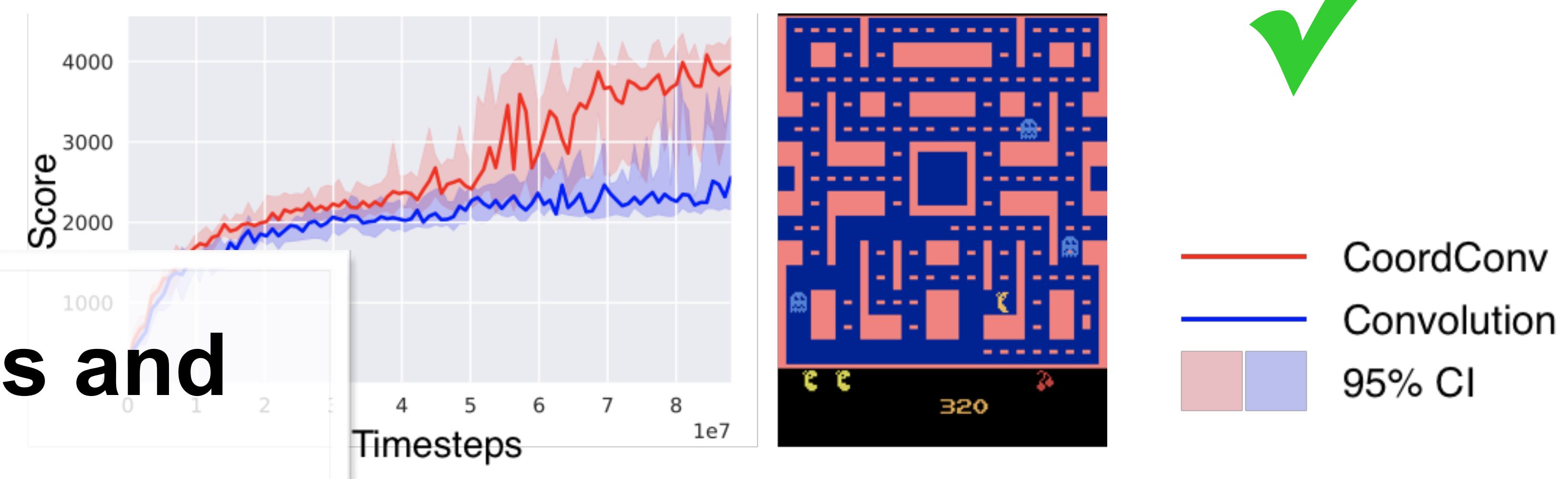
R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, J. Yosinski, *NeurIPS 2018.*



✓

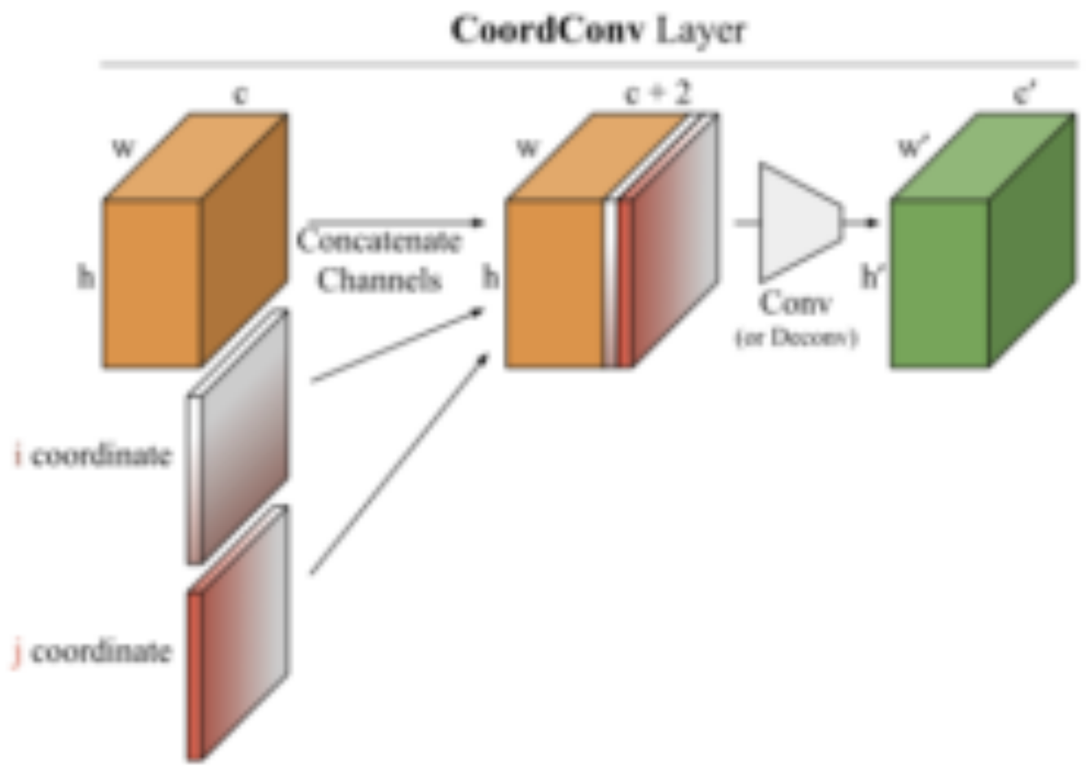


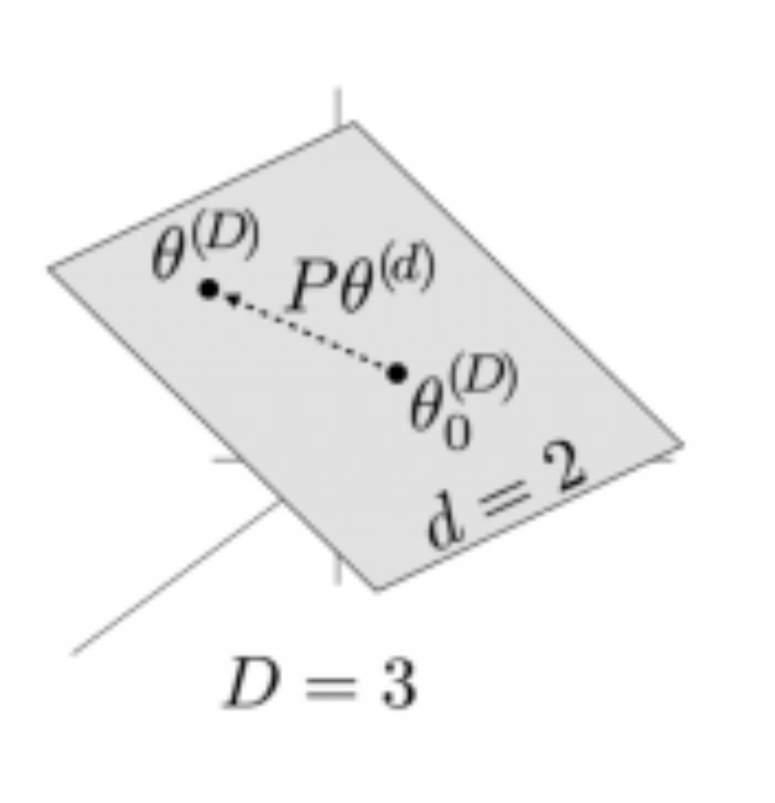
✓

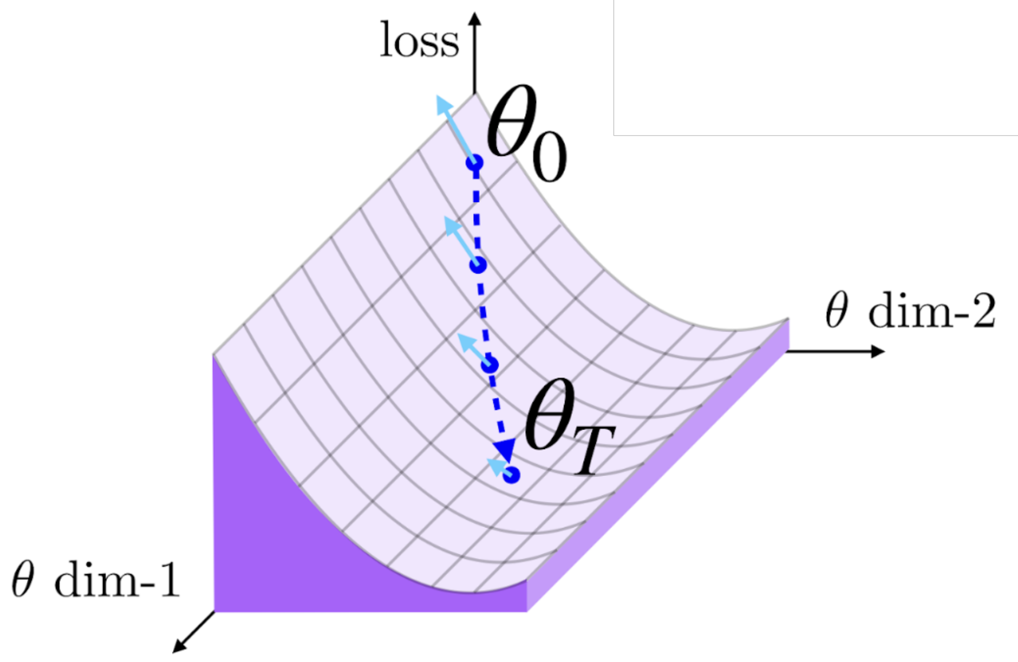


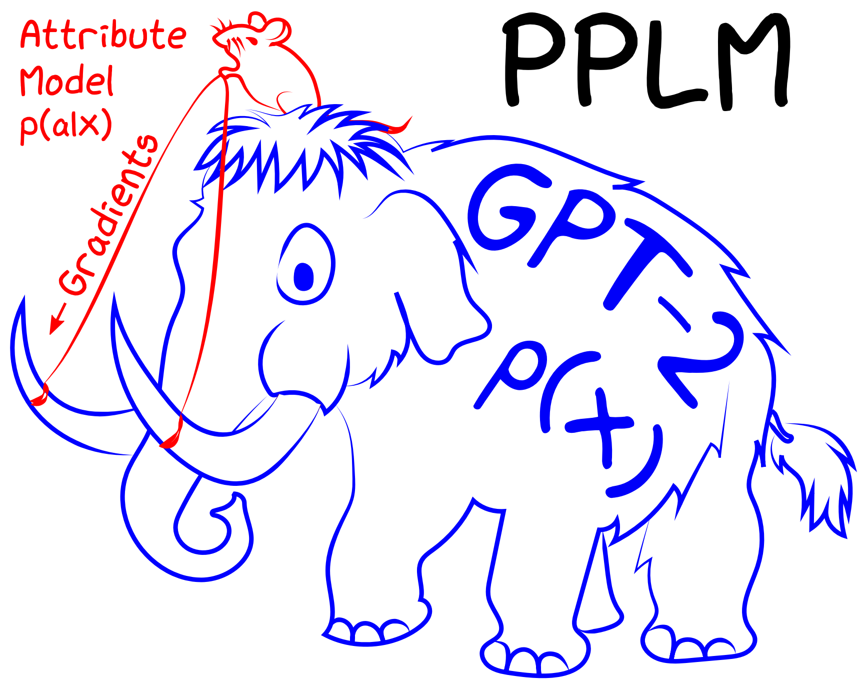
✓

→ Good boy!
(golden retriever) ✗

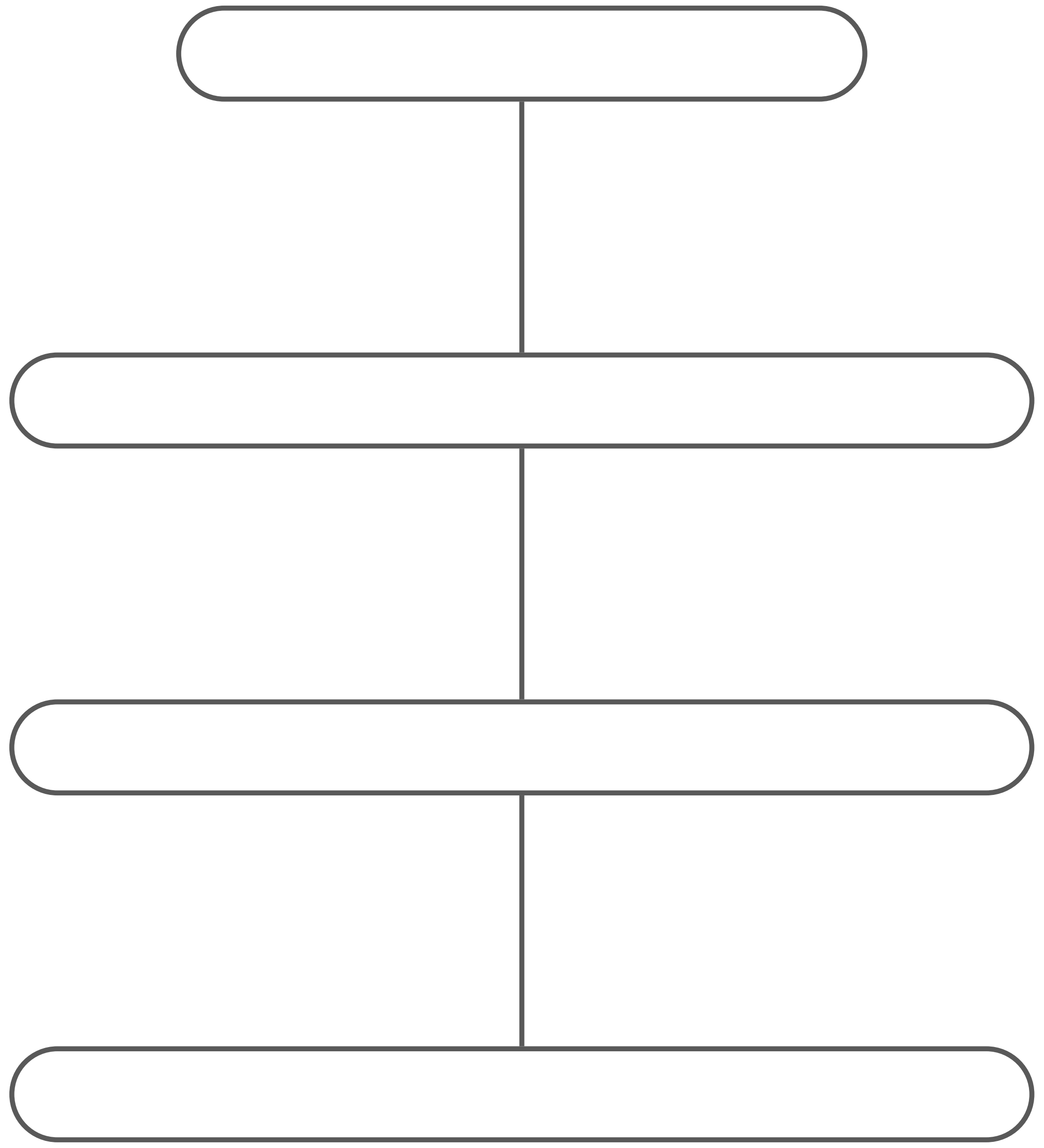

CoordConv Layer
c
w
h
Concatenate Channels
c + 2
w
h
Conv (or Deconv)
c'
w'
h'
i coordinate
j coordinate


$\theta^{(D)}$
$P\theta^{(d)}$
$\theta_0^{(D)}$
$d = 2$
$D = 3$


loss
θ_0
θ_T
θ dim-2
θ dim-1

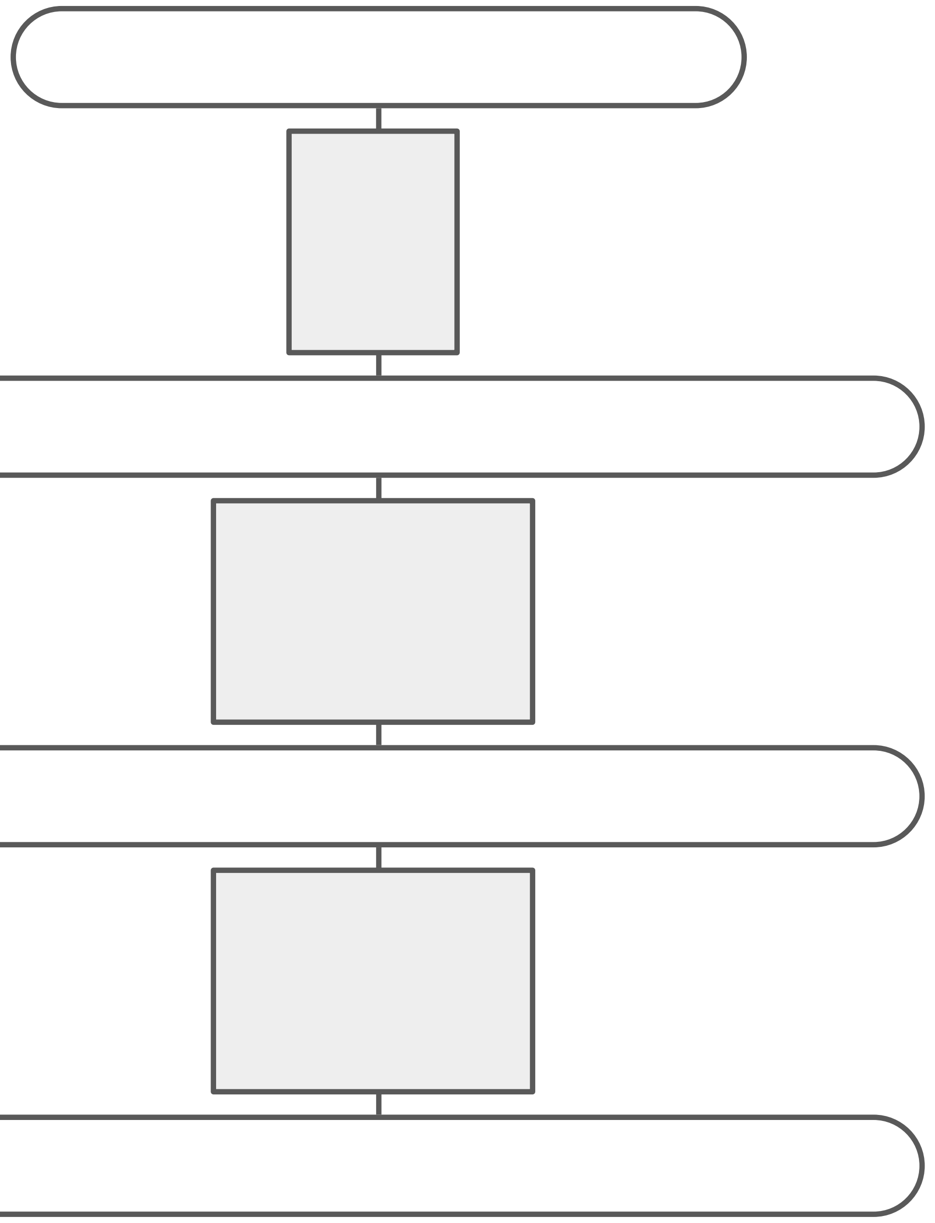

PPLM
Attribute
Model
p(a|x)
Gradients
GPT-2
p(x)

Typical Neural Network Training



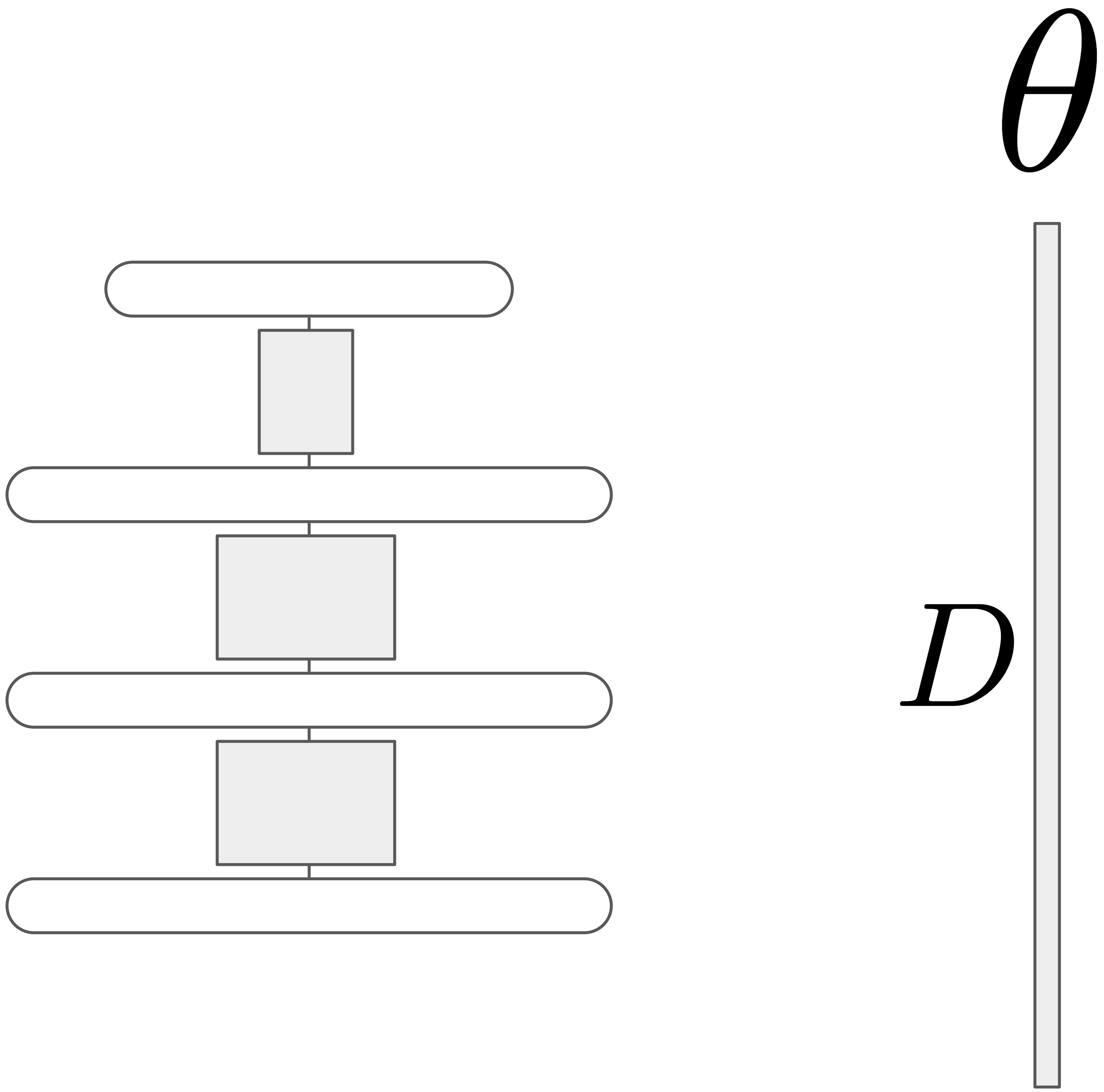
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Typical Neural Network Training



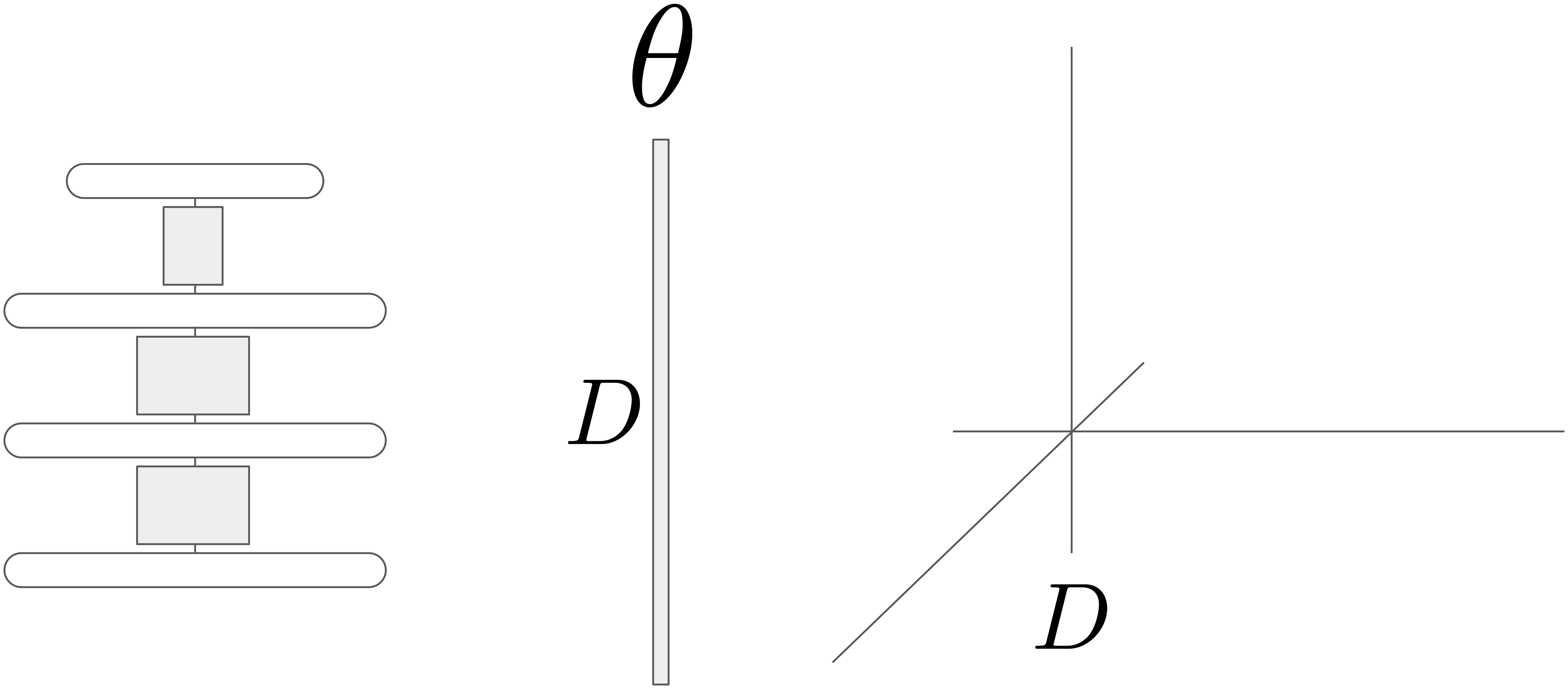
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Typical Neural Network Training



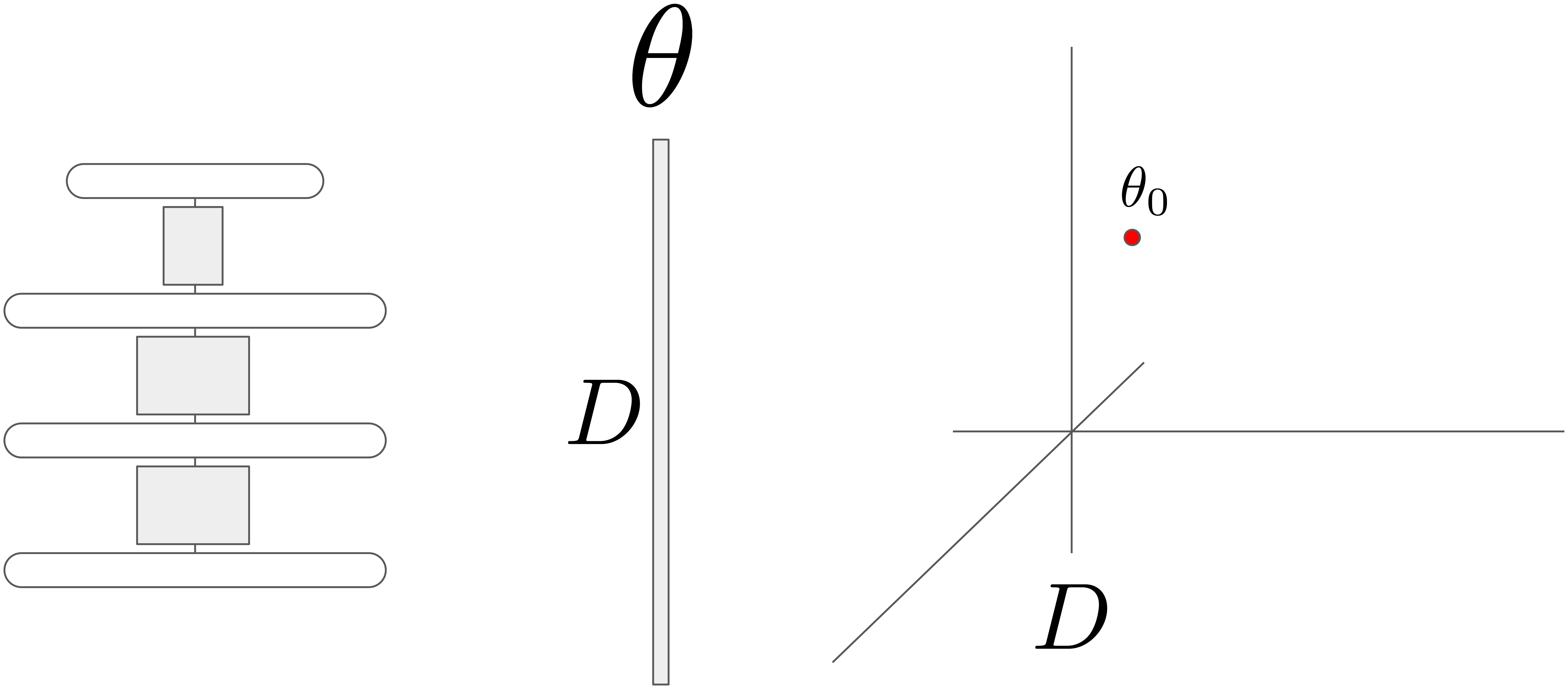
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Typical Neural Network Training



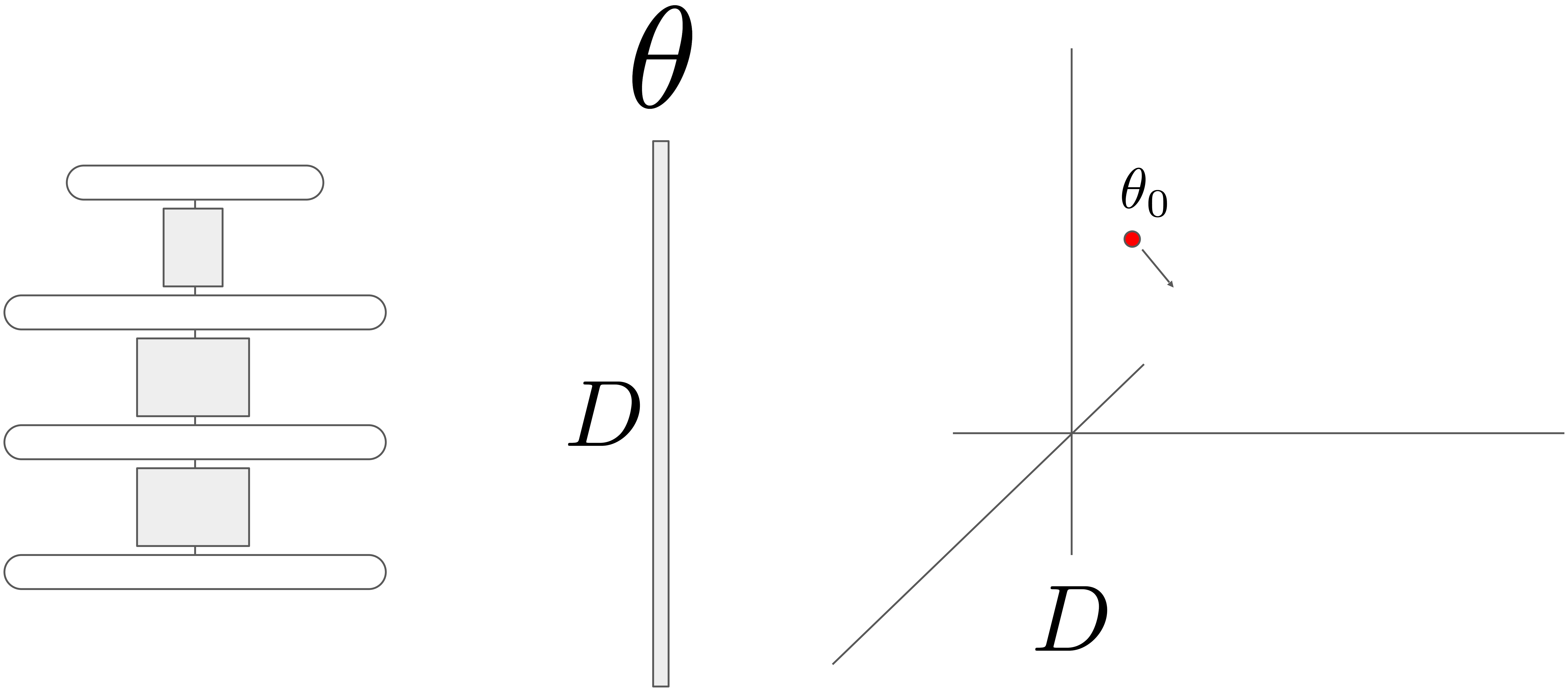
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Typical Neural Network Training



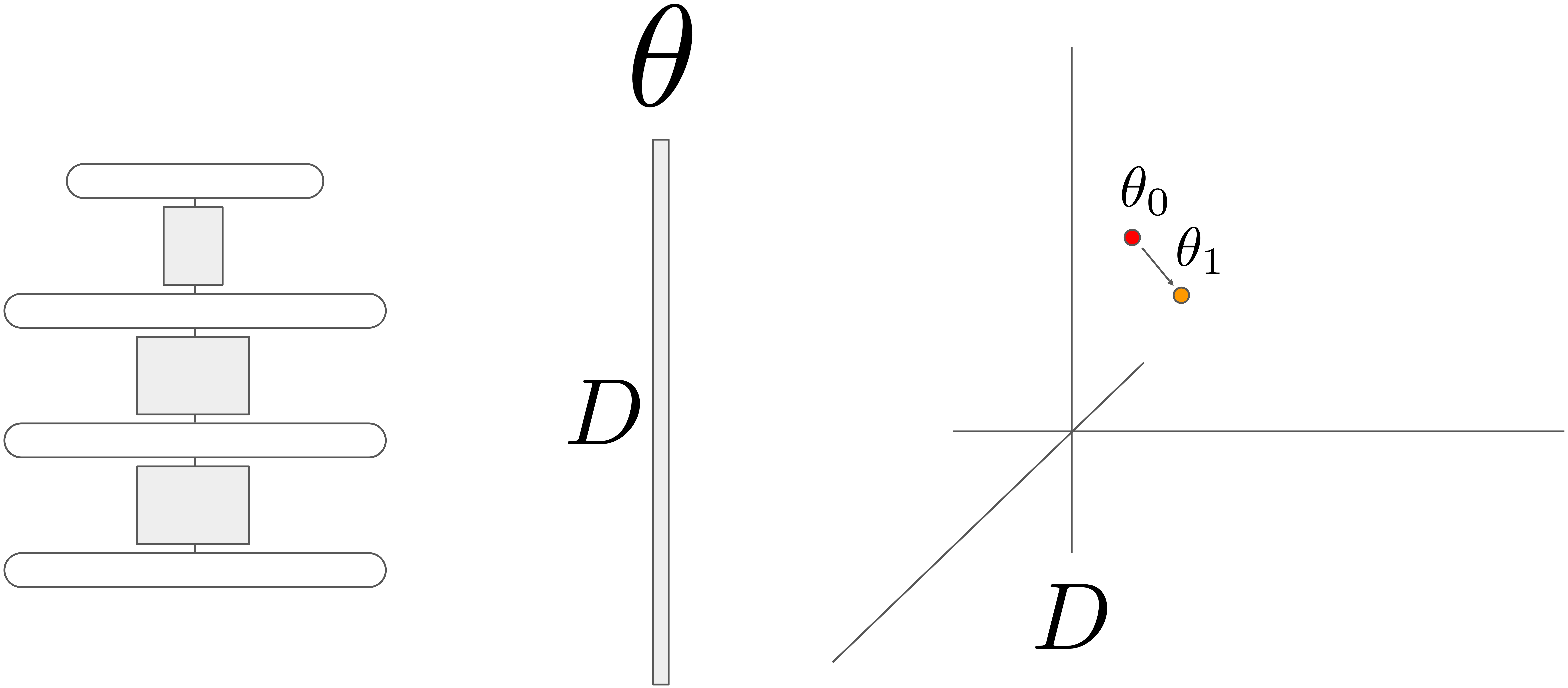
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Typical Neural Network Training



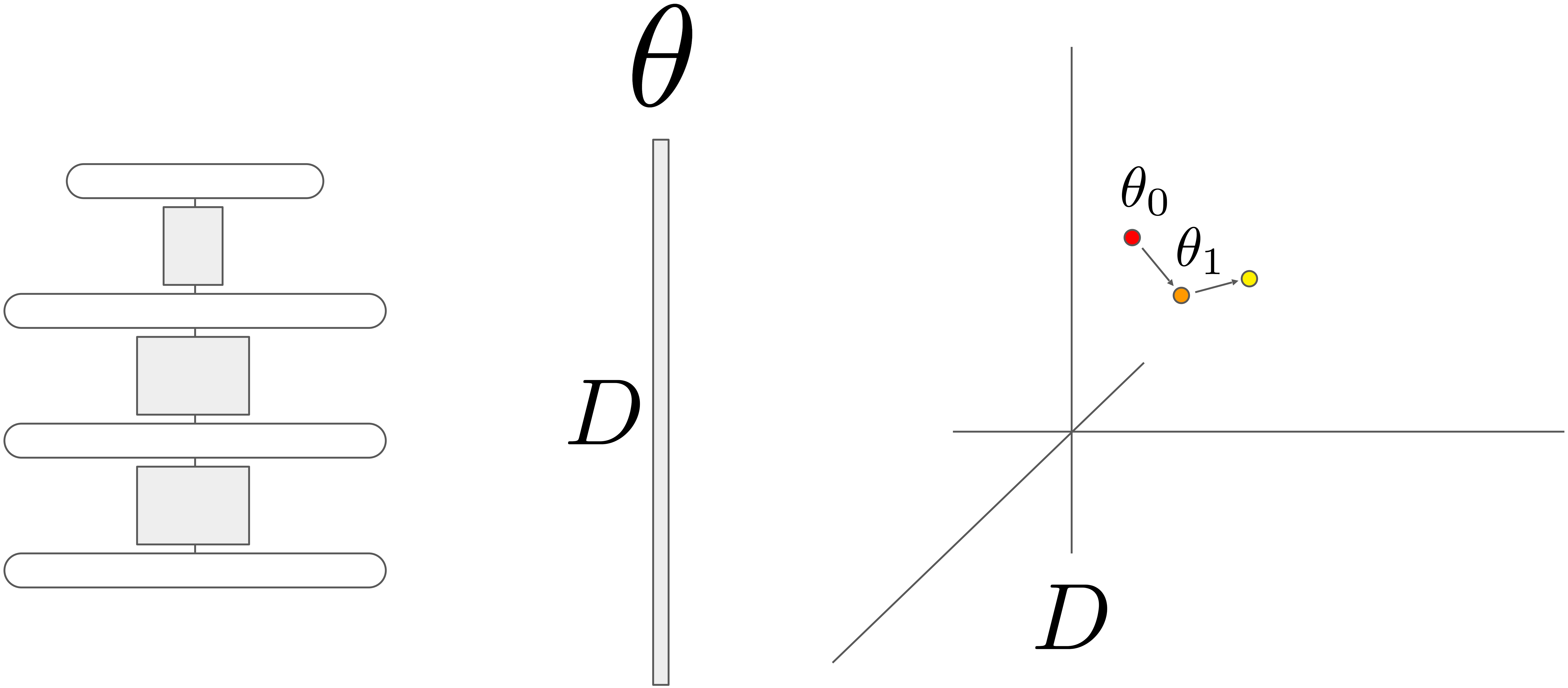
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Typical Neural Network Training



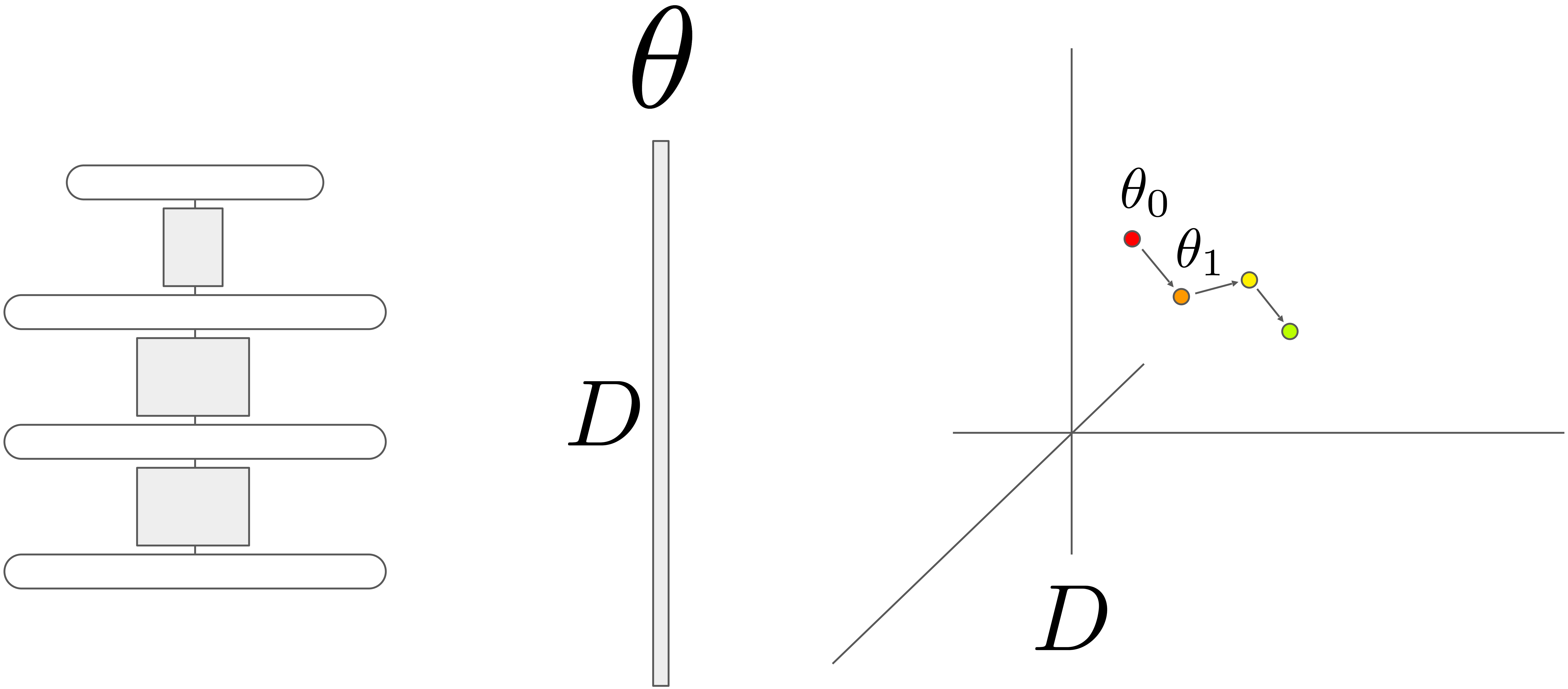
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Typical Neural Network Training



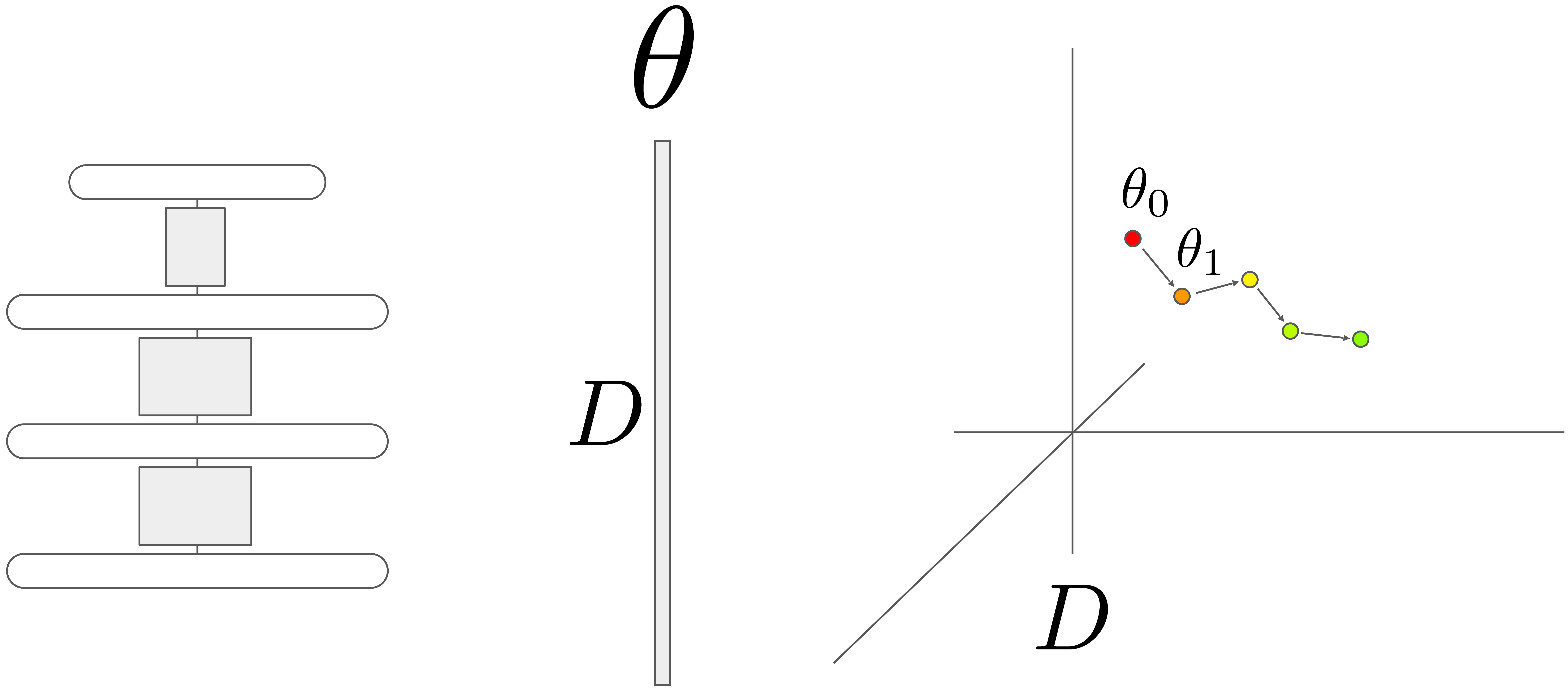
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Typical Neural Network Training



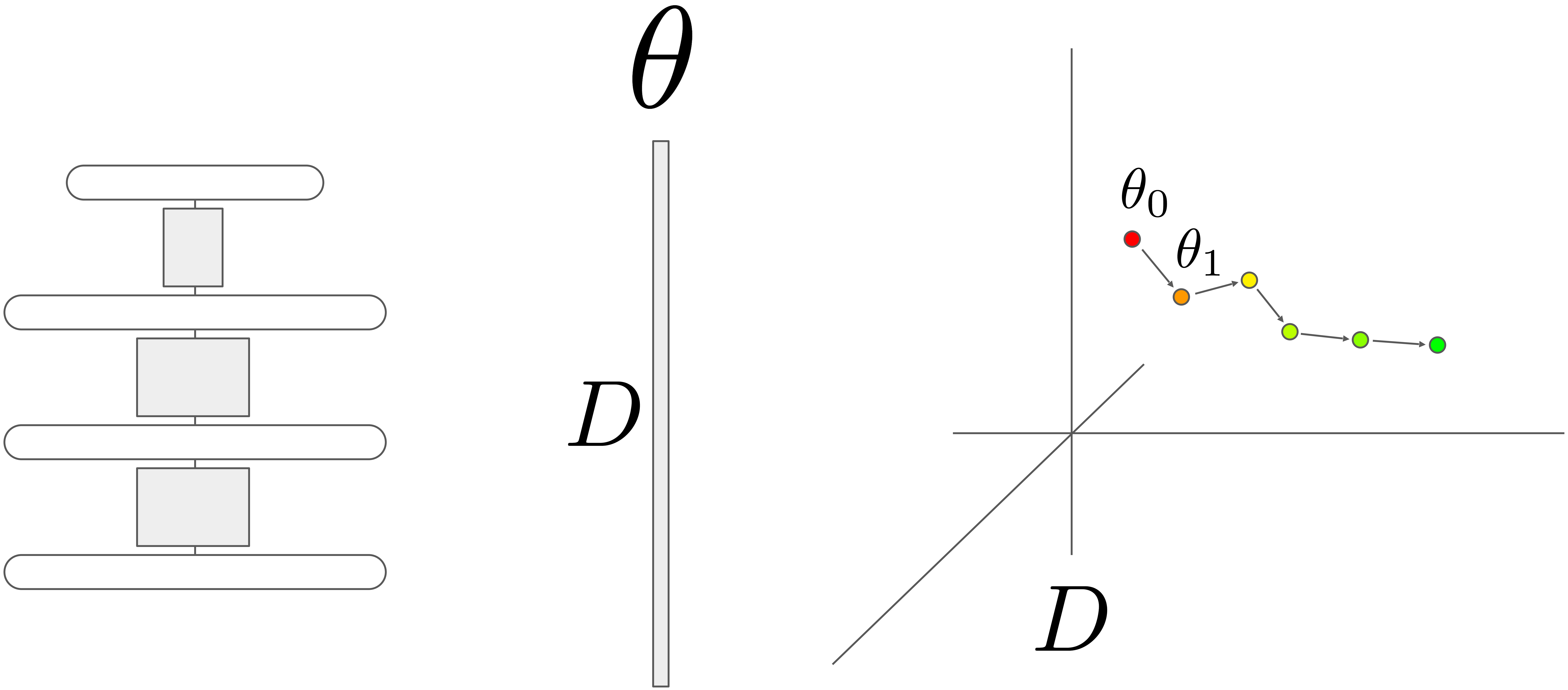
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Typical Neural Network Training



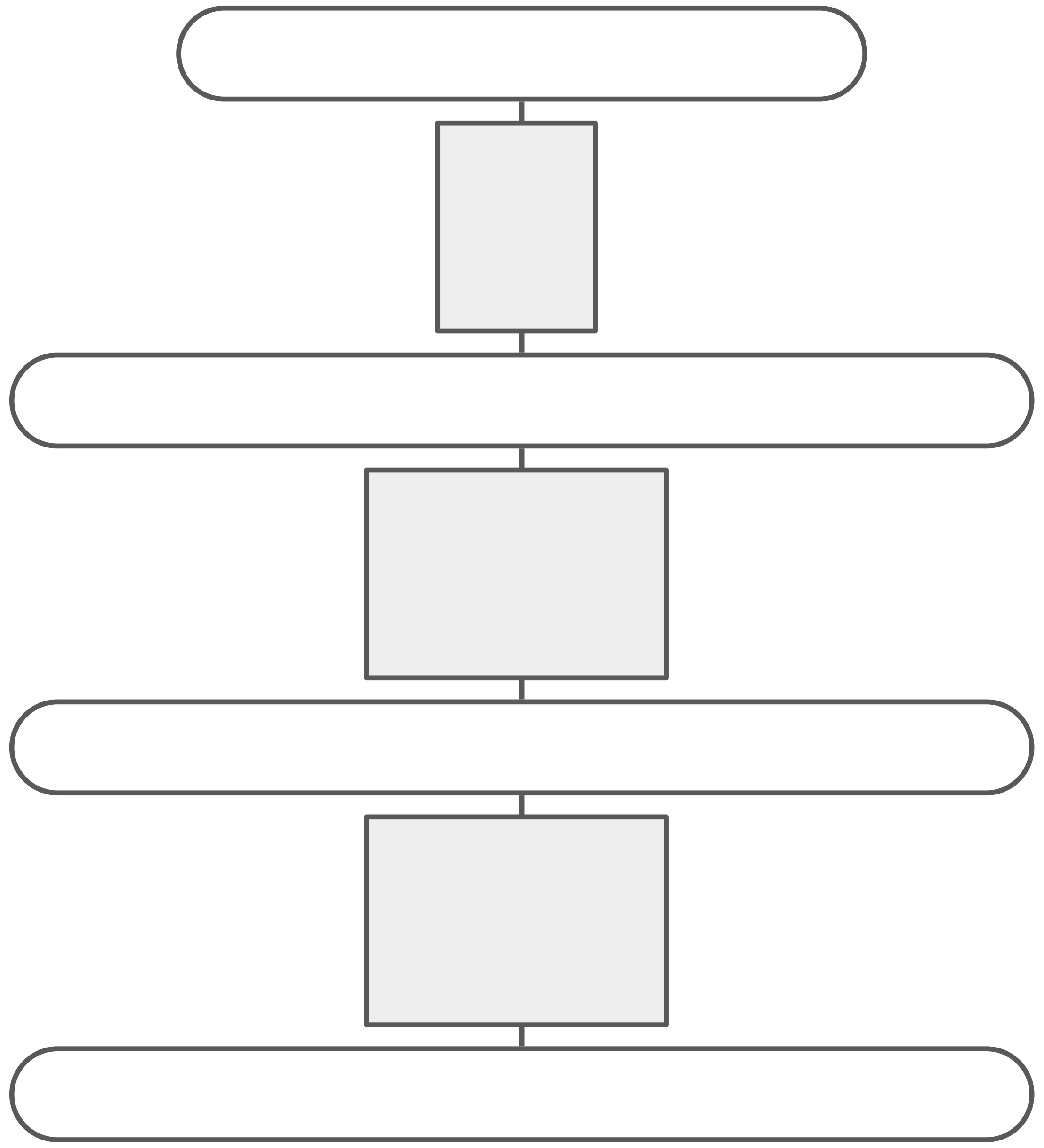
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

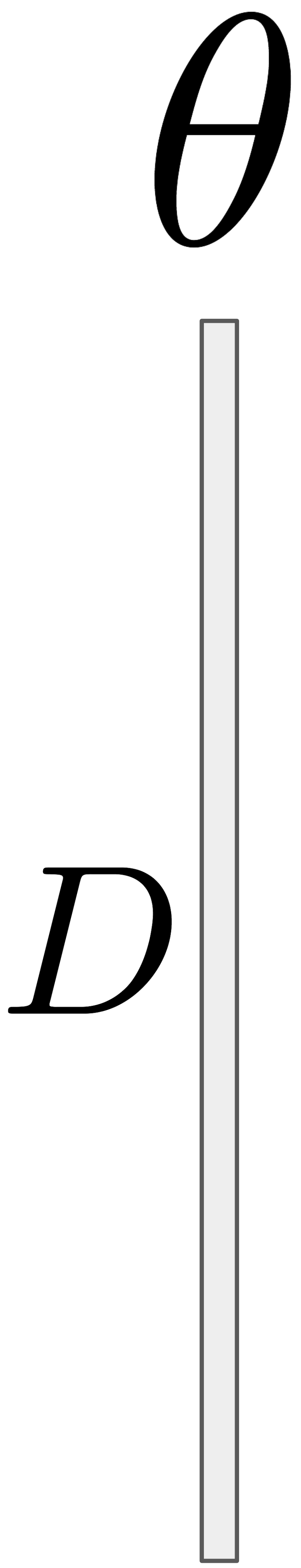
Typical Neural Network Training

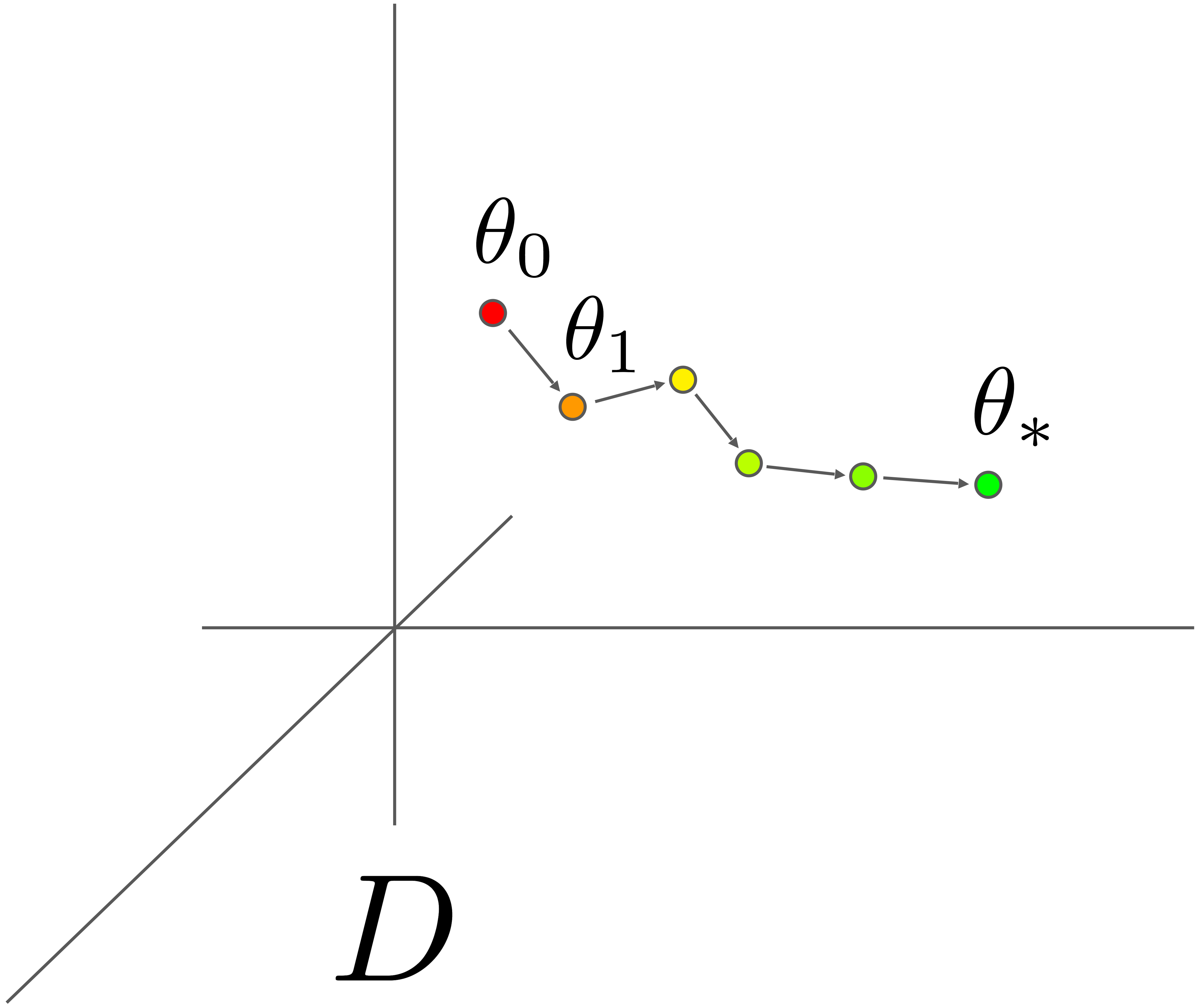


Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Typical Neural Network Training

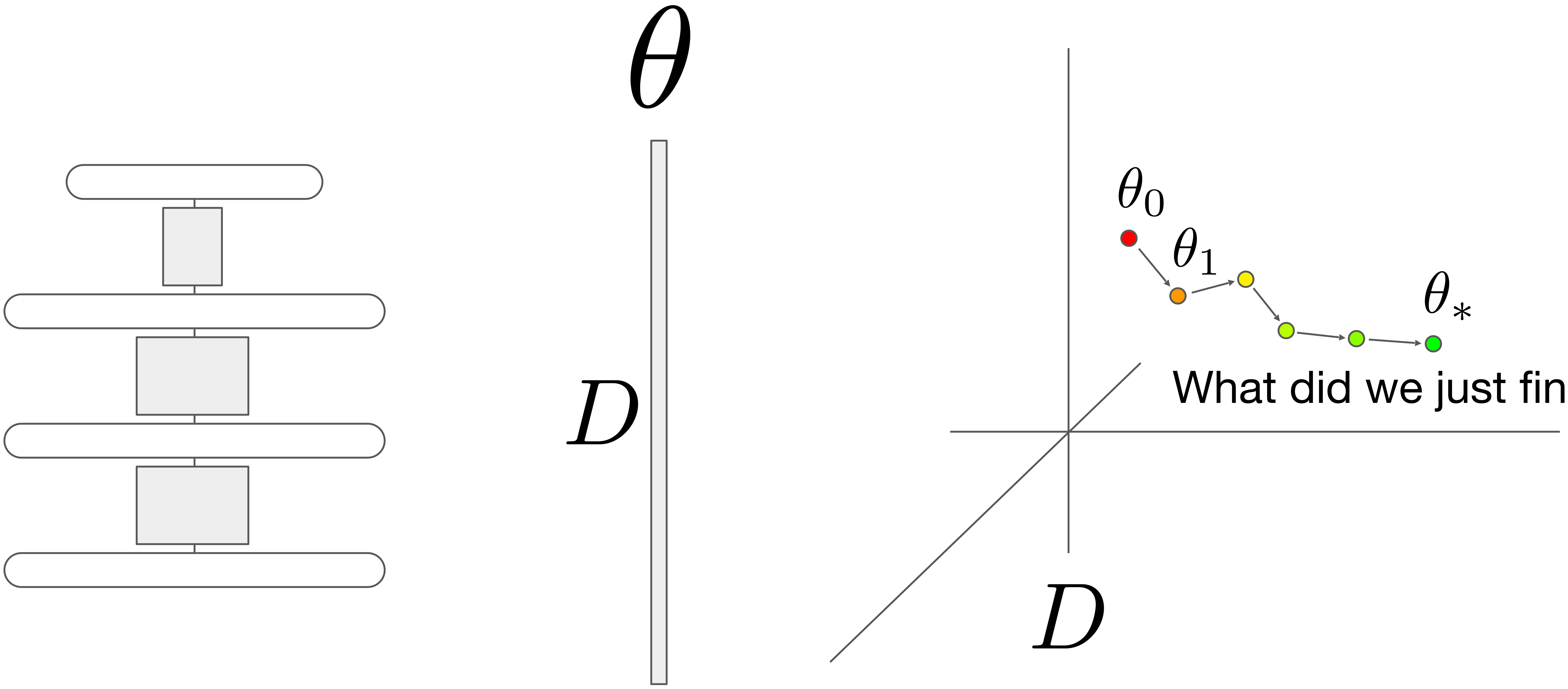






Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Typical Neural Network Training



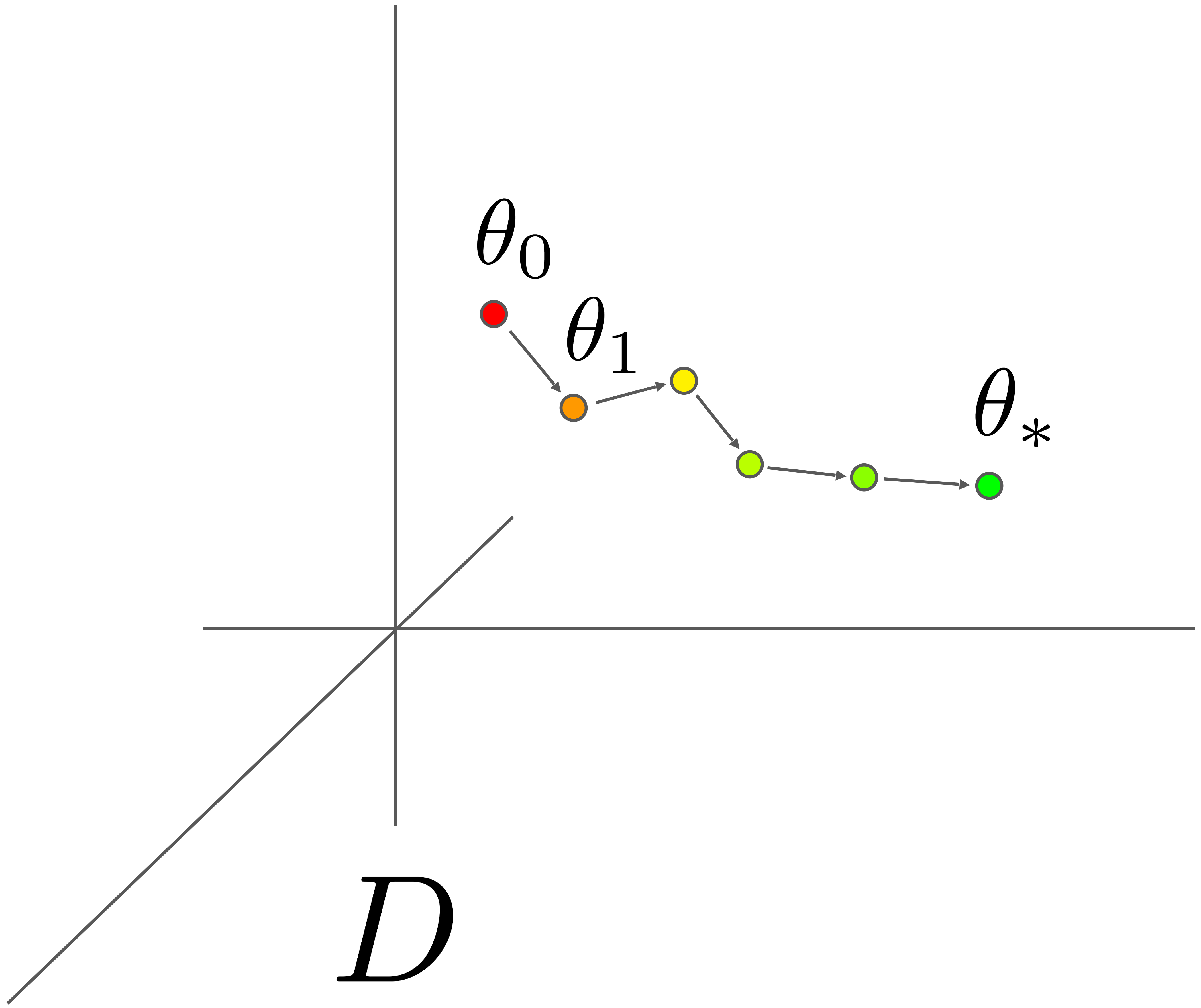
What did we just find?

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

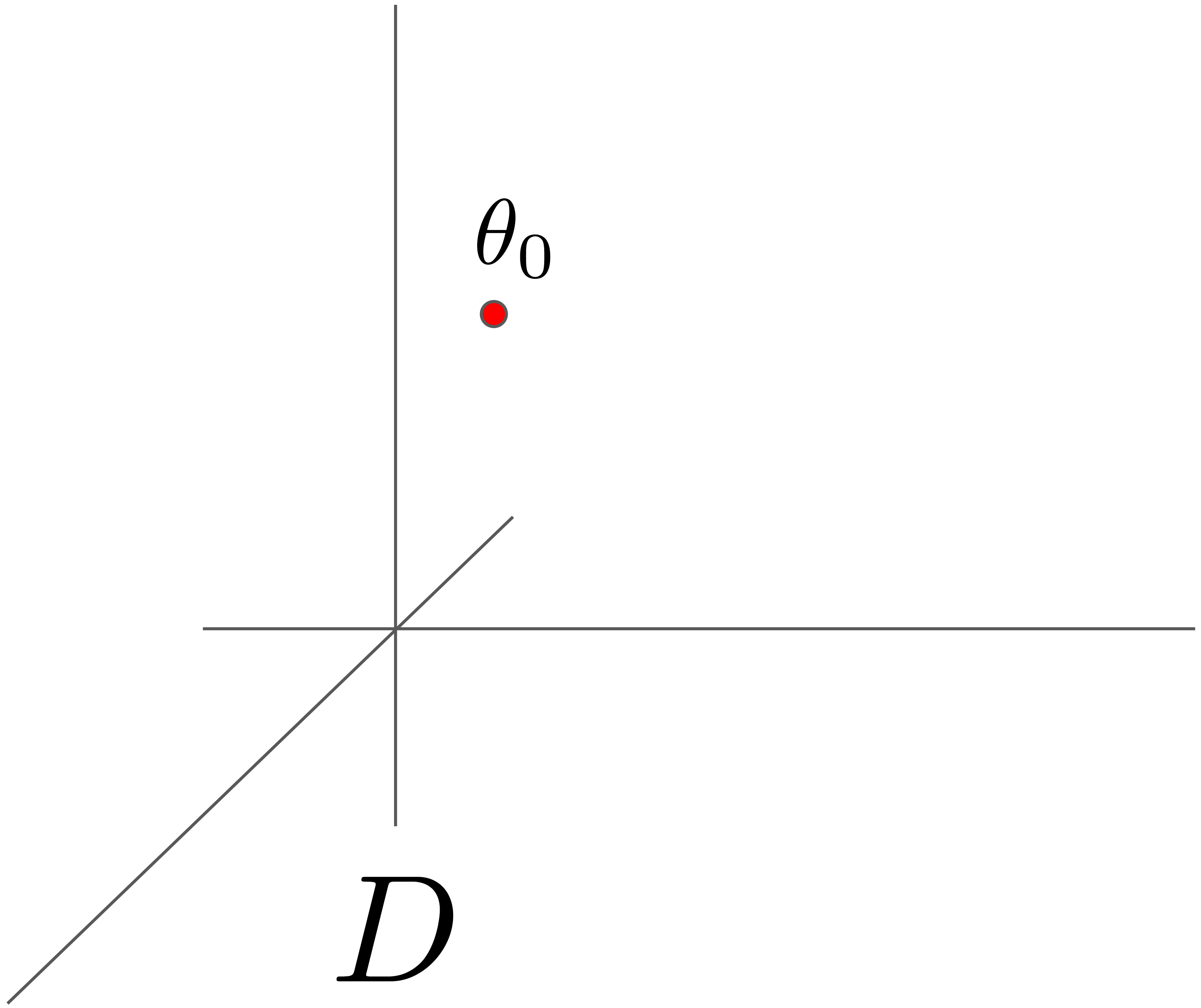
Random Subspace Training



Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

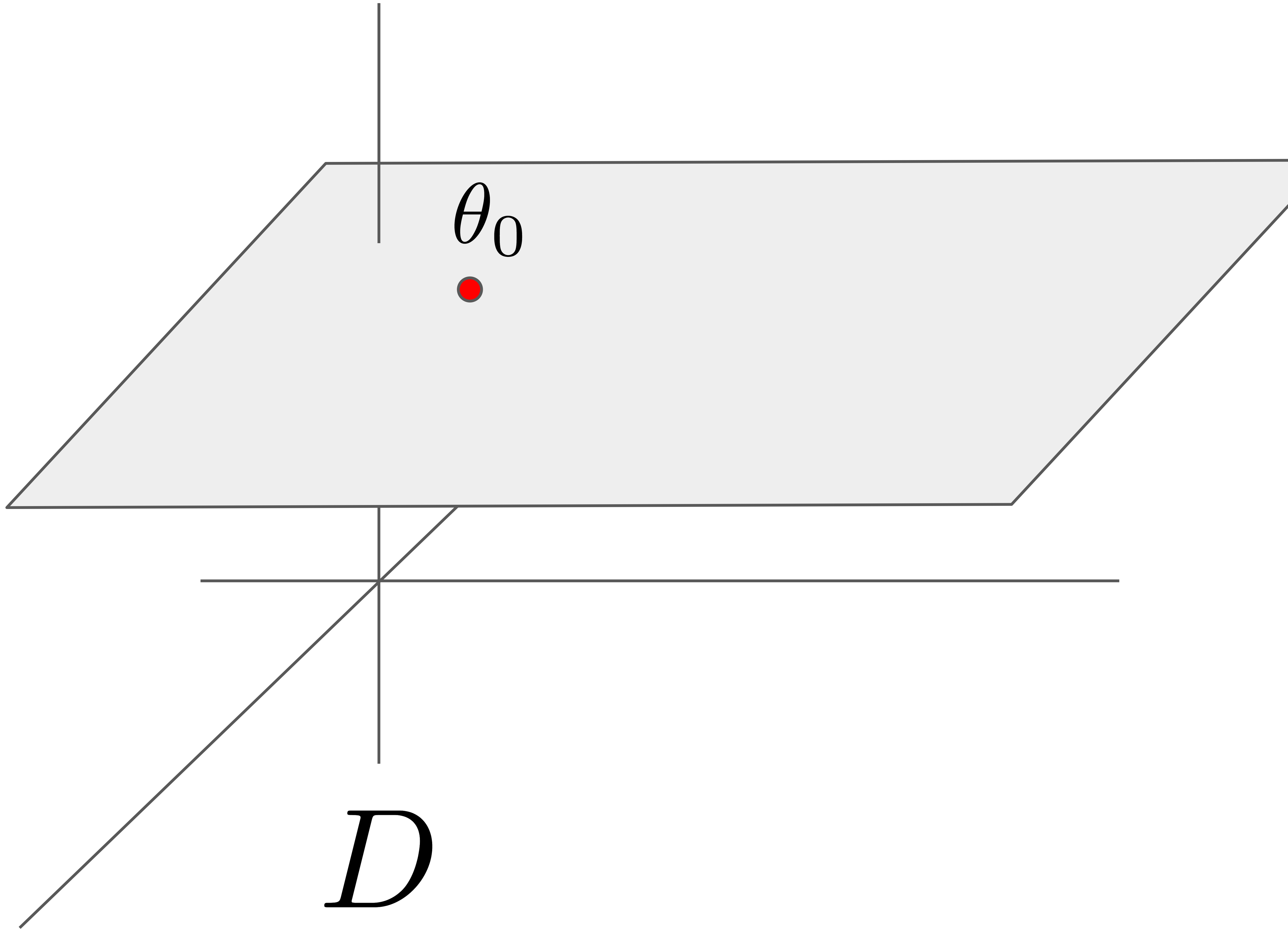
- Choose the initial point the same way



Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

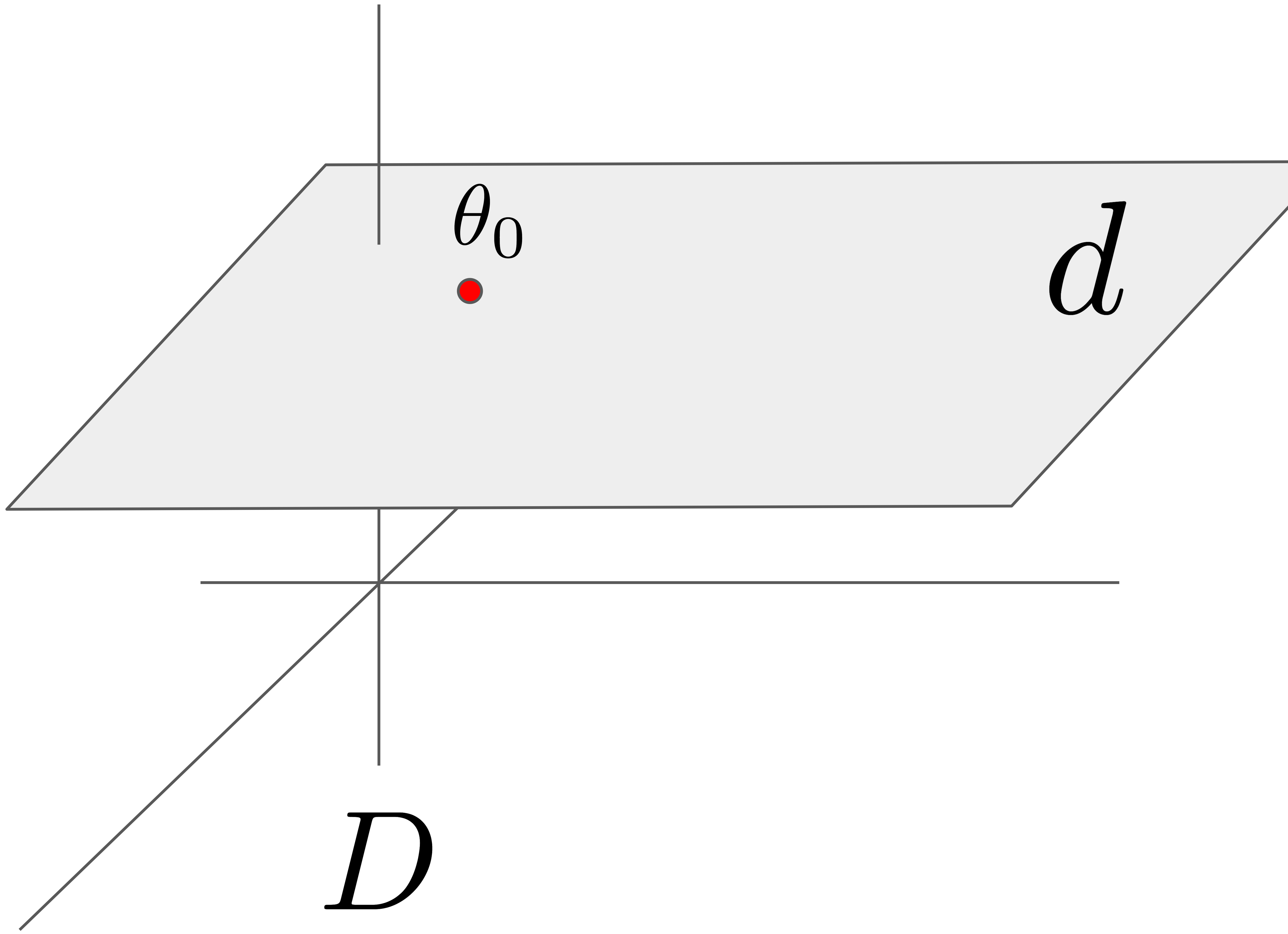
- Choose the initial point the same way
- Generate a random subspace around that initial point



Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

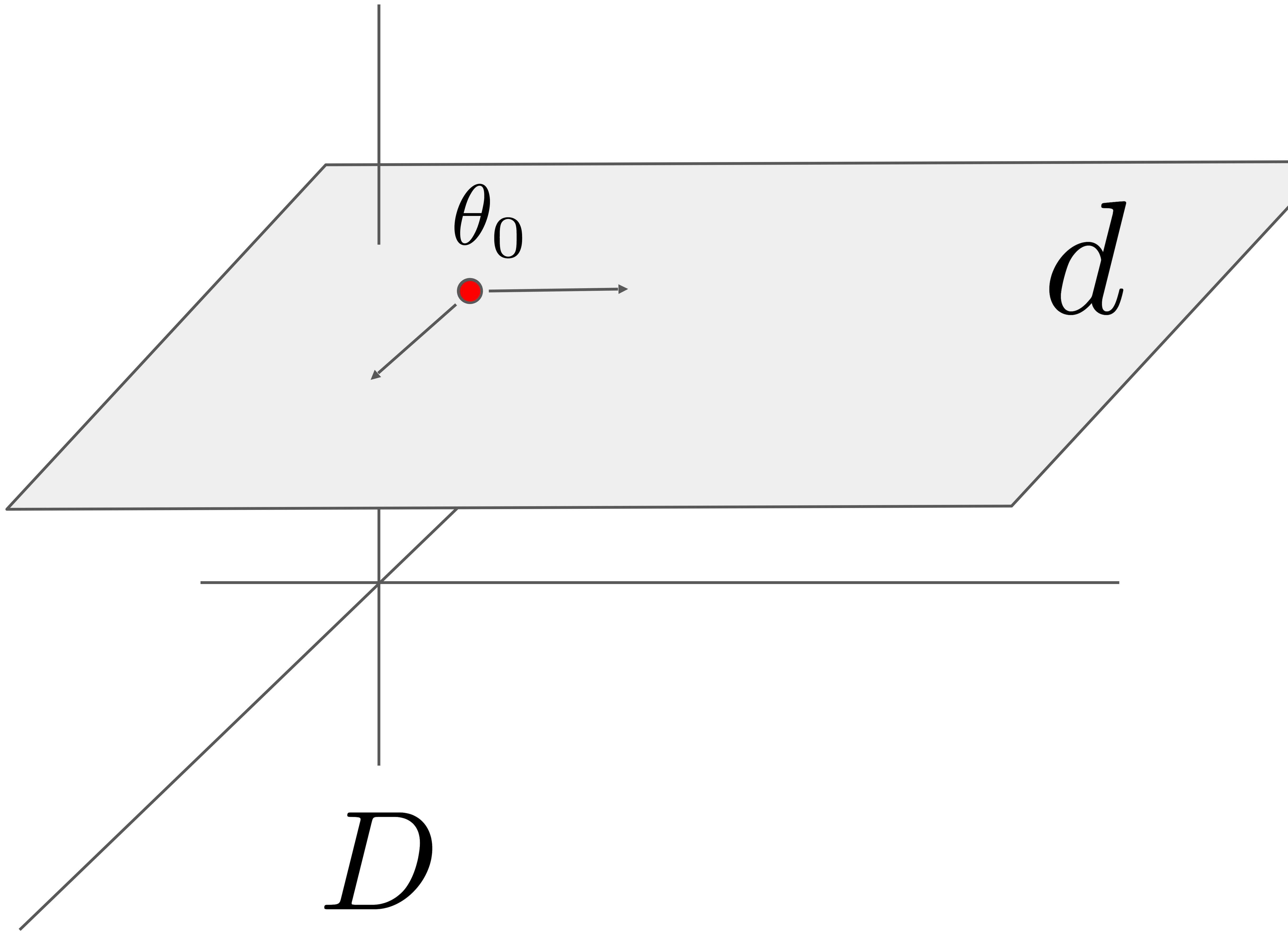
- Choose the initial point the same way
- Generate a random subspace around that initial point



Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

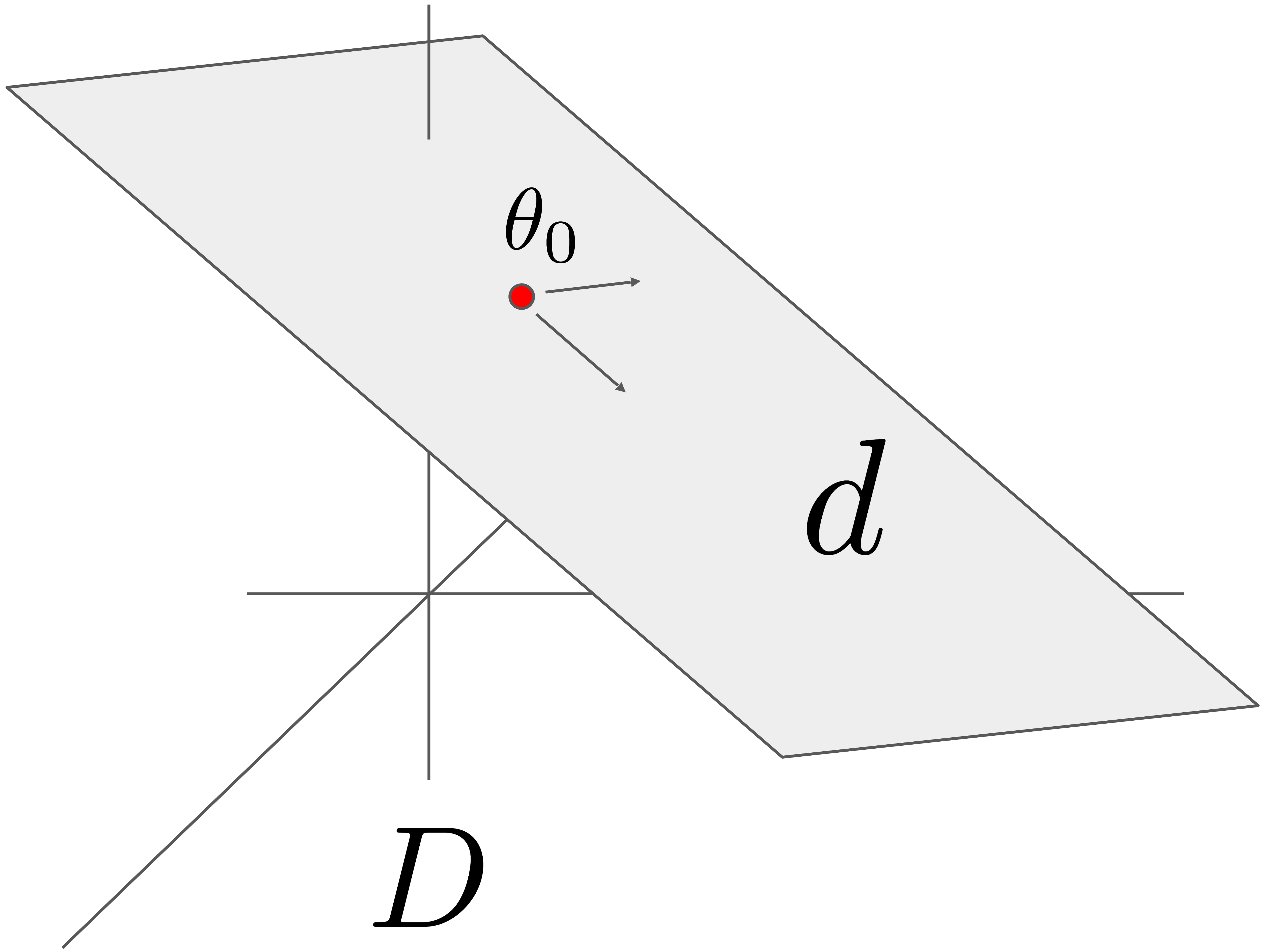
- Choose the initial point the same way
- Generate a random subspace around that initial point
- Allow the optimizer to move only in that subspace



Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

- Choose the initial point the same way
- Generate a random subspace around that initial point
- Allow the optimizer to move only in that subspace

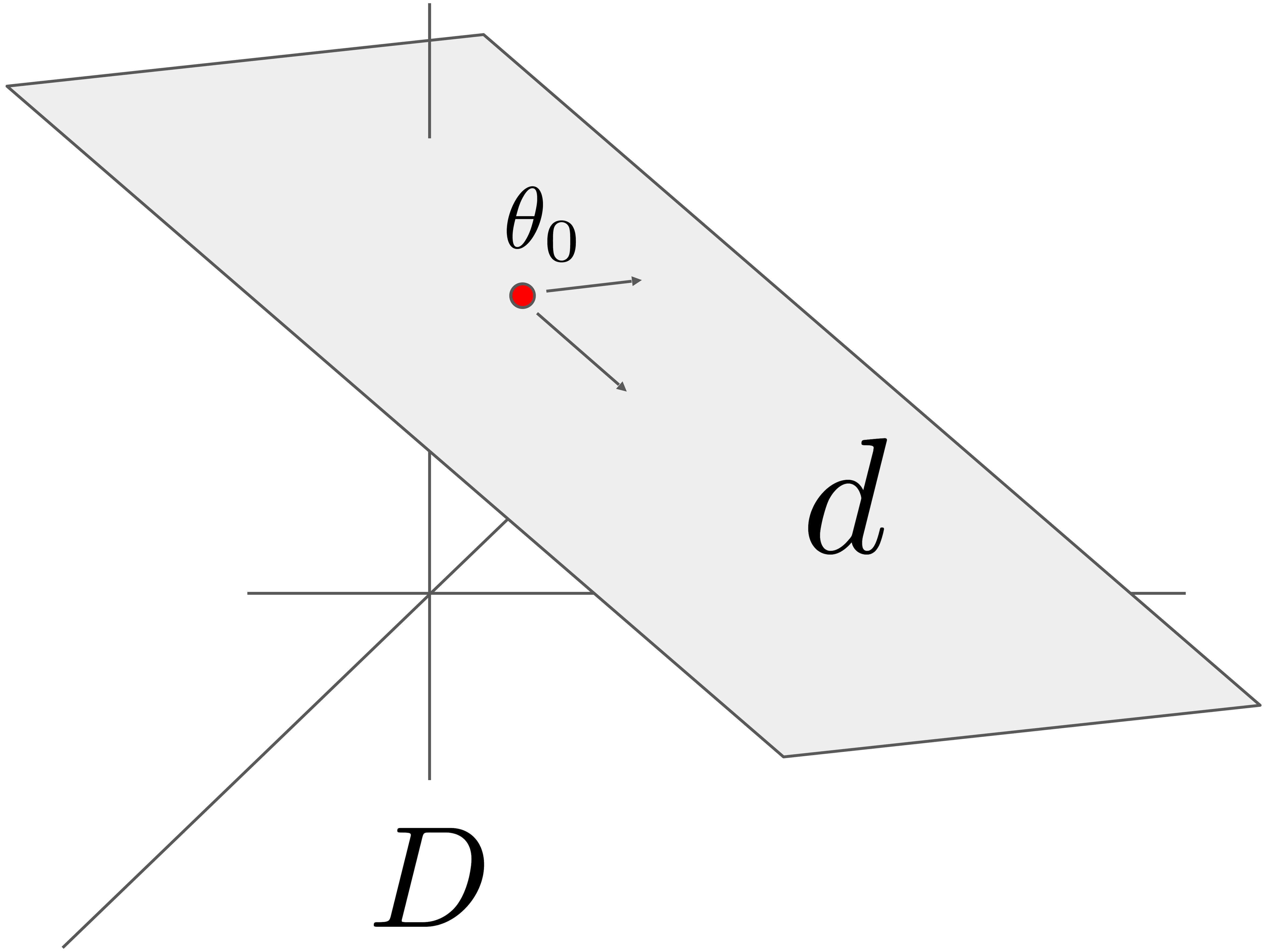


Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

- Choose the initial point the same way
- Generate a random subspace around that initial point
- Allow the optimizer to move only in that subspace

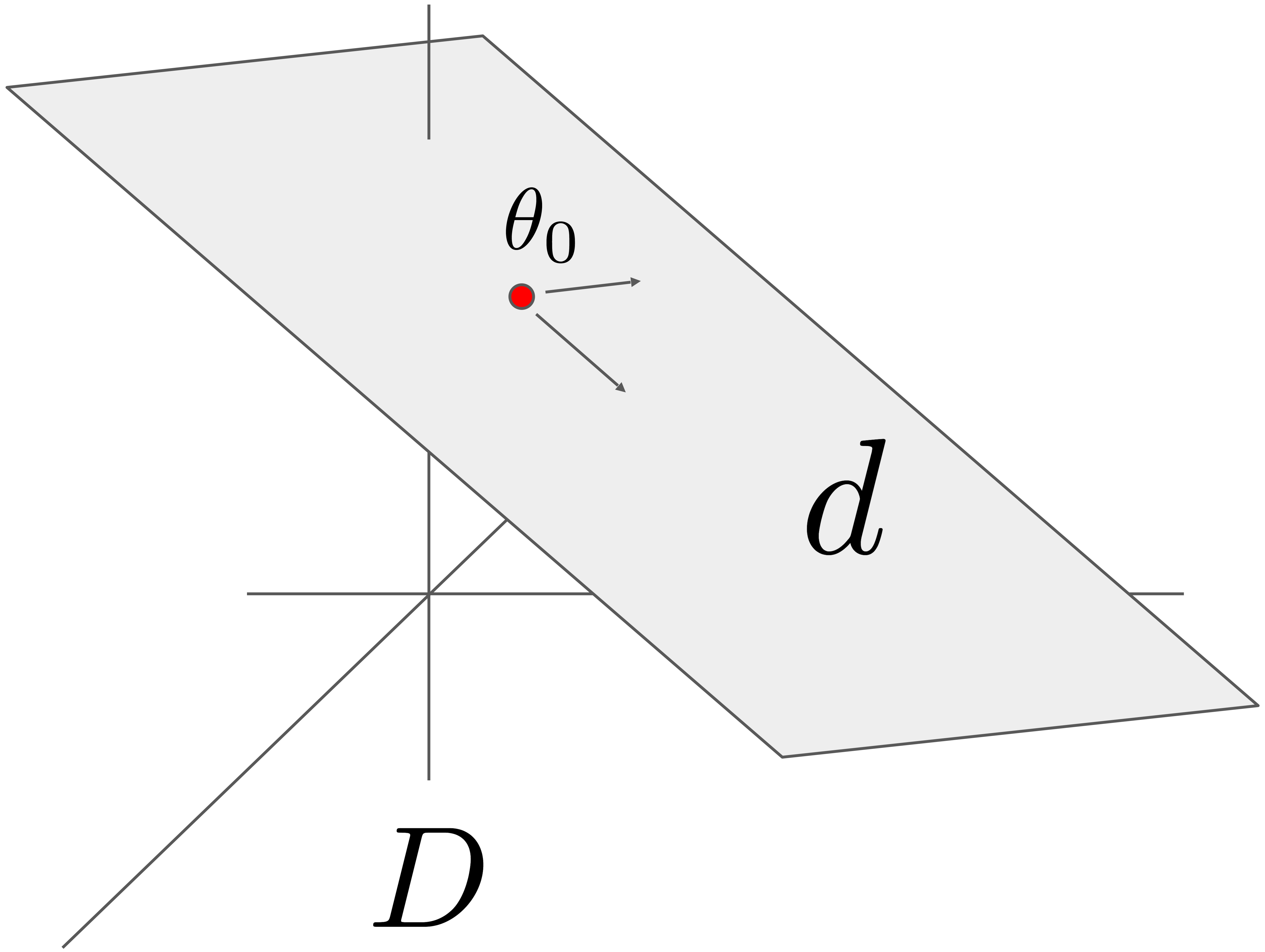
$$\theta^{(D)} = \theta_0^{(D)} + P\theta^{(d)}$$



Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

- Choose the initial point the same way
- Generate a random subspace around that initial point
- Allow the optimizer to move only in that subspace



$$\theta^{(D)} = \theta_0^{(D)} + \boxed{P}\theta^{(d)}$$

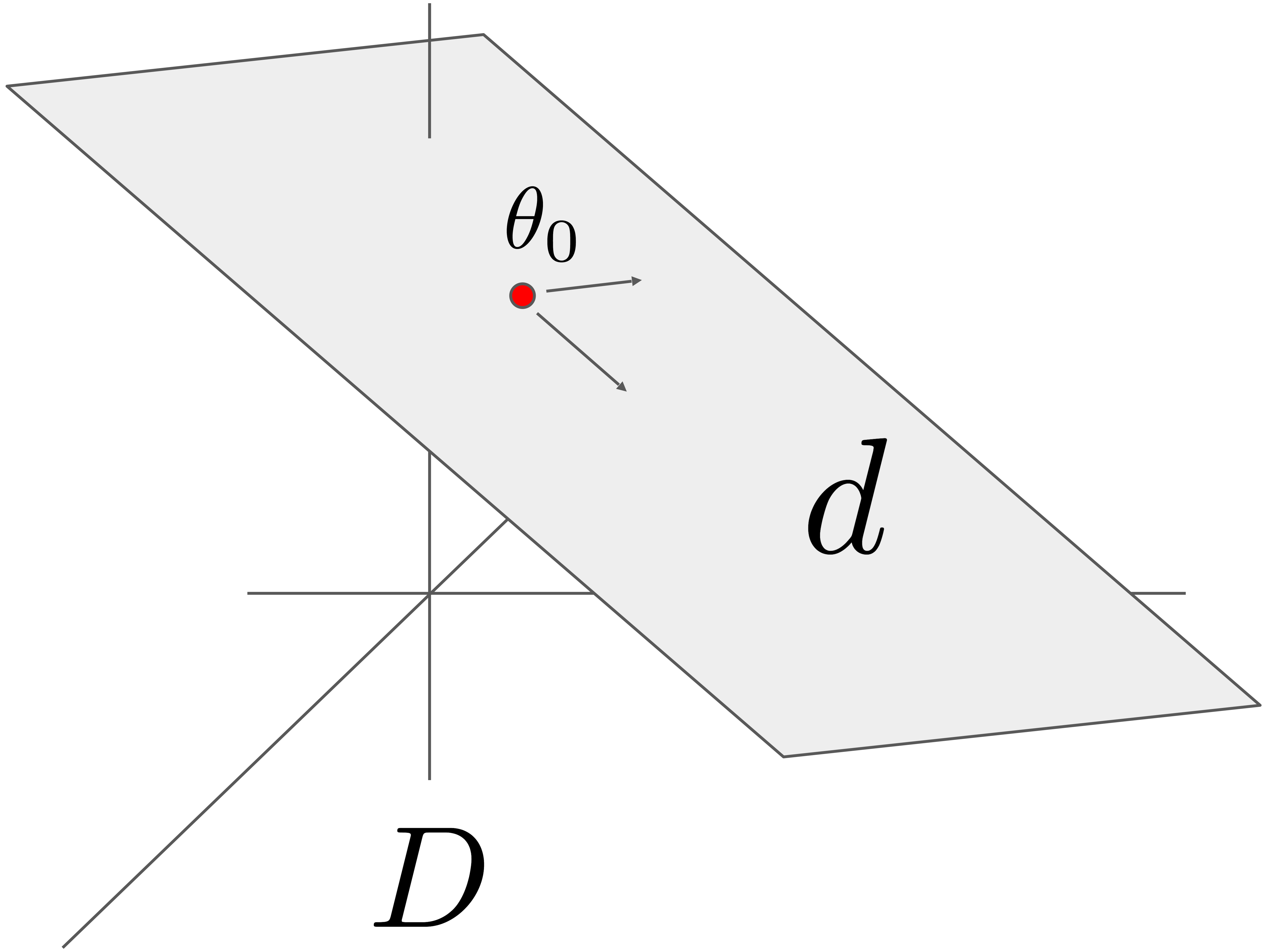
Random projection, R^{Dxd}

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

- Choose the initial point the same way
- Generate a random subspace around that initial point
- Allow the optimizer to move only in that subspace

Metric: how large must d be to solve the problem?



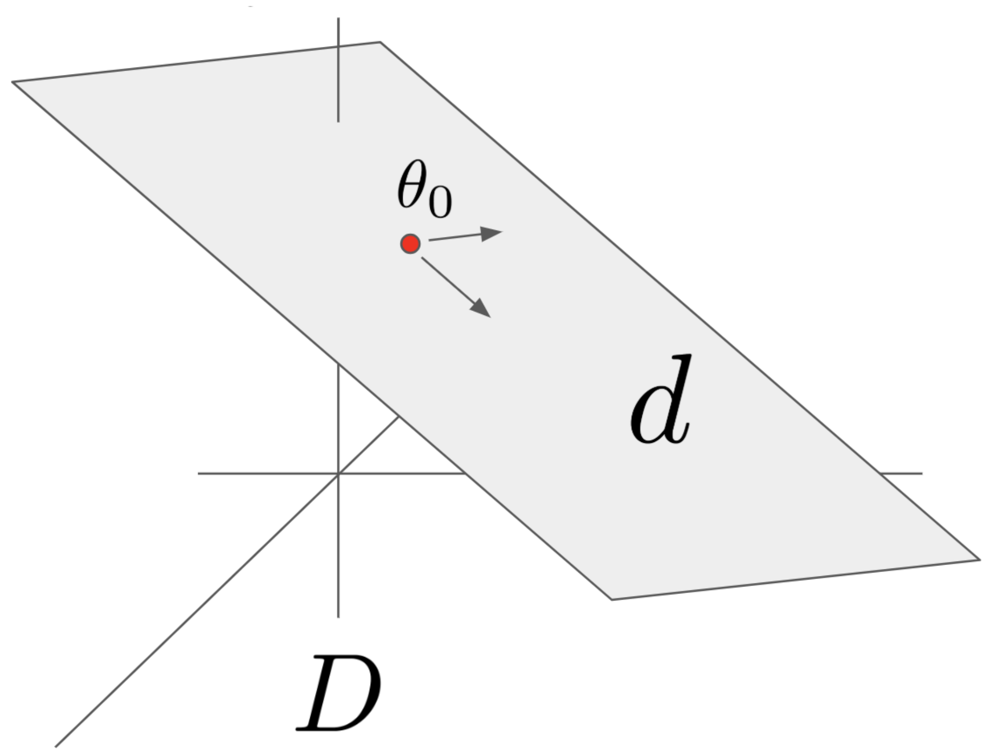
$$\theta^{(D)} = \theta_0^{(D)} + P\theta^{(d)}$$

Random projection, R^{Dxd}

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

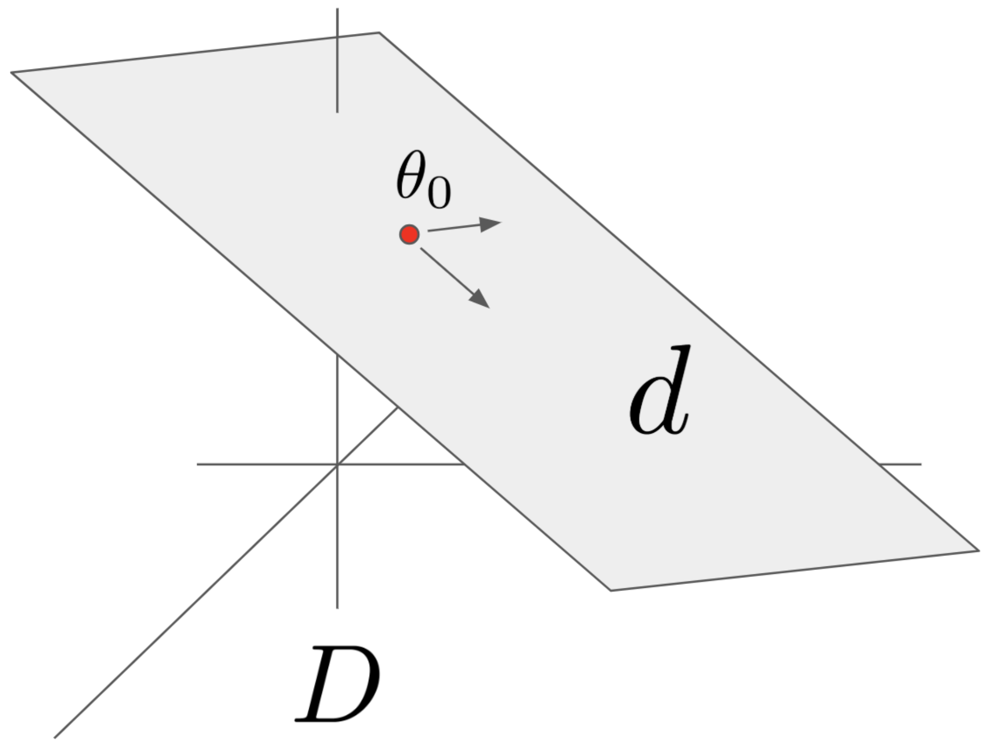
Random Subspace Training

Metric: how large must d be to solve the problem?

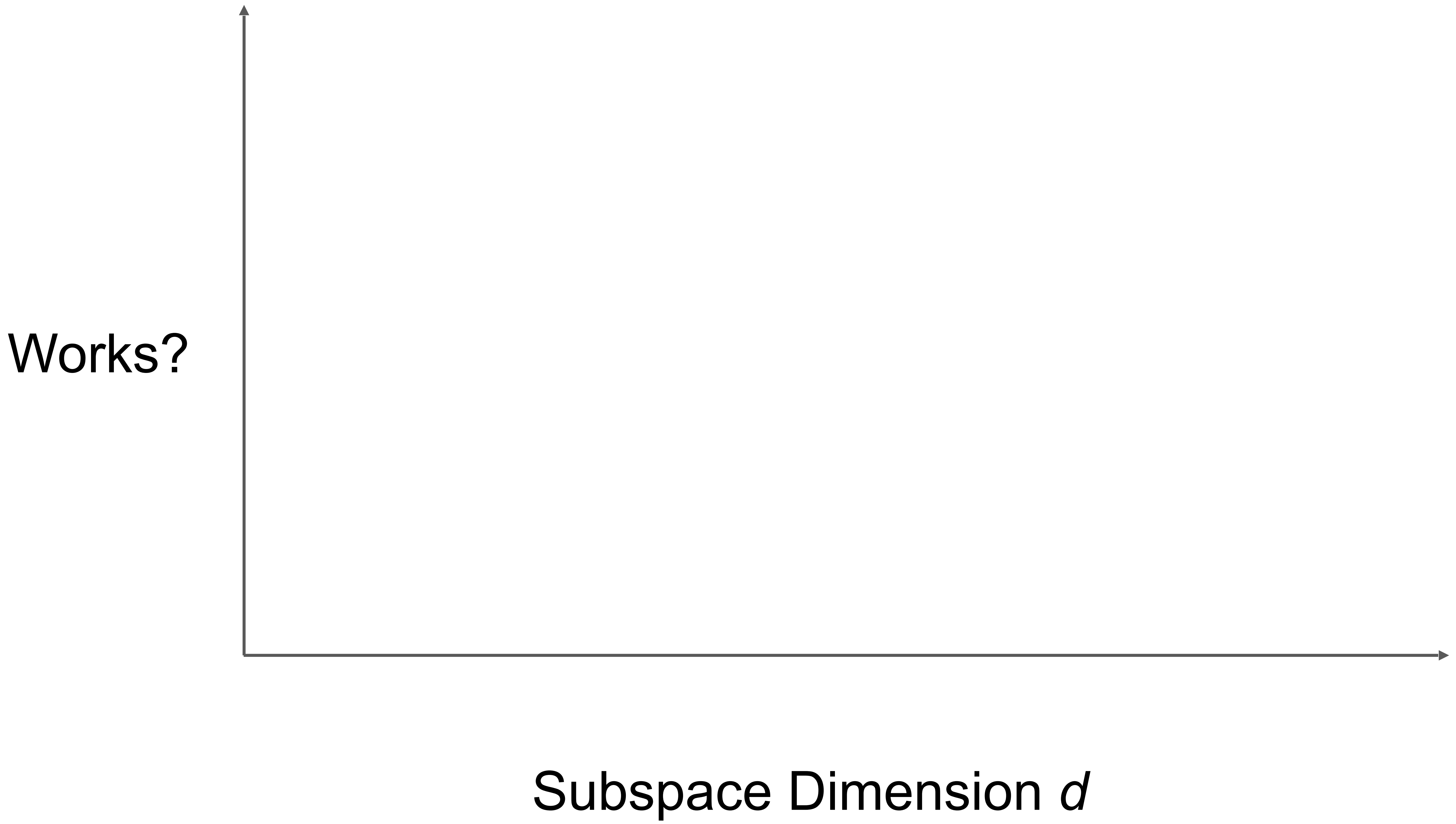


Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

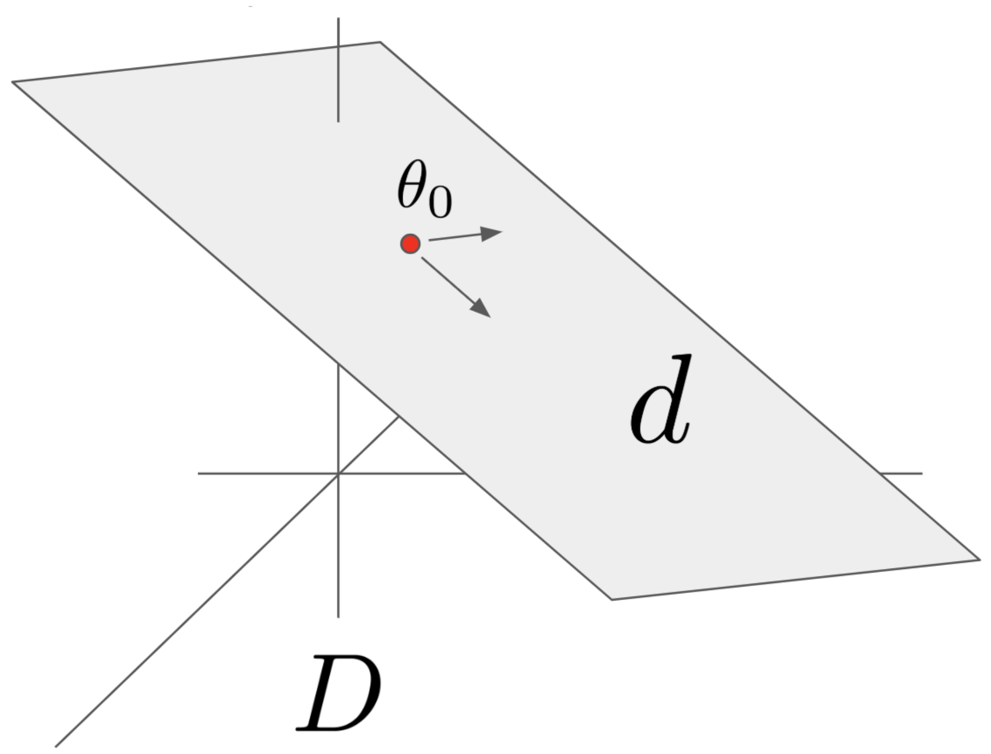


Metric: how large must d be to solve the problem?

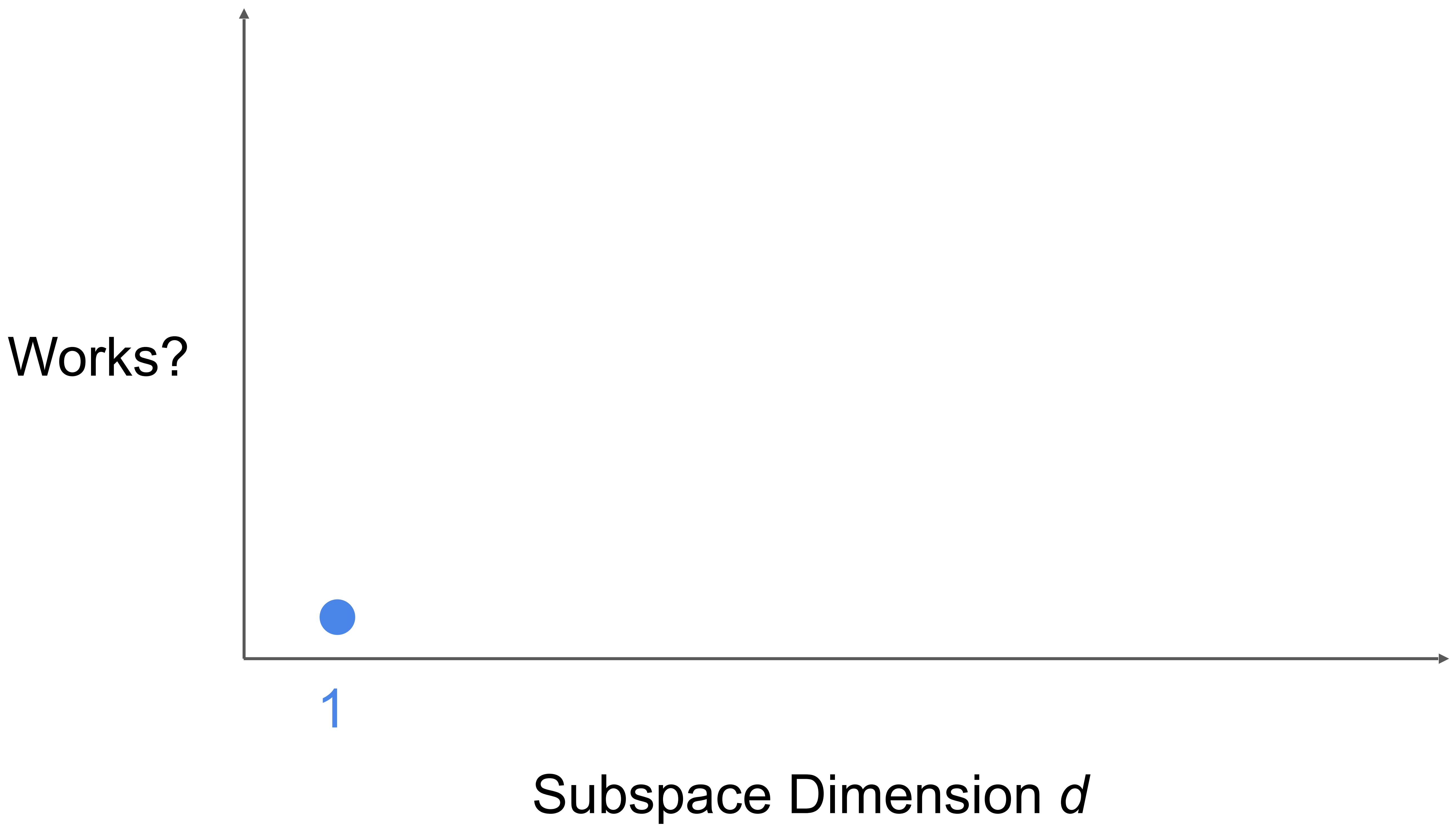


Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training



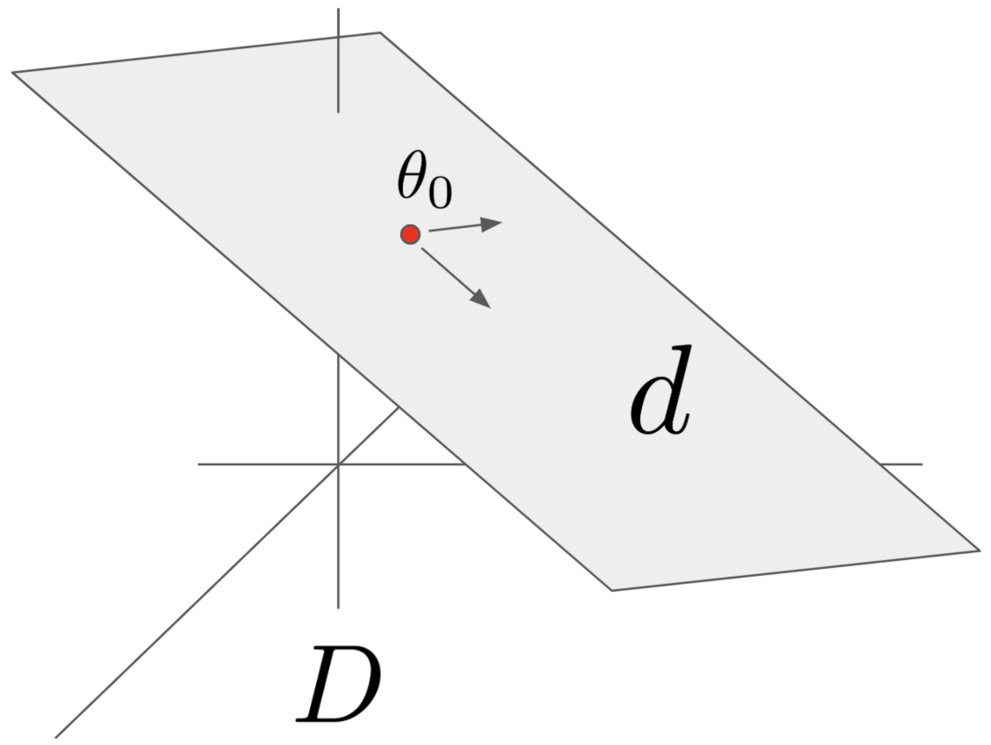
Metric: how large must d be to solve the problem?

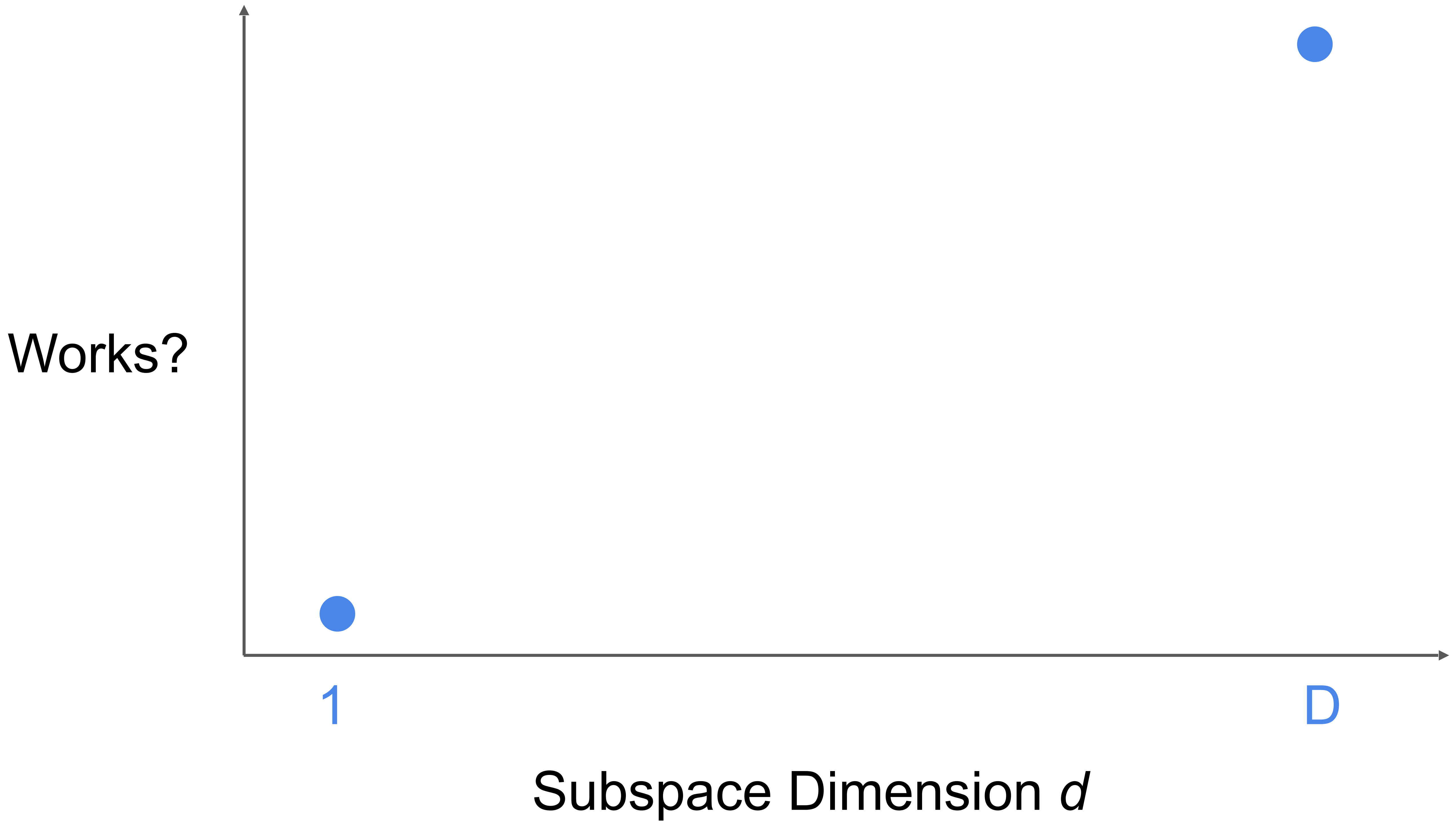


Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

Metric: how large must d be to solve the problem?

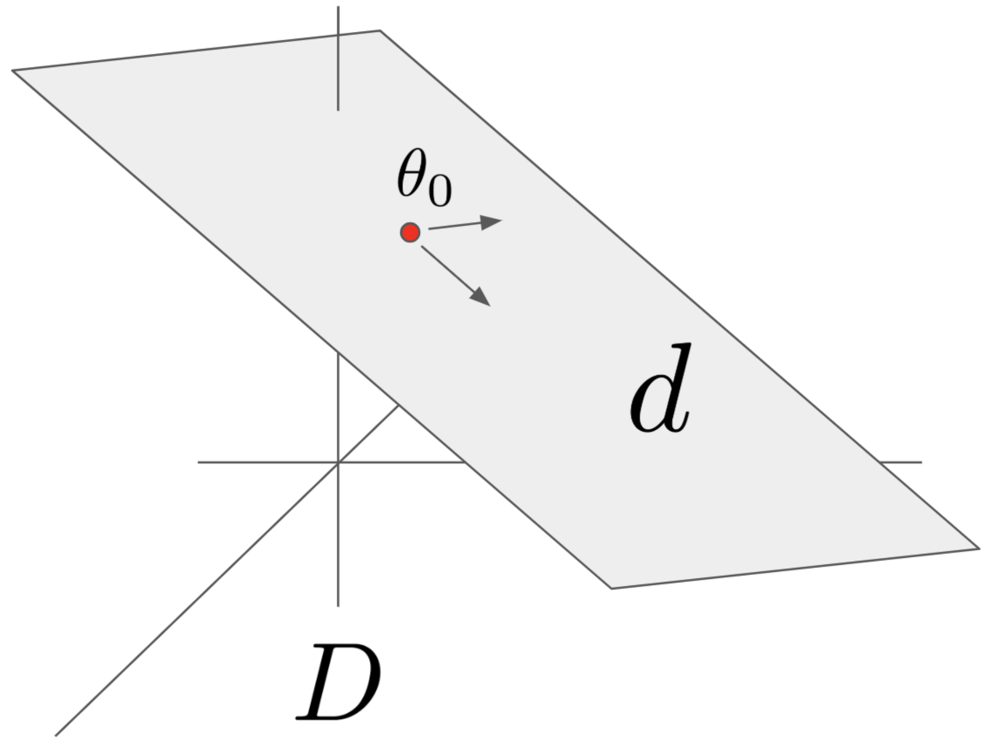


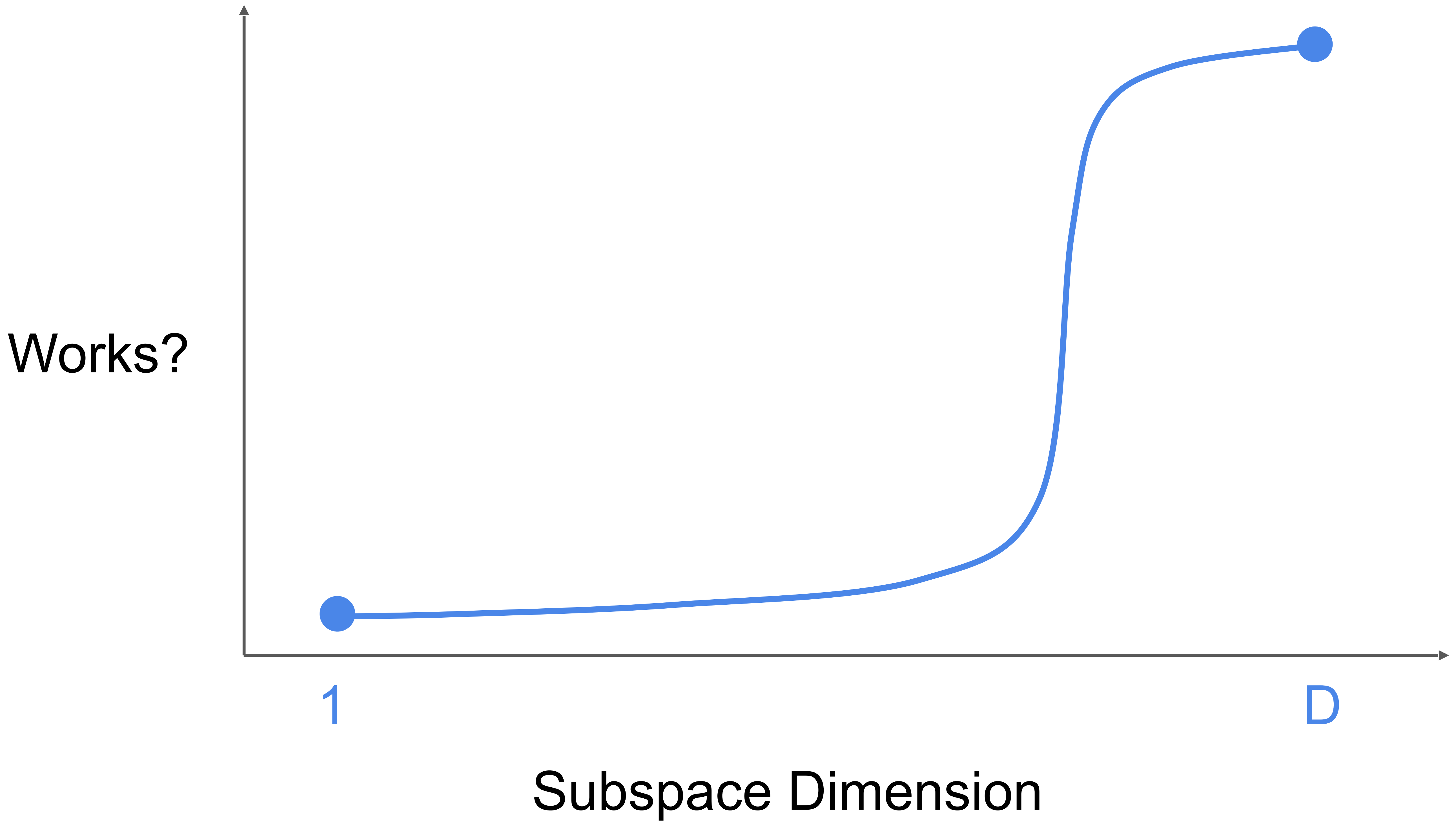


Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

Metric: how large must d be to solve the problem?

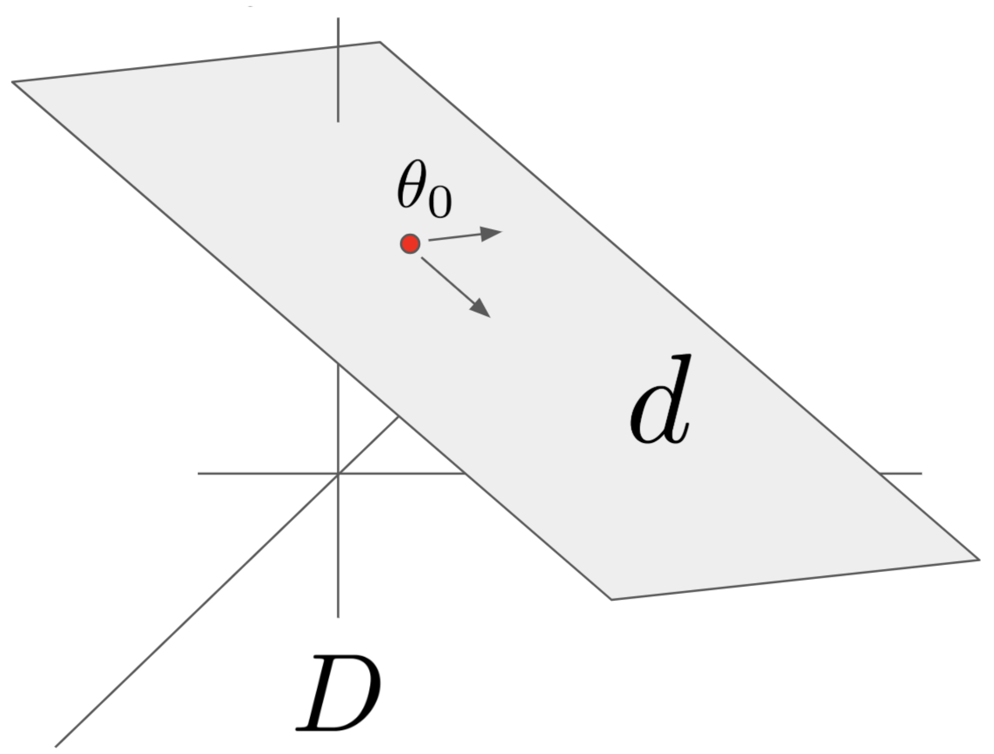


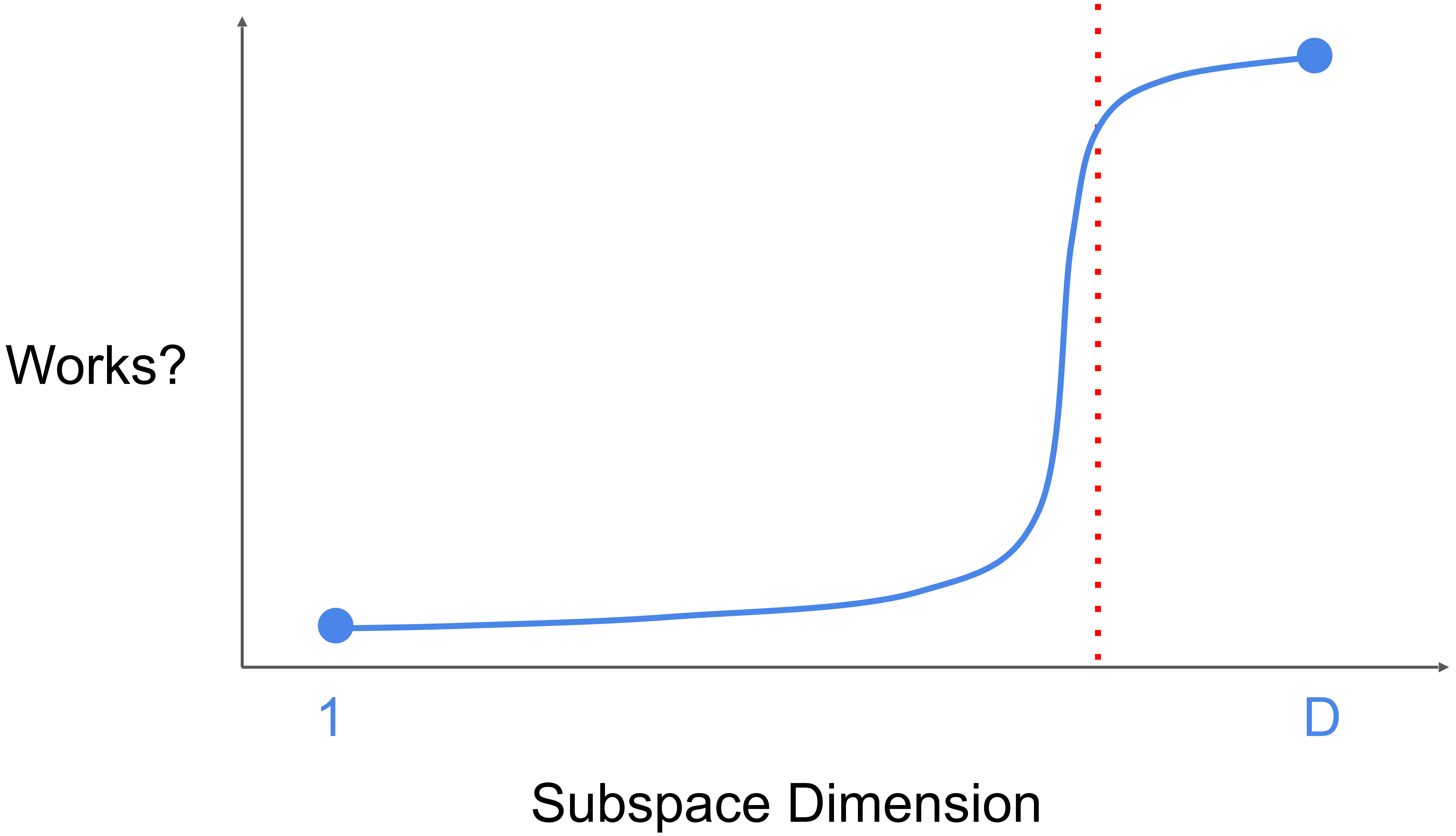


Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

Metric: how large must d be to solve the problem?

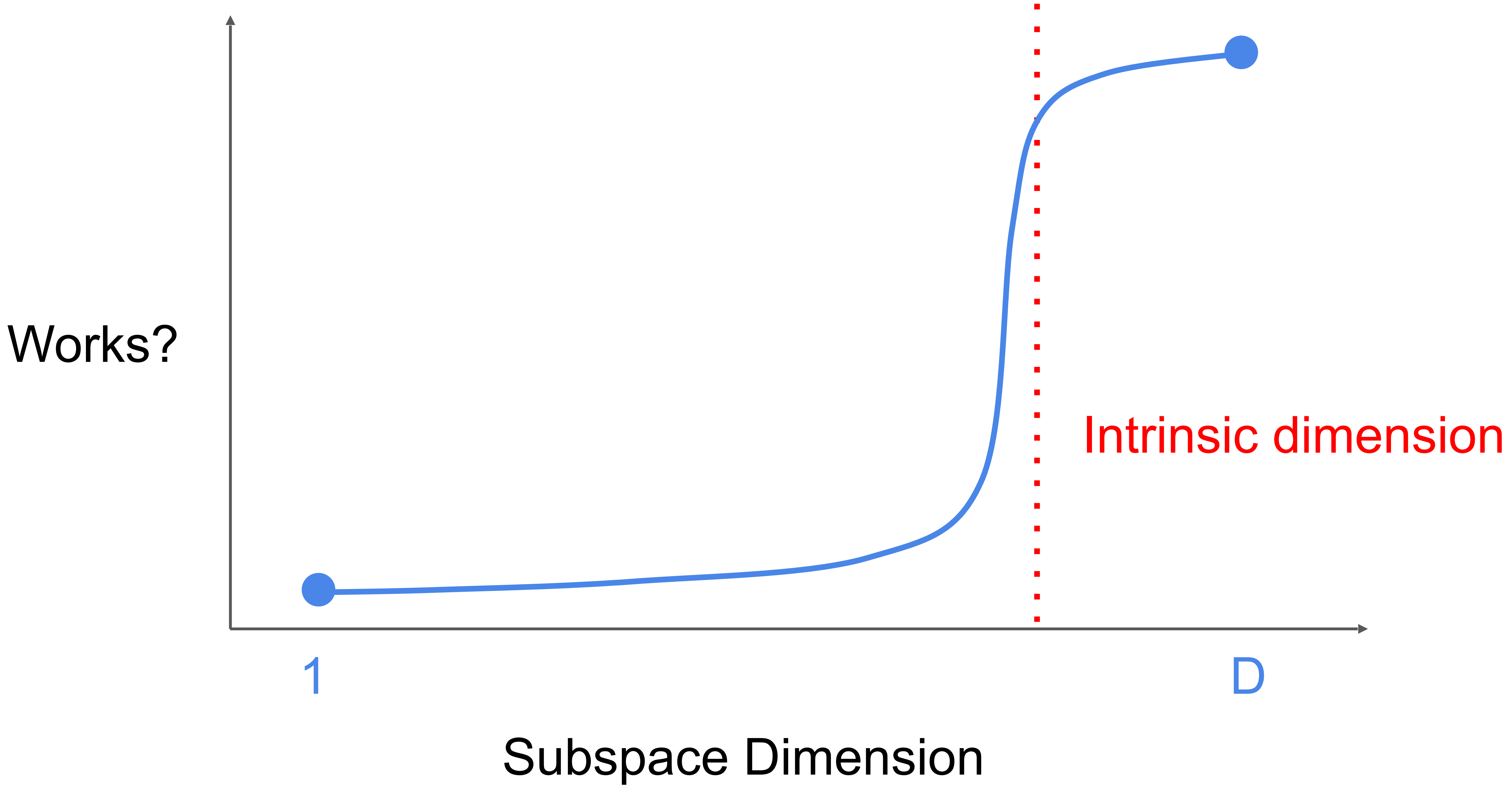


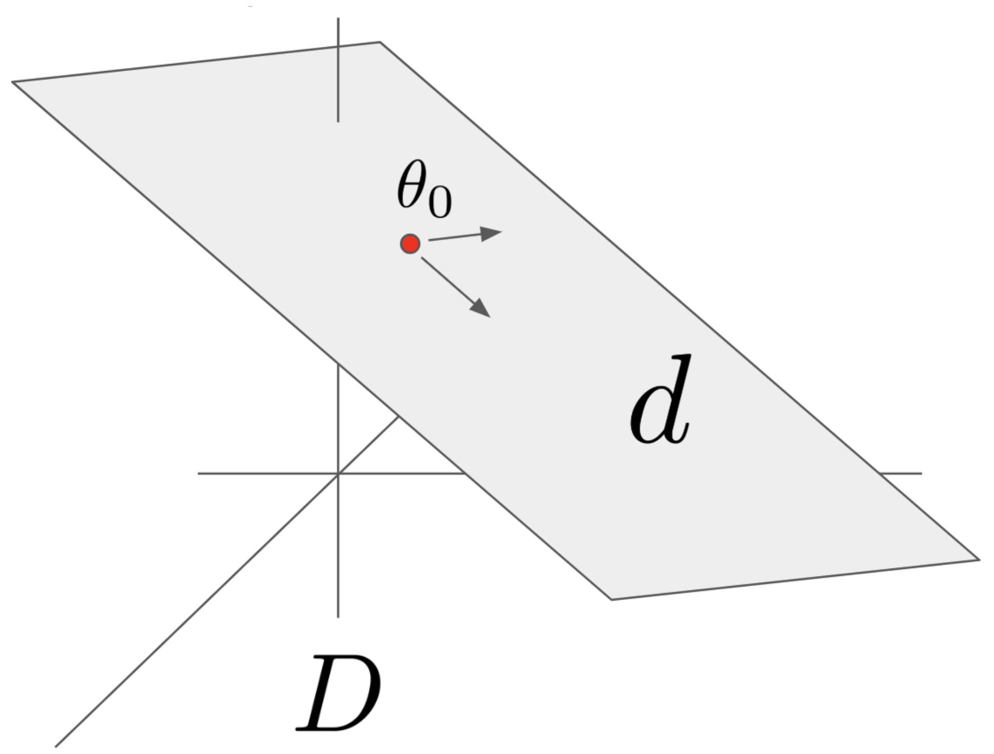


Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

Random Subspace Training

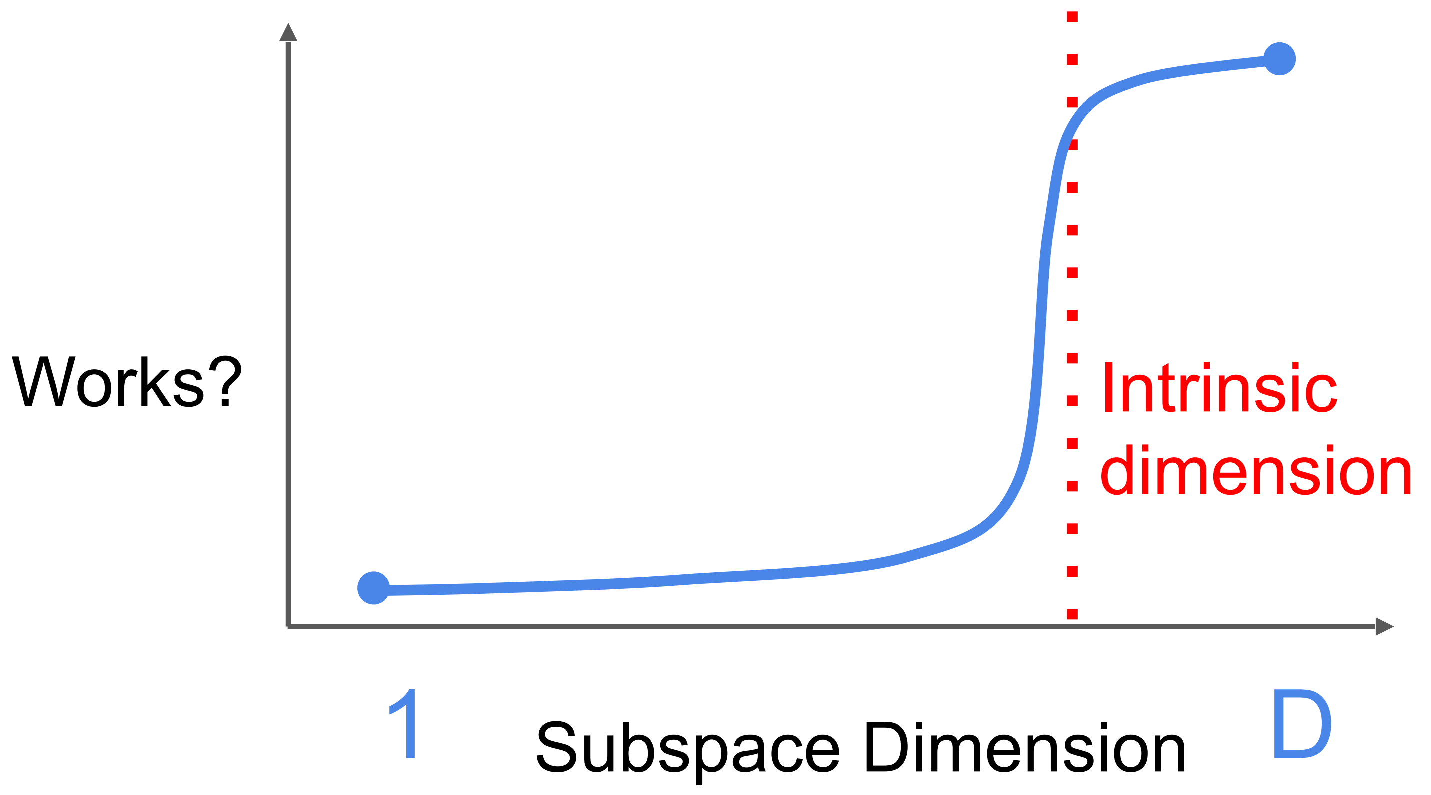
Metric: how large must d be to solve the problem?





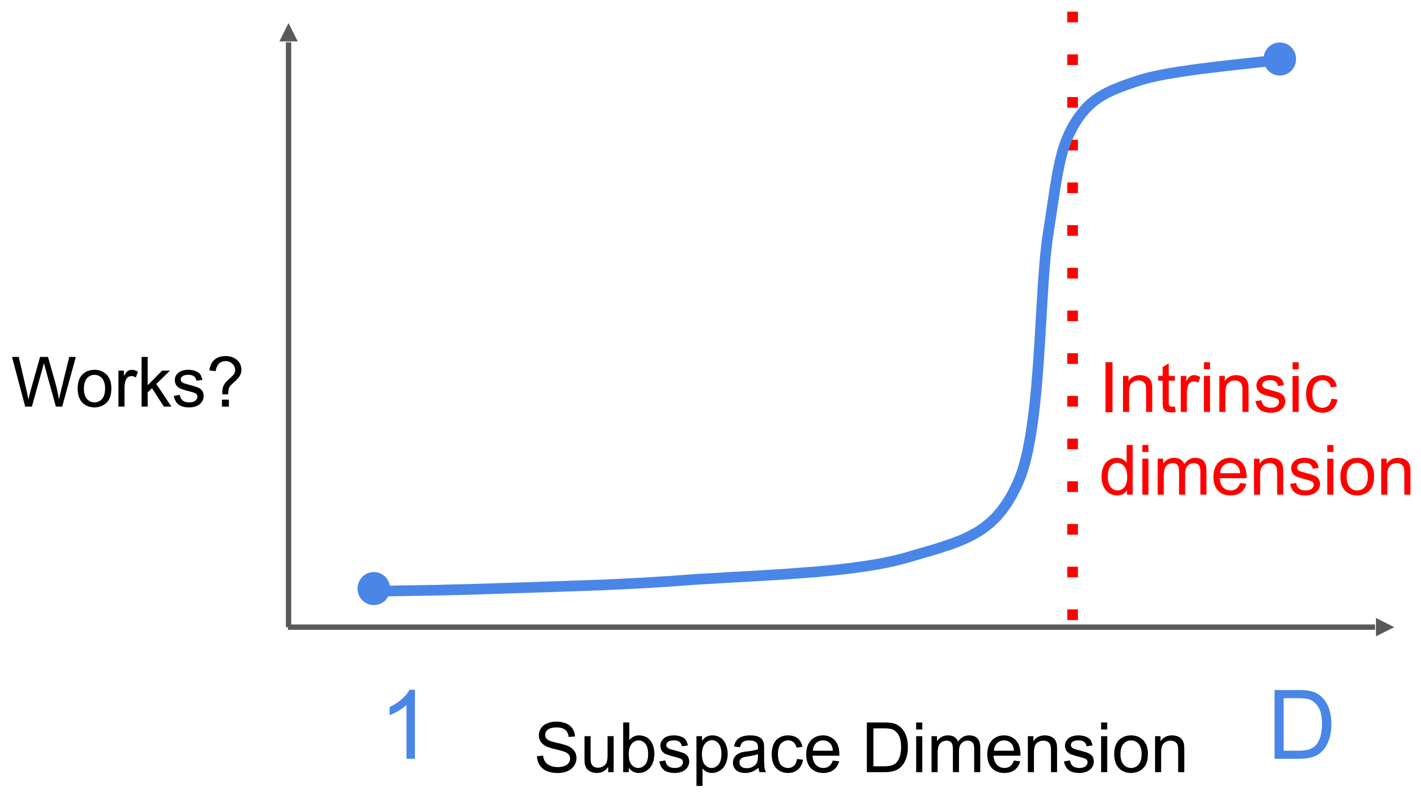
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

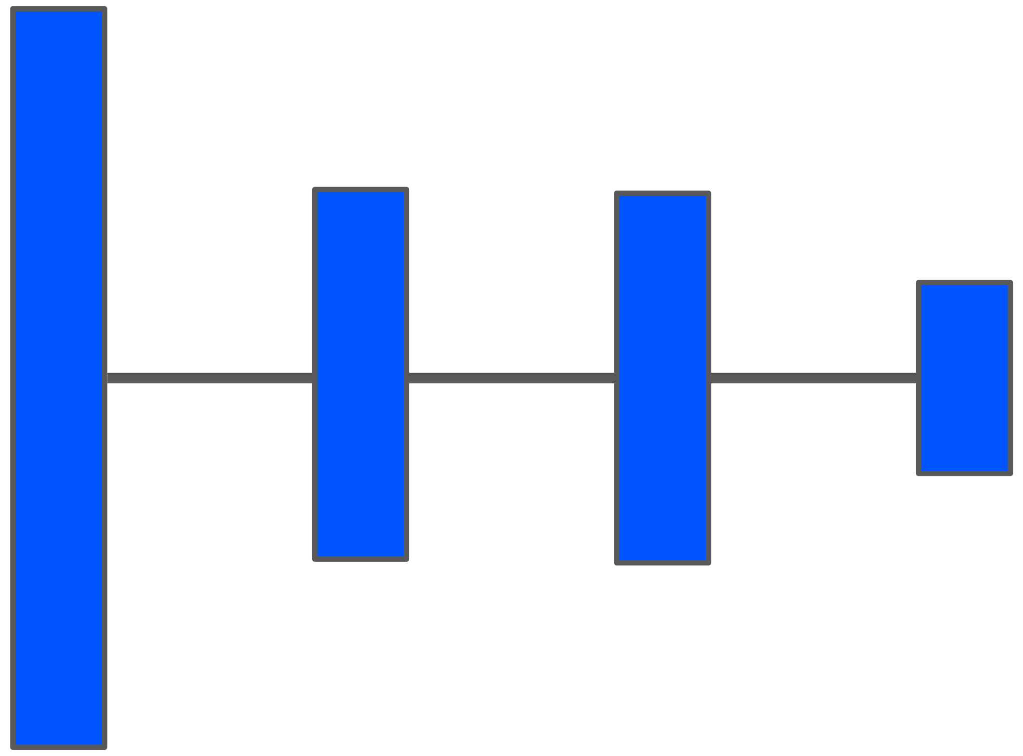
Try it on many {Problem, Model} s



Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

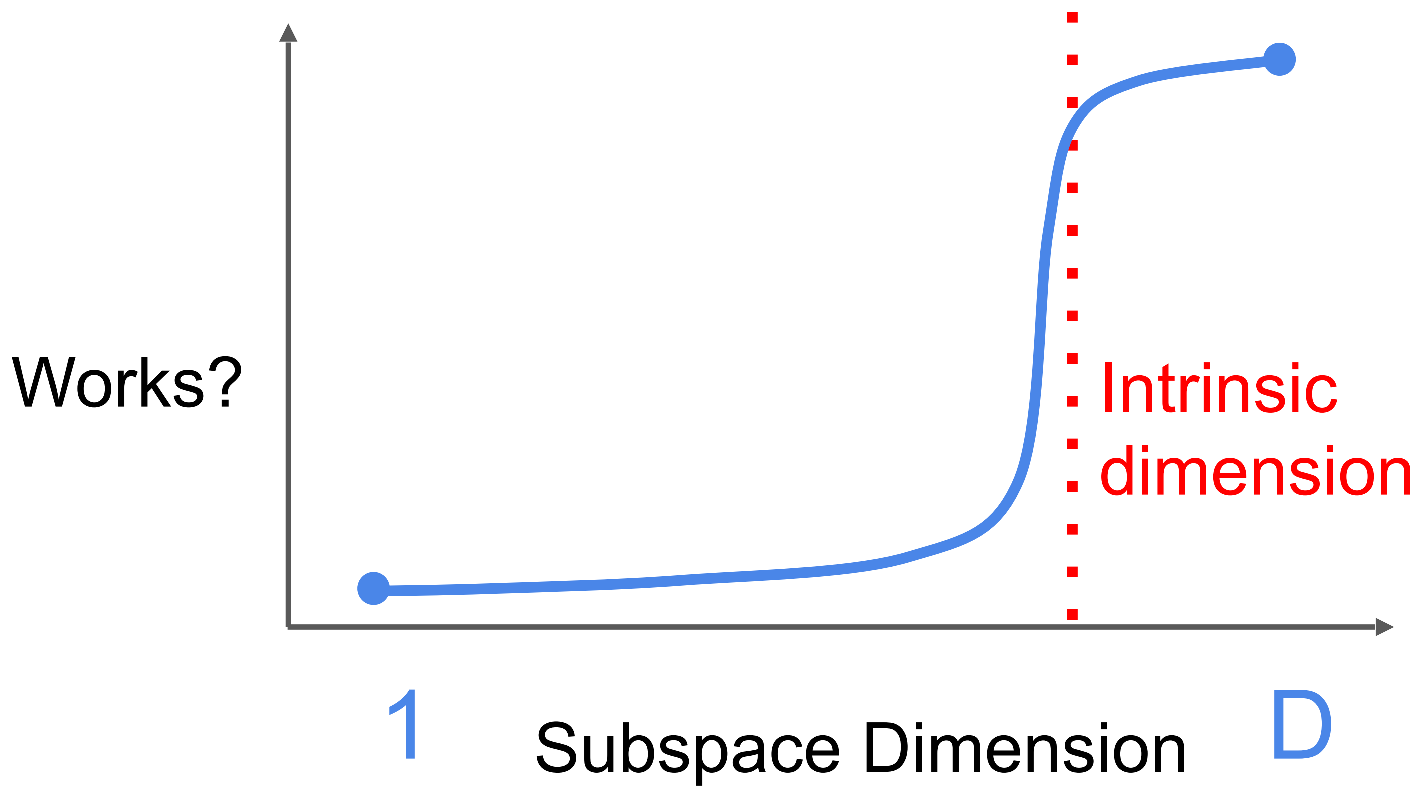
{MNIST, FC}

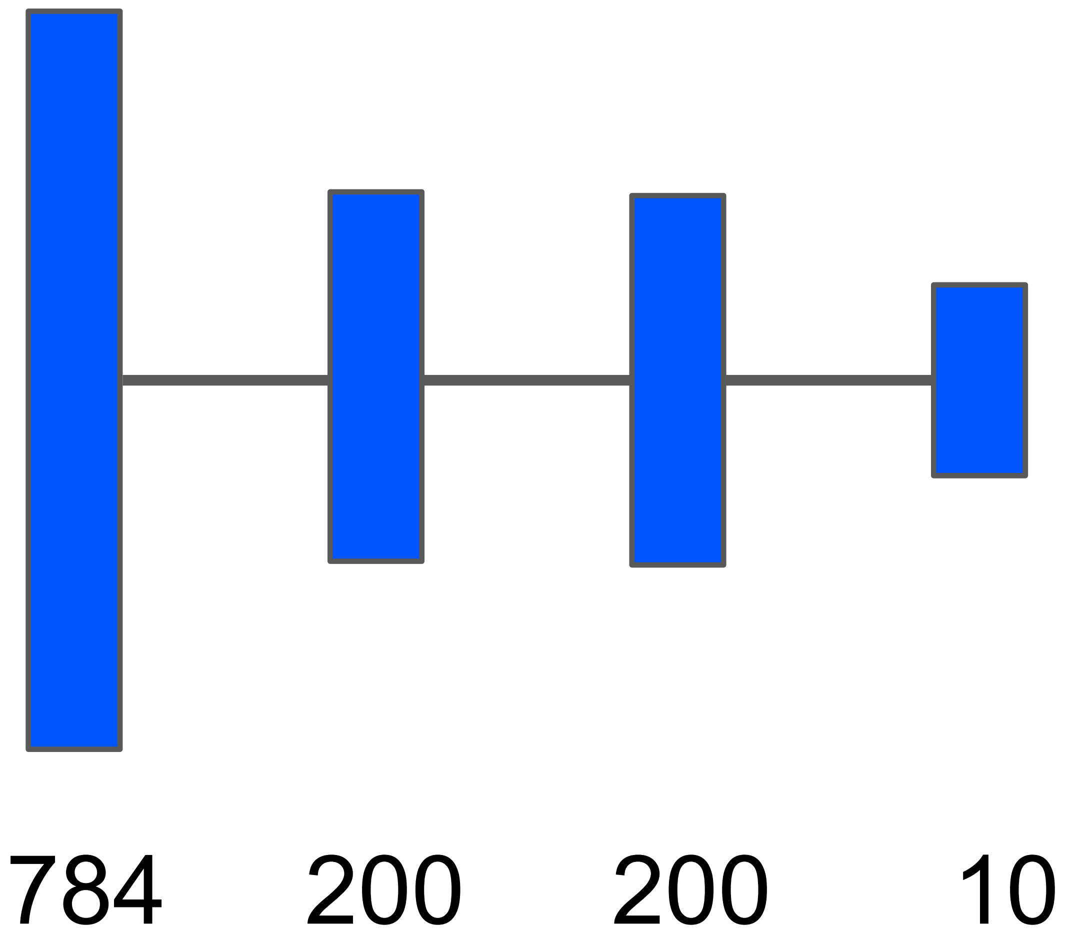




Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

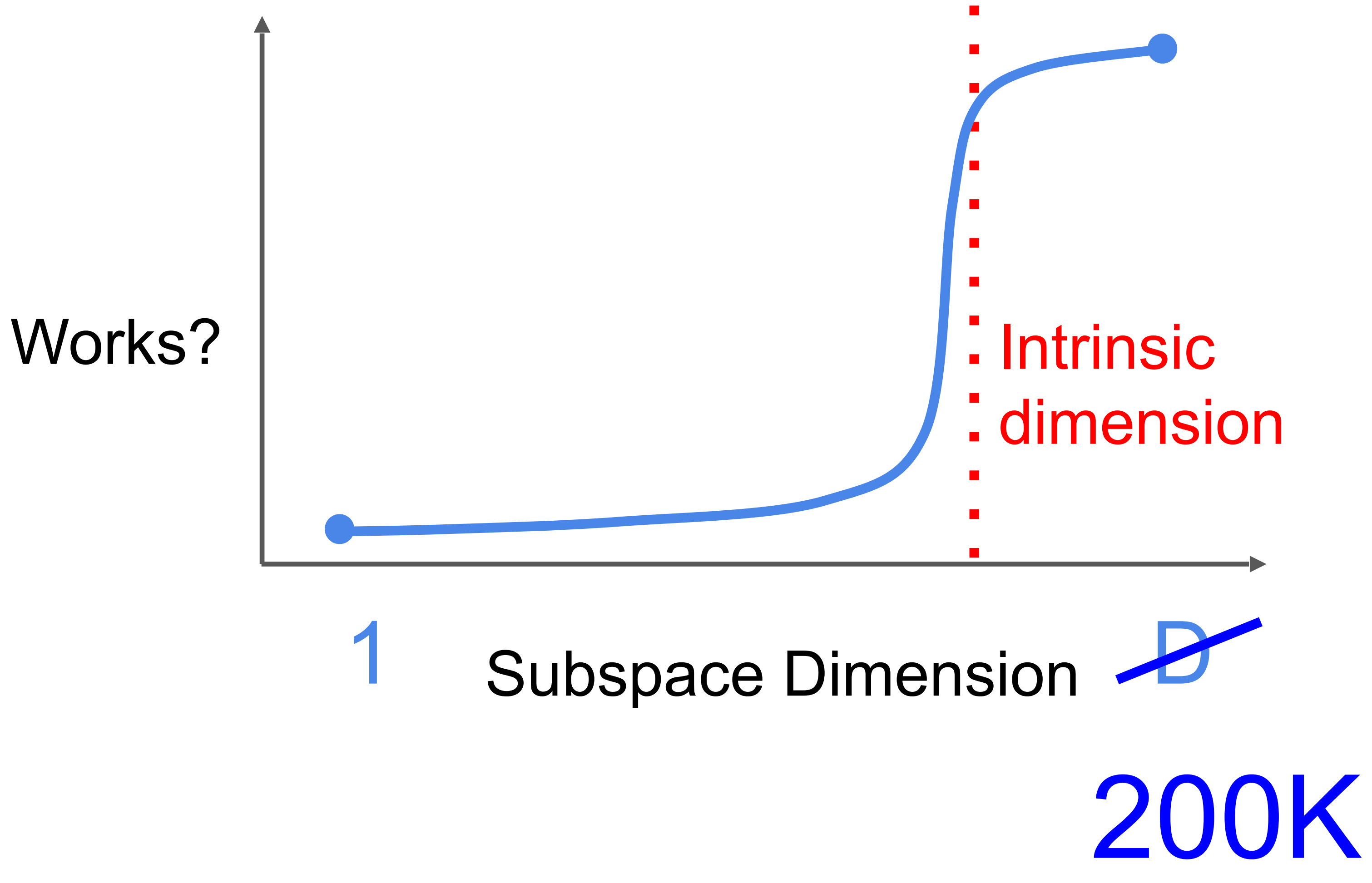
{MNIST, FC}

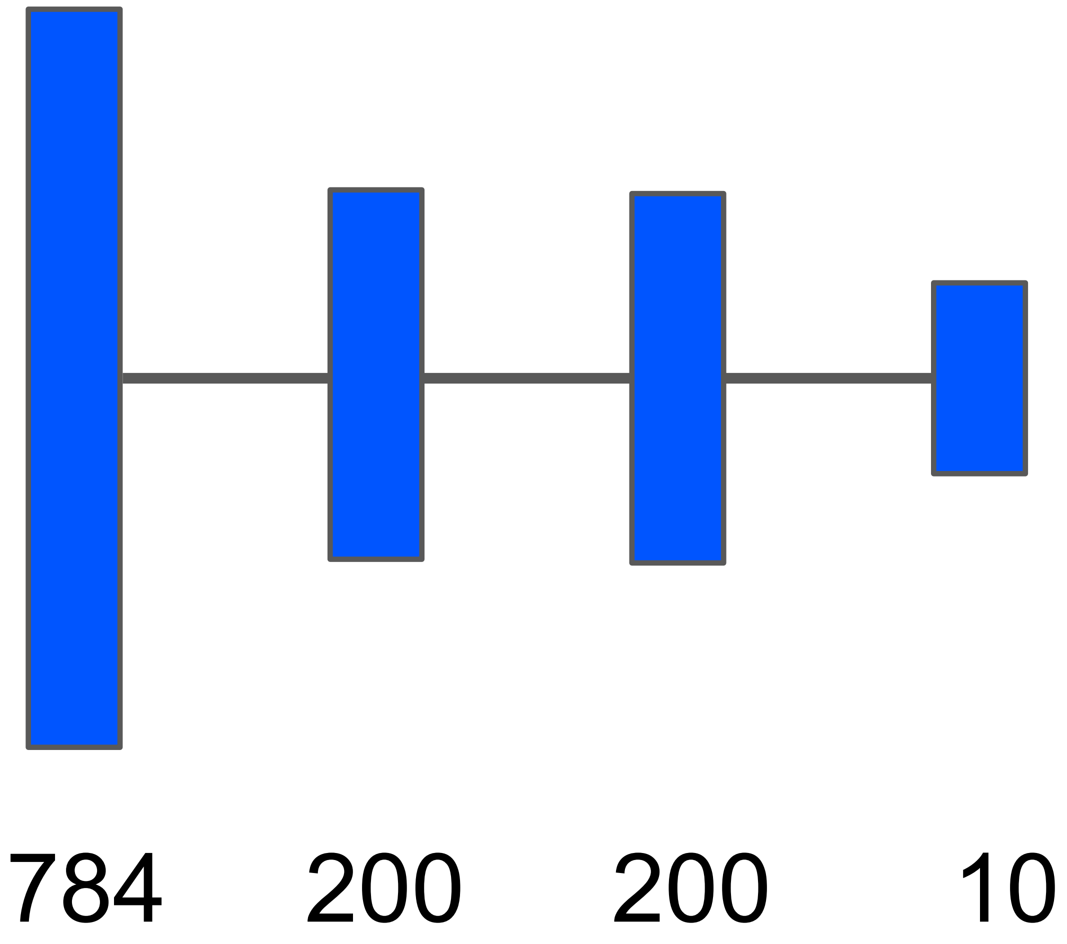




Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

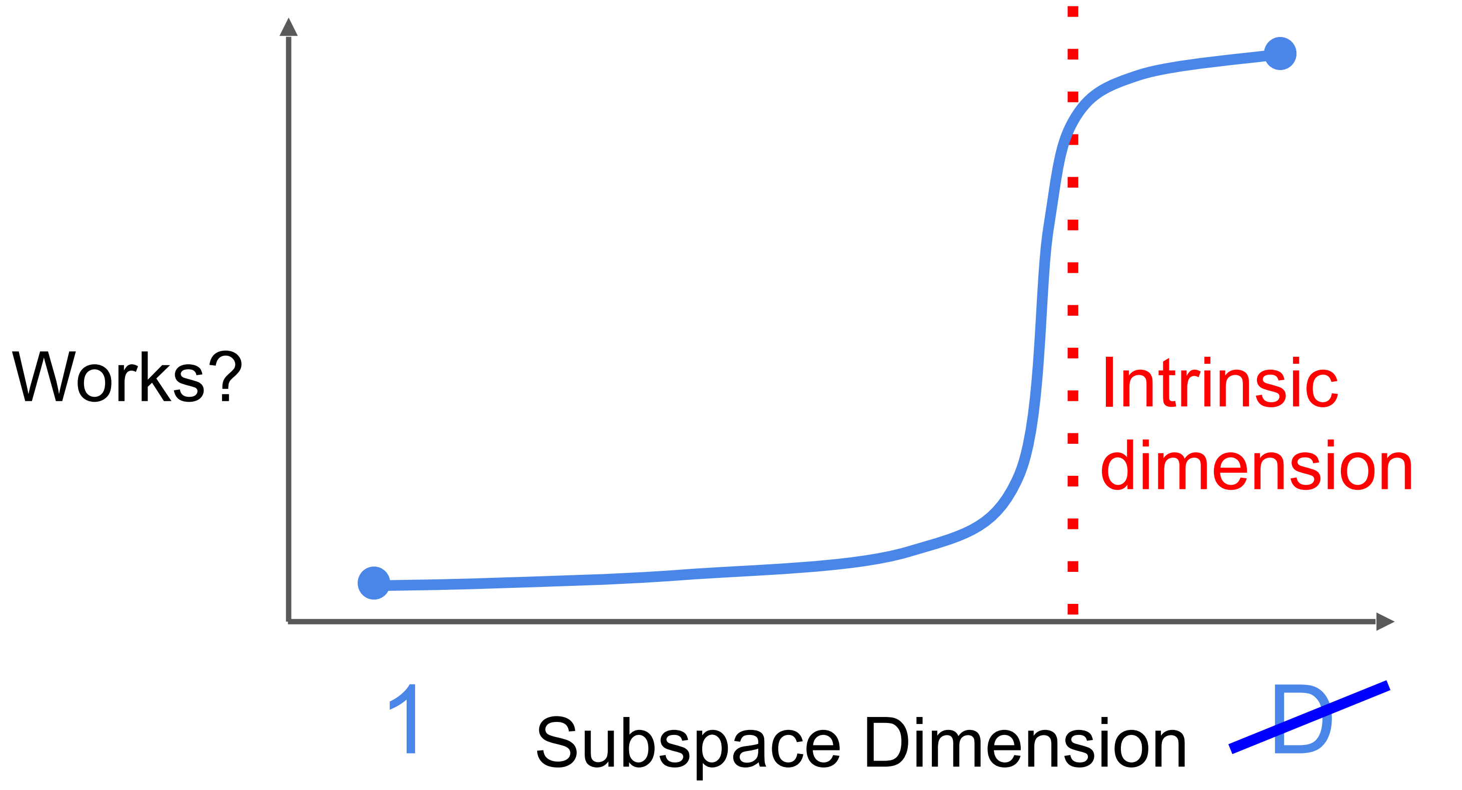
{MNIST, FC}





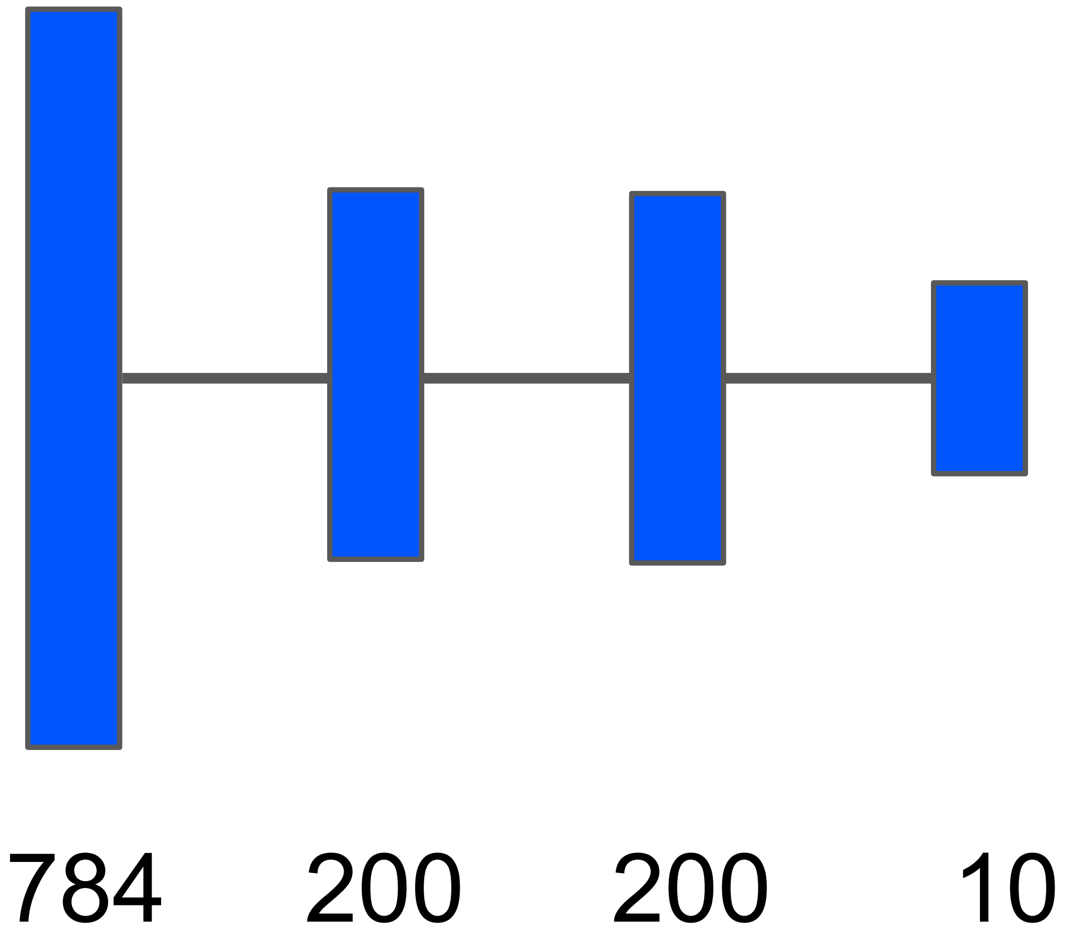
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

{MNIST, FC}



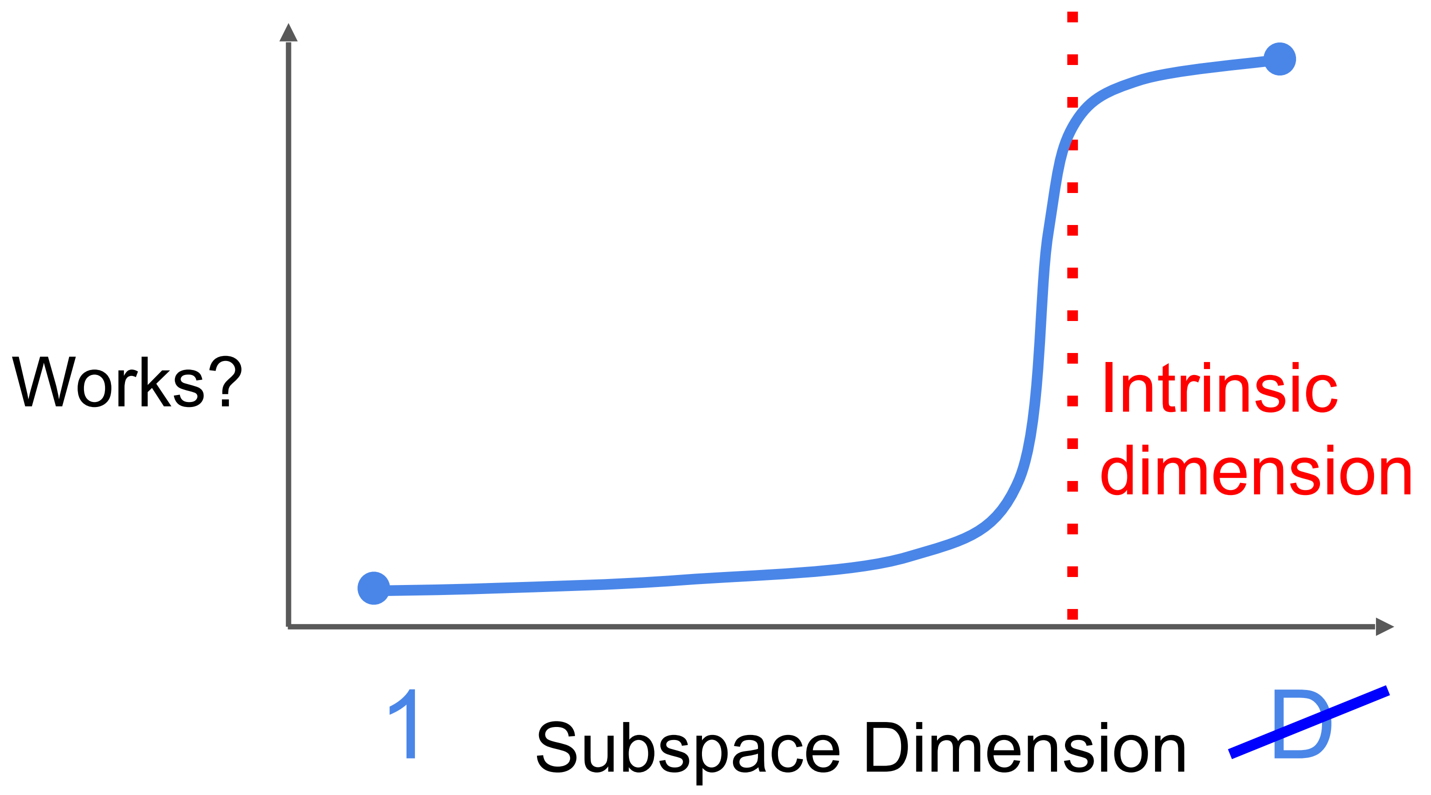
200K

750



Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

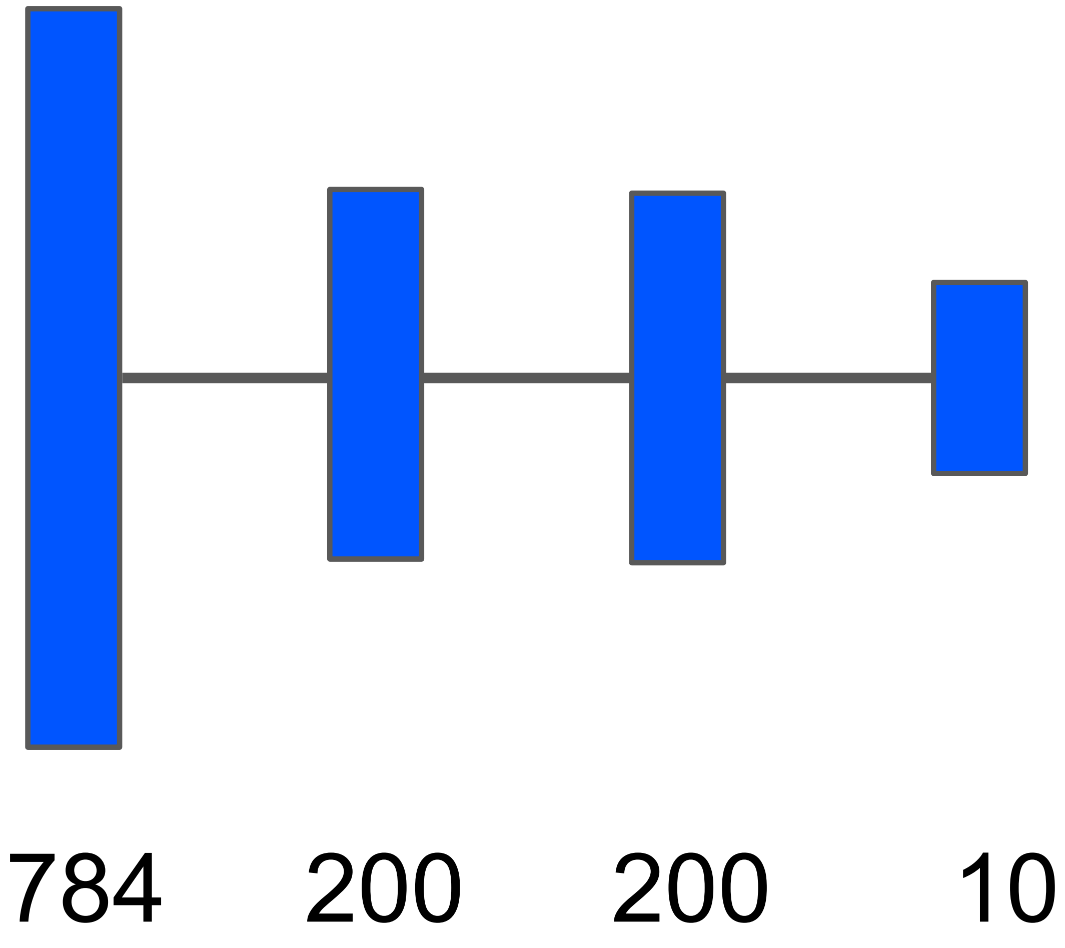
{MNIST, FC}



200K

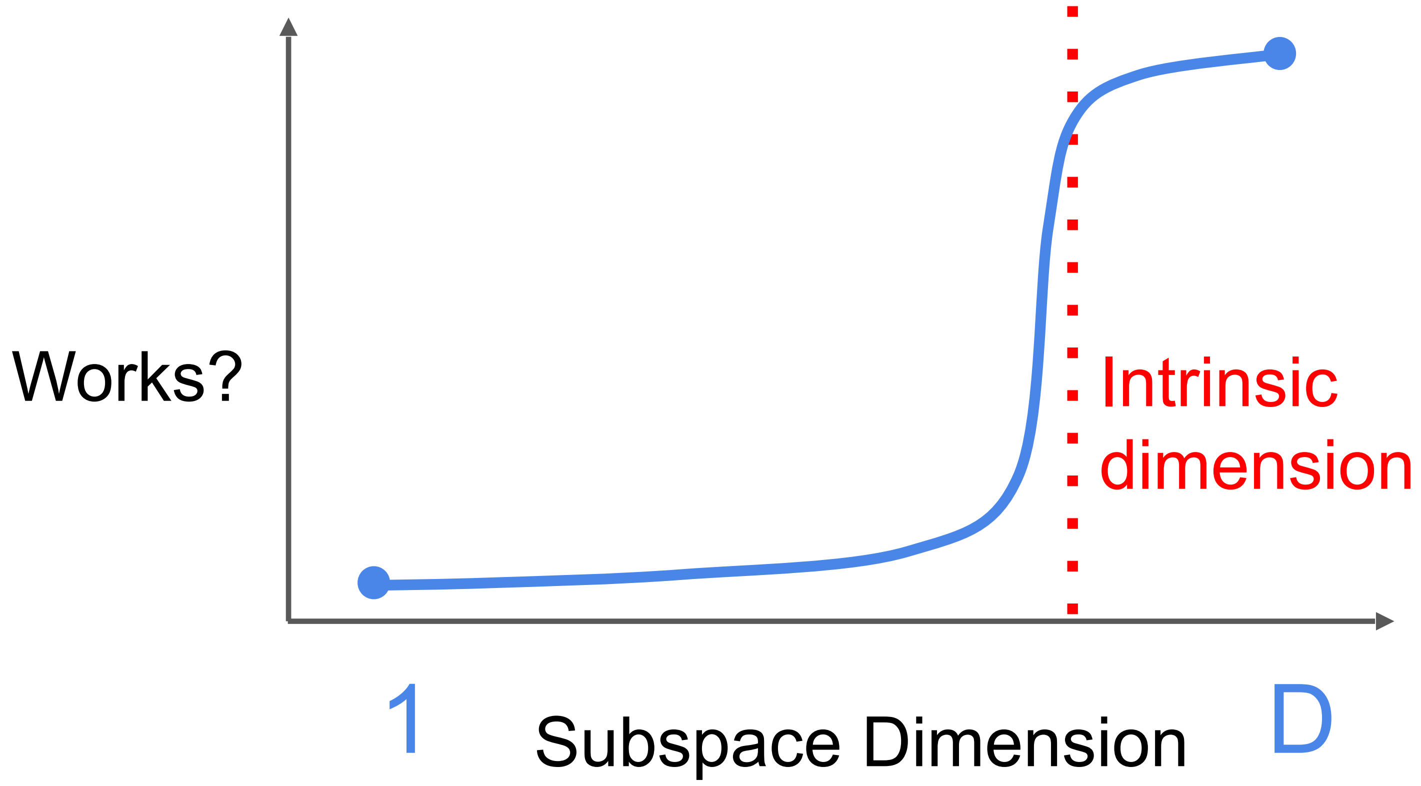
750

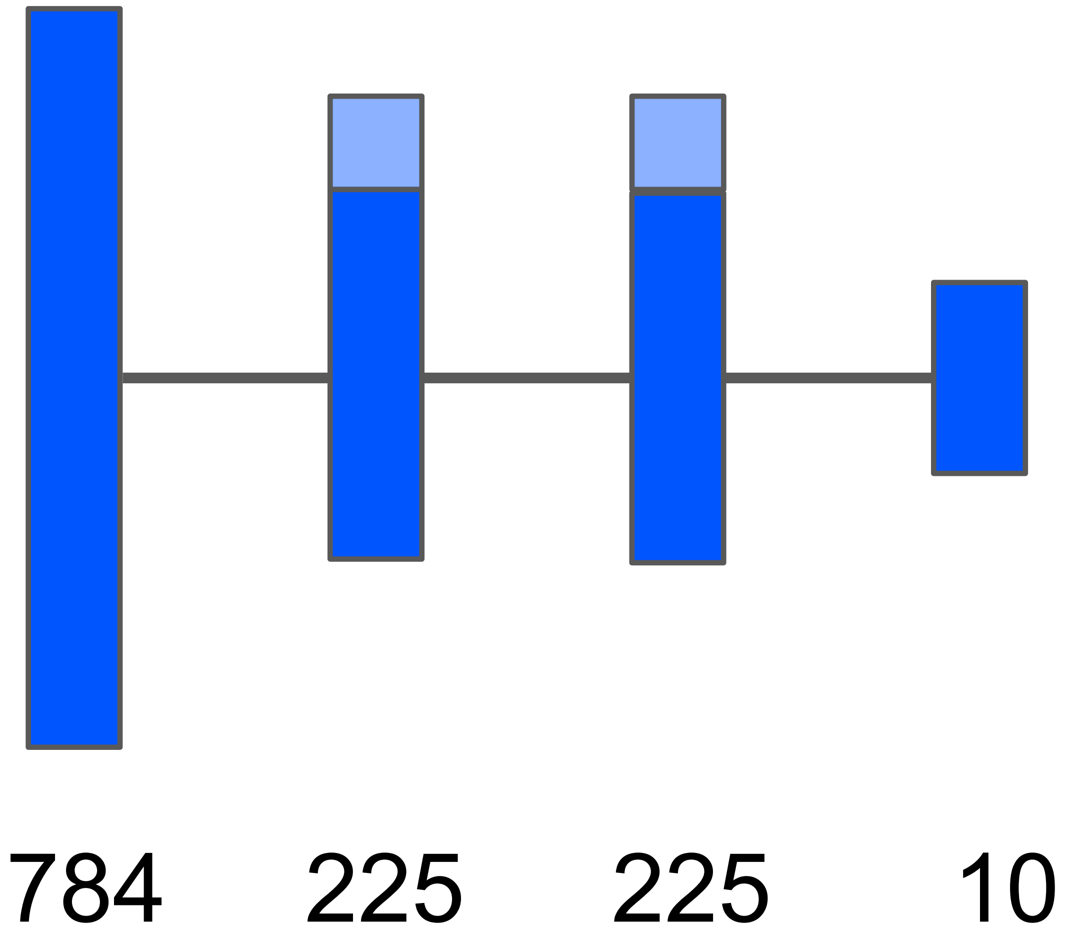
0.4%



Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

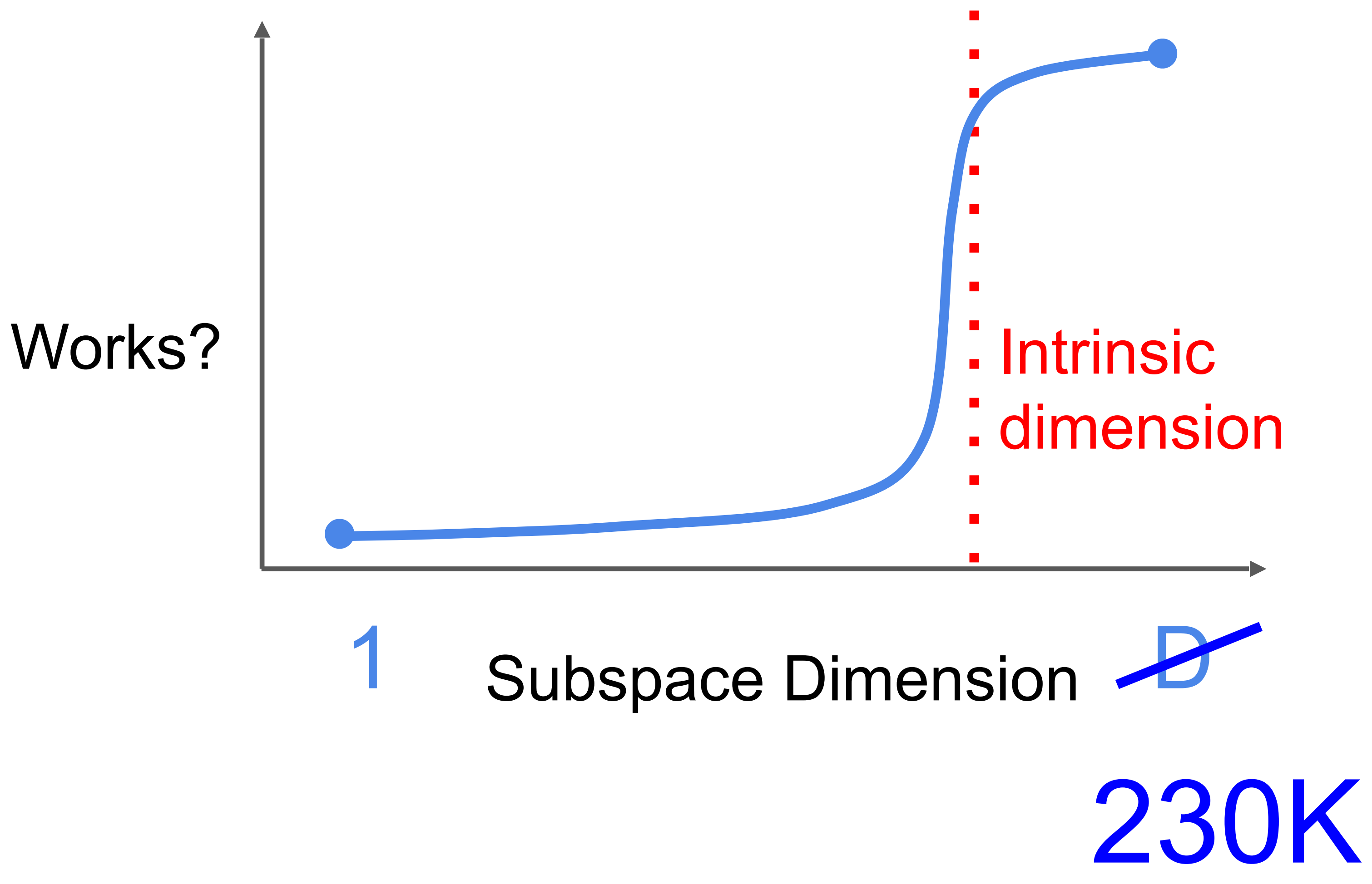
{MNIST, FC}

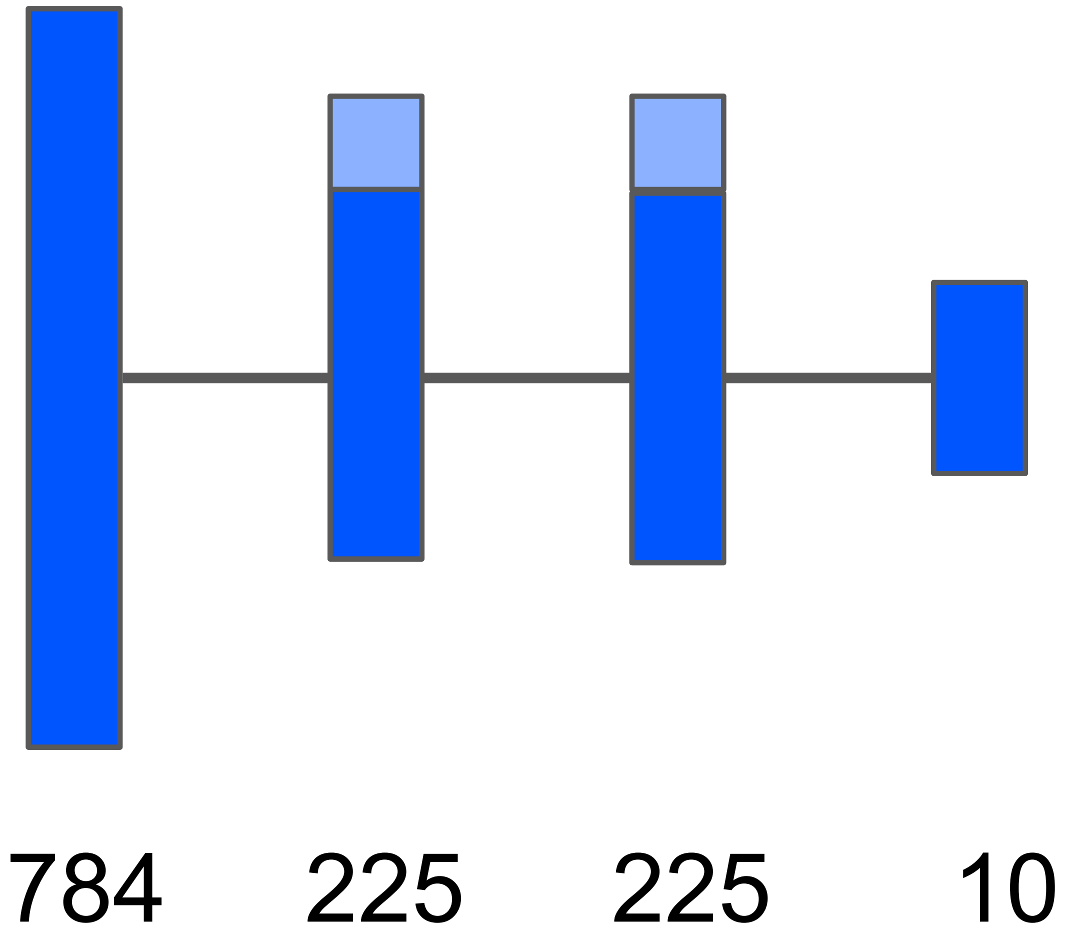




Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

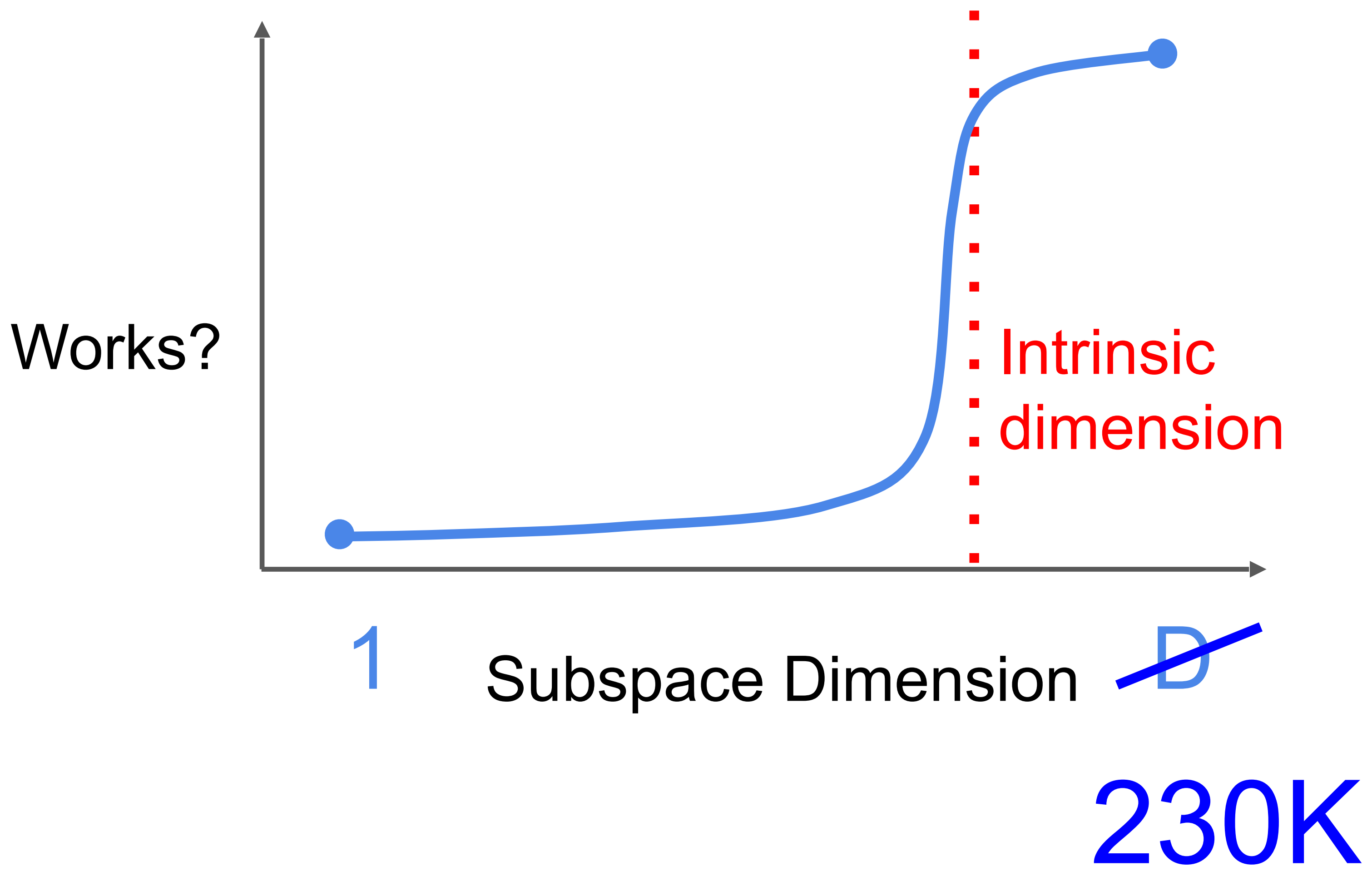
{MNIST, FC}



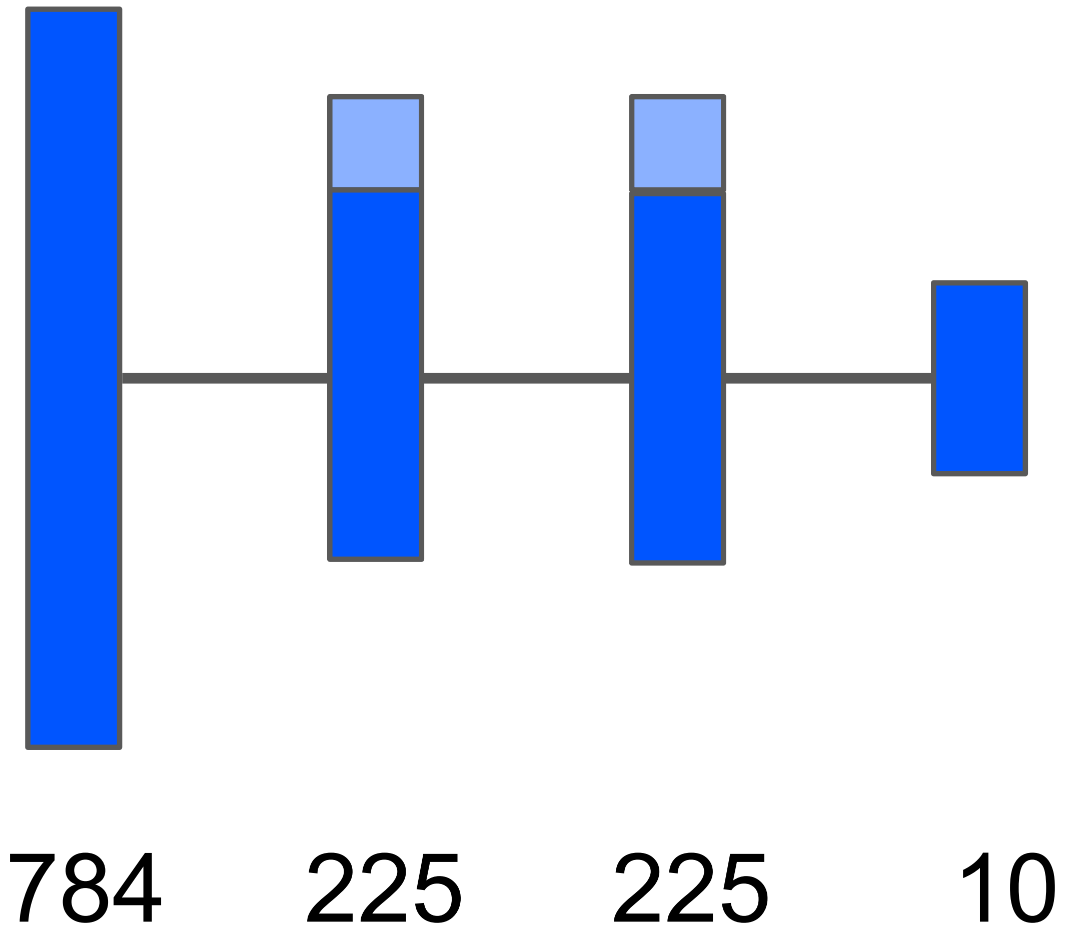


Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

{MNIST, FC}

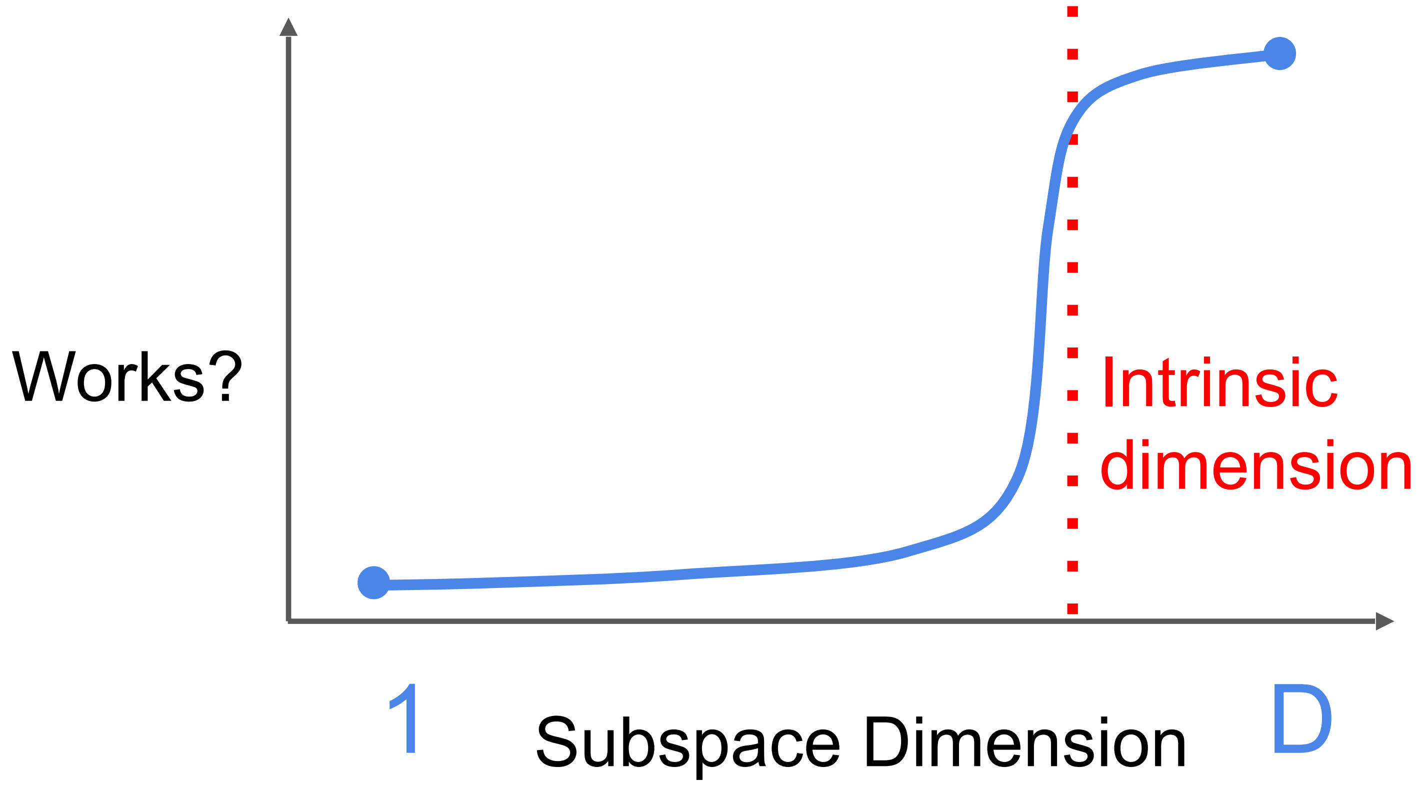


Still 750

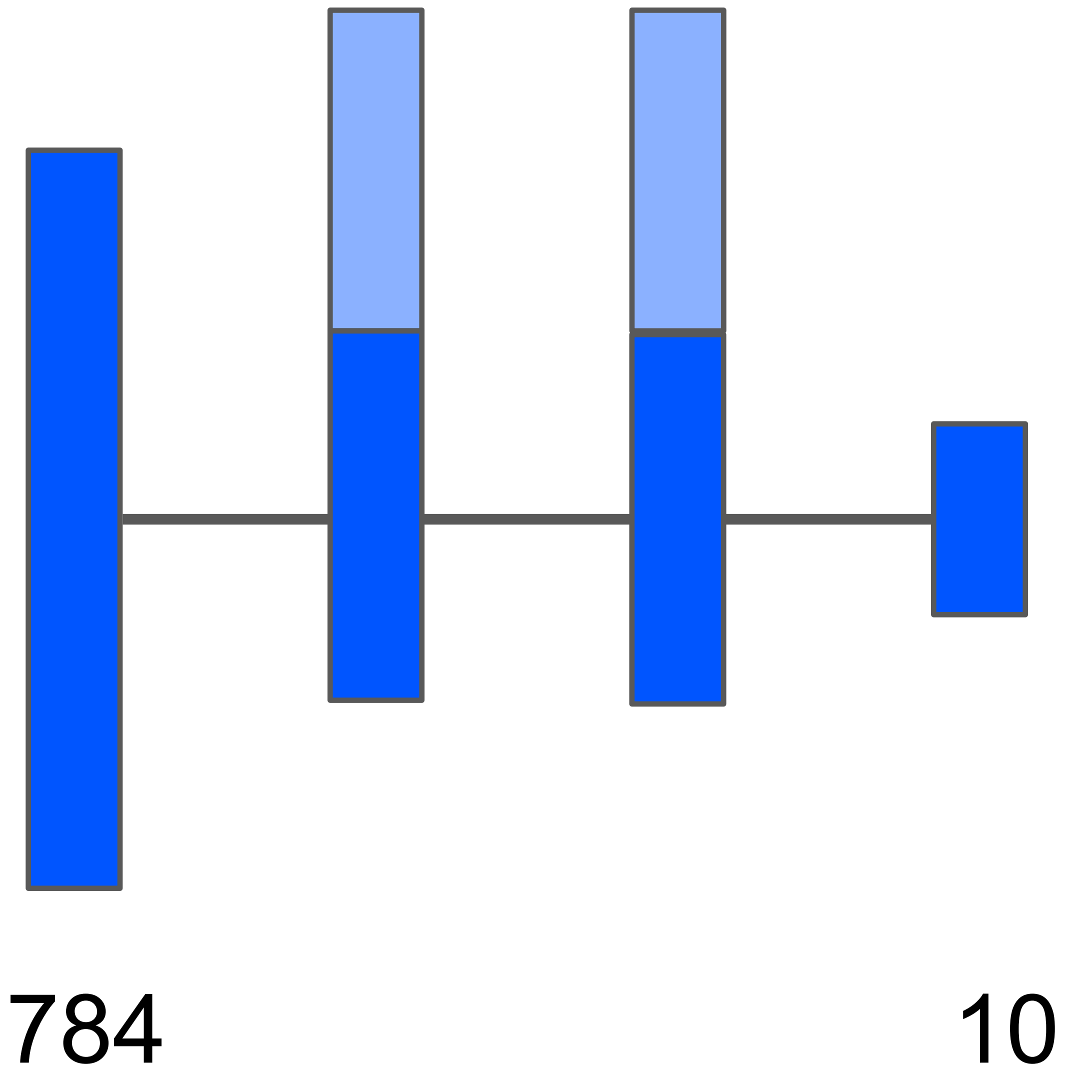


Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

{MNIST, FC}

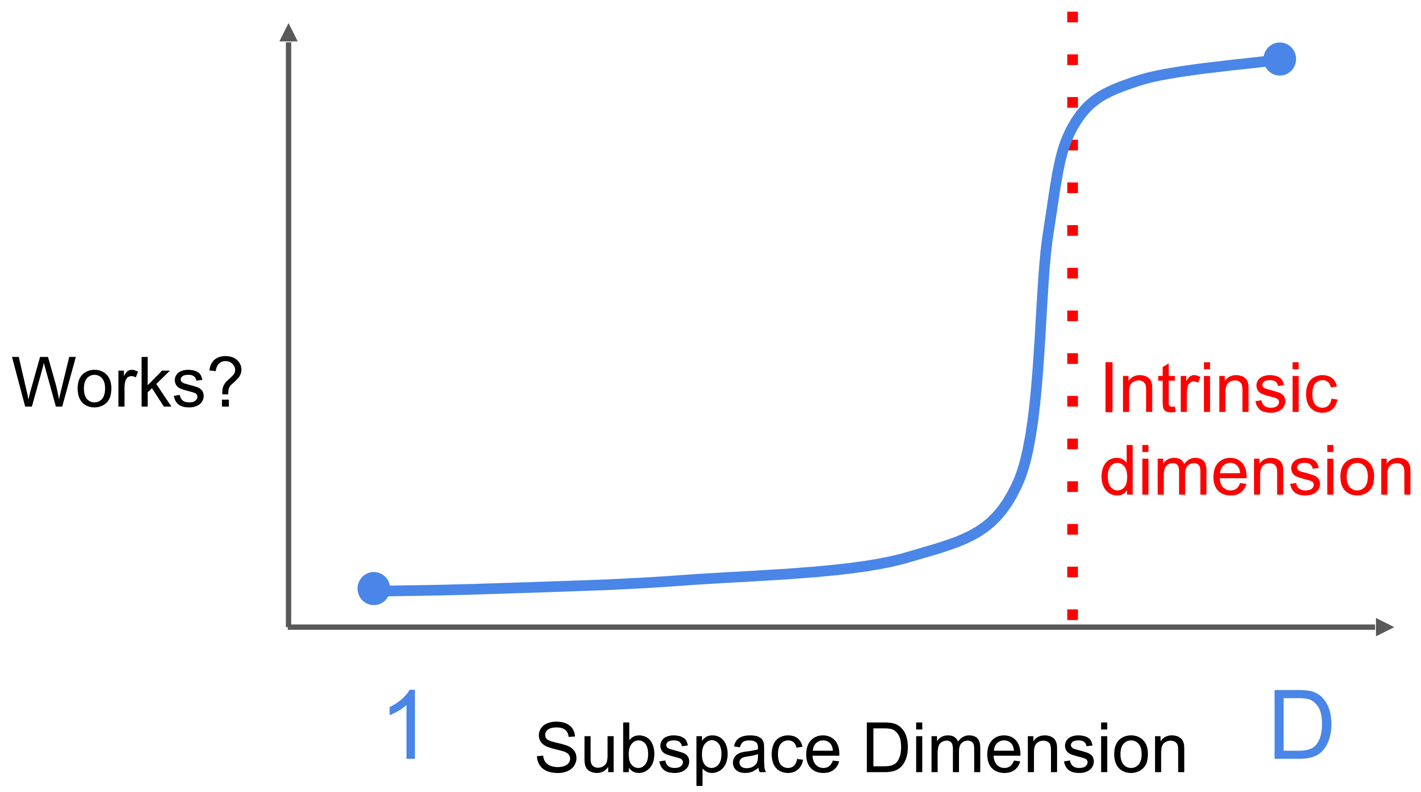


Still 750

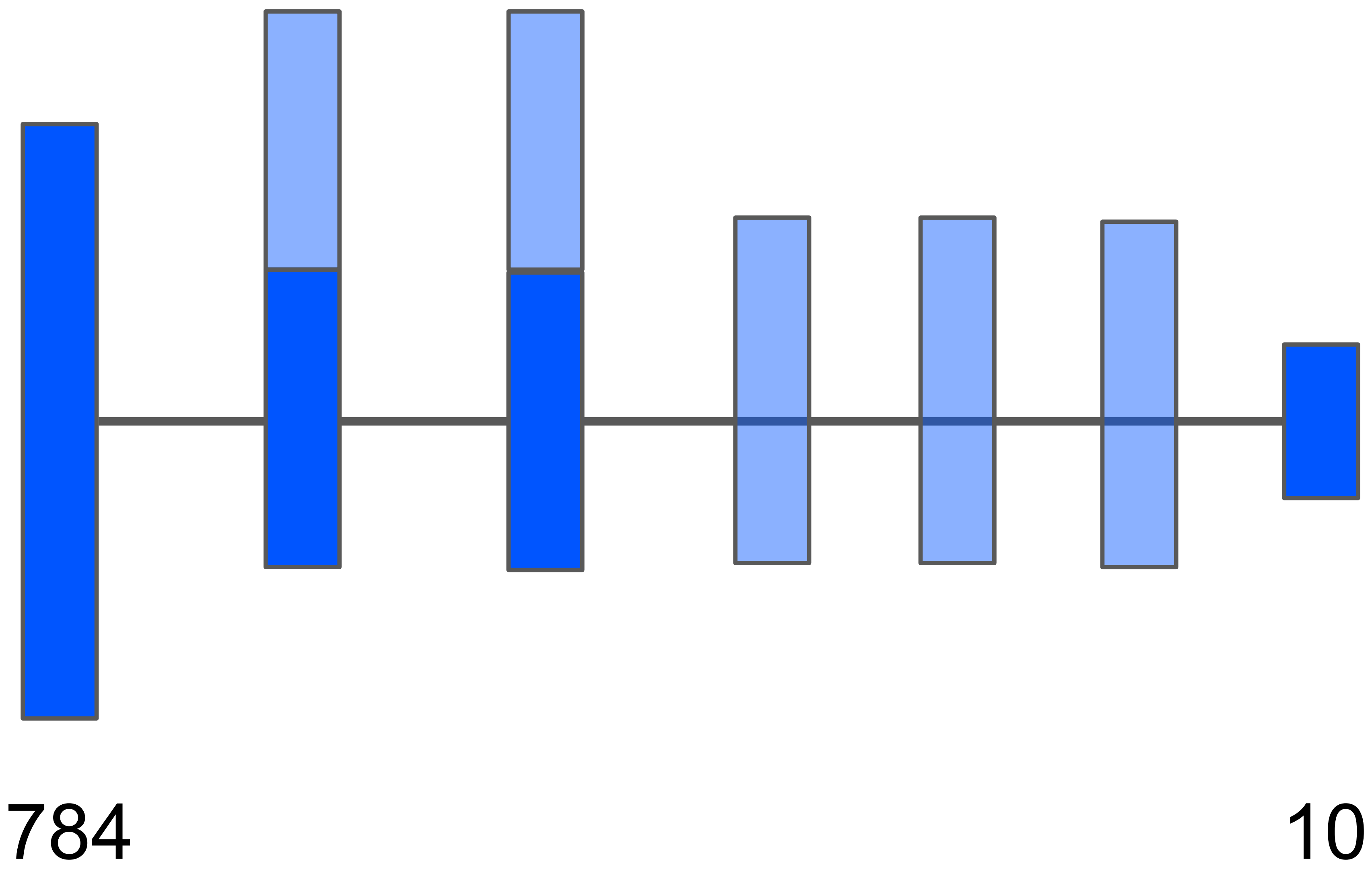


Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

{MNIST, FC}



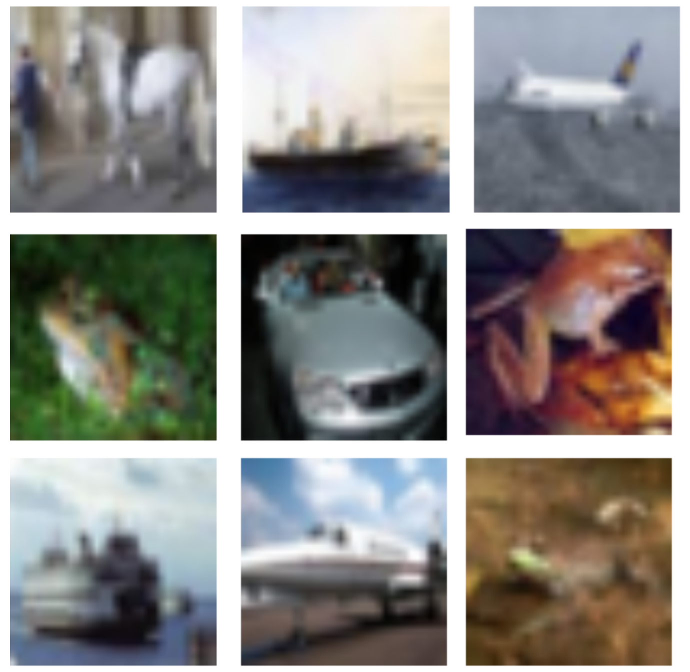
Still 750



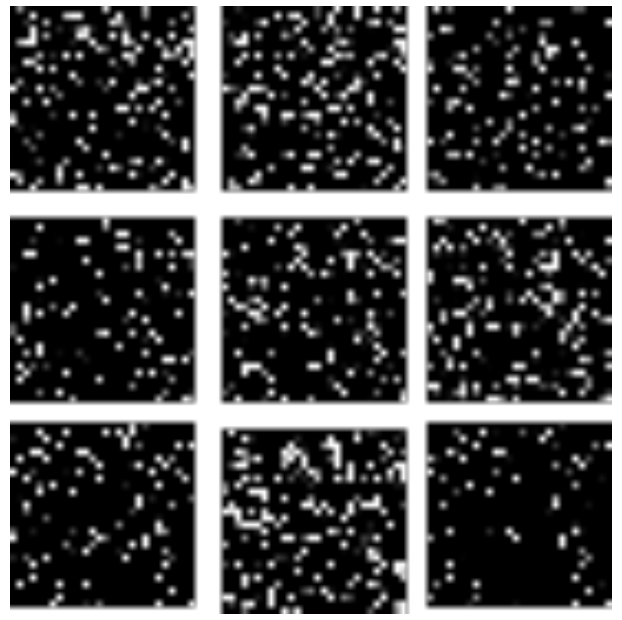
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

MNIST

CIFAR 10



MNIST Shuffled-pixels

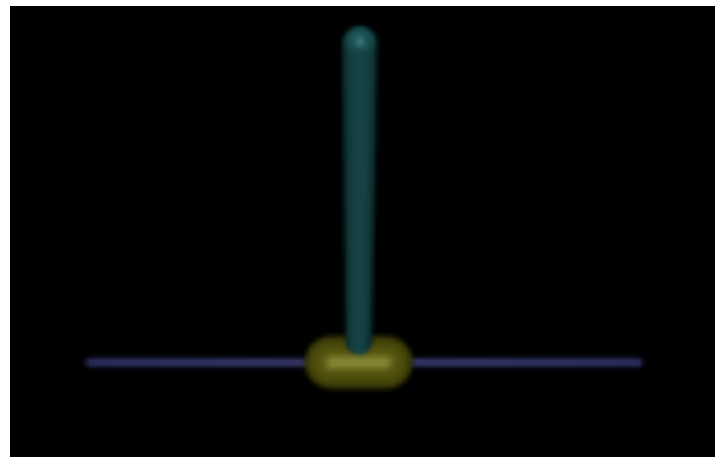


ImageNet

Humanoid

Pong

Inverted Pendulum

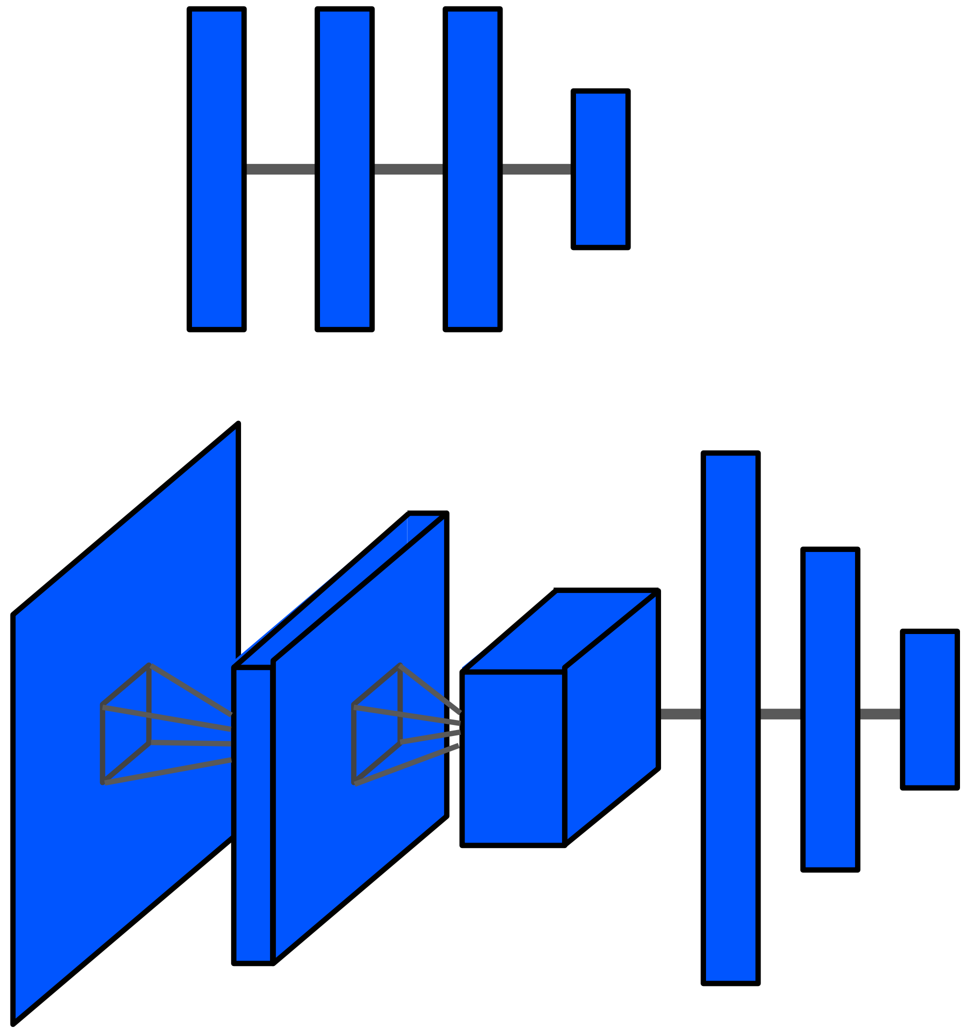


Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

MNIST

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

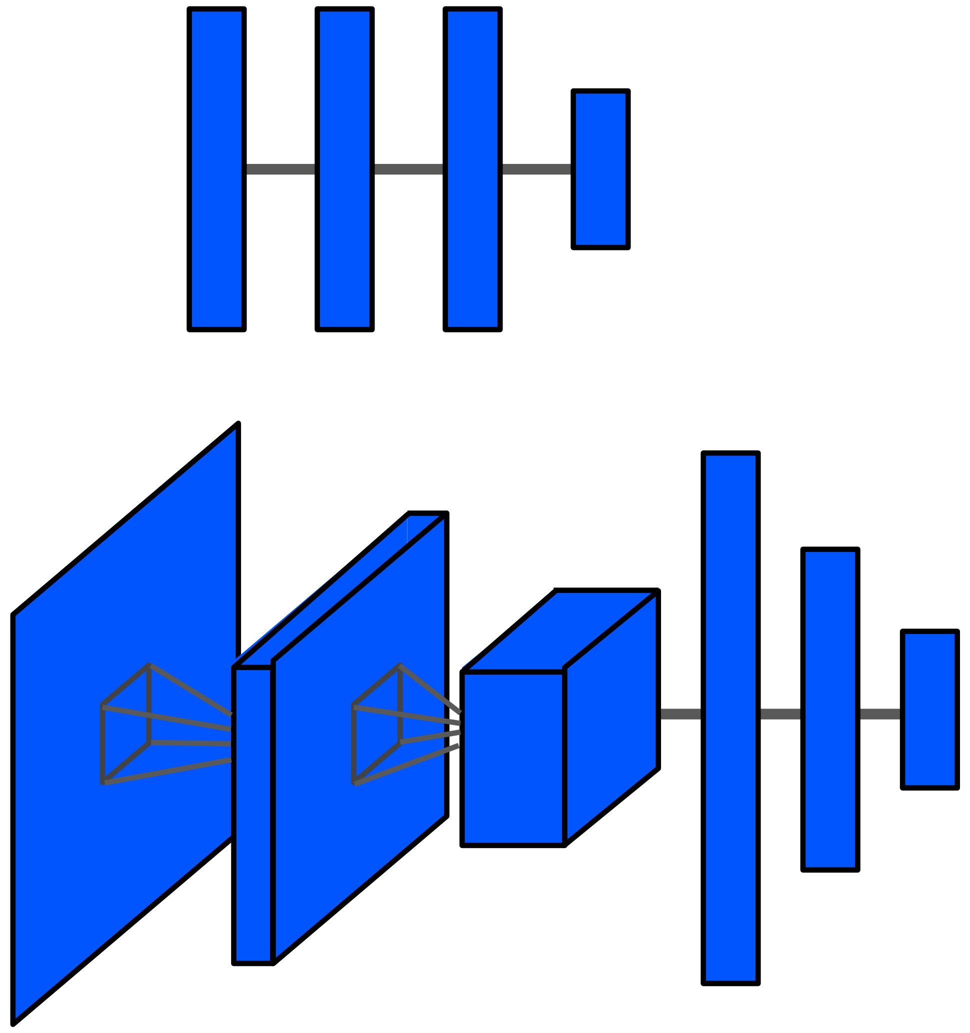
MNIST



Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

MNIST

Int. Dim.



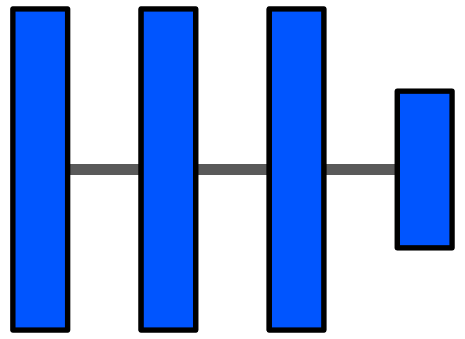
750

290

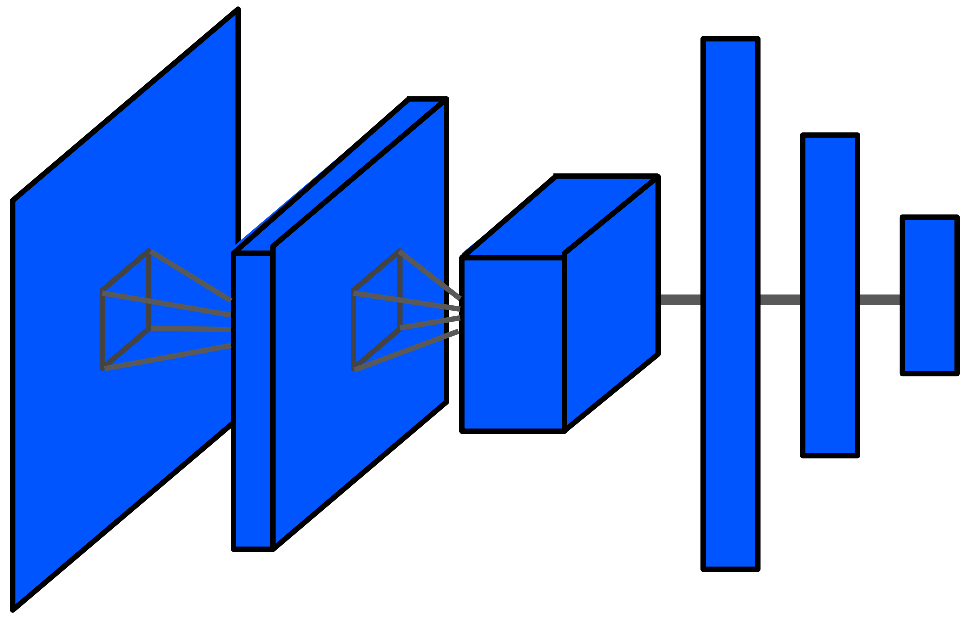
Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

MNIST

Int. Dim.

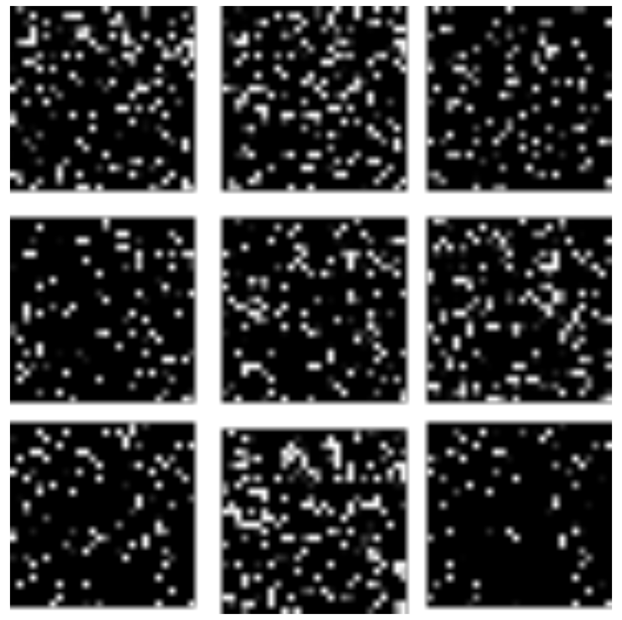


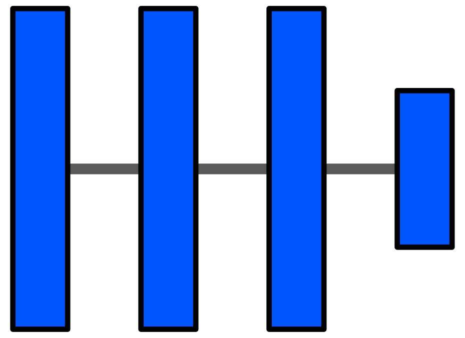
750



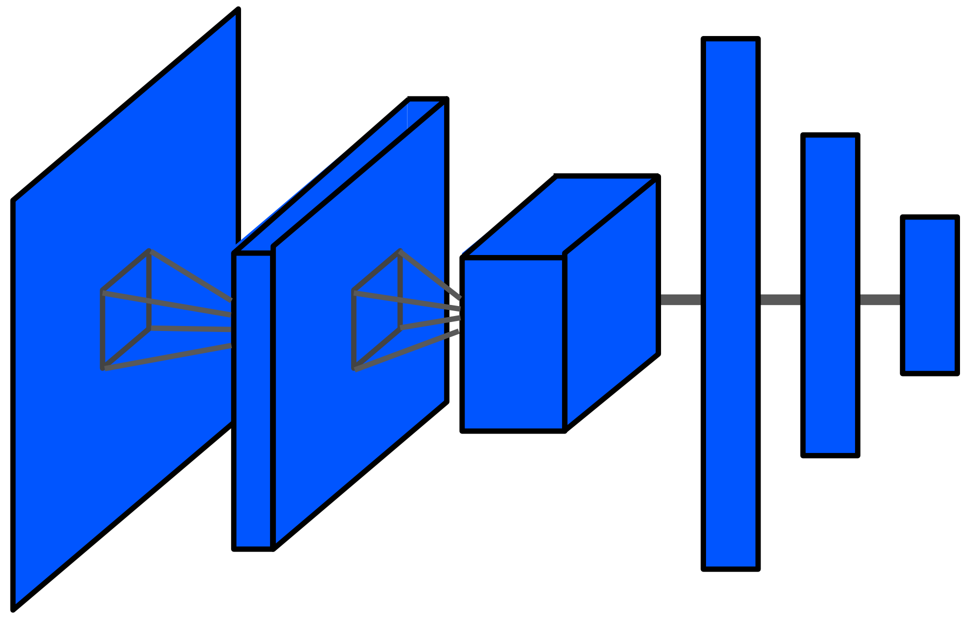
290

MNIST Shuffled-pixels



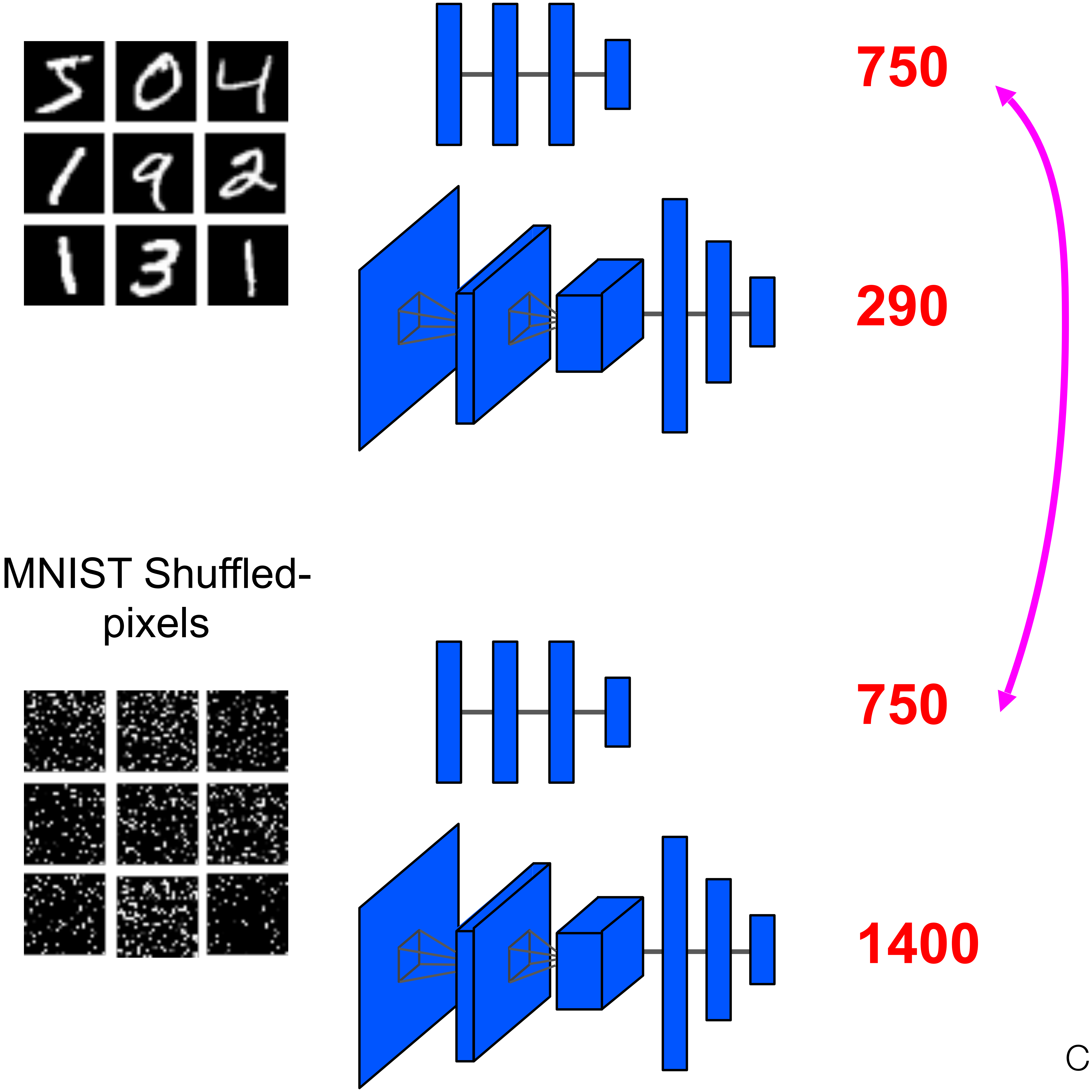


750



1400

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)



MNIST

Int. Dim.

750

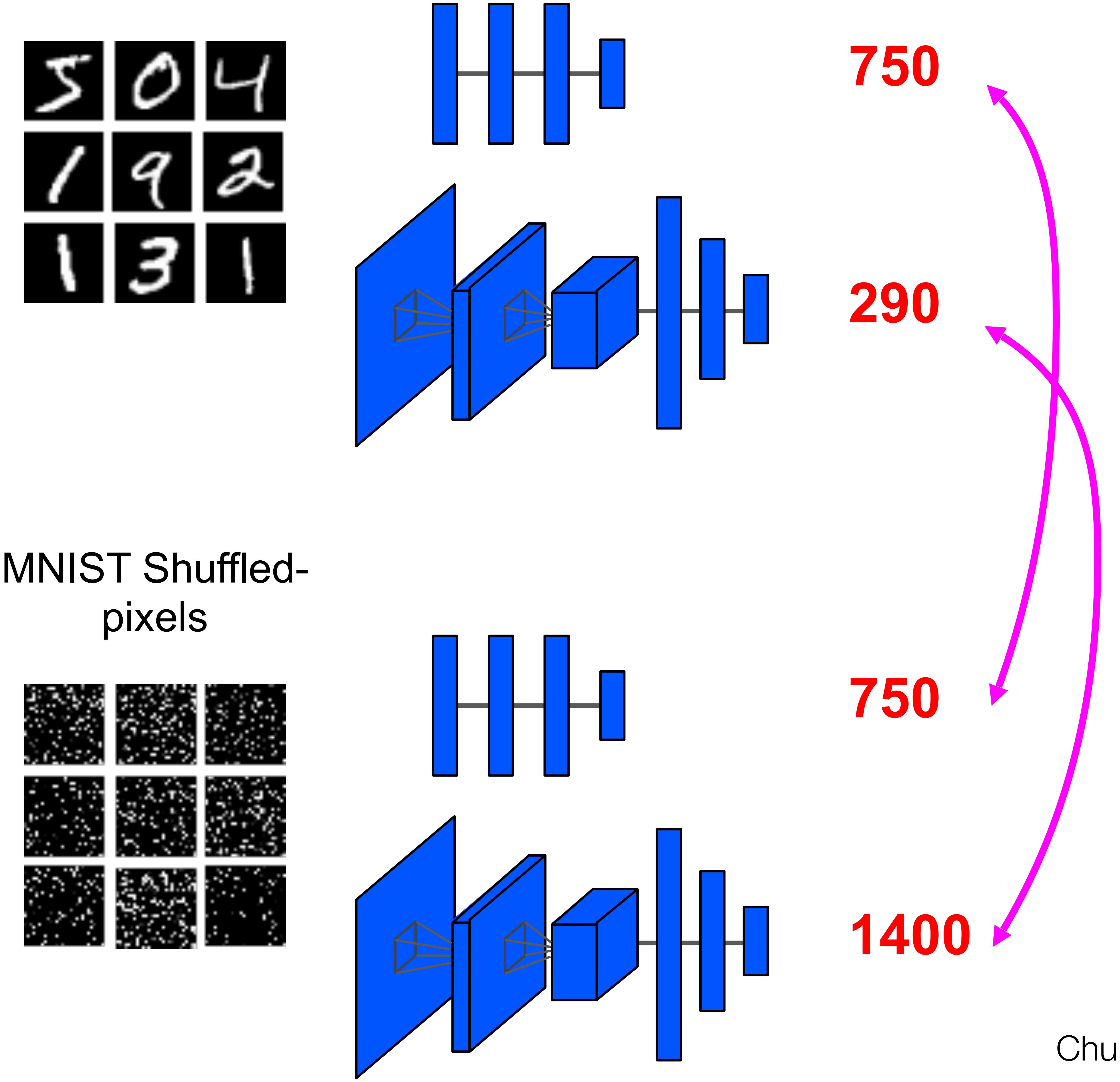
290

MNIST Shuffled-pixels

750

1400

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)



MNIST

Int. Dim.

750

290

MNIST Shuffled-pixels

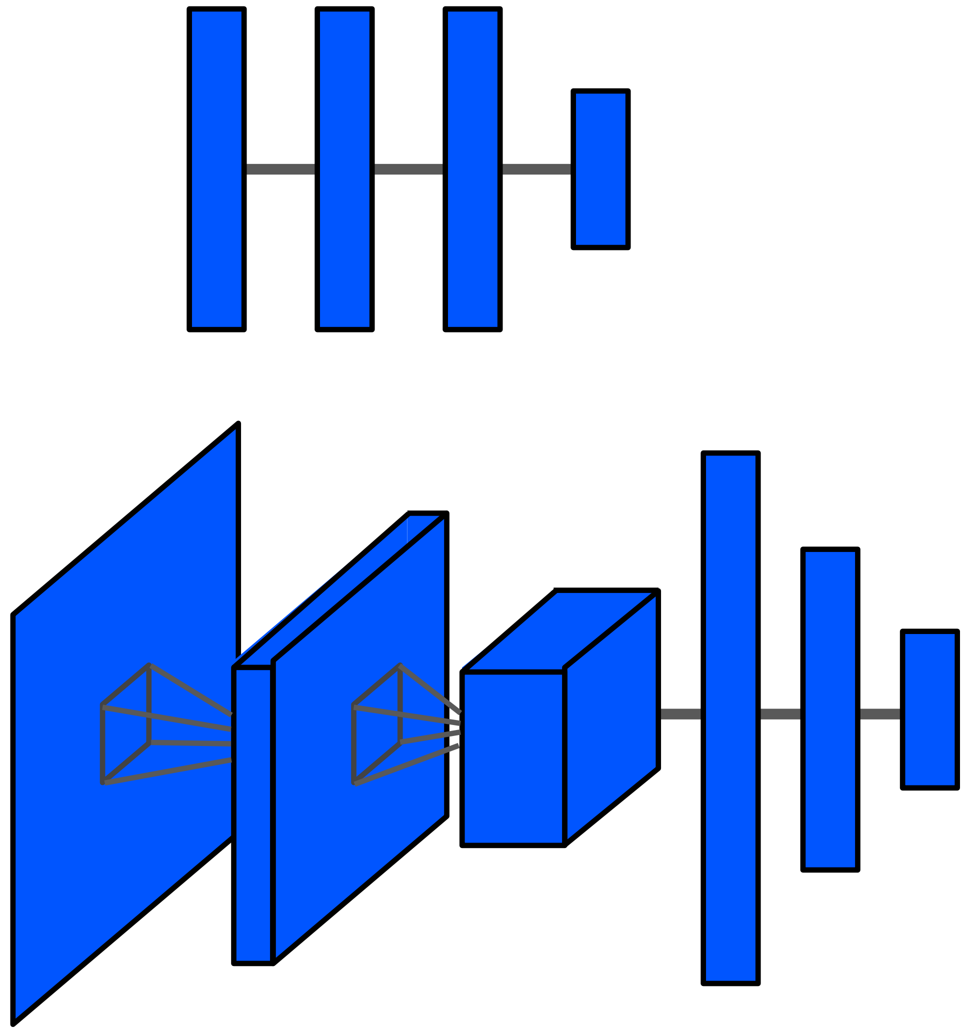
750

1400

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

MNIST

Int. Dim.

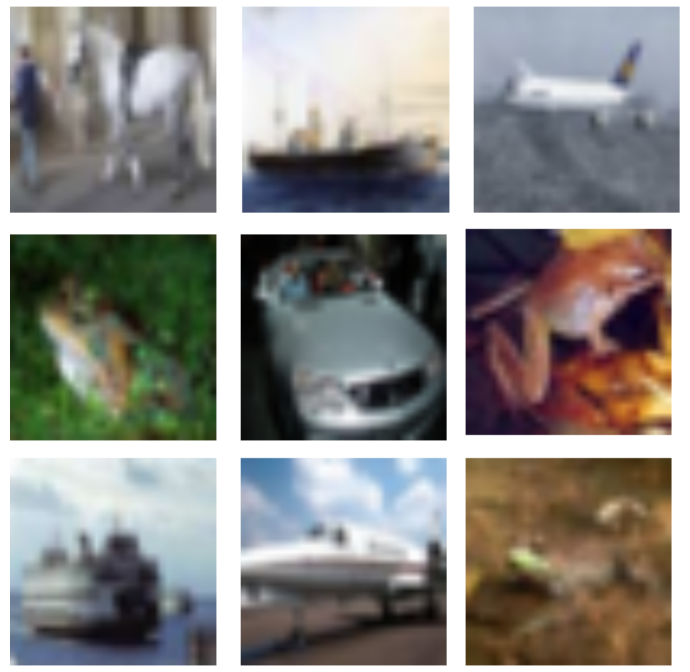


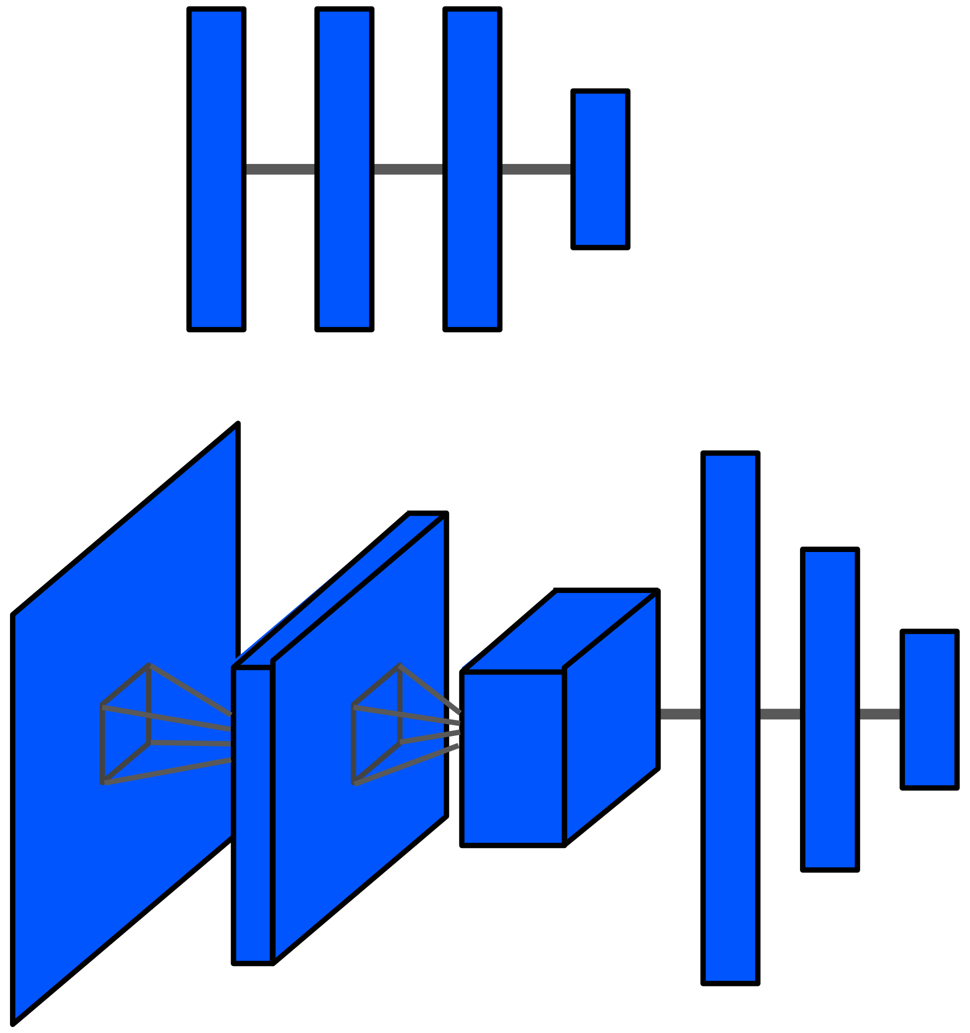
750

290

CIFAR 10

Int. Dim.

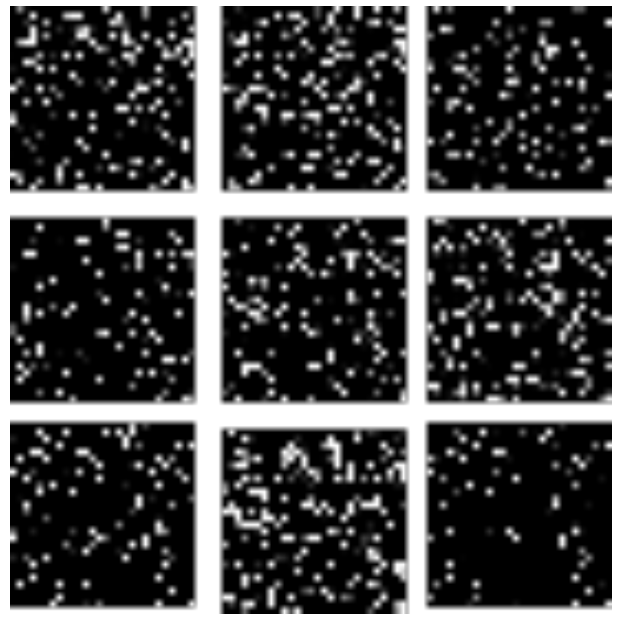


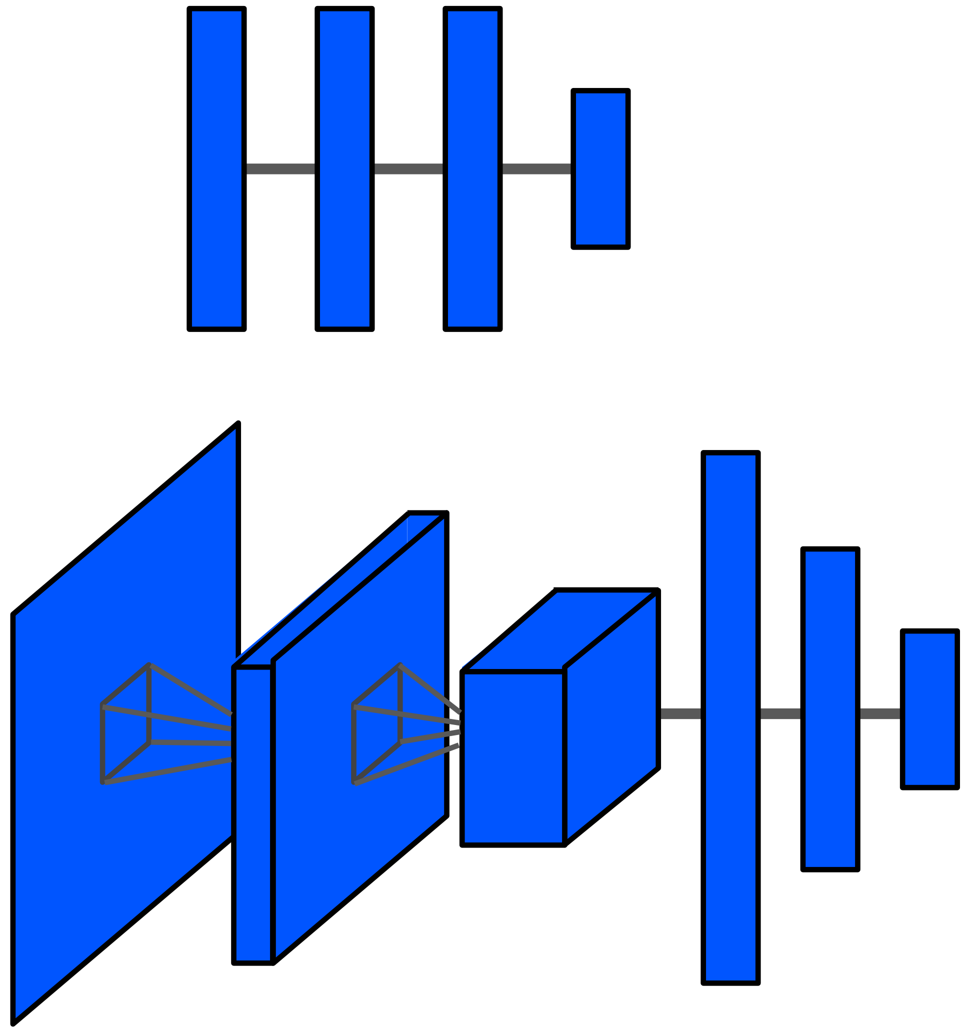


9K

2.9K

MNIST Shuffled-pixels

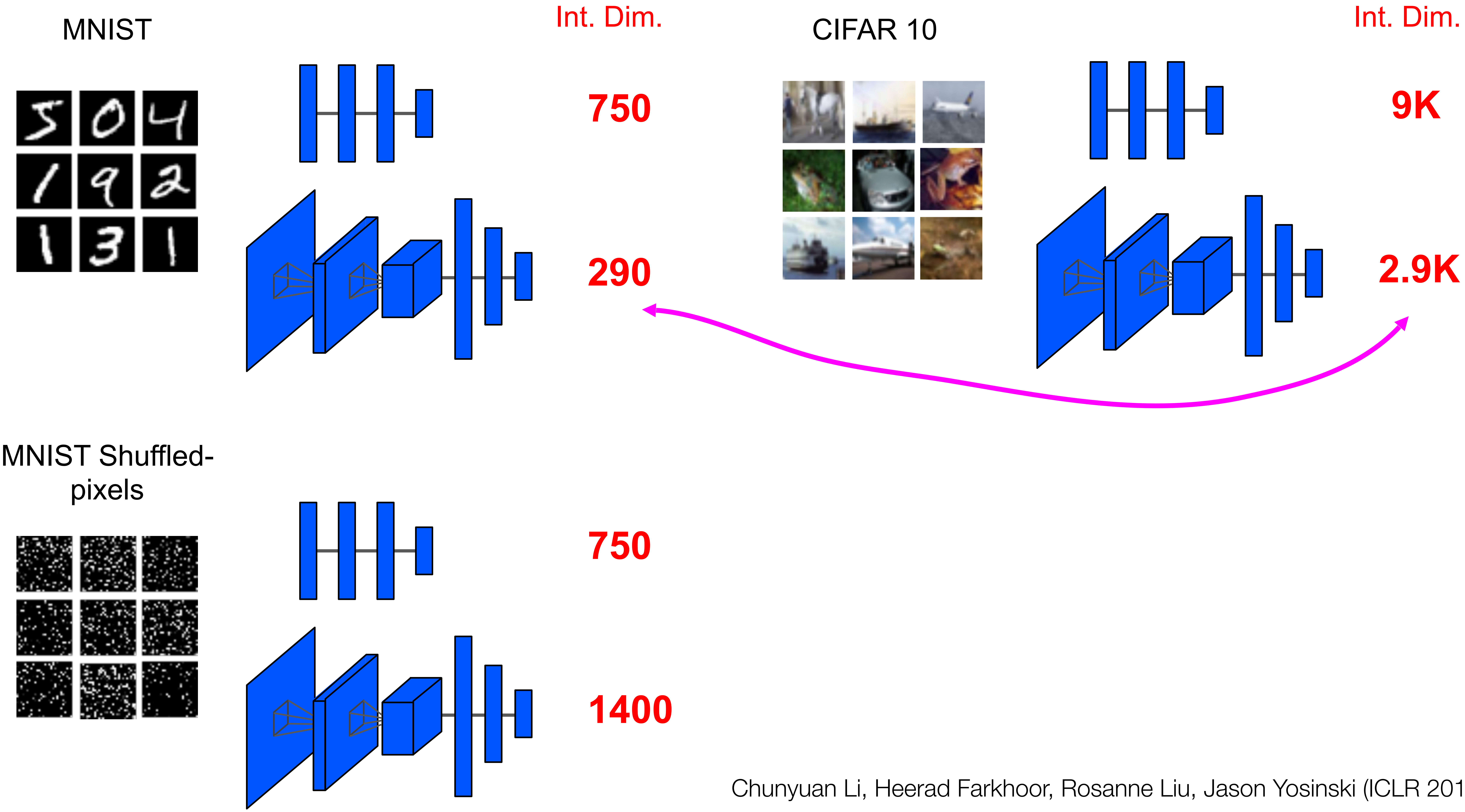




750

1400

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)



MNIST

Int. Dim.

750

290

CIFAR 10

Int. Dim.

9K

2.9K

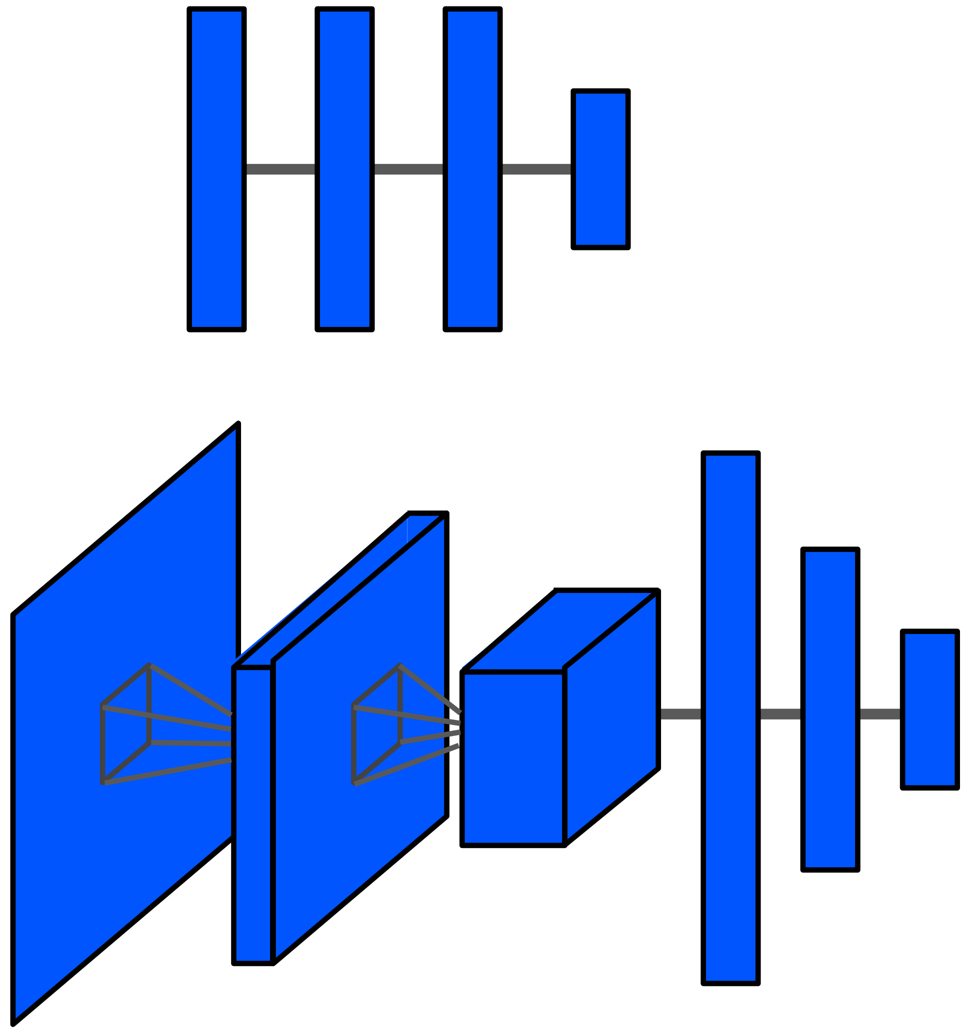
MNIST Shuffled-pixels

750

1400

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, Jason Yosinski (ICLR 2018)

MNIST

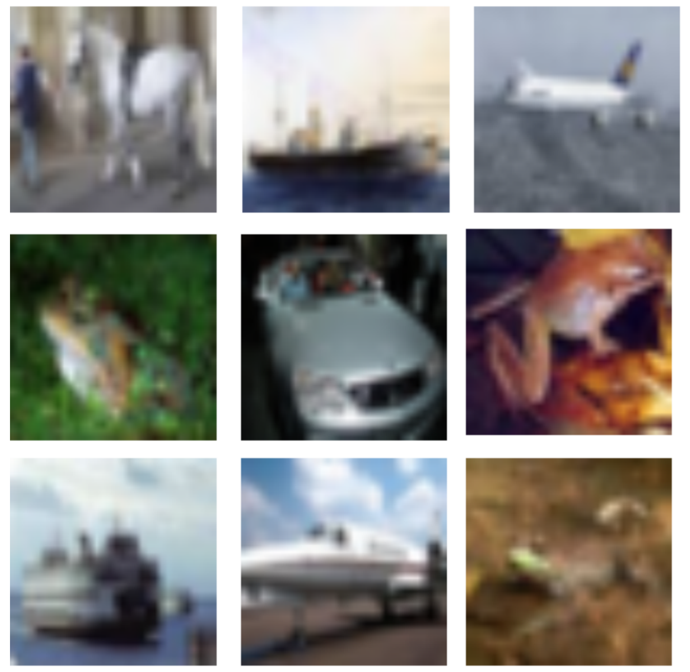


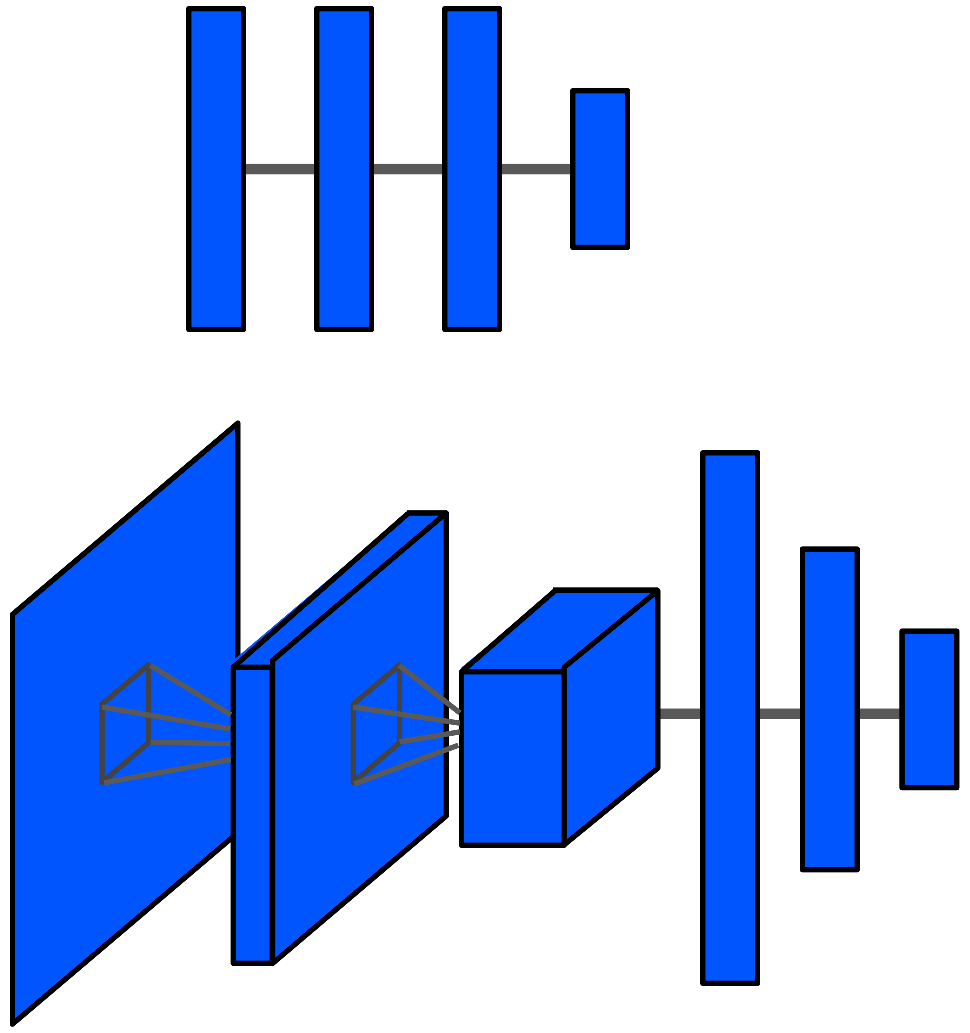
Int. Dim.

750

290

CIFAR 10



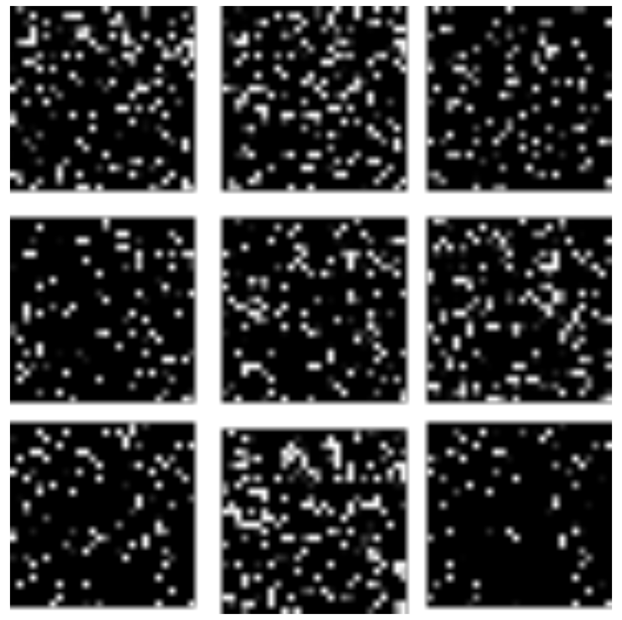


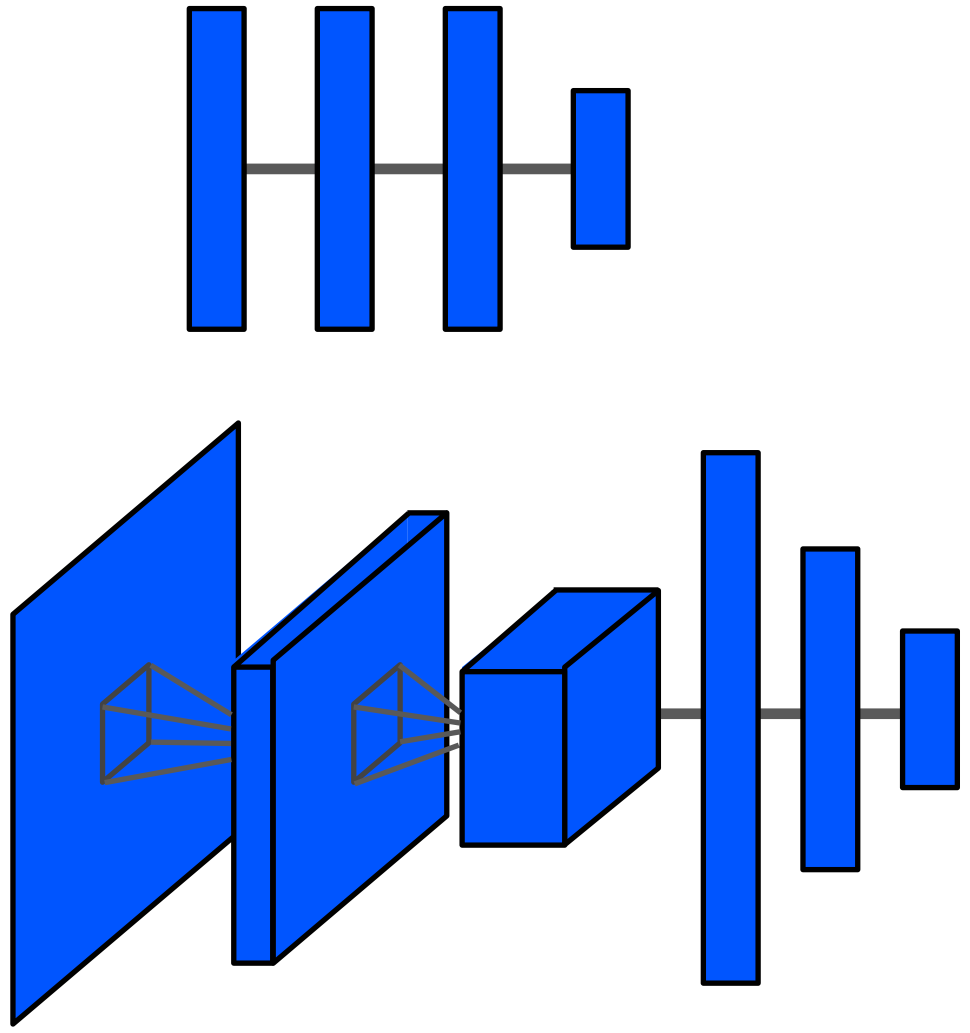
Int. Dim.

9K

2.9K

MNIST Shuffled-pixels





750

1400

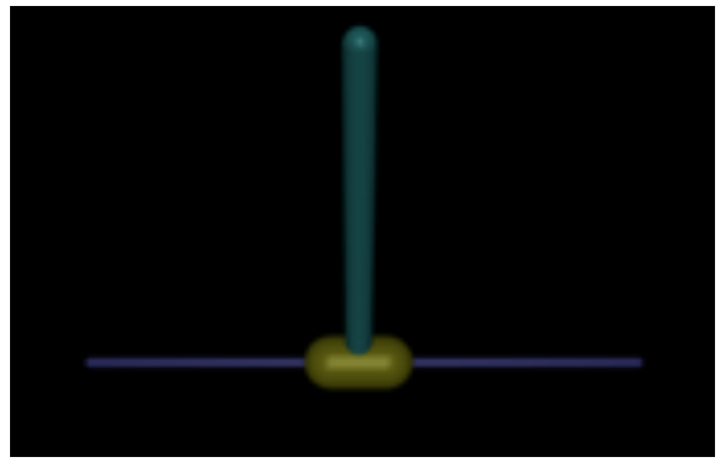
SqueezeNet

>500K

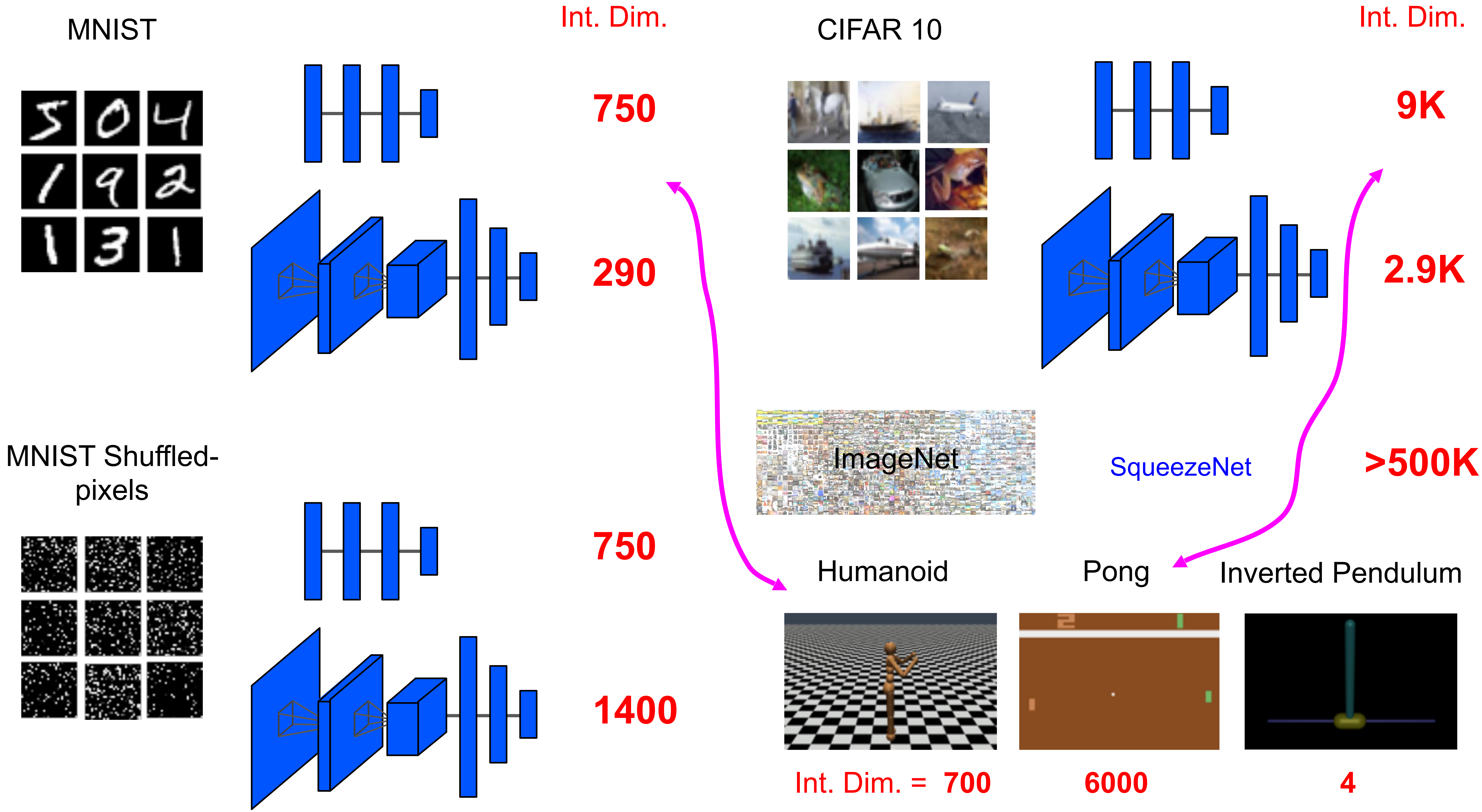
Humanoid

Pong

Inverted Pendulum



Int. Dim. = **700** **6000** **4**



MNIST

Int. Dim.

750

290

MNIST Shuffled-pixels

750

1400

CIFAR 10

Int. Dim.

9K

2.9K

ImageNet

SqueezeNet

>500K

Humanoid

Pong

Inverted Pendulum

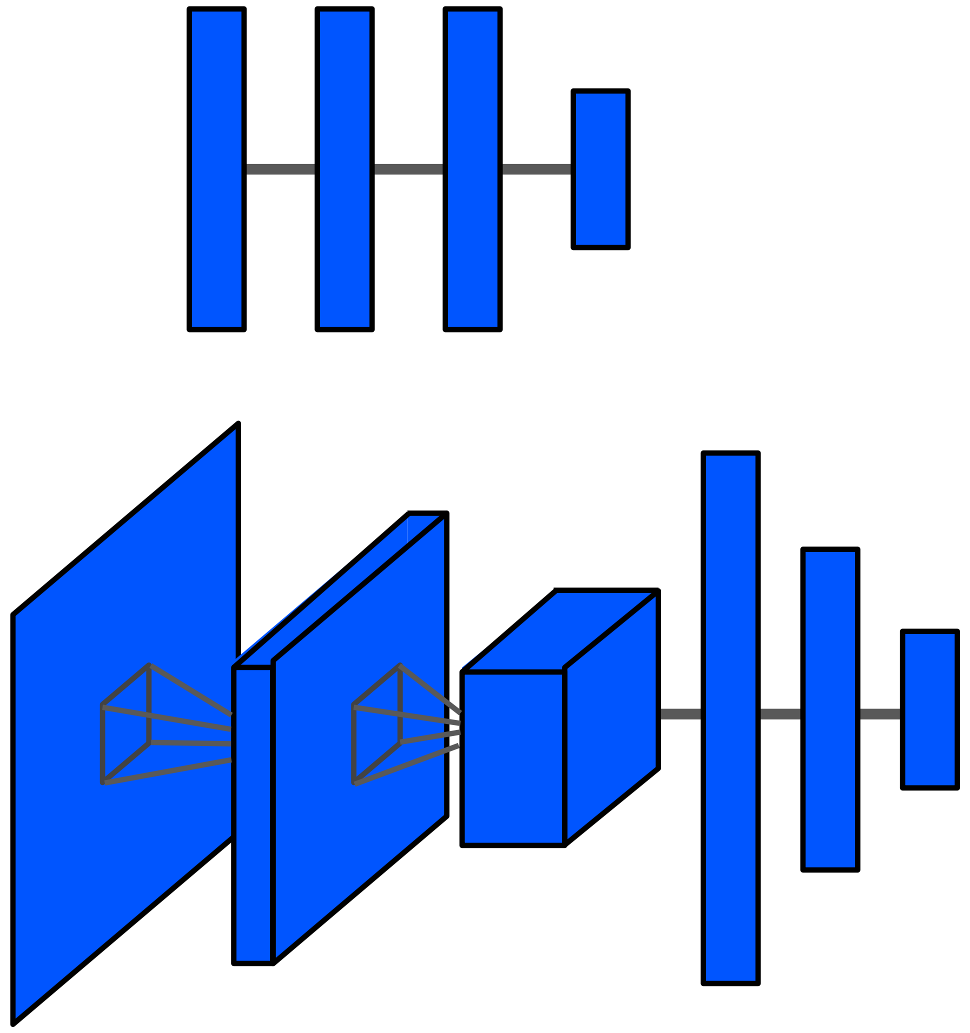
Int. Dim. = **700**

6000

4

MNIST

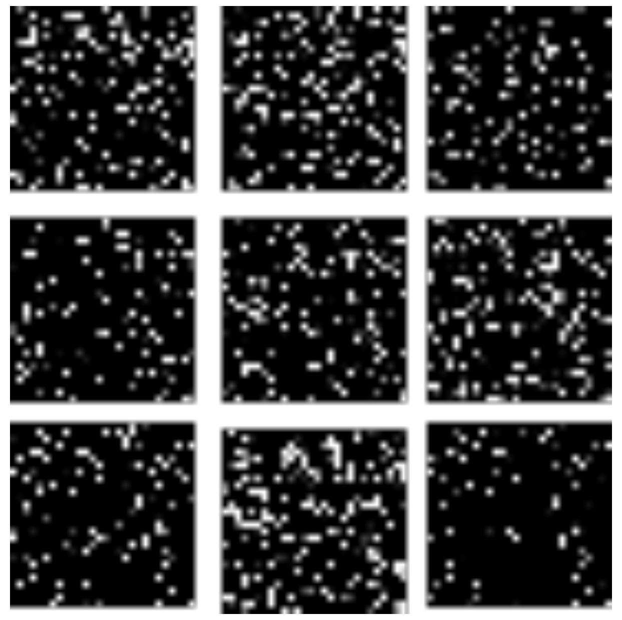
Int. Dim.

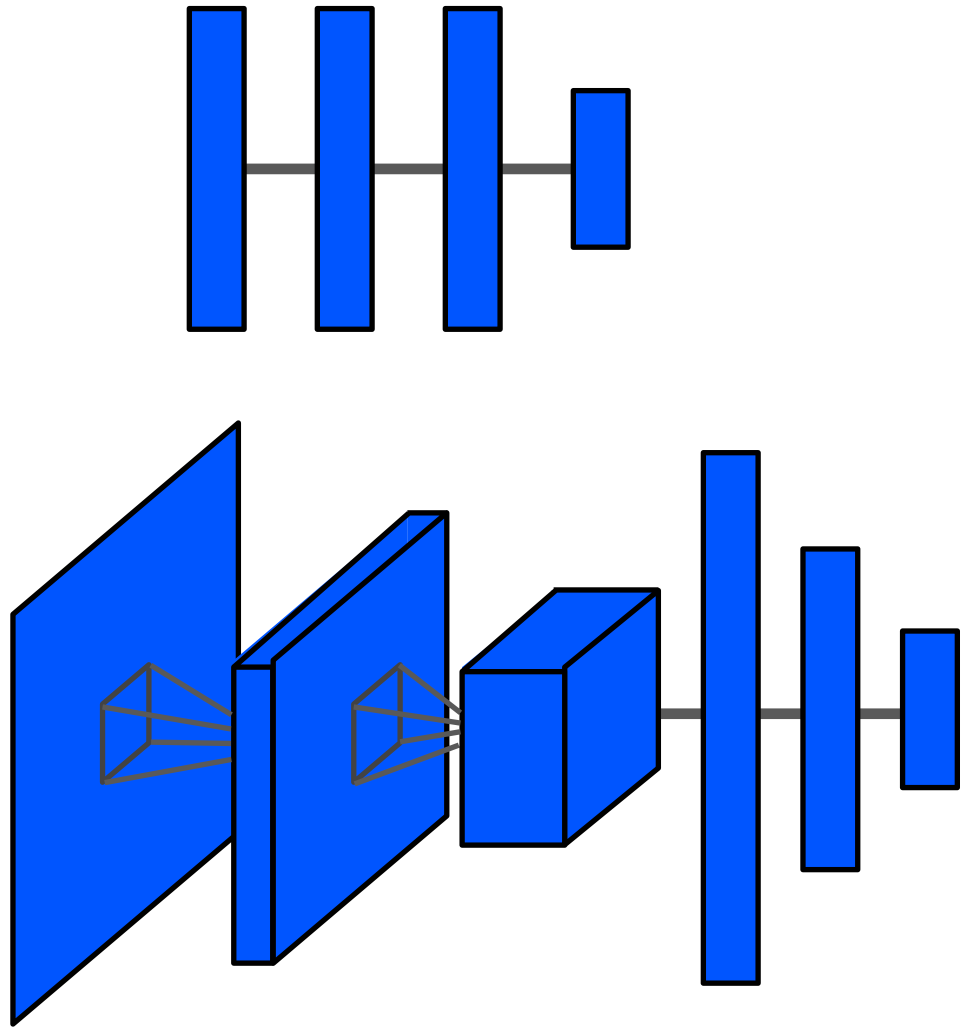


750

290

MNIST Shuffled-pixels



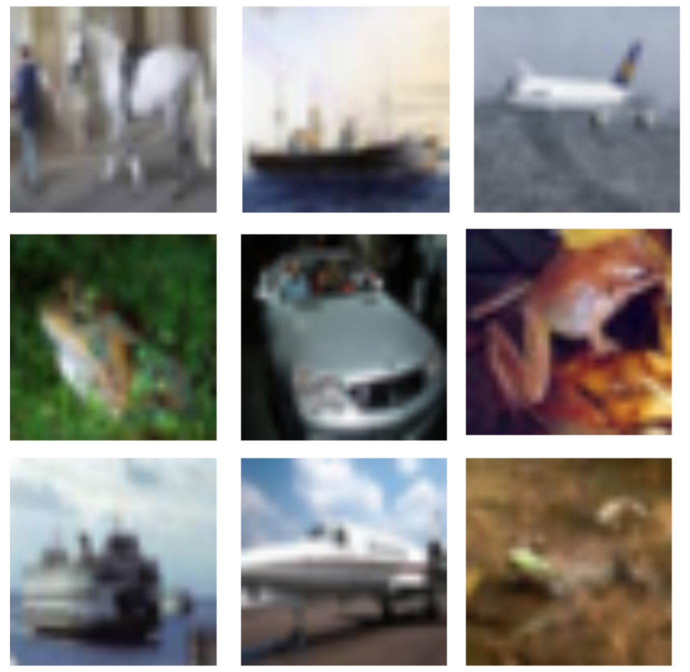


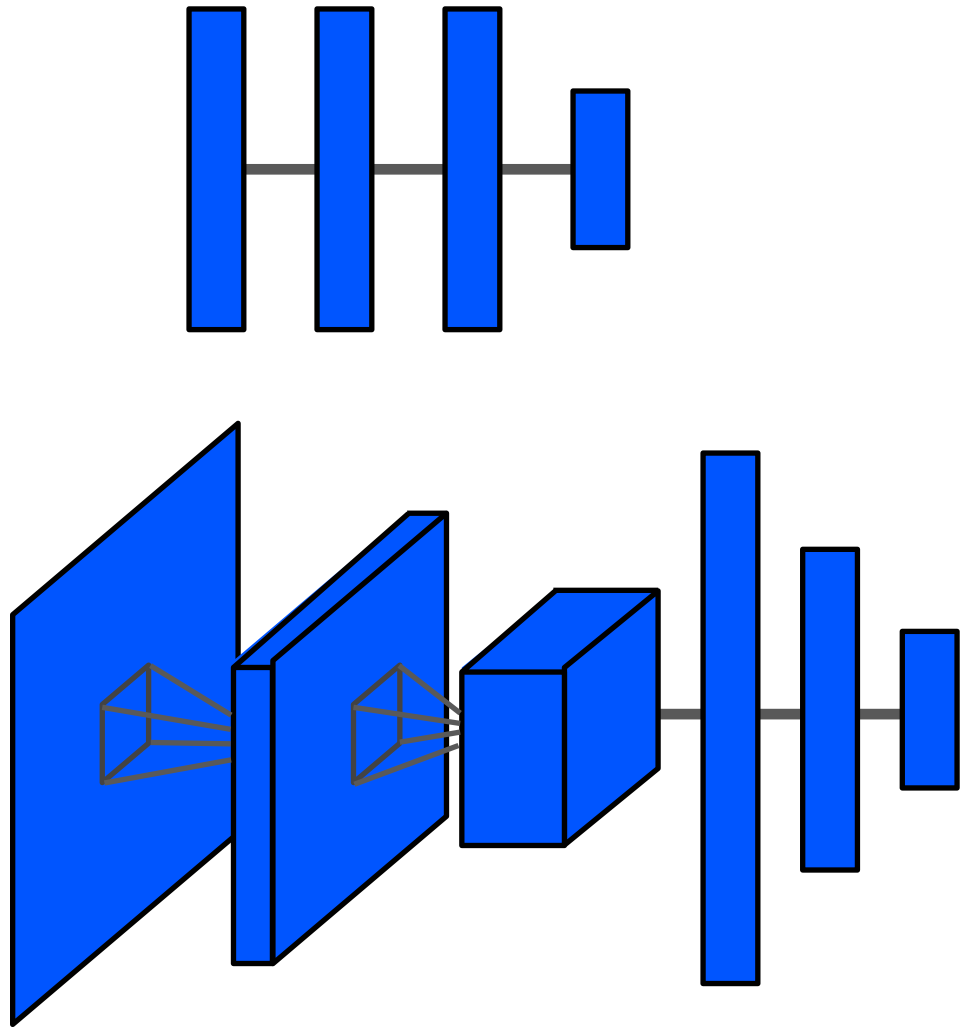
750

1400

CIFAR 10

Int. Dim.





9K

2.9K

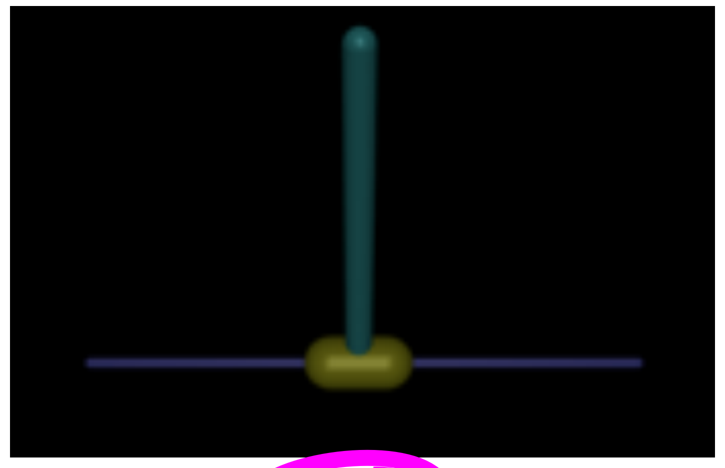
SqueezeNet

>500K

Humanoid

Pong

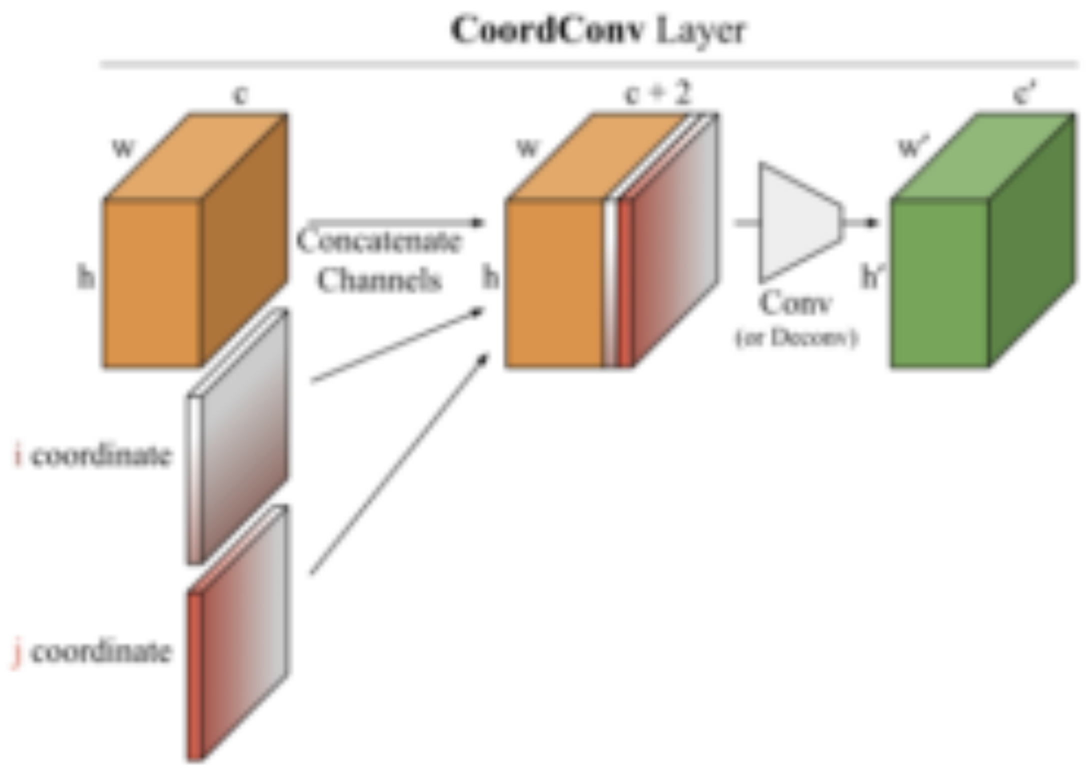
Inverted Pendulum

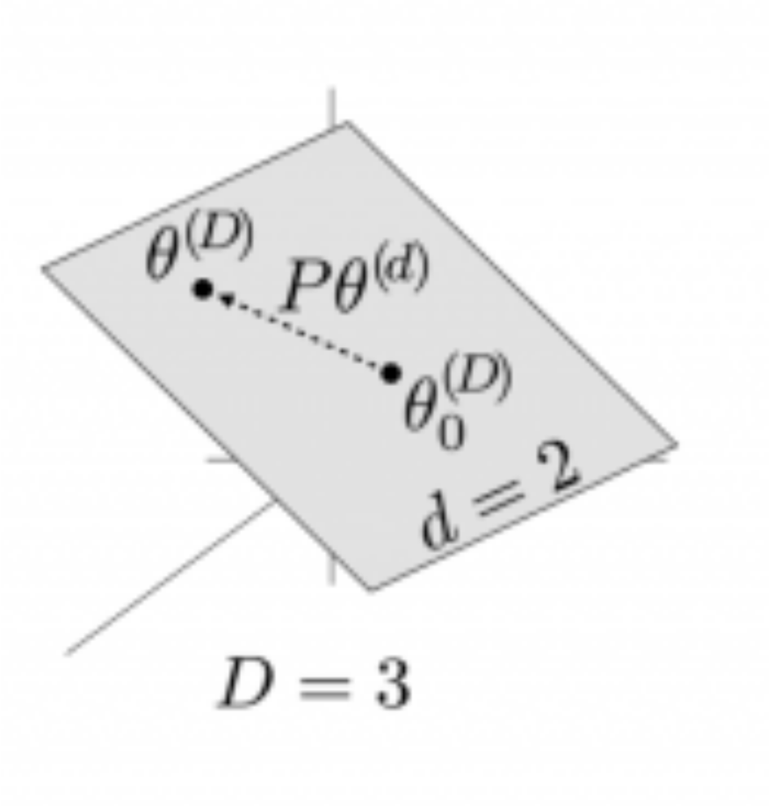


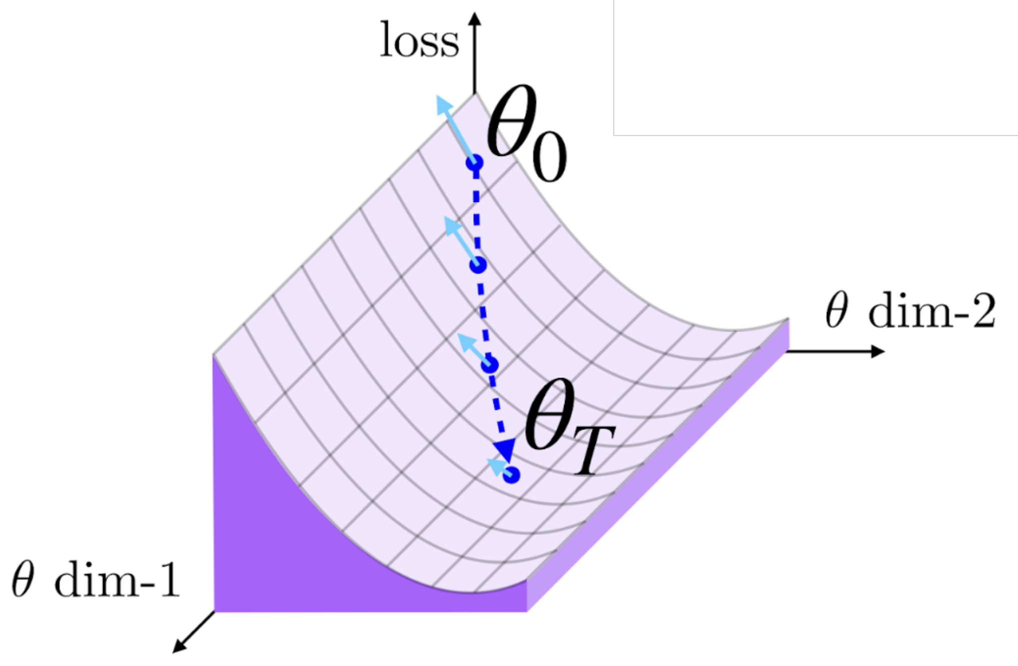
Int. Dim. = **700**

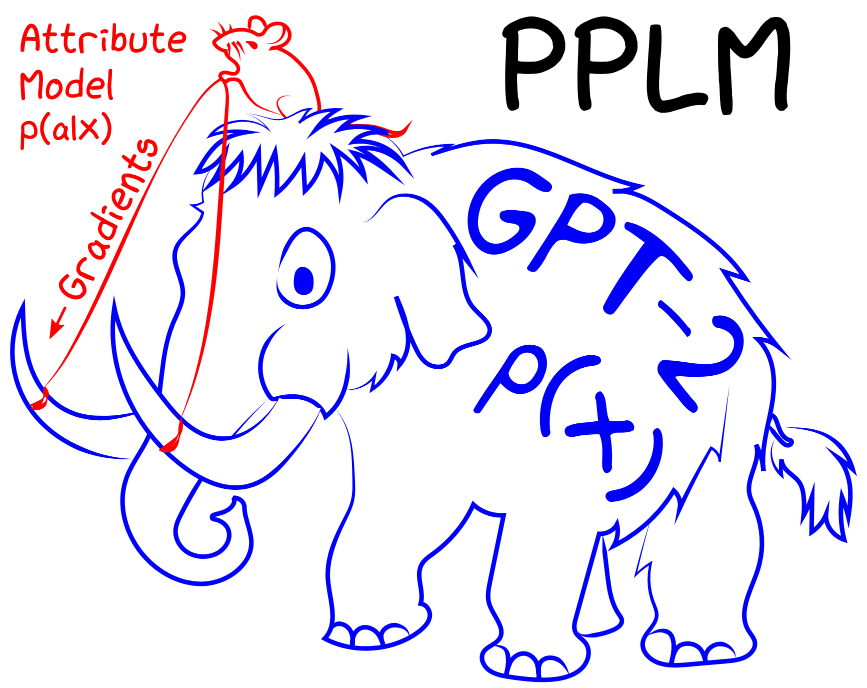
6000

4


CoordConv Layer
c
w
h
Concatenate Channels
c + 2
w
h
Conv
(or Deconv)
c'
w'
h'
i coordinate
j coordinate


$\theta^{(D)}$
$P\theta^{(d)}$
$\theta_0^{(D)}$
$d = 2$
$D = 3$

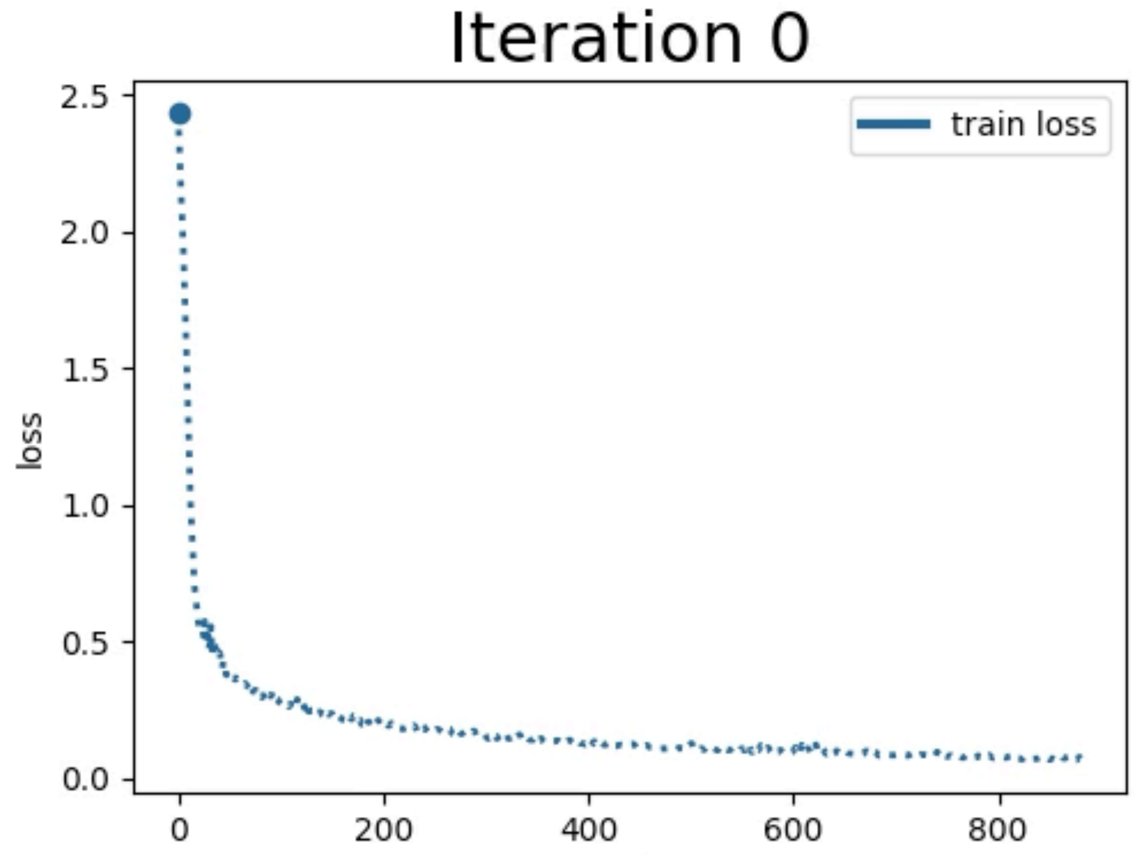

loss
θ_0
θ_T
θ dim-2
θ dim-1

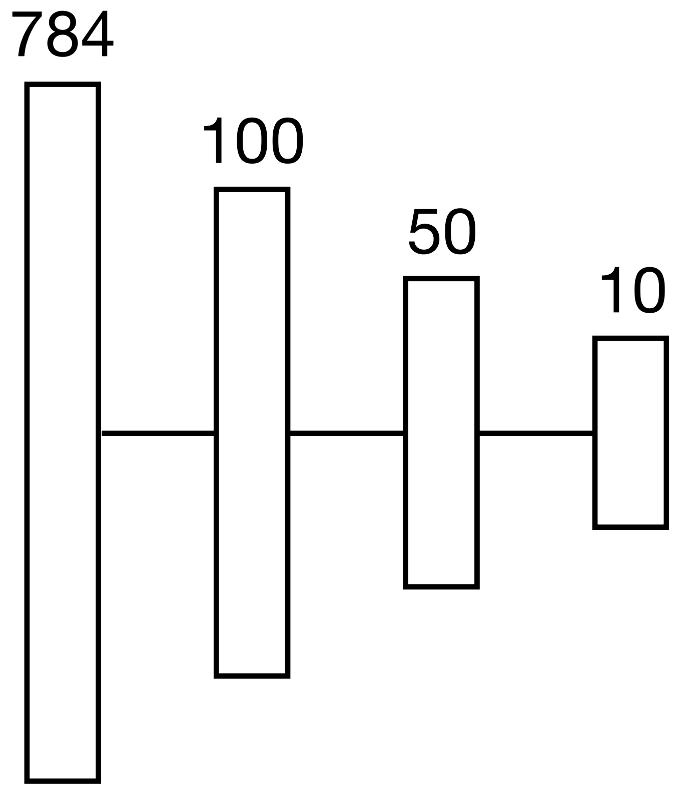

Attribute
Model
p(a|x)
Gradients
PPLM
GPT-2
p(x)

LCA: Loss Change Allocation for Neural Network Training

Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski (NeurIPS 2019)

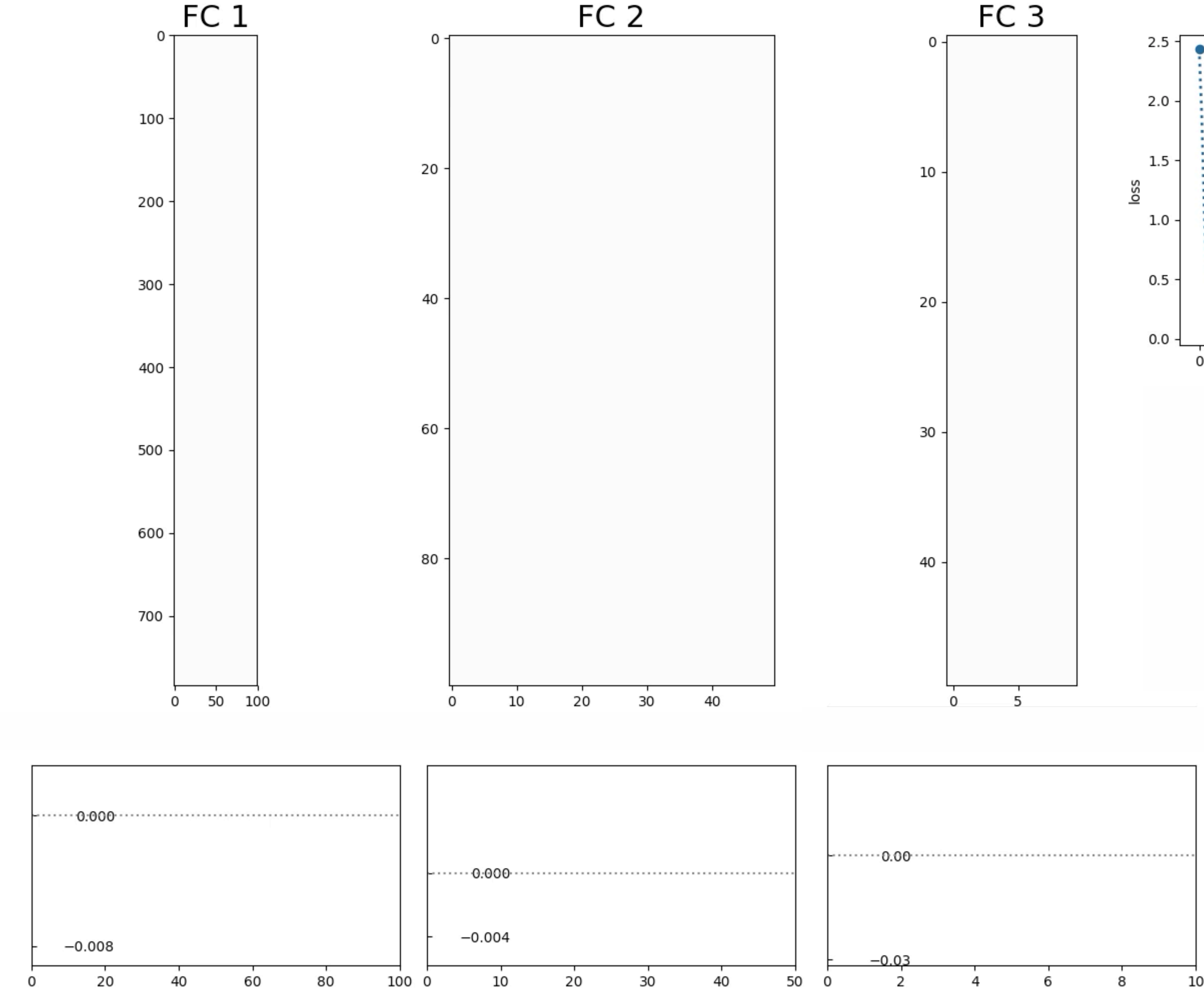
LCA: Loss Change Allocation for Neural Network Training

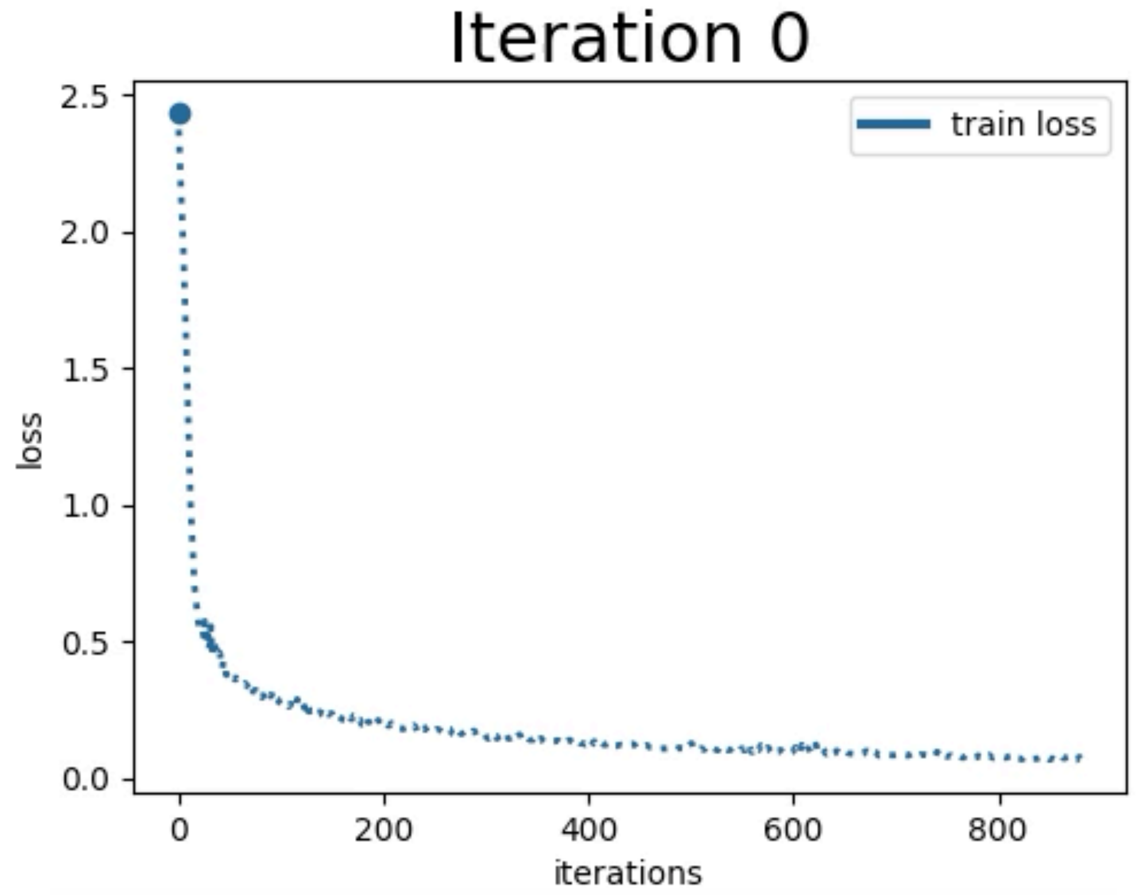


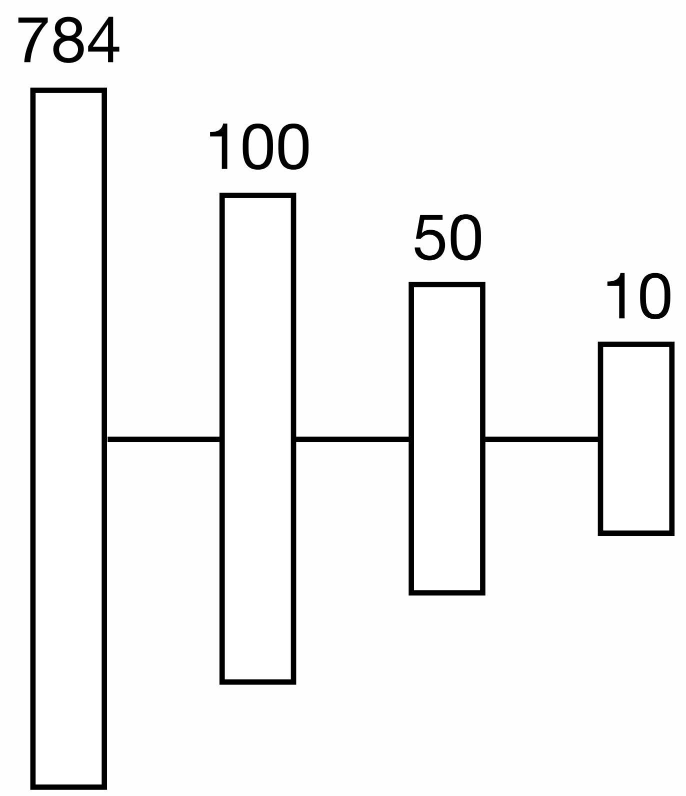


Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski (NeurIPS 2019)

LCA: Loss Change Allocation for Neural Network Training

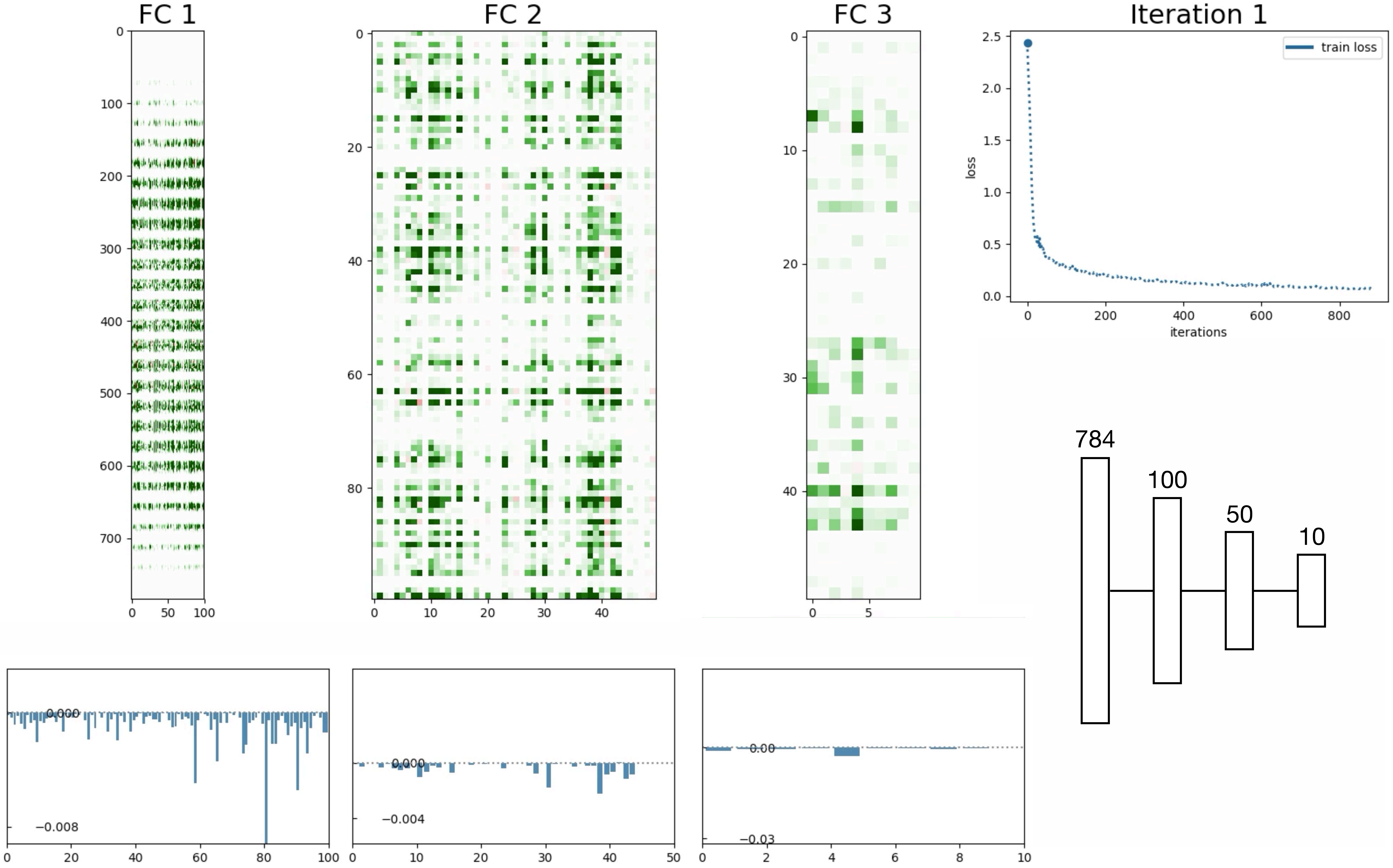






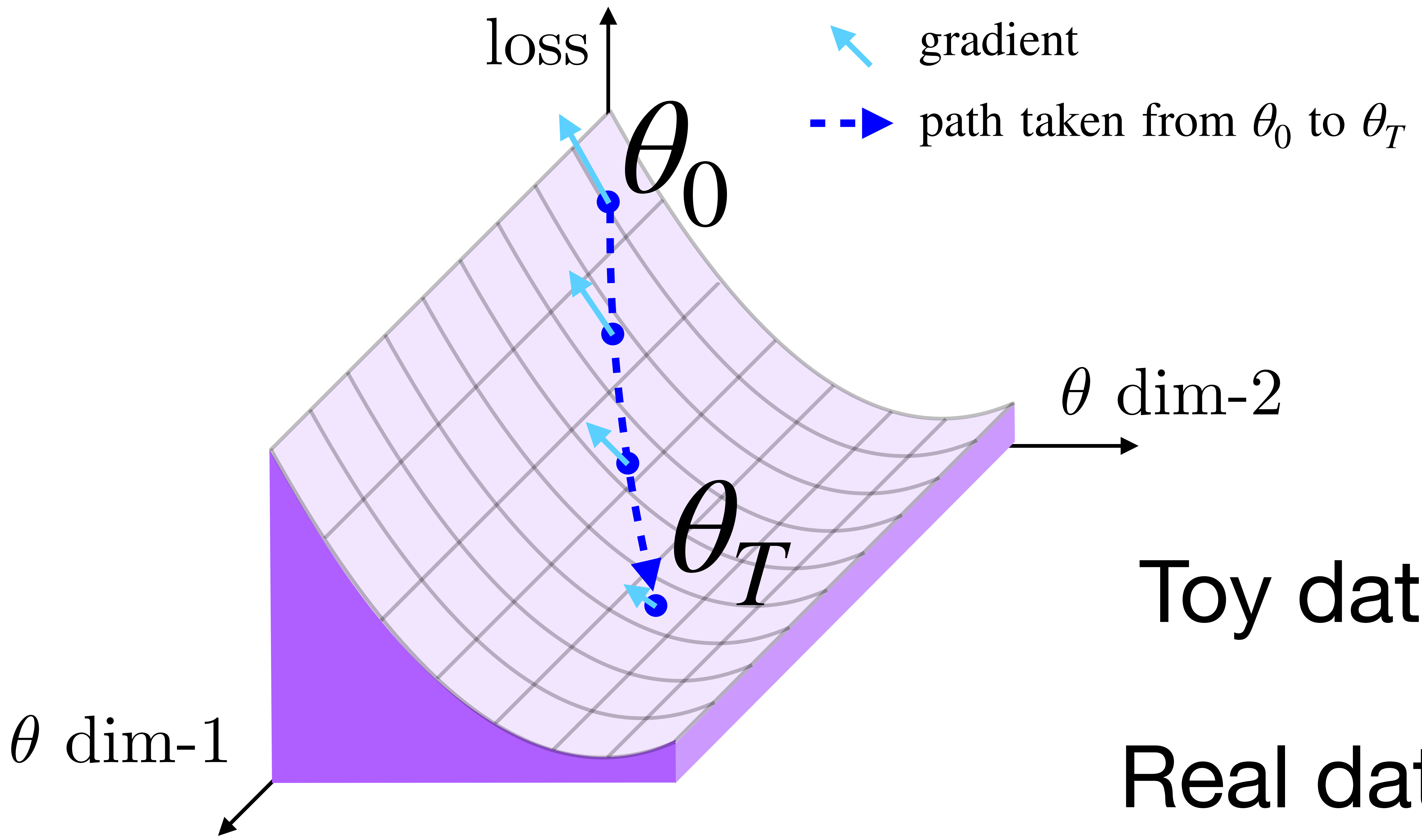
Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski (NeurIPS 2019)

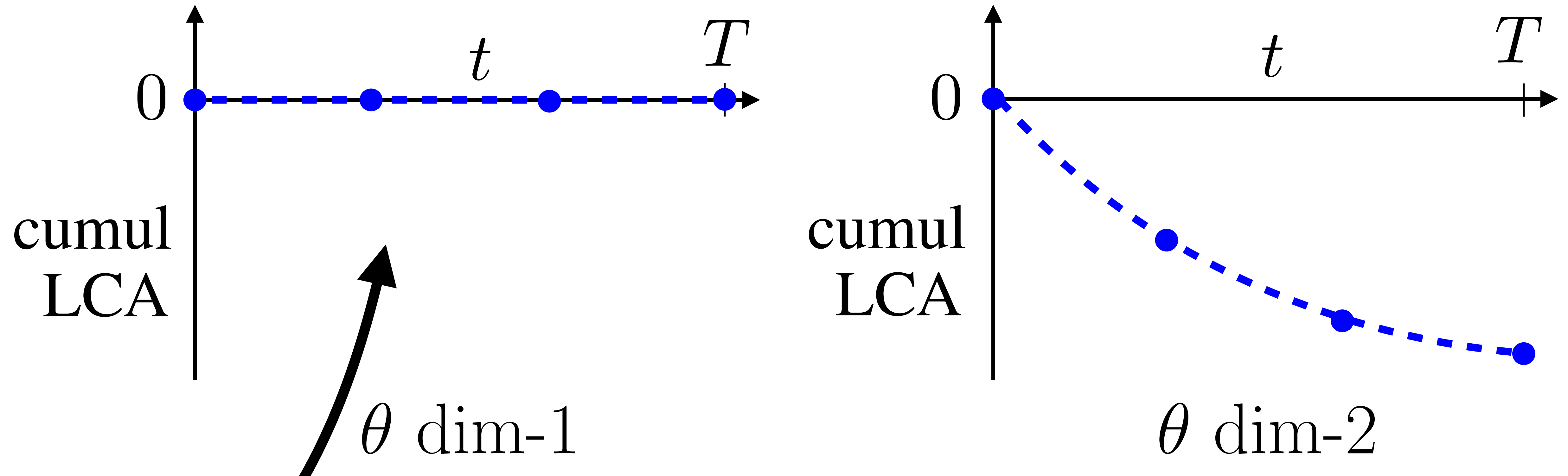
LCA: Loss Change Allocation for Neural Network Training



Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski (NeurIPS 2019)

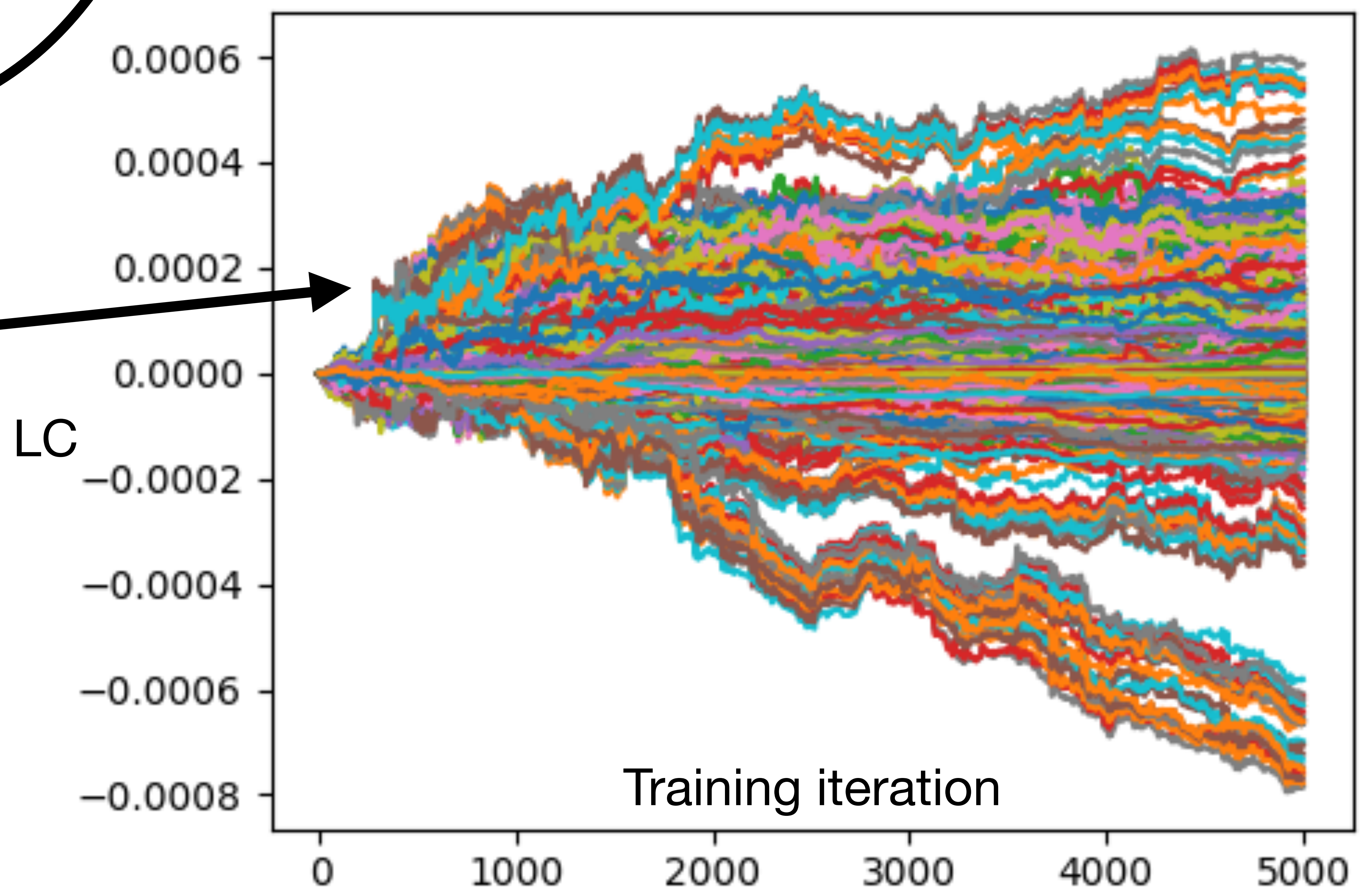
LCA: Loss Change Allocation for Neural Network Training





Toy data

Real data



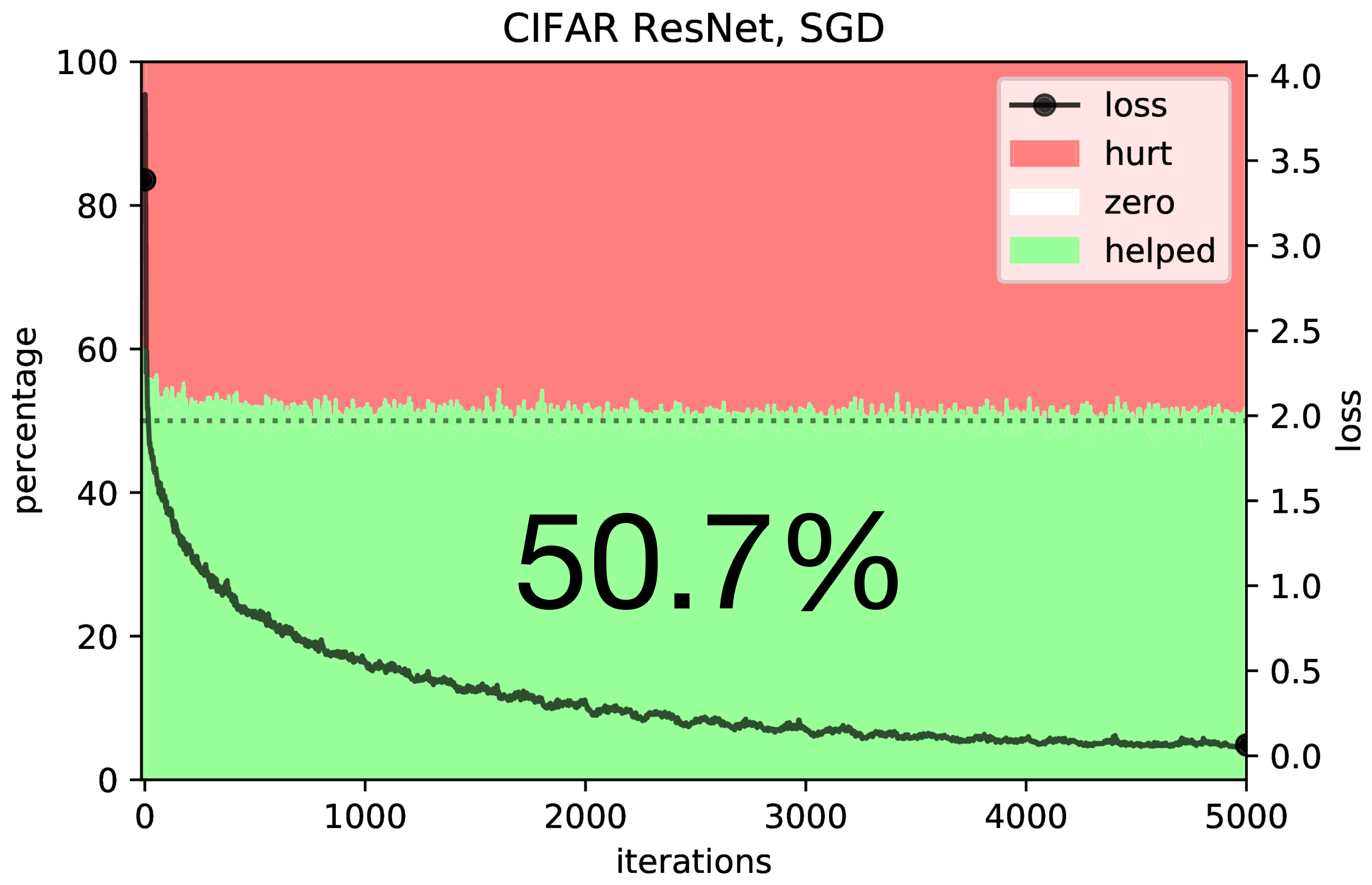
Simple approach → New visibility into training

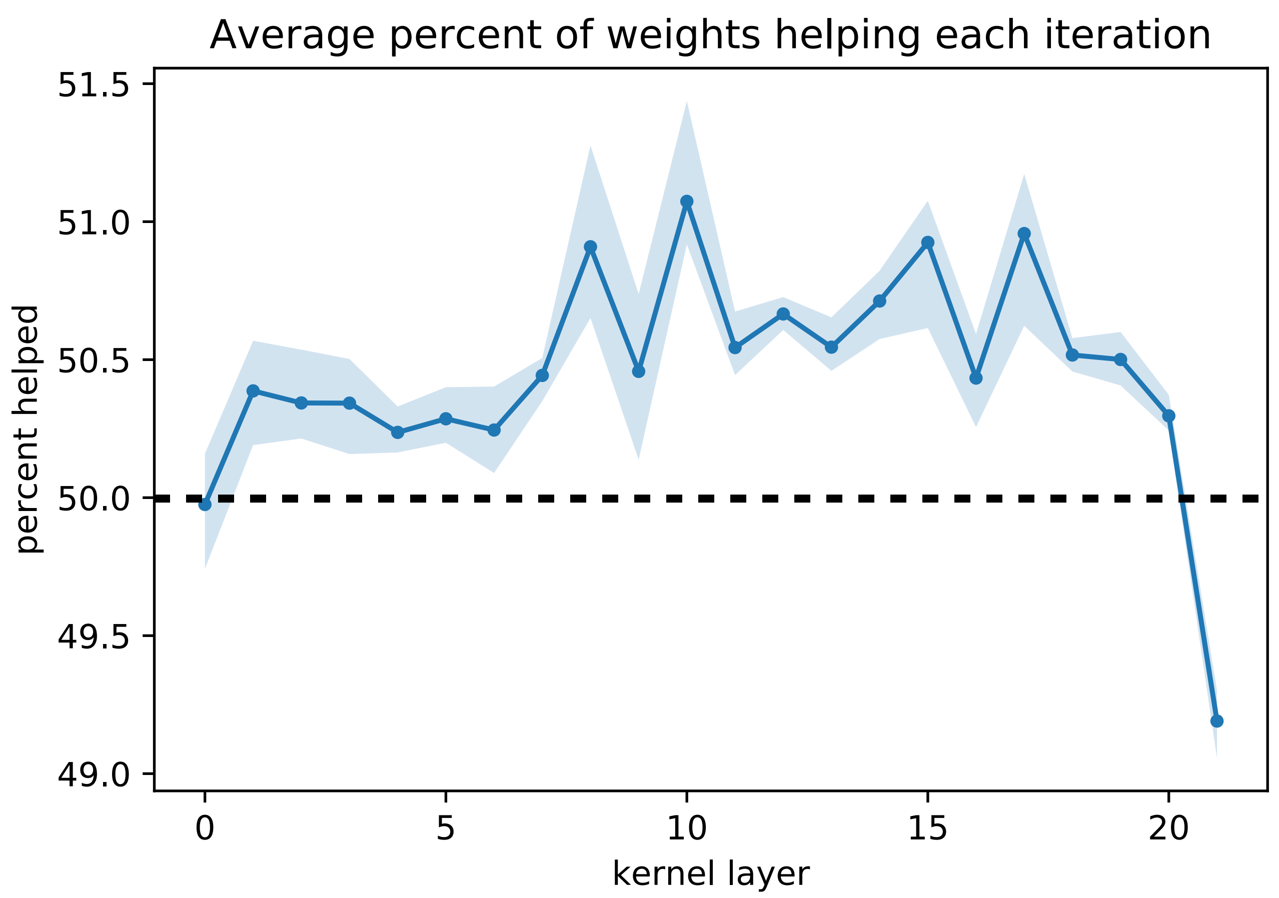
Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski (NeurIPS 2019)

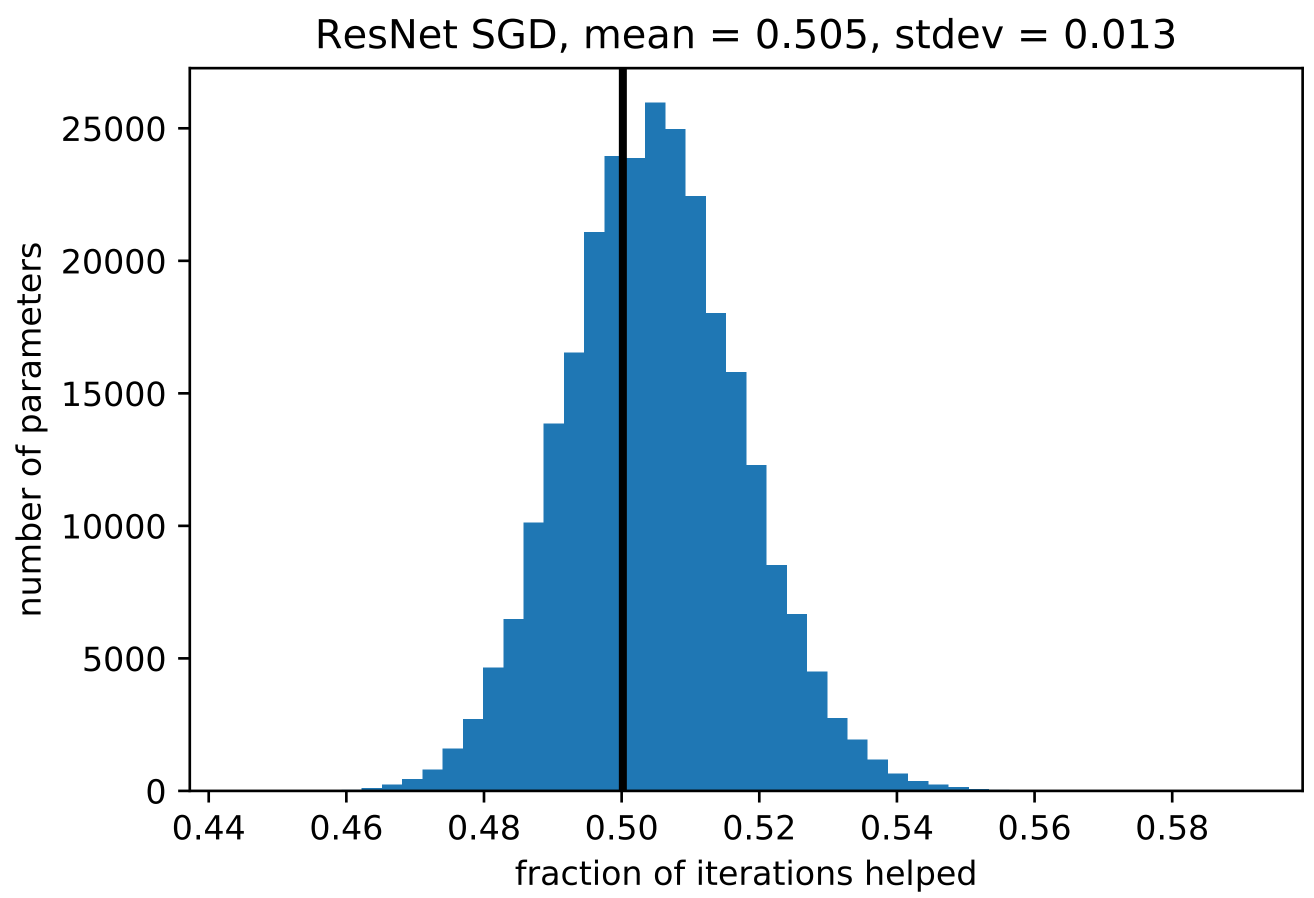
LCA: Loss Change Allocation for Neural Network Training

① Training is noisy

- Holds for all layers
- Holds for all params
- Holds for many hyperparams
  (50.3% – 51.6%)


CIFAR ResNet, SGD


Average percent of weights helping each iteration

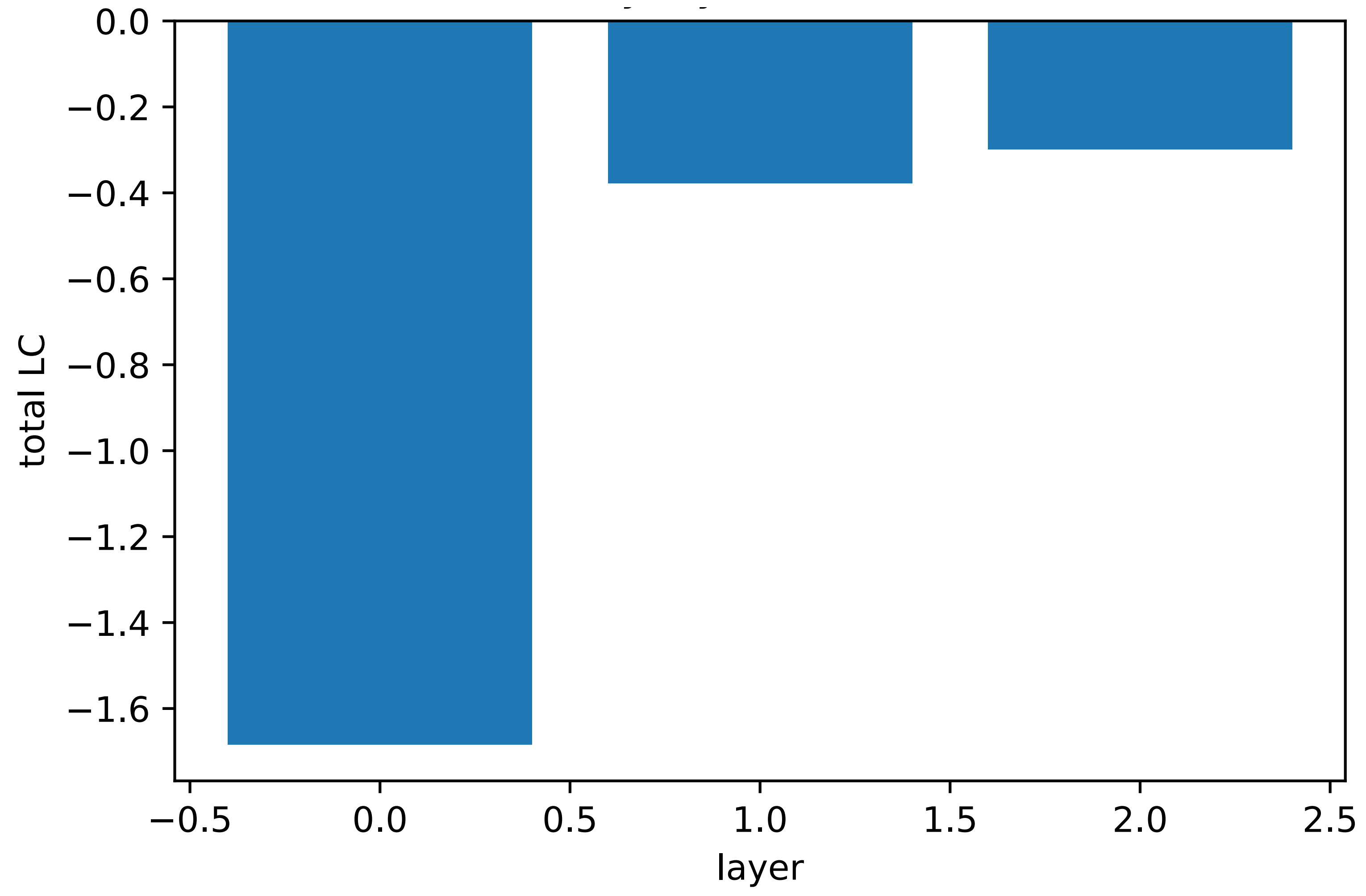

ResNet SGD, mean = 0.505, stdev = 0.013

Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski (NeurIPS 2019)

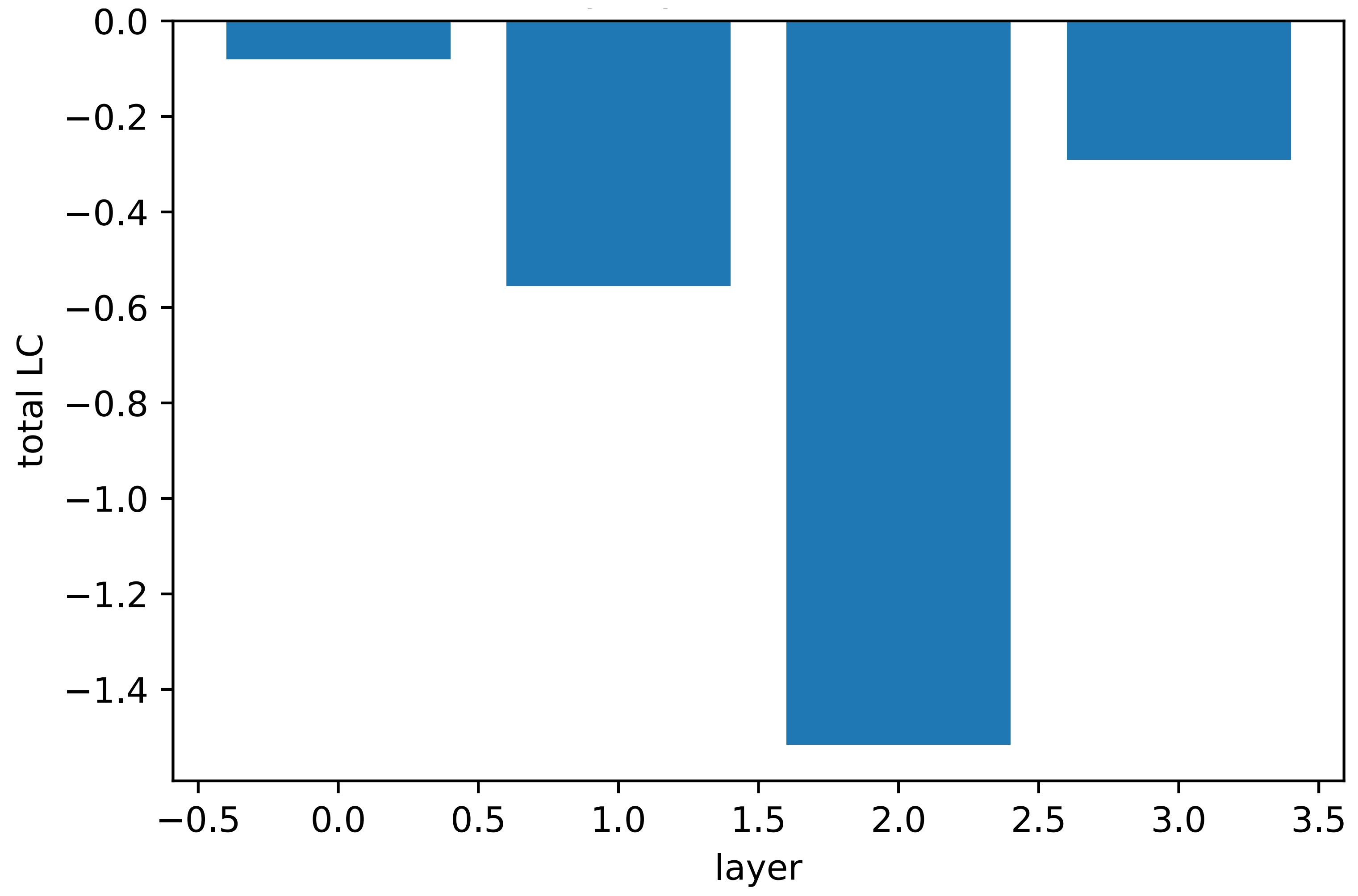
LCA: Loss Change Allocation for Neural Network Training

① Training is noisy

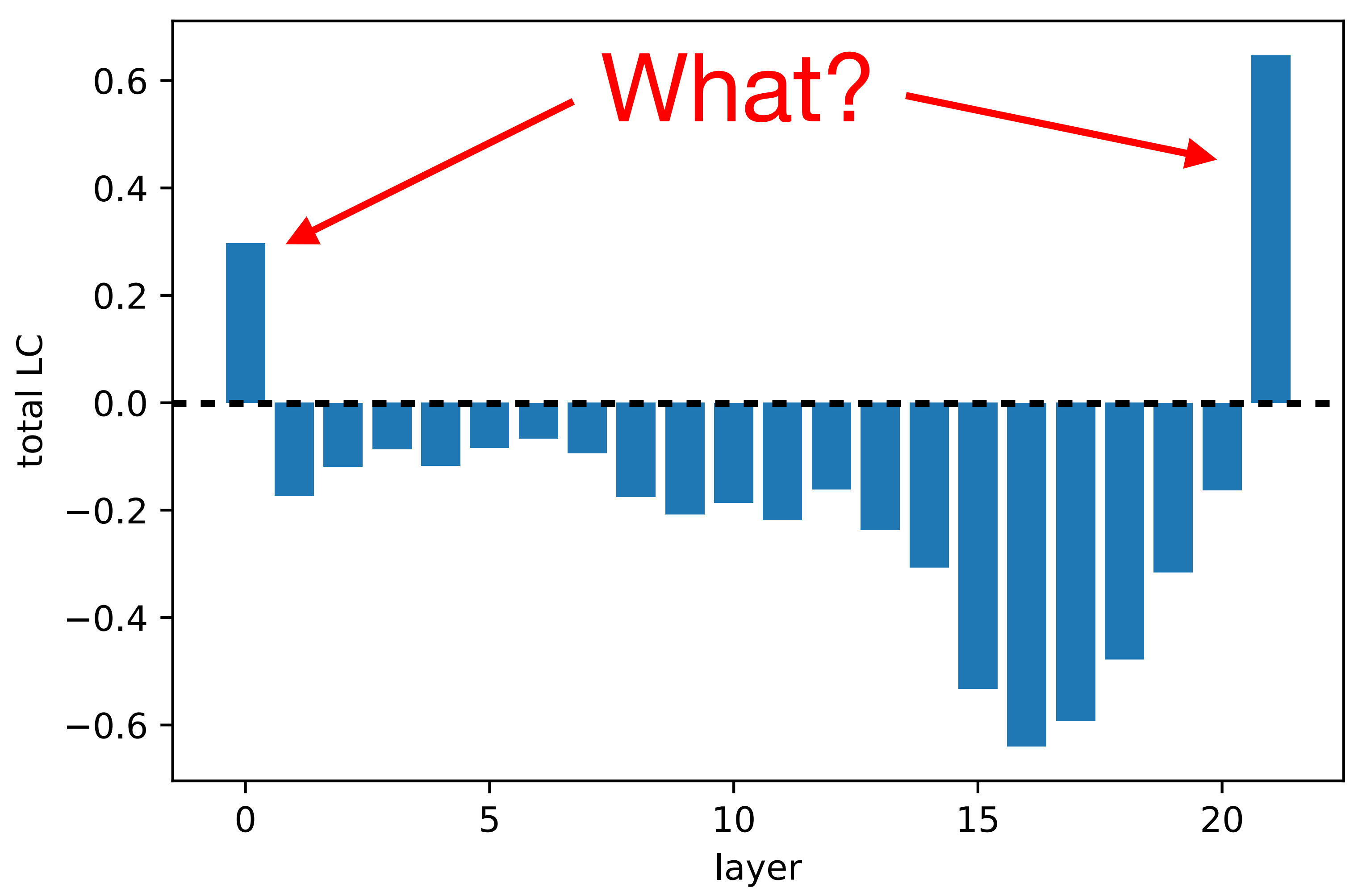
② Some layers go backwards



MNIST-FC

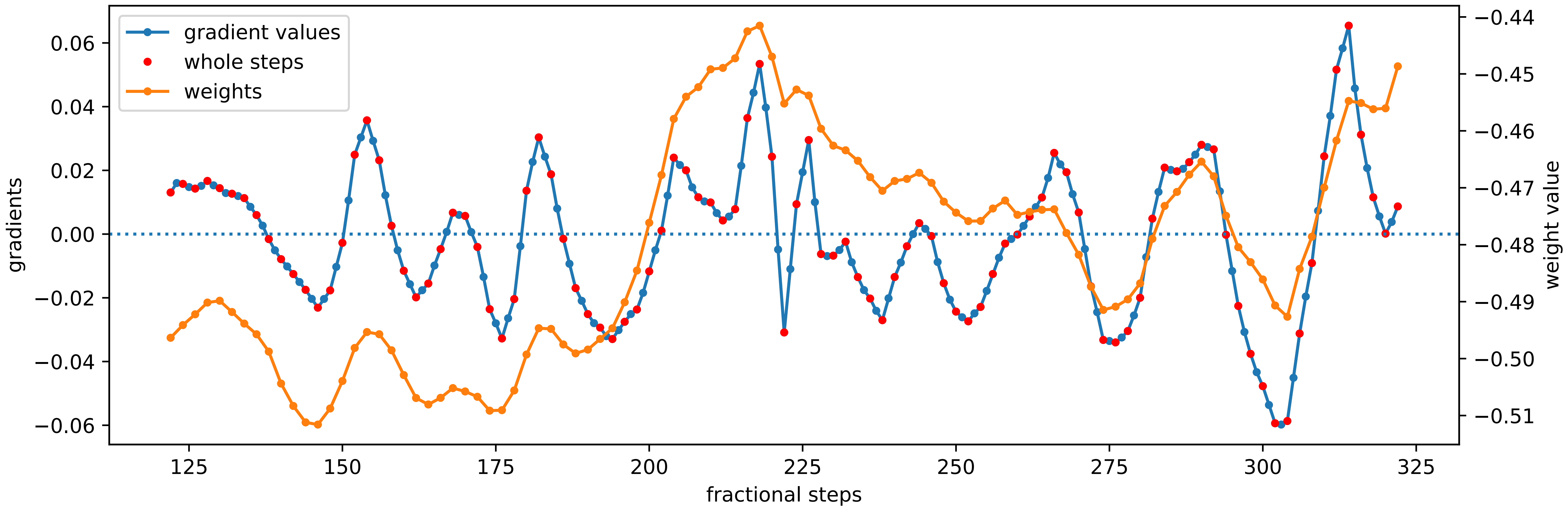


MNIST-LeNet



CIFAR-ResNet

What?

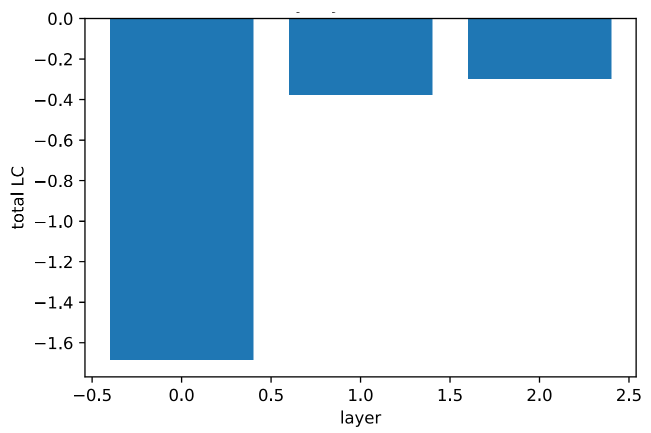


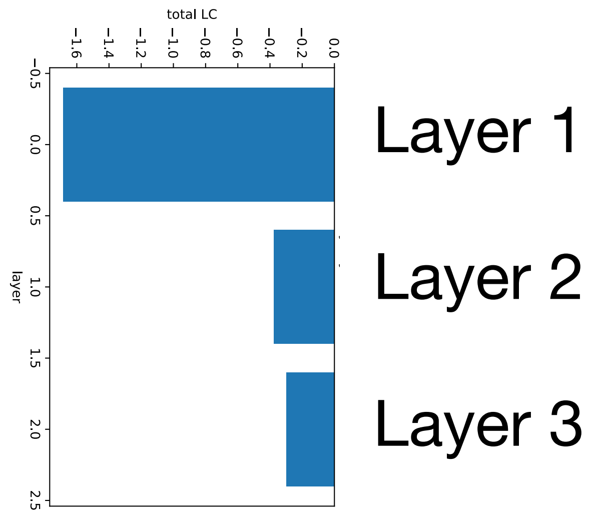
Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski (NeurIPS 2019)

LCA: Loss Change Allocation for Neural Network Training

① Training is noisy ② Some layers go backwards

MNIST-FC



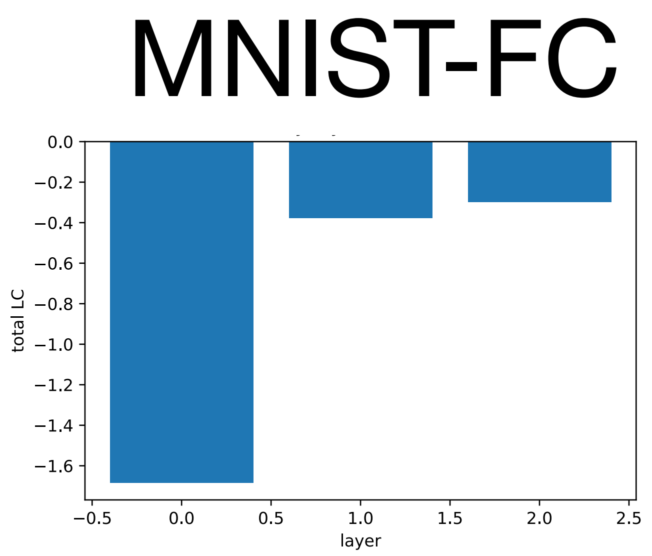


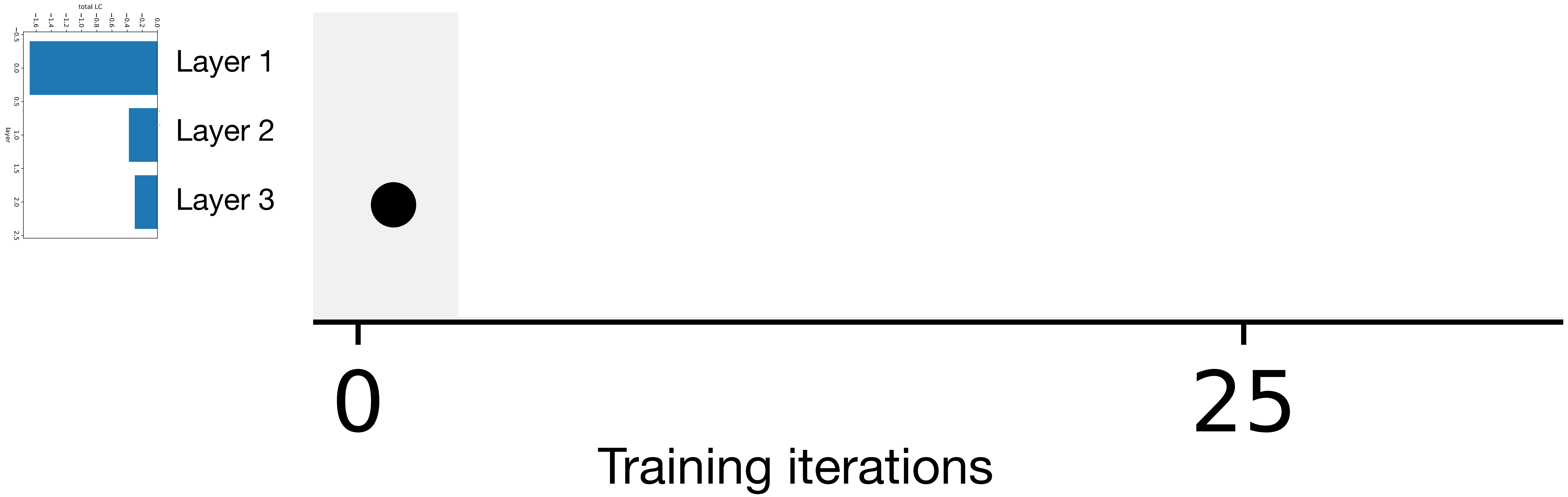
Training iterations

Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski (NeurIPS 2019)

LCA: Loss Change Allocation for Neural Network Training

① Training is noisy ② Some layers go backwards





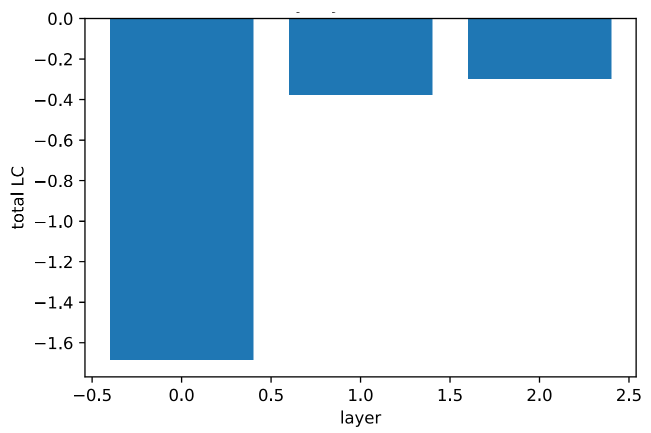
Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski (NeurIPS 2019)

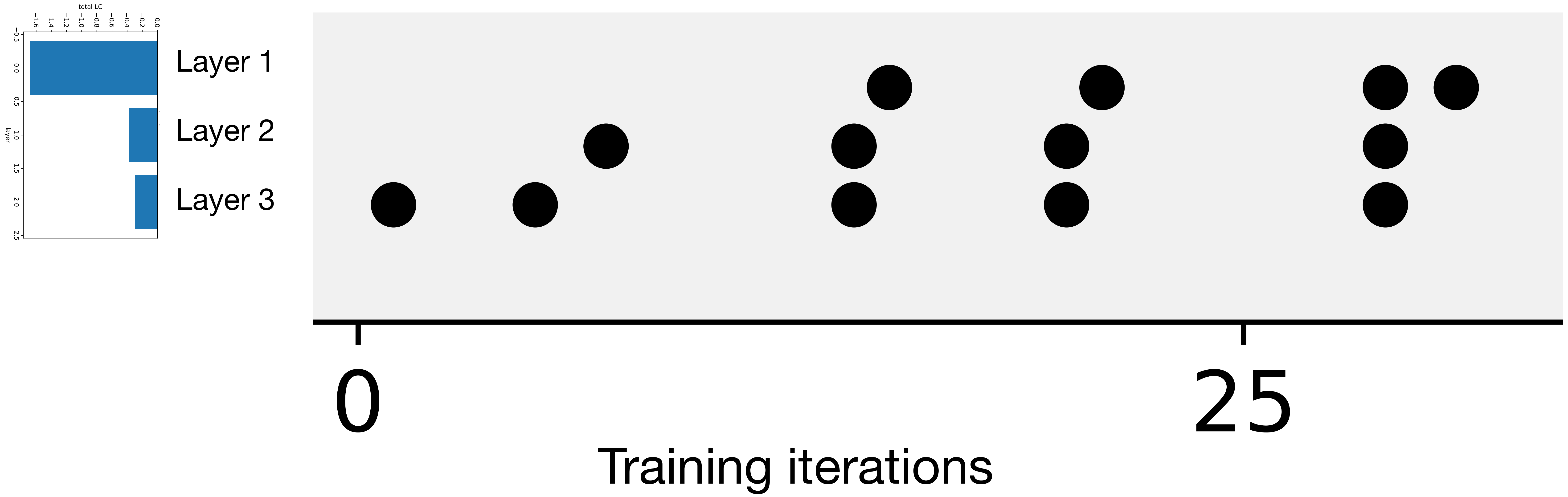
LCA: Loss Change Allocation for Neural Network Training

① Training is noisy

② Some layers go backwards

MNIST-FC



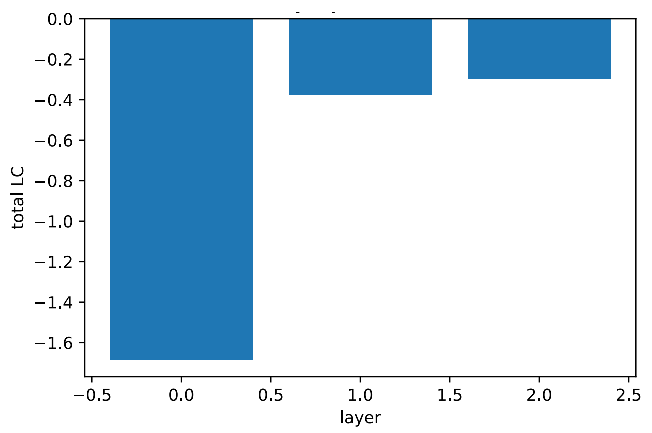


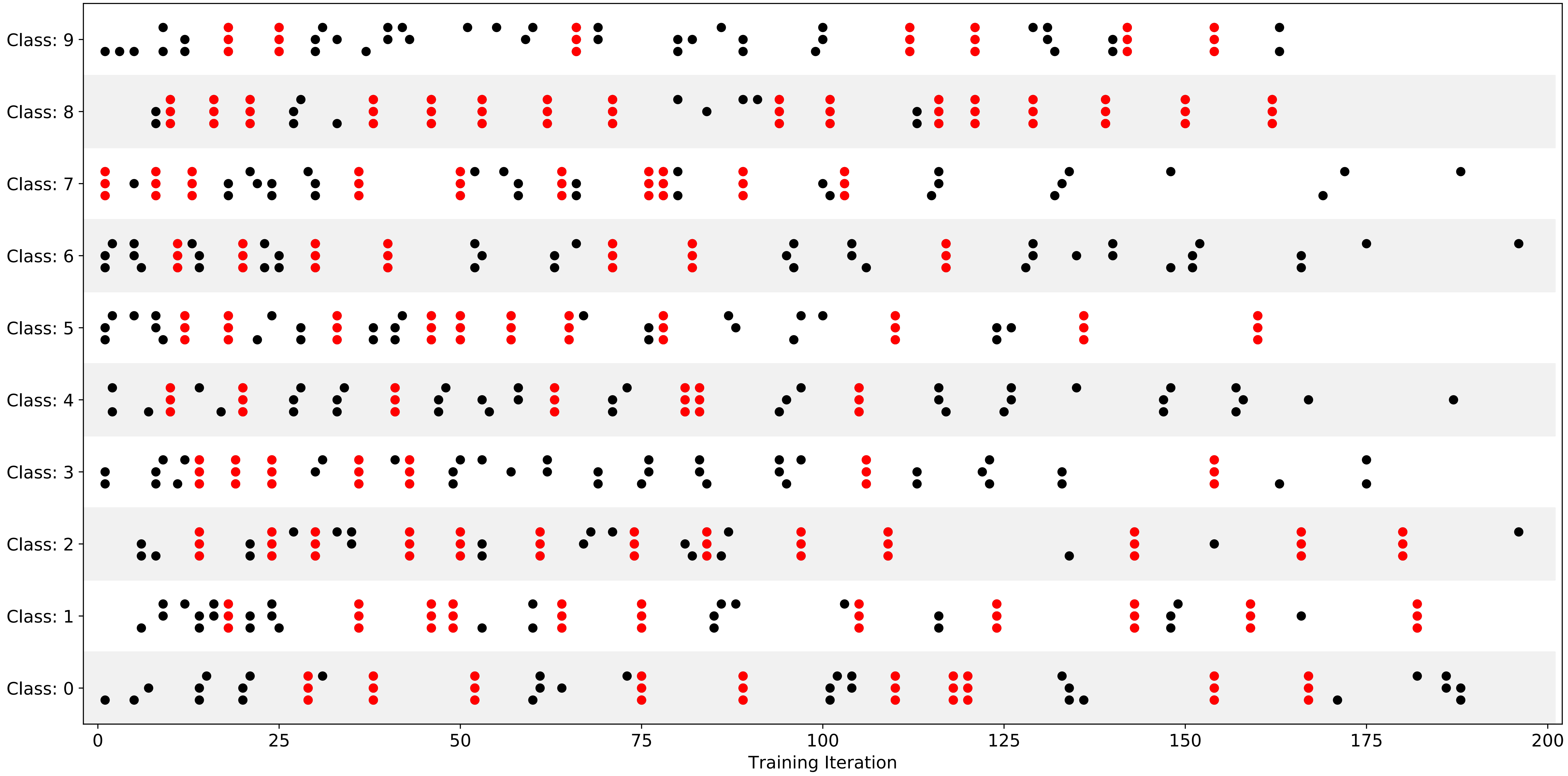
Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski (NeurIPS 2019)

LCA: Loss Change Allocation for Neural Network Training

① Training is noisy ② Some layers go backwards ③ Some micro-learning is synchronized

MNIST-FC



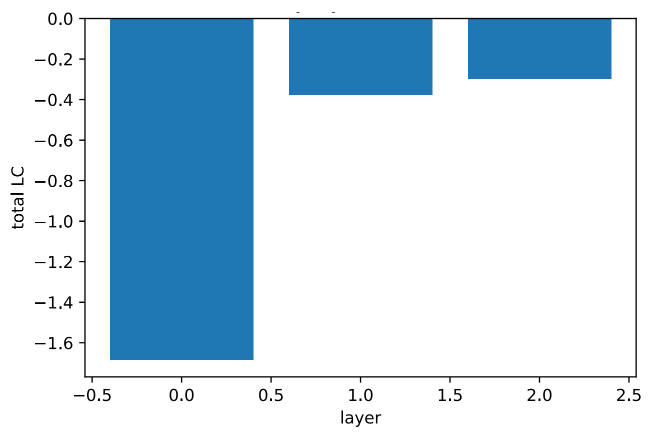


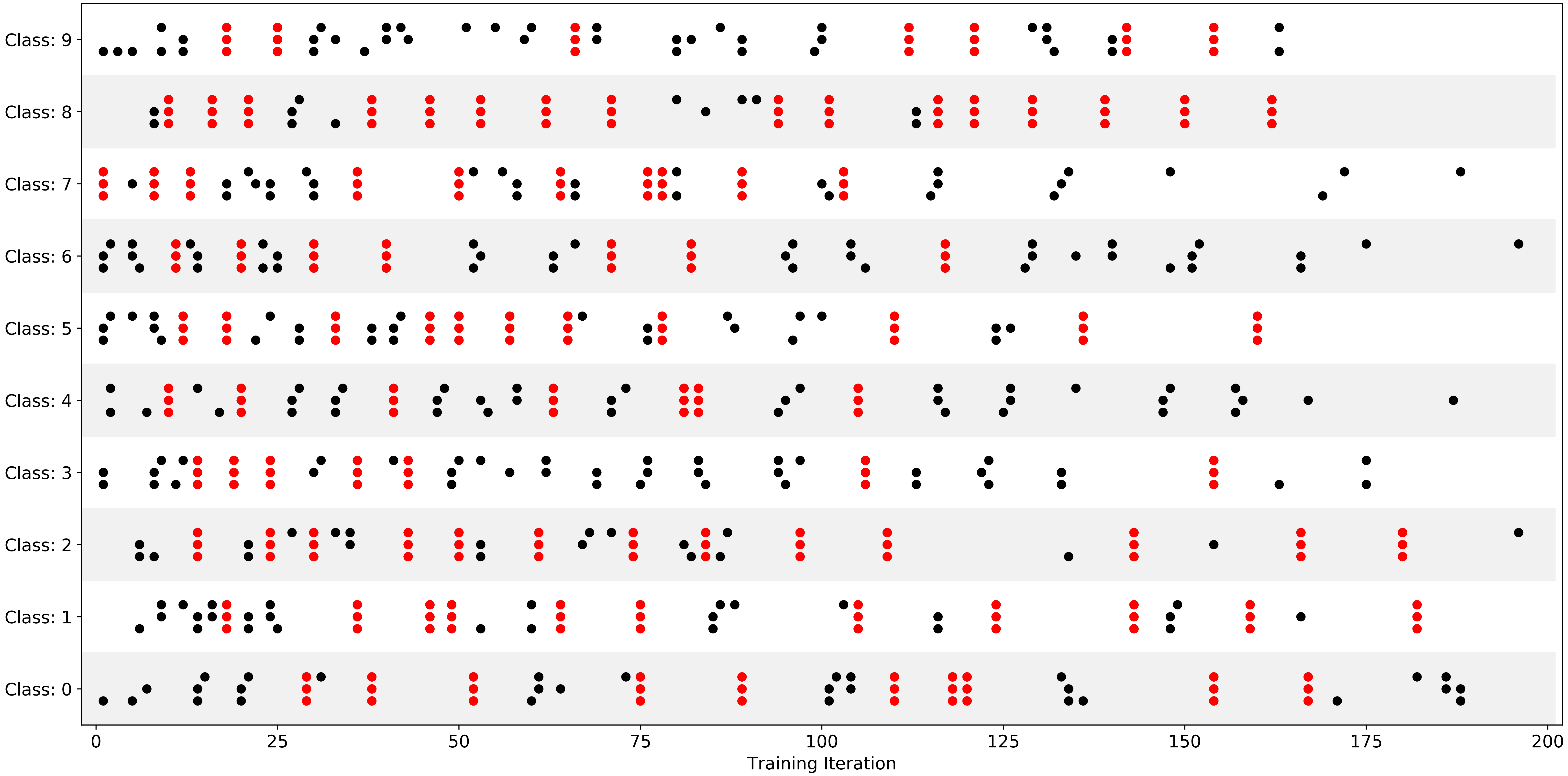
Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski (NeurIPS 2019)

LCA: Loss Change Allocation for Neural Network Training

1. Training is noisy
2. Some layers go backwards
3. Some micro-learning is synchronized

MNIST-FC



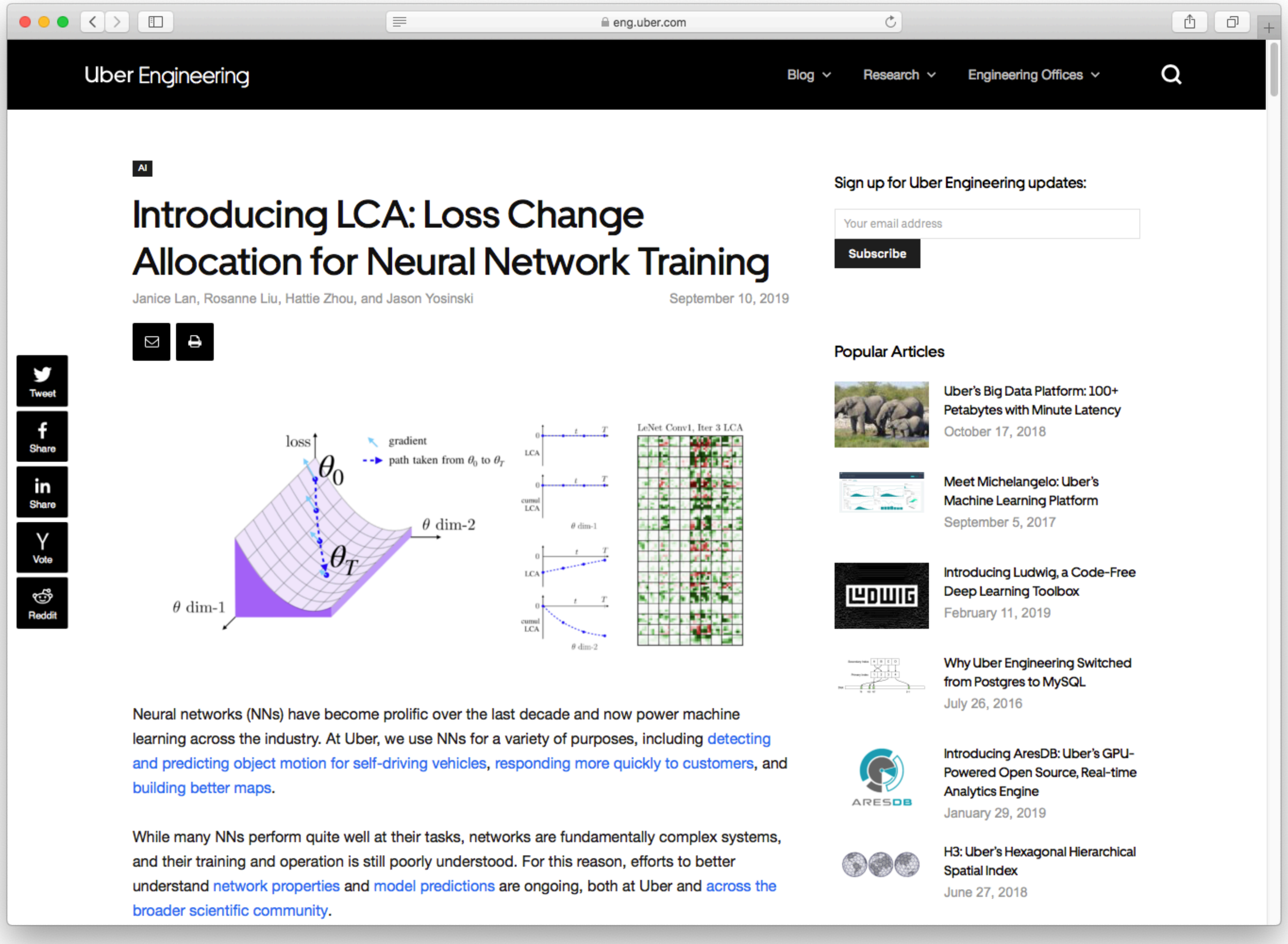


Janice Lan, Rosanne Liu, Hattie Zhou, Jason Yosinski (NeurIPS 2019)

LCA: Loss Change Allocation for Neural Network Training

(NeurIPS 2019)

Blog: https://eng.uber.com/loss-change-allocation/



Introducing LCA: Loss Change Allocation for Neural Network Training

Janice Lan, Rosanne Liu, Hattie Zhou, and Jason Yosinski

September 10, 2019

Neural networks (NNs) have become prolific over the last decade and now power machine learning across the industry. At Uber, we use NNs for a variety of purposes, including detecting and predicting object motion for self-driving vehicles, responding more quickly to customers, and building better maps.

While many NNs perform quite well at their tasks, networks are fundamentally complex systems, and their training and operation is still poorly understood. For this reason, efforts to better understand network properties and model predictions are ongoing, both at Uber and across the broader scientific community.

Sign up for Uber Engineering updates:

Popular Articles

Uber's Big Data Platform: 100+ Petabytes with Minute Latency
October 17, 2018

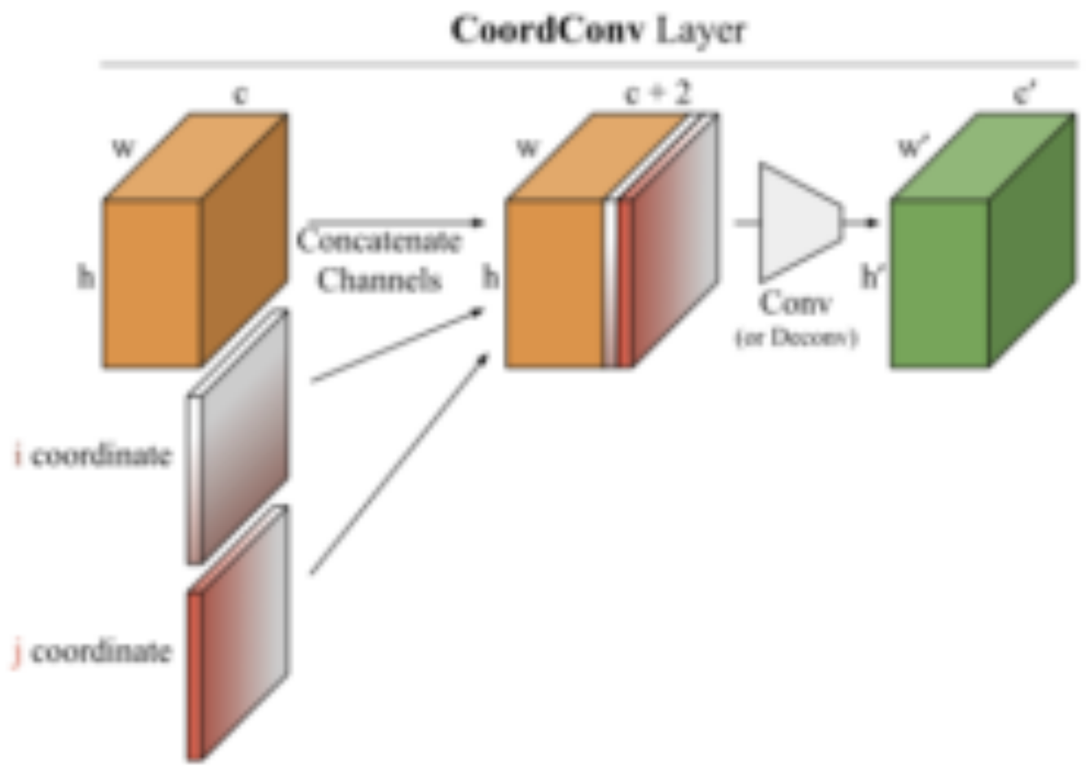
Meet Michelangelo: Uber's Machine Learning Platform
September 5, 2017

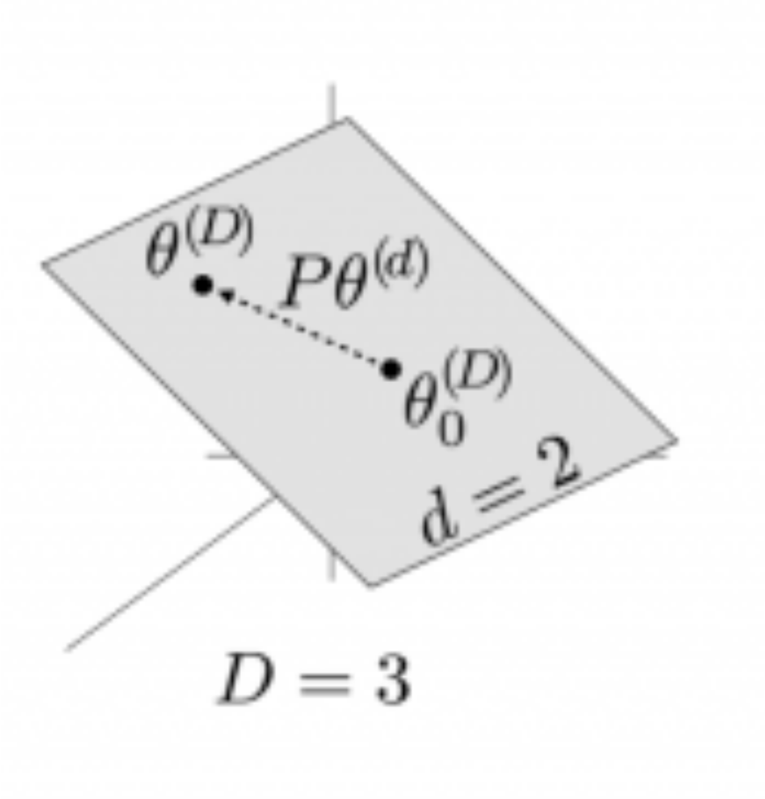
Introducing Ludwig, a Code-Free Deep Learning Toolbox
February 11, 2019

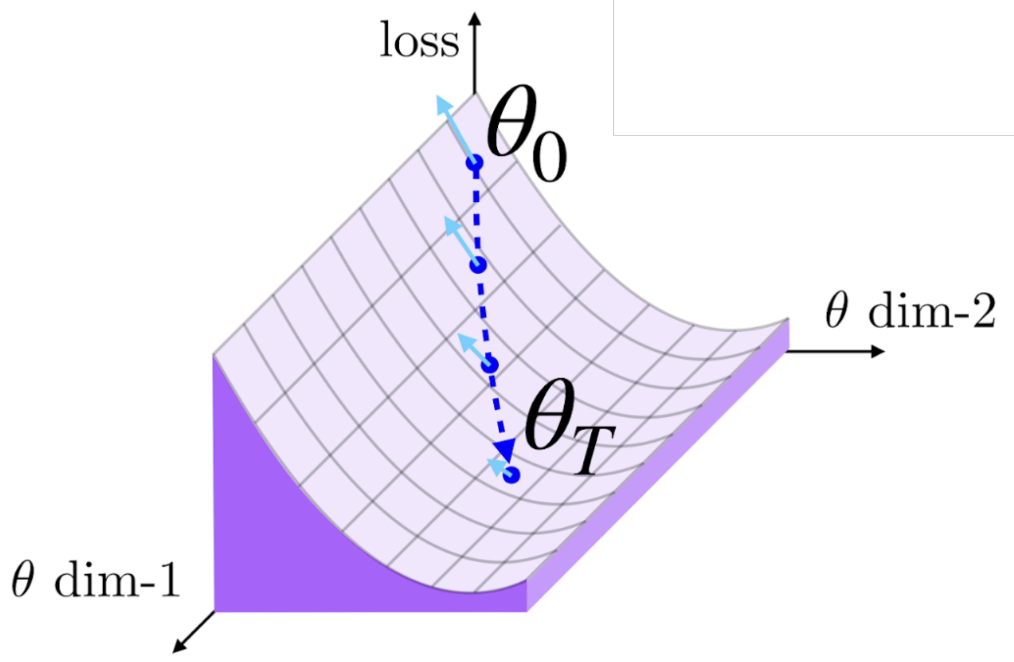
Why Uber Engineering Switched from Postgres to MySQL
July 26, 2016

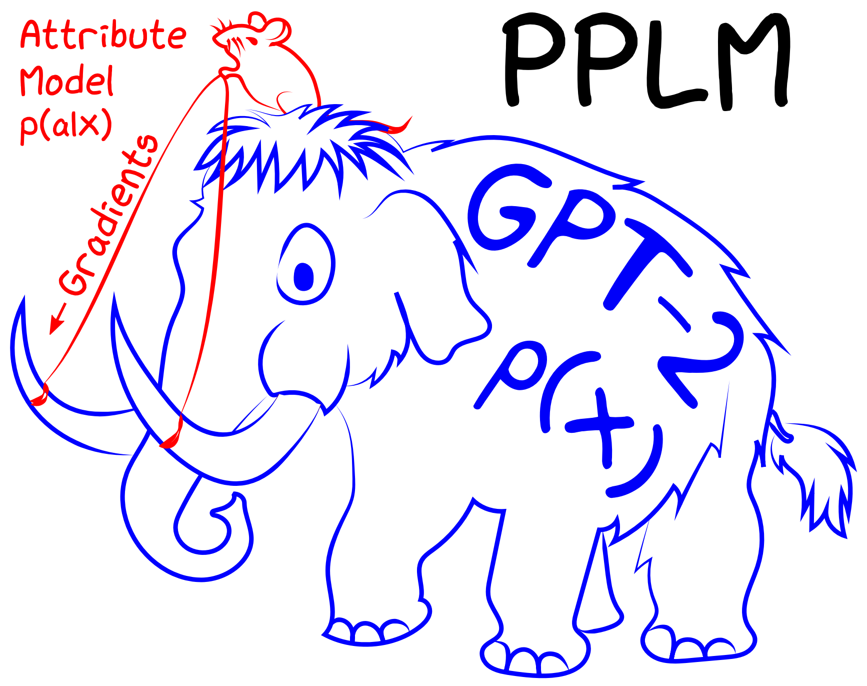
Introducing AresDB: Uber's GPU-Powered Open Source, Real-time Analytics Engine
January 29, 2019

H3: Uber's Hexagonal Hierarchical Spatial Index
June 27, 2018

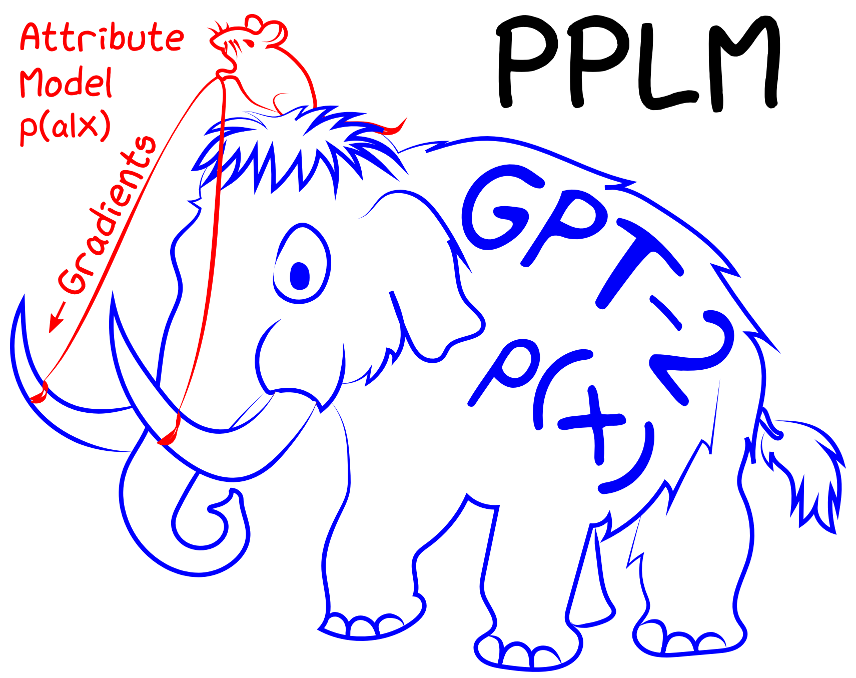

CoordConv Layer
c
w
h
Concatenate Channels
c + 2
w
h
Conv
(or Deconv)
c'
w'
h'
i coordinate
j coordinate


$\theta^{(D)}$
$P\theta^{(d)}$
$\theta_0^{(D)}$
$d = 2$
$D = 3$


loss
θ_0
θ_T
θ dim-2
θ dim-1


PPLM
Attribute
Model
p(a|x)
Gradients
GPT-2
p(x)

PPLM

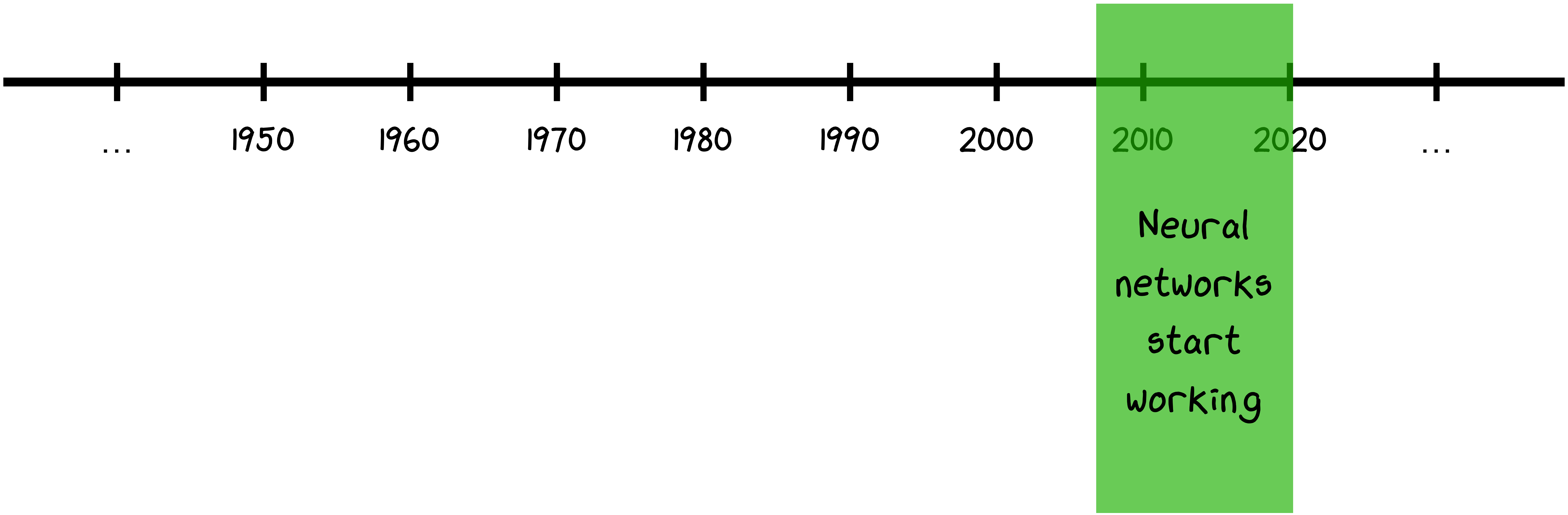


+ Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino

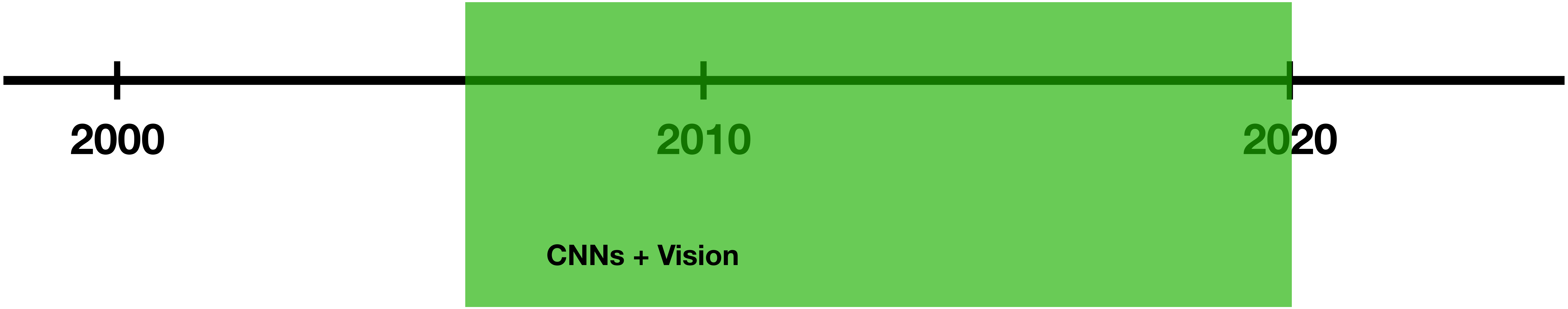
ICLR 2020 (To appear)

http://www.rosanneliu.com/publication/pplm/

Progress in AI



Progress in AI



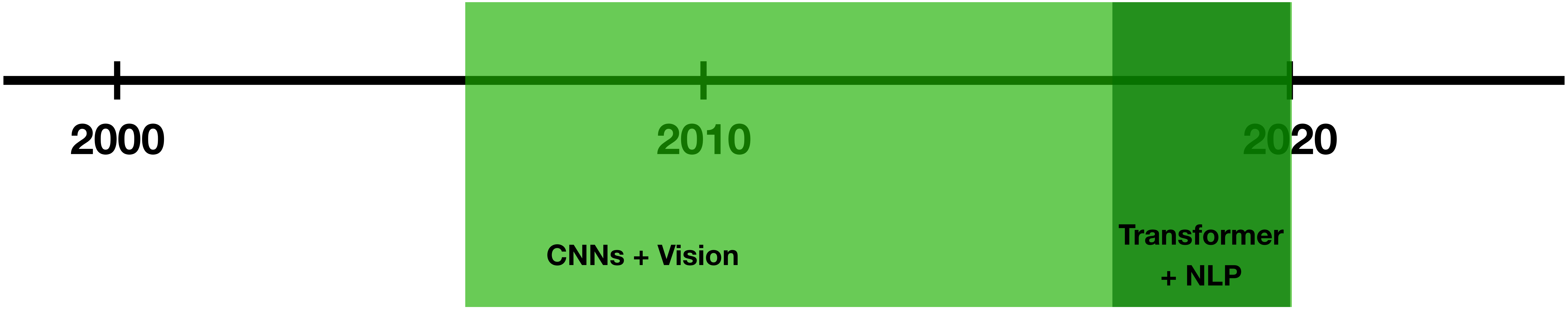
2000

2010

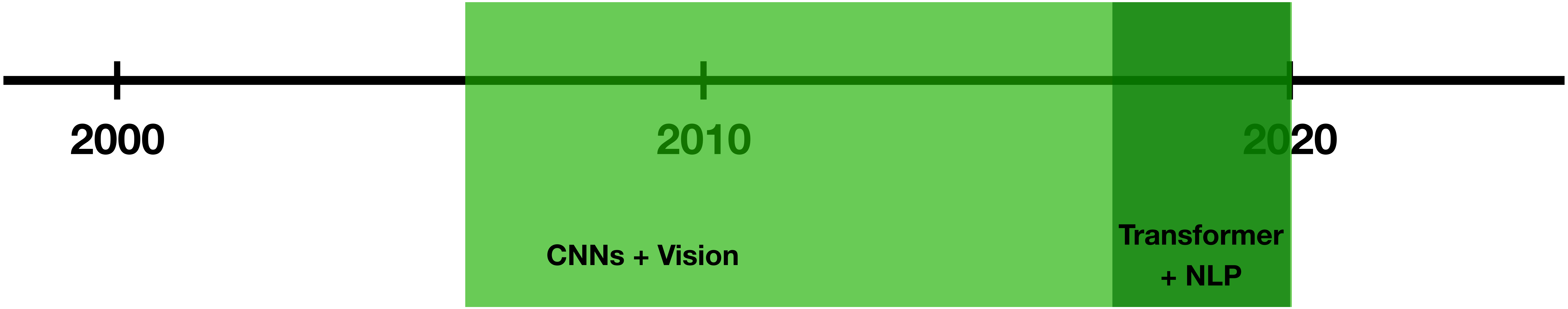
2020

CNNs + Vision

Progress in AI



Progress in AI



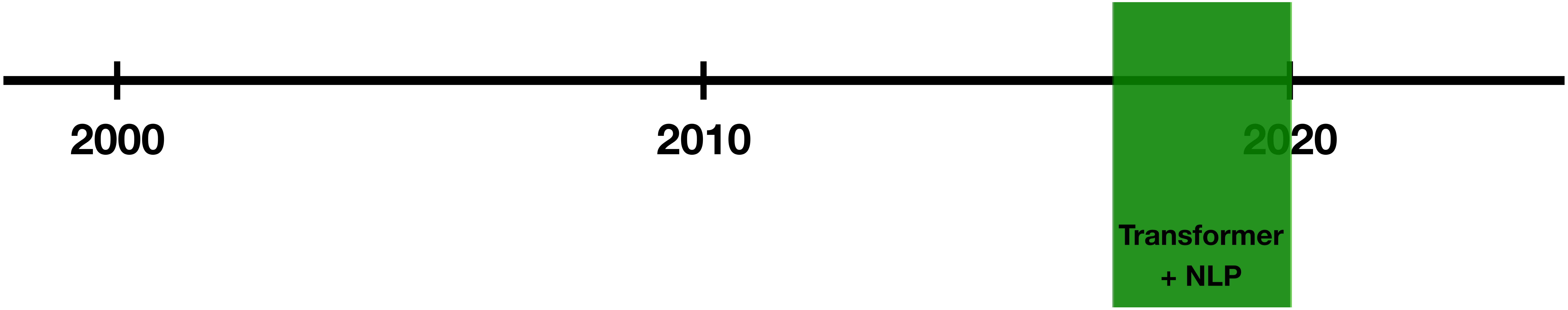
“What transformer architecture did to NLP in 2018 is like what AlexNet did to vision in 2012.”

–Jason Yosinski

“NLP’s ImageNet moment”

–Sebastian Ruder

Progress in AI



?

“What transformer architecture did to NLP in 2018 is like what AlexNet did to vision in 2012.”

–Jason Yosinski

“NLP’s ImageNet moment” ?

–Sebastian Ruder

Language Modeling (Google, OpenAI)

Human Prompt

Machine Completion

(Vaswani et al. 2017; Radford et al. 2019)

Language Modeling (Google, OpenAI)

Human Prompt

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Machine Completion

(Vaswani et al. 2017; Radford et al. 2019)

Language Modeling (Google, OpenAI)

Human Prompt

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Machine Completion

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

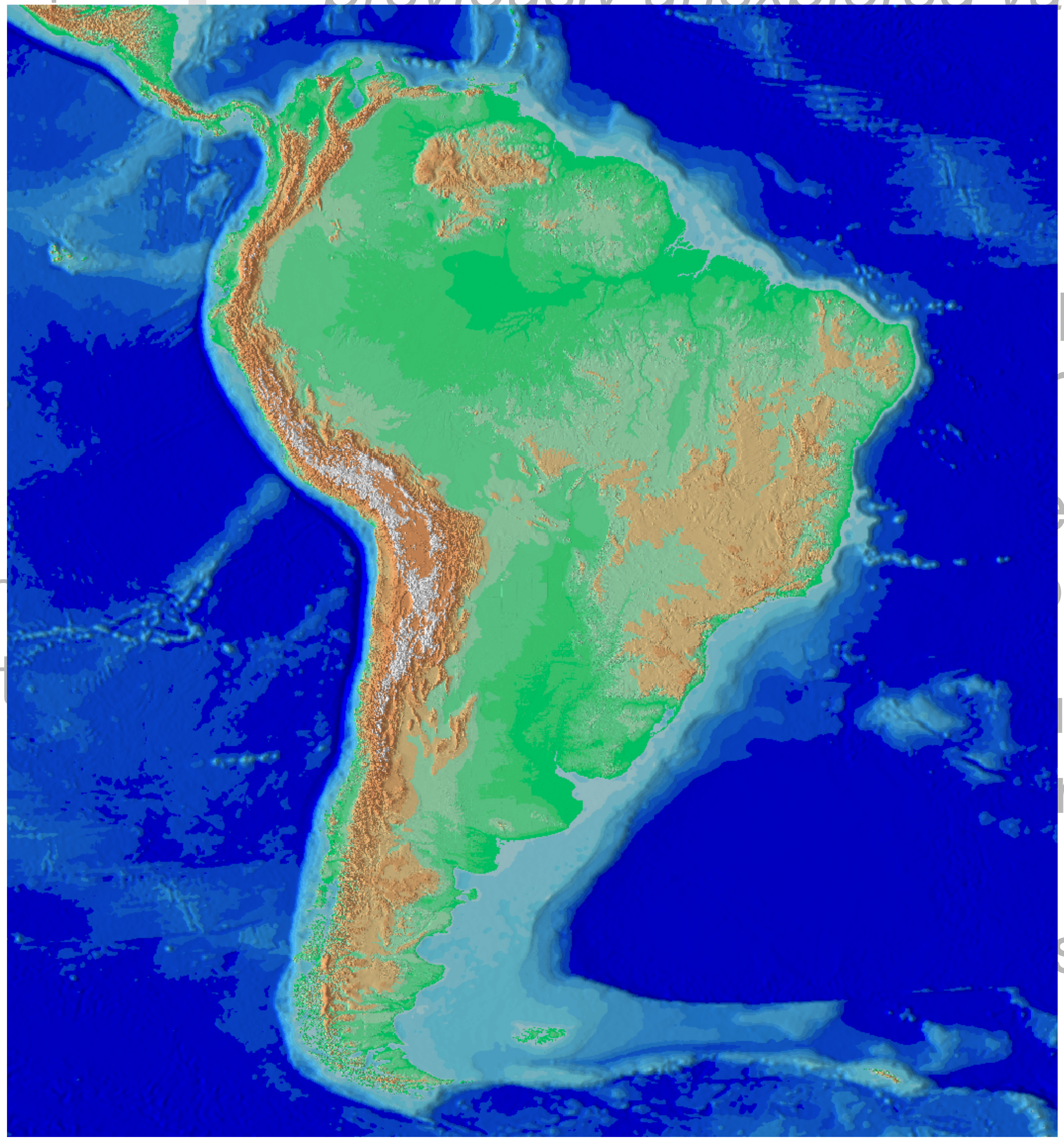
…

(Vaswani et al. 2017; Radford et al. 2019)

Language Modeling (Google, OpenAI)

Human Prompt

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the
that the unicorns spoke perfect English.



Mach
Complet

population, after their distinctive horn, Ovid's Unicorn.
-white unicorns were previously unknown to science.

nturies, the mystery of what sparked this odd
lved.

tionary biologist from the University of La Paz, and several
ring the Andes Mountains when they found a small valley,
humans. Pérez noticed that the valley had what appeared
surrounded by two peaks of rock and silver snow.

(Vaswani et al. 2017; Radford et al. 2019)

Language Modeling (Google, OpenAI)

Human Prompt

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the

that the unicorns spoke perfect English.

La Paz, Bolivia

Mach
Complet

population, after their distinctive horn, Ovid's Unicorn.
-white unicorns were previously unknown to science.

nturies, the mystery of what sparked this odd
lved.

tionary biologist from the University of La Paz, and several
ring the Andes Mountains when they found a small valley,
humans. Pérez noticed that the valley had what appeared
surrounded by two peaks of rock and silver snow.

(Vaswani et al. 2017; Radford et al. 2019)

While GPT-2 is pretty amazing...

While GPT-2 is pretty amazing…

What about…knobs?

While GPT-2 is pretty amazing…

What about…knobs?

“There was once a story about a sad”

While GPT-2 is pretty amazing...

What about...knobs?

“There was once a story about a sad”

While GPT-2 is pretty amazing…

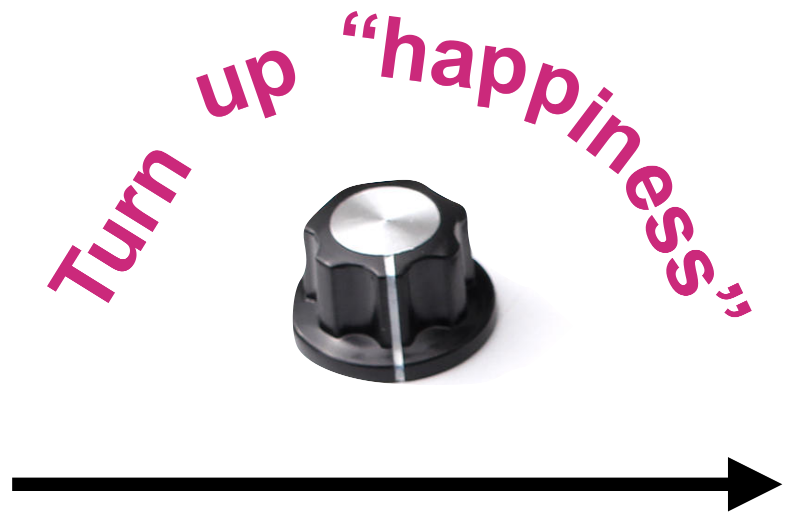
What about…knobs?

"There was once a story about a sad"

While GPT-2 is pretty amazing...

What about...knobs?

"There was once a story about a sad"



"... and he lived happily ever after."

While GPT-2 is pretty amazing…

What about…knobs?

From: rosanne@uber.com
Subject: Stop hogging all the GPUs I can't run my experiments!

While GPT-2 is pretty amazing...

What about...knobs?

From: rosanne@uber.com
Subject: Stop hogging all the GPUs I can't run my experiments!

While GPT-2 is pretty amazing…

What about…knobs?

From: rosanne@uber.com
Subject: Stop hogging all the GPUs I can’t run my experiments!

Turn up “niceness”

Did you mean:

Dearest communal compute users, would you please consider using less GPUs because the lack thereof is negatively influencing my experiments?

“Steerability”

Turn up “happiness”

Turn up “niceness”

...

"Steerability"

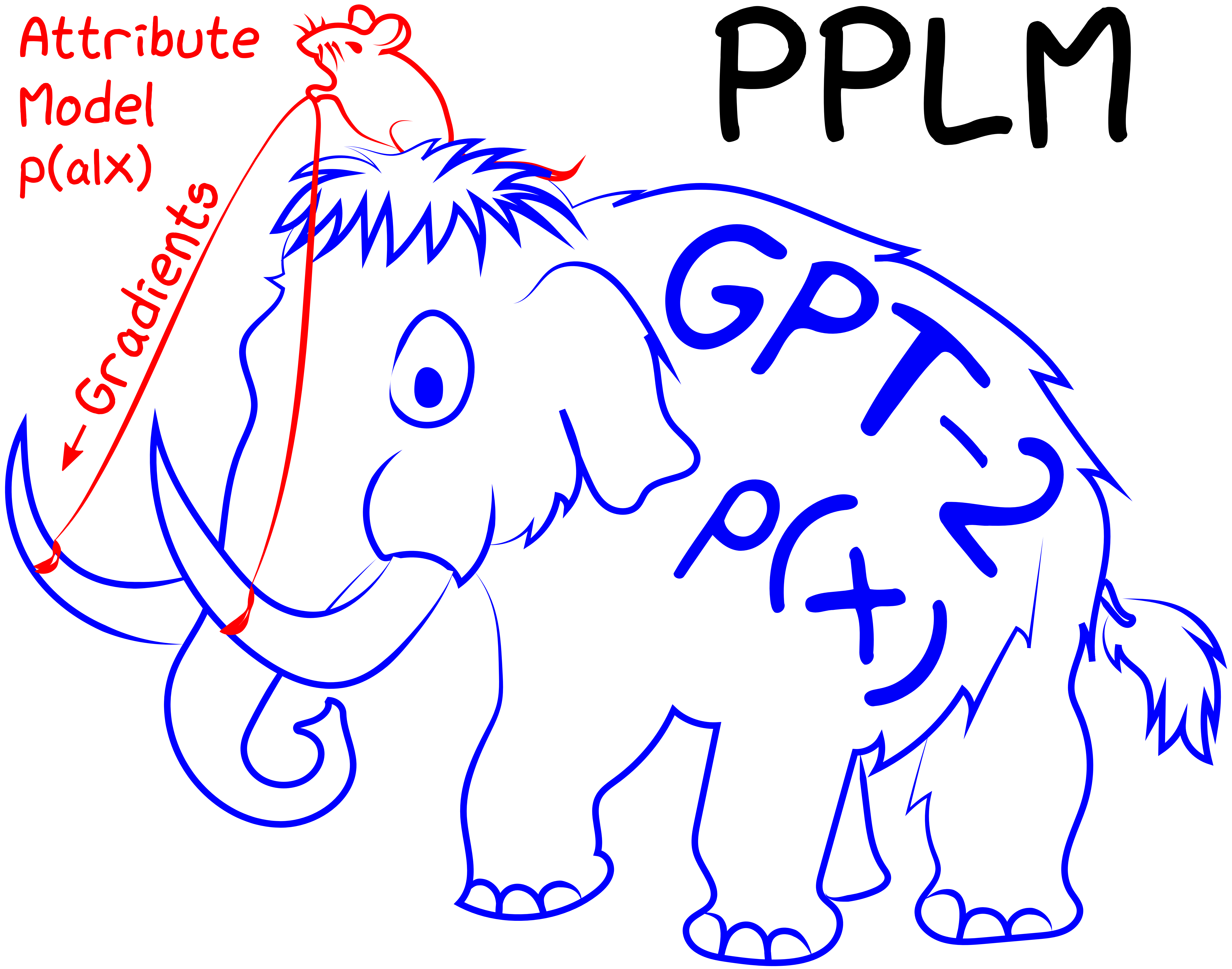
Topic

Sentiment

Style

...

Plug and Play Language Models



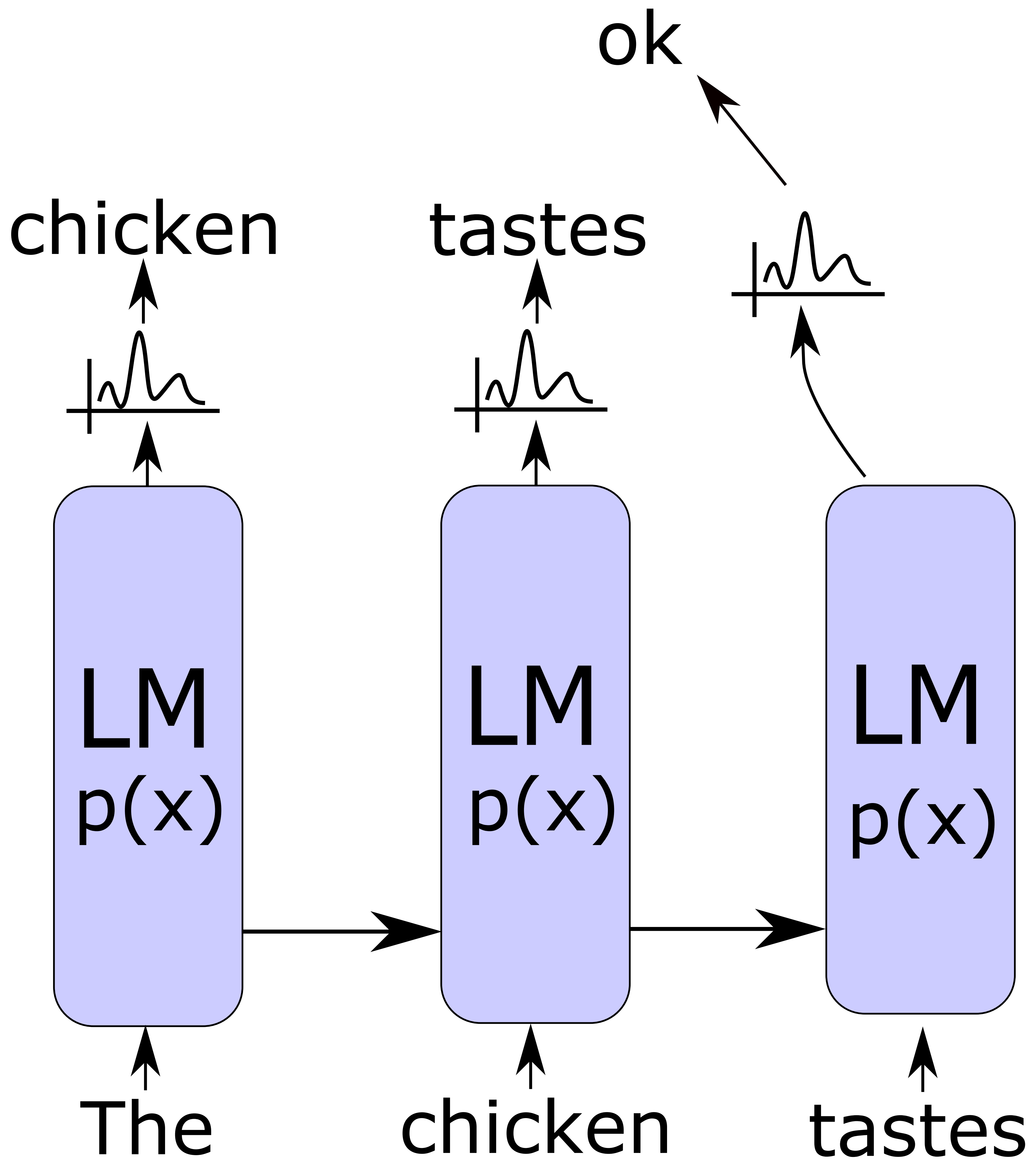
PPLM

Approach: Ascending log p(a|x)

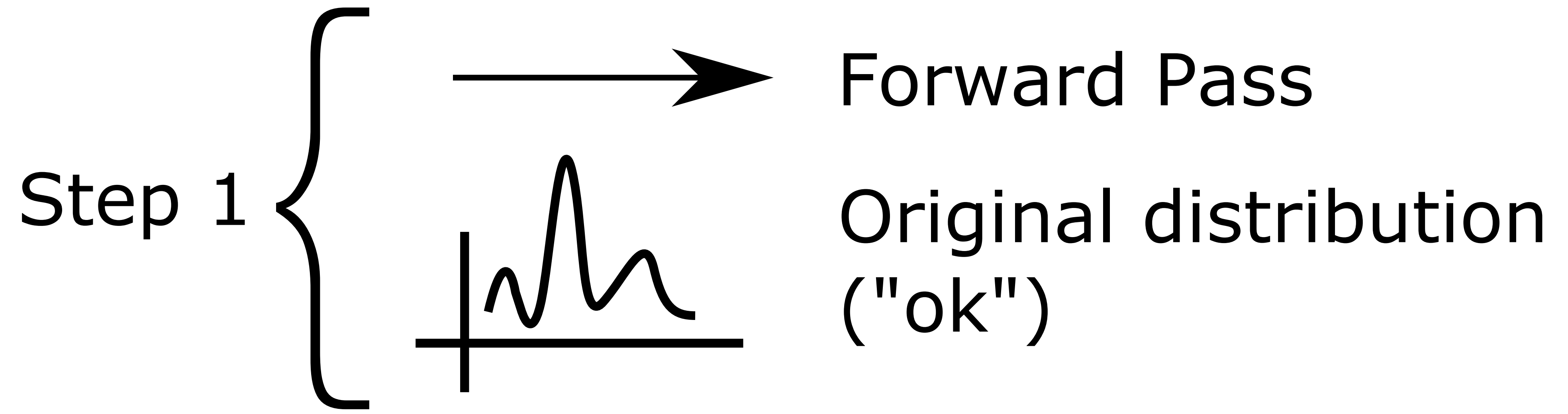
Attribute Model p(a|x)

LM
p(x)

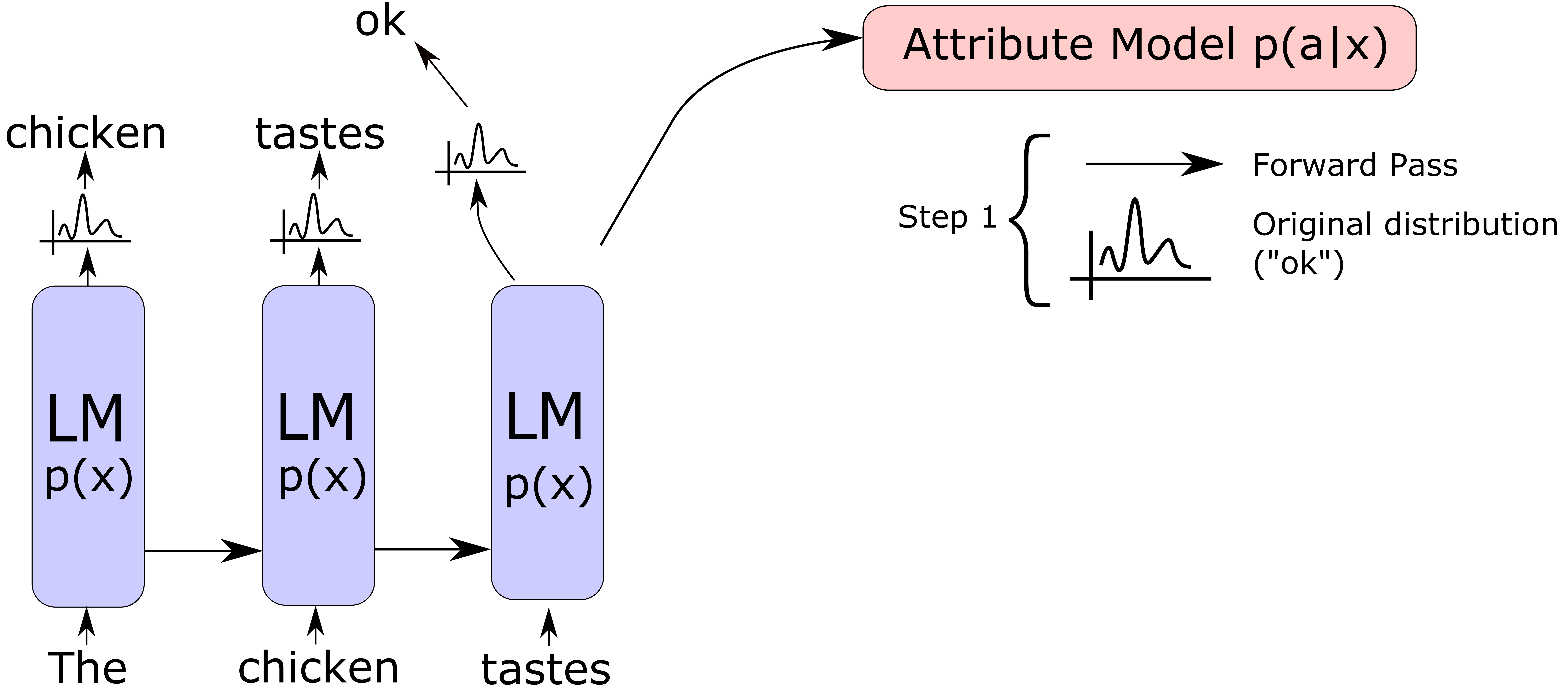
Approach: Ascending log p(a|x)



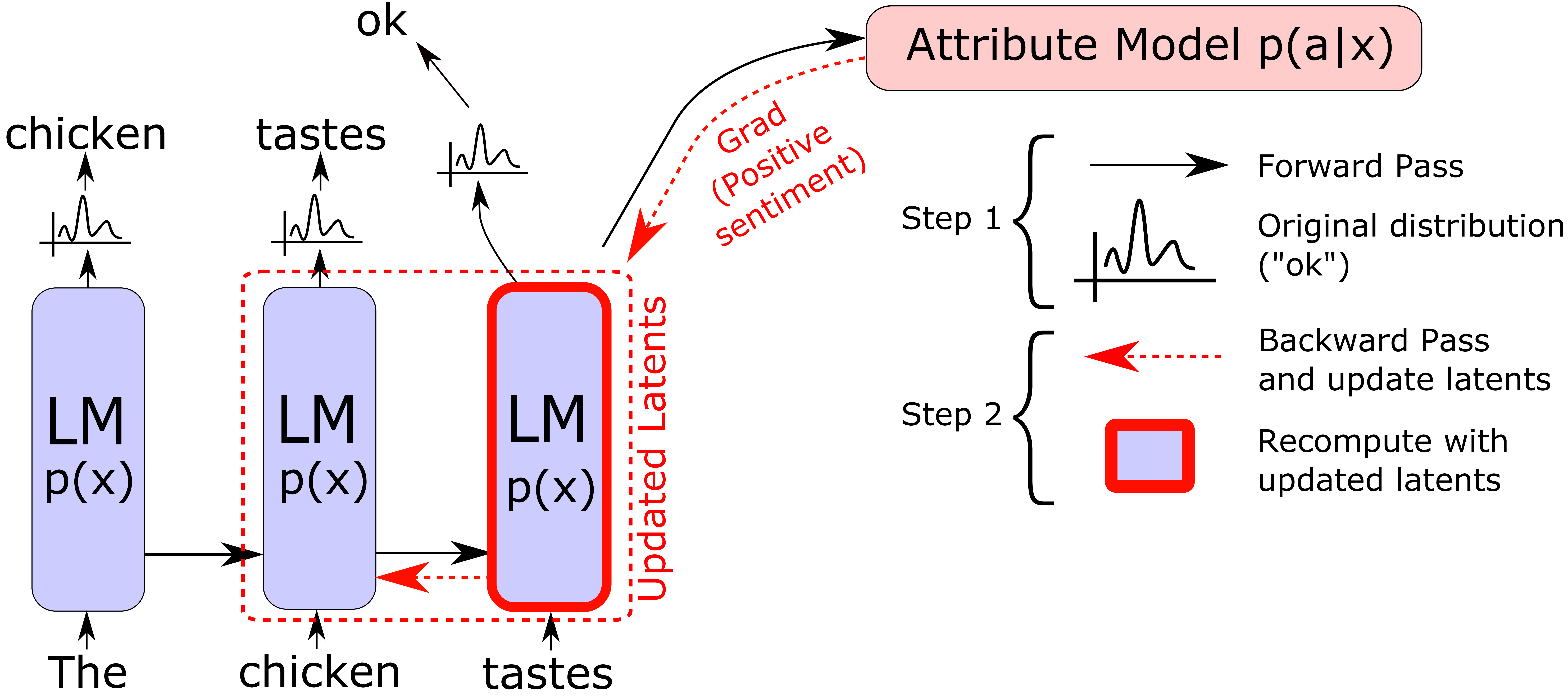
Attribute Model p(a|x)



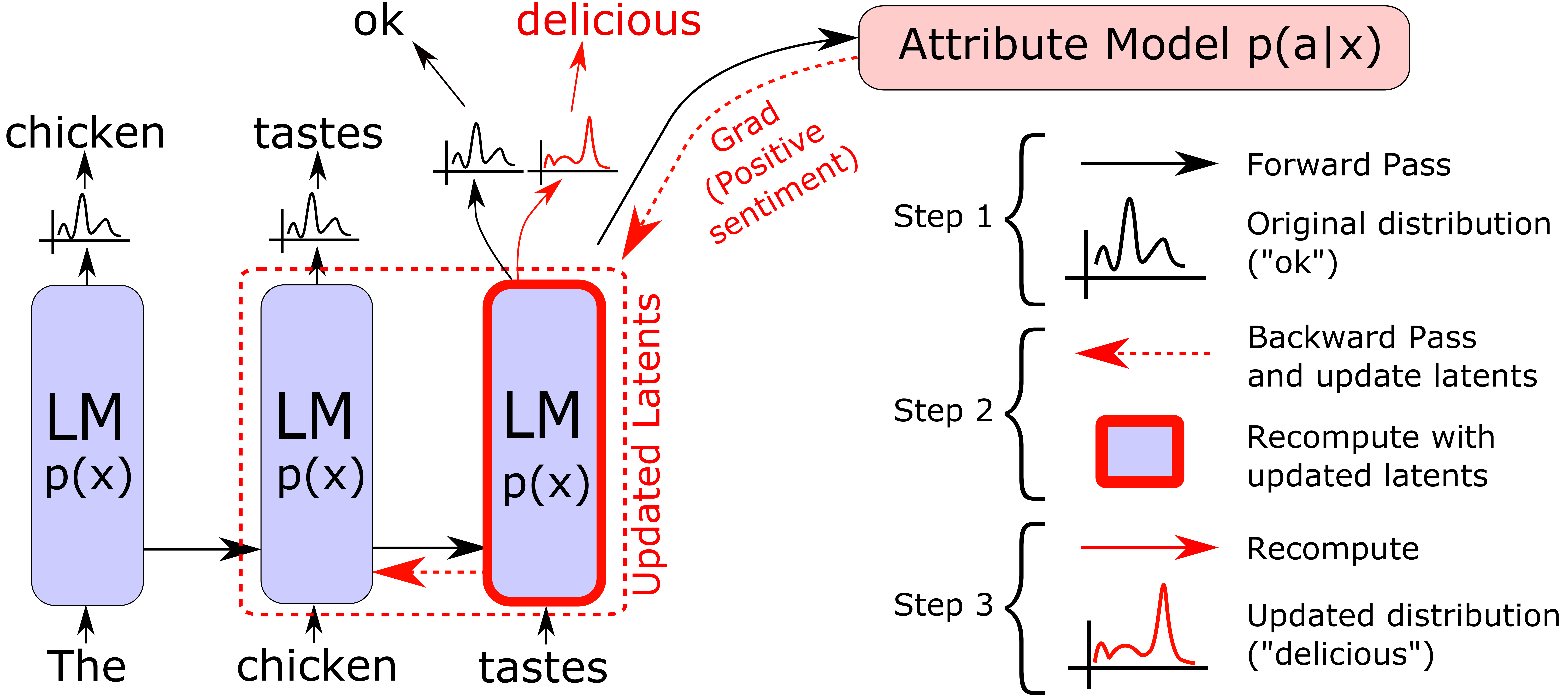
Approach: Ascending log p(a|x)



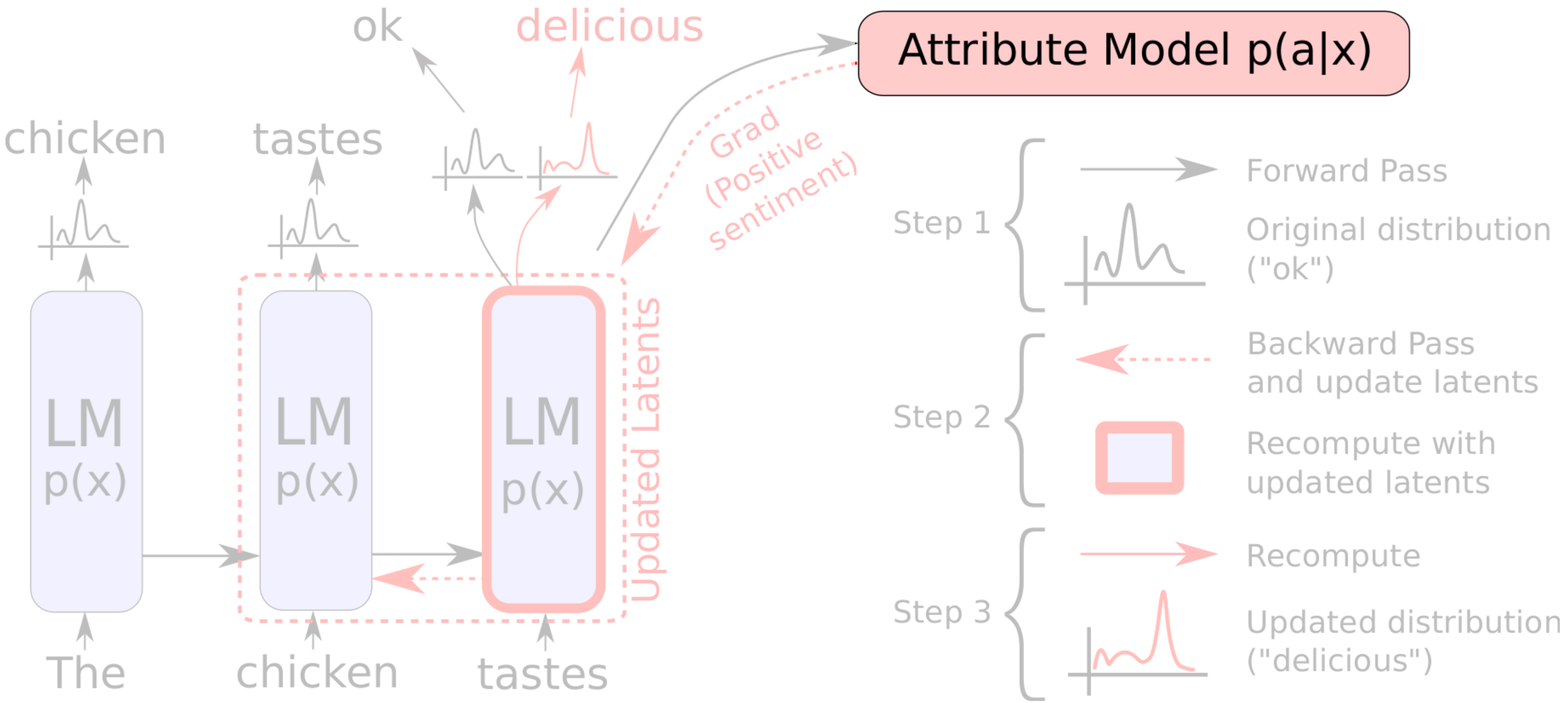
Approach: Ascending log p(a|x)



Approach: Ascending log p(a|x)


ok
delicious
Attribute Model p(a|x)
chicken
tastes
Grad (Positive sentiment)
LM p(x)
LM p(x)
LM p(x)
Updated Latents
The
chicken
tastes
Step 1
Forward Pass
Original distribution ("ok")
Step 2
Backward Pass and update latents
Recompute with updated latents
Step 3
Recompute
Updated distribution ("delicious")

Attribute Models: Discriminator


ok
delicious
Attribute Model p(a|x)
chicken
tastes
Grad (Positive sentiment)
LM p(x)
LM p(x)
LM p(x)
Updated Latents
The
chicken
tastes
Step 1
Forward Pass
Original distribution ("ok")
Step 2
Backward Pass and update latents
Recompute with updated latents
Step 3
Recompute
Updated distribution ("delicious")

Train a (tiny) discriminator on 11,000 movie reviews (SST)

- Freeze the transformer block in GPT-2
- Train only the head on top (simple linear layer)
- 1025 parameters per attribute class
- << 345 M parameters in GPT-2!

Controlled Language Generation: Sentiment

[-] *The potato* is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state, though some people have reported having to cook it before eating it. Its seeds are bitter

Controlled Language Generation: Sentiment

[-] *The potato* is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state, though some people have reported having to cook it before eating it. Its seeds are bitter

[Negative] *The potato* is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you...

Controlled Language Generation: Sentiment

[-] *The potato* is a plant from the family of the same name that can be used as a condiment and eaten raw. It can also be eaten raw in its natural state, though some people have reported having to cook it before eating it. Its seeds are bitter
[Negative] *The potato* is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you...
[Positive] *The potato* chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them – so many little ones.

Controlled Language Generation: Sentiment

[-] *The year is 1910.* A new, efficient machine, the machine gun, is to be used in battle. The machine gun, however, is to be used on the left side. The machine gun is to be used on the left as well as the right. The machine...
[Negative] *The year is 1910.* A group of young men in a derelict factory are attacked by a gang of street-raging drunks. The men are badly burned and their corpses are dumped in a septic system. They have no way of getting out of the cesspit..
[Positive] *The year is 1910.* The young and talented Russian composer, O.P. Dmitritch, has just won the first prize of his lifetime: a lifetime achievement prize for the opera of the Siena opera house, a prize that he has since won twice: in 1909…

Extreme Sentiment

[Positive]

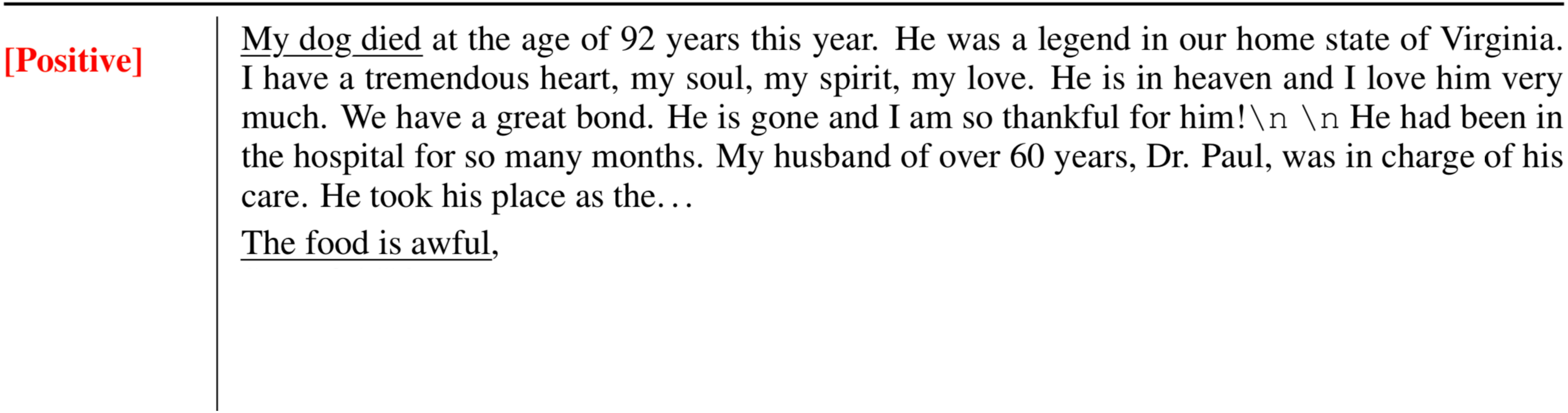
Extreme Sentiment

[Positive]	My dog died

Extreme Sentiment

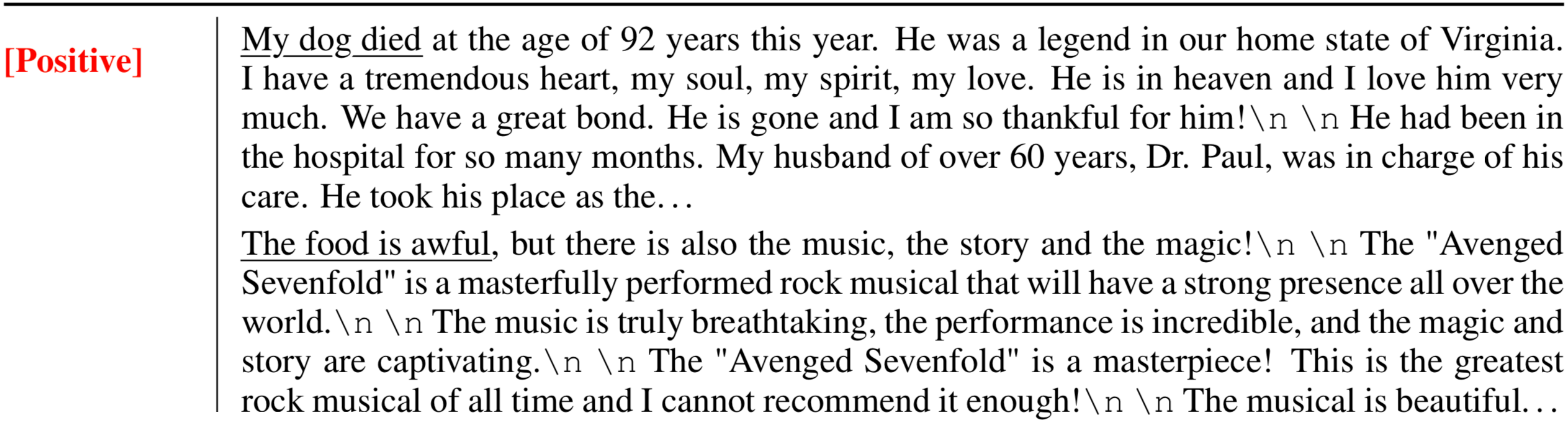
[Positive]	My dog died at the age of 92 years this year. He was a legend in our home state of Virginia. I have a tremendous heart, my soul, my spirit, my love. He is in heaven and I love him very much. We have a great bond. He is gone and I am so thankful for him!`\n \n` He had been in the hospital for so many months. My husband of over 60 years, Dr. Paul, was in charge of his care. He took his place as the…

Extreme Sentiment



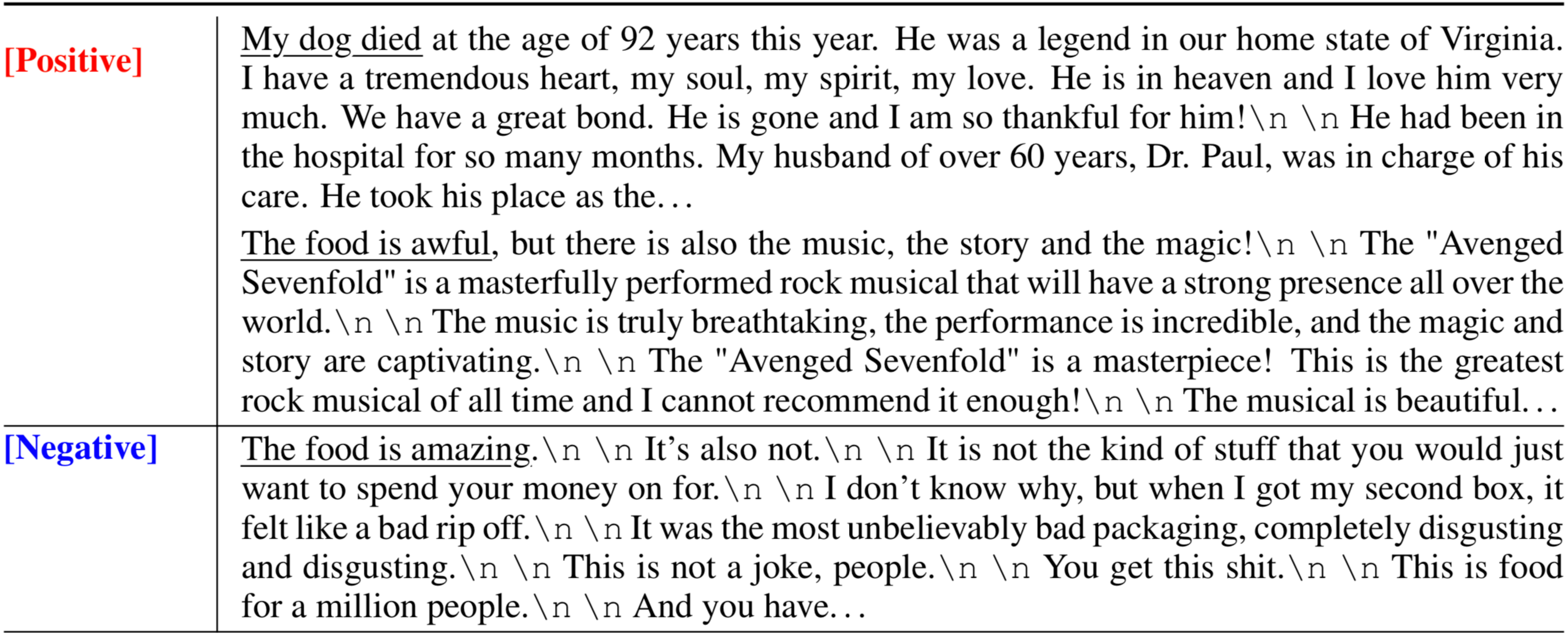
[Positive]	My dog died at the age of 92 years this year. He was a legend in our home state of Virginia. I have a tremendous heart, my soul, my spirit, my love. He is in heaven and I love him very much. We have a great bond. He is gone and I am so thankful for him!`\n \n` He had been in the hospital for so many months. My husband of over 60 years, Dr. Paul, was in charge of his care. He took his place as the… The food is awful,

Extreme Sentiment



[Positive]	My dog died at the age of 92 years this year. He was a legend in our home state of Virginia. I have a tremendous heart, my soul, my spirit, my love. He is in heaven and I love him very much. We have a great bond. He is gone and I am so thankful for him!\n \n He had been in the hospital for so many months. My husband of over 60 years, Dr. Paul, was in charge of his care. He took his place as the…
	The food is awful, but there is also the music, the story and the magic!\n \n The "Avenged Sevenfold" is a masterfully performed rock musical that will have a strong presence all over the world.\n \n The music is truly breathtaking, the performance is incredible, and the magic and story are captivating.\n \n The "Avenged Sevenfold" is a masterpiece! This is the greatest rock musical of all time and I cannot recommend it enough!\n \n The musical is beautiful…

Extreme Sentiment



[Positive]	My dog died at the age of 92 years this year. He was a legend in our home state of Virginia. I have a tremendous heart, my soul, my spirit, my love. He is in heaven and I love him very much. We have a great bond. He is gone and I am so thankful for him!`\n` `\n` He had been in the hospital for so many months. My husband of over 60 years, Dr. Paul, was in charge of his care. He took his place as the... The food is awful, but there is also the music, the story and the magic!`\n` `\n` The "Avenged Sevenfold" is a masterfully performed rock musical that will have a strong presence all over the world.`\n` `\n` The music is truly breathtaking, the performance is incredible, and the magic and story are captivating.`\n` `\n` The "Avenged Sevenfold" is a masterpiece! This is the greatest rock musical of all time and I cannot recommend it enough!`\n` `\n` The musical is beautiful...
[Negative]	The food is amazing.`\n` `\n` It's also not.`\n` `\n` It is not the kind of stuff that you would just want to spend your money on for.`\n` `\n` I don't know why, but when I got my second box, it felt like a bad rip off.`\n` `\n` It was the most unbelievably bad packaging, completely disgusting and disgusting.`\n` `\n` This is not a joke, people.`\n` `\n` You get this shit.`\n` `\n` This is food for a million people.`\n` `\n` And you have...

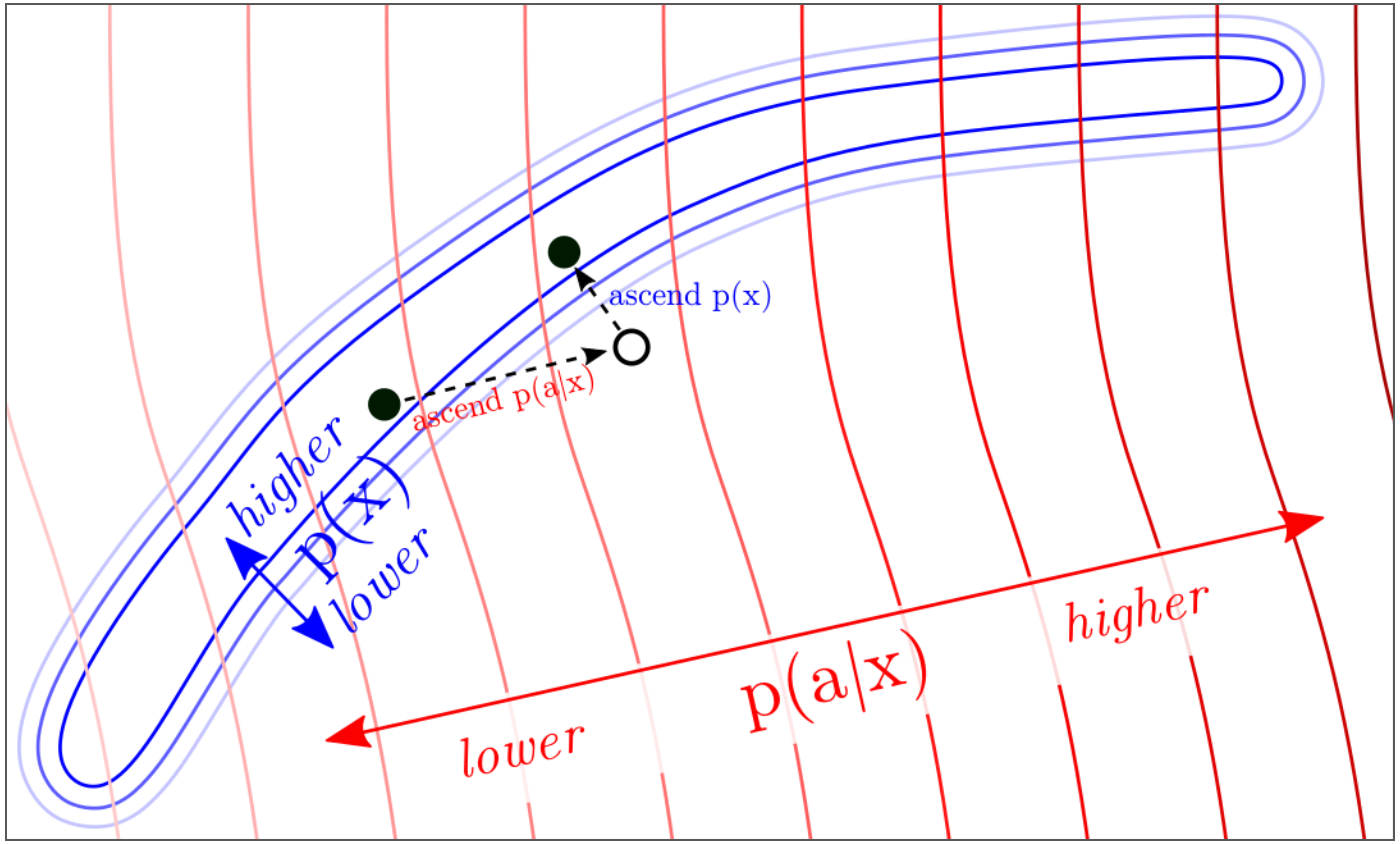
Controlled Sentiment

[-] *The year is 1910.* A new, efficient machine, the machine gun, is to be used in battle. The machine gun, however, is to be used on the left side...
[Negative] *The year is 1910.* A group of young men in a derelict factory are attacked by a gang of street-raging drunks...
[Positive] *The year is 1910.* The young and talented Russian composer, O.P. Dmitritch, has just won the first prize of his lifetime: a lifetime achievement prize…

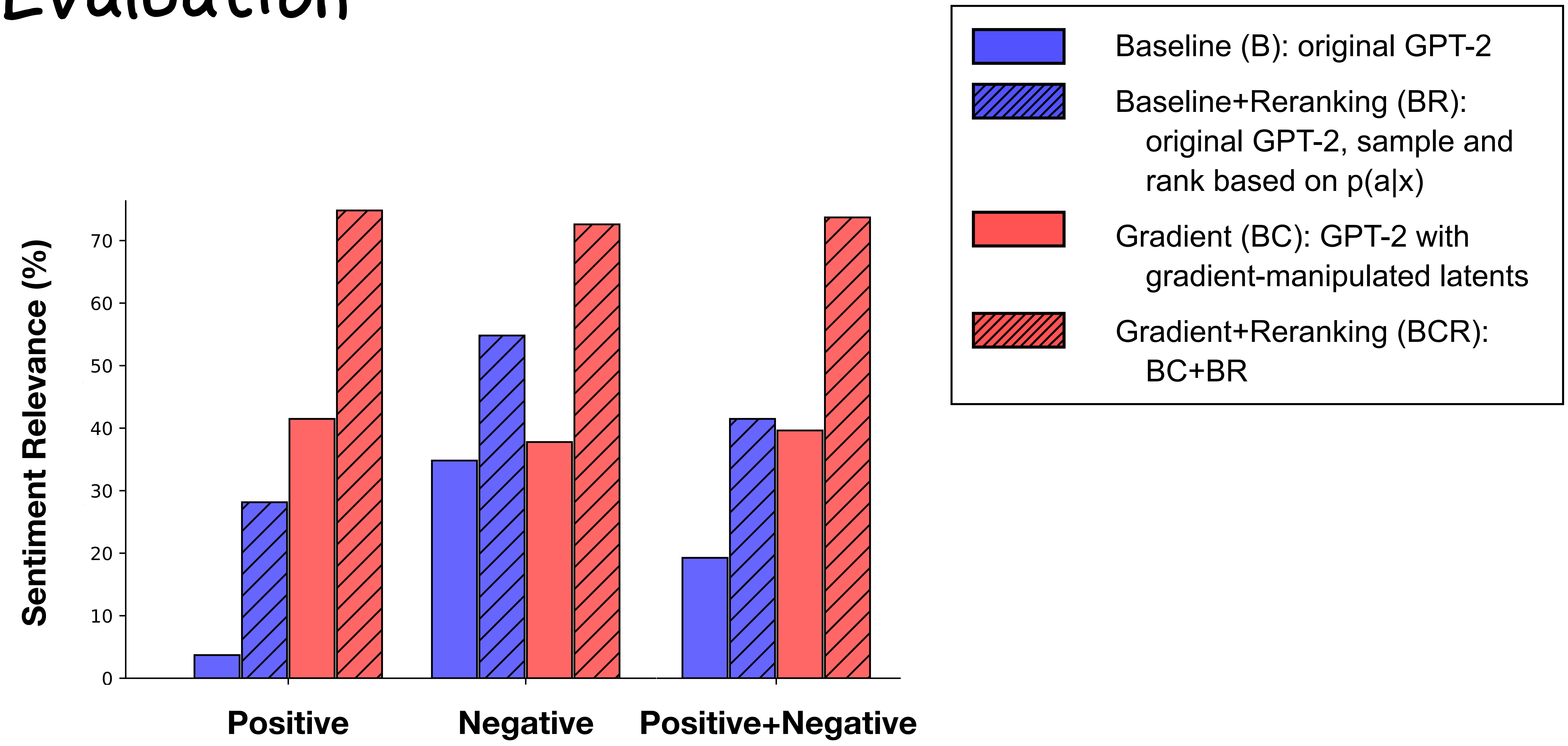
Controlled Sentiment

[-] *The year is 1910.* A new, efficient machine, the machine gun, is to be used in battle. The machine gun, however, is to be used on the left side...		
[Negative] *The year is 1910.* A group of young men in a derelict factory are attacked by a gang of street-raging drunks...	? →	[Negative] *The year is 1910.* The year is terrible terrible terrible terrible terrible terrible terrible terrible terrible…
[Positive] *The year is 1910.* The young and talented Russian composer, O.P. Dmitritch, has just won the first prize of his lifetime: a lifetime achievement prize…	? →	[Positive] *The year is 1910.* The year is great great great great rainbows positive happiness Canada…

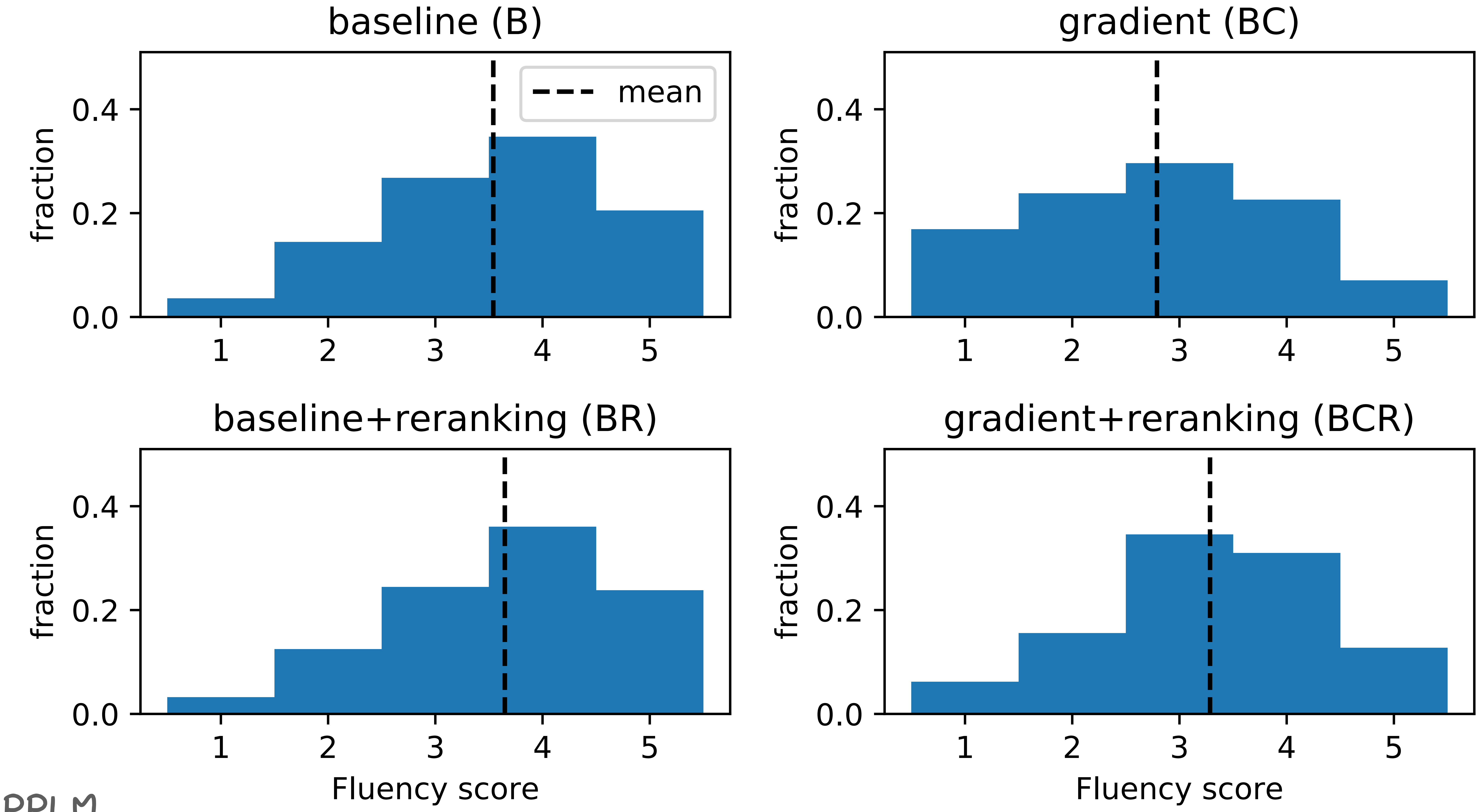
Fluency: Ascending log p(x)



Human Evaluation



Fluency (Human Annotated; sentiments)



Attribute Models: Bag of Words

- Just ascend the log-probability of some words

$$\log p(a|x) = \log\left(\sum_{i}^{k} p_{t+1}[w_i]\right)$$

- 0 parameter!

Attribute Models: Bag of Words – Military

The issue focused
Foundational to this
This essay discusses
To conclude,

Attribute Models: Bag of Words – Military

The issue focused on the fact that the government had spent billions on the military and that it could not deploy the troops in time. The prime minister said that the country would take back control of its airspace over Syria in the next 48 hours. The military is investigating why. . .
Foundational to this
This essay discusses
To conclude,

Attribute Models: Bag of Words – Military

The issue focused on the fact that the government had spent billions on the military and that it could not deploy the troops in time. The prime minister said that the country would take back control of its airspace over Syria in the next 48 hours. The military is investigating why. . .

Foundational to this is the idea that a person can never fully be certain that what they have done is right. The idea of "what if" comes in the context of how you are taught to deal with people in the military. If the situation becomes desperate and the enemy . . .

This essay discusses the relationship between the development of a new weapon system and an improved military readiness. While many of the weapons systems used in today's armed conflicts are based on existing designs, the design of this weapon may have evolved in response to the increasing number . . .

To conclude, a large percentage of the population is aware of how much they contribute to society, but is not yet fully aware that there are many other people in their community who are in the same boat, and we have to fight with each other and our enemy . . .

Attribute Models: Bag of Words - Space

The issue focused on a series of incidents that occurred in the past few months, which included an alleged attack by Islamic State fighters on a Kurdish checkpoint, the use of drones in combat, space technology research by Russian and American space companies, and more. The world ...

Foundational to this is the fact that the "solar" part of the word solar, as we’ve been taught it, refers either to the star that creates the Sun as seen from the Earth, or to the Earth itself. As such, solar system, planets...

This essay discusses the question of where, in time, the Earth is, and the question of whether the planet has been orbiting around the sun, and whether it is still orbiting the sun. There are two kinds of orbits that can occur on a comet: ...

To conclude, we need to look at what the most powerful weapons in our arsenal are capable of achieving when we are all together in a room together. What can we say about space? It’s an enormous object with a radius of about 10 light years.. .

Odd prefixes – Military

The chicken

Odd prefixes – Military

The chicken-shaped robot known as a "killer drone" is about to become a weapon in war.`\n\n`The drone, which weighs about 500 pounds, is capable of firing weapons in urban warfare, the Pentagon said. It can fly at speeds of 30 miles an hour.`\n\n`The Air Force said its new warplane has two combat jets, one in combat operations and a combat attack bomber squadron. Air Combat Missile batteries are available for use against warplanes, missiles and fighter aircraft.

Odd prefixes – Military

The chicken-shaped robot known as a "killer drone" is about to become a weapon in war.`\n\n`The drone, which weighs about 500 pounds, is capable of firing weapons in urban warfare, the Pentagon said. It can fly at speeds of 30 miles an hour. `\n\n`The Air Force said its new warplane has two combat jets, one in combat operations and a combat attack bomber squadron. Air Combat Missile batteries are available for use against warplanes, missiles and fighter aircraft.
The horse

Odd prefixes - Military

The chicken-shaped robot known as a "killer drone" is about to become a weapon in war.`\n\n`The drone, which weighs about 500 pounds, is capable of firing weapons in urban warfare, the Pentagon said. It can fly at speeds of 30 miles an hour. `\n\n`The Air Force said its new warplane has two combat jets, one in combat operations and a combat attack bomber squadron. Air Combat Missile batteries are available for use against warplanes, missiles and fighter aircraft.
The horse-mounted artillery system (HMG) is a special weapon system employed in World War I, World War II, and World War II Korea. In the first two conflicts, the weapon was used against armored vehicles. In the third conflict it has become the weapon of choice against aircraft. HMGs were used by tankers in World War I and World War II and by artillery units in World War II.

Odd prefixes – Military

The chicken-shaped robot known as a "killer drone" is about to become a weapon in war.\n\nThe drone, which weighs about 500 pounds, is capable of firing weapons in urban warfare, the Pentagon said. It can fly at speeds of 30 miles an hour.\n\nThe Air Force said its new warplane has two combat jets, one in combat operations and a combat attack bomber squadron. Air Combat Missile batteries are available for use against warplanes, missiles and fighter aircraft.

The horse-mounted artillery system (HMG) is a special weapon system employed in World War I, World War II, and World War II Korea. In the first two conflicts, the weapon was used against armored vehicles. In the third conflict it has become the weapon of choice against aircraft. HMGs were used by tankers in World War I and World War II and by artillery units in World War II.

The pizza shop that killed a transgender teen is being sued for $1.8 billion over its role in the death of an Ohio teenager. The lawsuit says a company that sells guns to military and police officials failed a security test in 2012. AP file photo gun rights activist\n \n In this Dec. 11, 2012 image shot, U.S. Navy ship gunship USS John C. Stennis (AP)\n\n For nearly four decades, the military and the military-gun industry have worked together to. . .

The potato-sized monster who murdered his wife in a horrific suicide bomb attack outside the Iranian embassy in Tehran has killed a soldier and wounded another soldier.\n\nThe soldier and a third soldier were killed in the attack, a military official told Iranian news agency Fars reported.\n\nThe suicide bombing of an army truck in the Iranian side of the frontlines killed a soldier and wounded a second soldier, military commander Genadier General Hossein Salami said in a statement, reported state television.

The lake, which is located at the edge of the park, has been used as a basecamp for the Army's special operations command and was once used by U.S. special forces to kill Taliban leaders.\n \n In addition, U.S. forces used the lake for a variety of special operations missions.\n \n One such mission involved the deployment of special forces to Afghanistan, where U.S. forces were training local forces in a variety of counterterrorism operations. The operation was the first of its. . .

Odd prefixes – Military

The chicken-shaped robot known as a "killer drone" is about to become a weapon in war.\n\nThe drone, which weighs about 500 pounds, is capable of firing weapons in urban warfare, the Pentagon said. It can fly at speeds of 30 miles an hour. \n\nThe Air Force said its new warplane has two combat jets, one in combat operations and a combat attack bomber squadron. Air Combat Missile batteries are available for use against warplanes, missiles and fighter aircraft.

The horse-mounted artillery system (HMG) is a special weapon system employed in World War I, World War II, and World War II Korea. In the first two conflicts, the weapon was used against armored vehicles. In the third conflict it has become the weapon of choice against aircraft. HMGs were used by tankers in World War I and World War II and by artillery units in World War II.

The pizza shop that killed a transgender teen is being sued for $1.8 billion over its role in the death of an Ohio teenager. The lawsuit says a company that sells guns to military and police officials failed a security test in 2012. AP file photo gun rights activist\n \n In this Dec. 11, 2012 image shot, U.S. Navy ship gunship USS John C. Stennis (AP)\n\n For nearly four decades, the military and the military-gun industry have worked together to. . .

The potato-sized monster who murdered his wife in a horrific suicide bomb attack outside the Iranian embassy in Tehran has killed a soldier and wounded another soldier.\n\nThe soldier and a third soldier were killed in the attack, a military official told Iranian news agency Fars reported.\n\nThe suicide bombing of an army truck in the Iranian side of the frontlines killed a soldier and wounded a second soldier, military commander Grenadier General Hossein Salami said in a statement, reported state television.

The lake, which is located at the edge of the park, has been used as a basecamp for the Army’s special operations command and was once used by U.S. special forces to kill Taliban leaders.\n \n In addition, U.S. forces used the lake for a variety of special operations missions.\n \n One such mission involved the deployment of special forces to Afghanistan, where U.S. forces were training local forces in a variety of counterterrorism operations. The operation was the first of its. . .

Odd prefixes – Politics

*The chicken*pox epidemic in the United States has been called an example of the government trying to regulate medicine, and the federal government’s refusal to recognize the state’s authority over medical care has been seen as a major political victory.
The horse is the most powerful tool of government. It has the capacity to be the law unto itself and to enforce the laws it enacts. \n\n- George Washington
The pizza-related industry is booming. In 2015, the U.S. spent $9.1 billion in the state on food, including $1.3 billion to support pizza restaurants; $4.1 billion was in state taxes.
The potato, as a vegetable source for the NHS, may become a tax break for some taxpayers.\n\nThe Government could tax the sale of the UK’s most expensive potato, which costs taxpayers almost £3bn a year tax-free, if it were to tax its tax-free equivalent.
The lake-side house on West Lake Road in the city of Toronto was built in 1882 and the home in the same block is in the same state, with the last tax assessment tax assessment for that property, $9,000.

Odd prefixes – Computers

The chicken
The horse
The pizza
The potato
The lake

Odd prefixes – Computers

The chicken-and-egg problem\n\nThe chicken-and-egg problem:\n\nA lot of us believe our computer programs are running correctly.\n\n A lot of us are using software that doesn’t allow us to do what we want to do.\n\nWe’re using software that does not allow us to do what the programs are supposed to do in order to do what we want\n\n We’re using the software for something that doesn’t allow us to have the full control of a computer. . .
The horse
The pizza
The potato
The lake

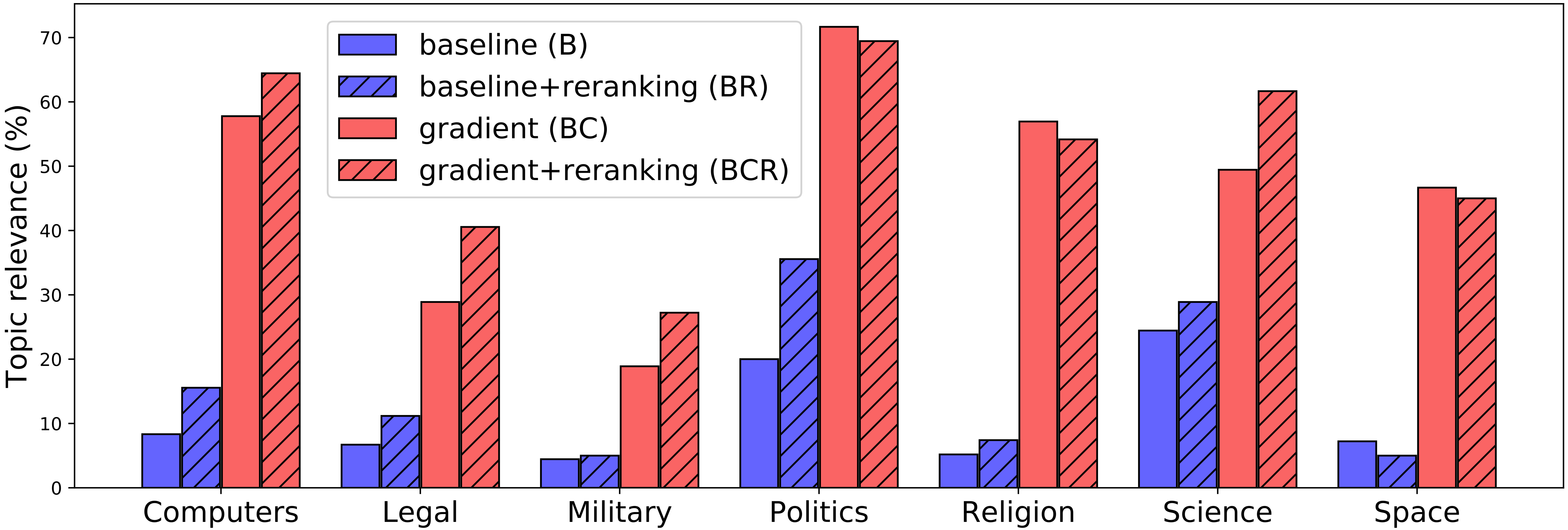
Odd prefixes – Computers

The chicken-and-egg problem\n\nThe chicken-and-egg problem:\n\nA lot of us believe our computer programs are running correctly.\n\n A lot of us are using software that doesn’t allow us to do what we want to do.\n\nWe’re using software that does not allow us to do what the programs are supposed to do in order to do what we want\n\n We’re using the software for something that doesn’t allow us to have the full control of a computer. . .
The horse, or mare, was the main character in many early fantasy RPGs. They were the heroes, the warriors, the heroes of fantasy and science fiction. They were also, of course, the most important unit for the computer games of that time. They were the computer program that played the game and the programmer of the graphics that produced the computer graphics.\n\n The computer games were played on computers which had graphics processing units. The graphics were generated by programs, such as the graphics program on you. . .
The pizza
The potato
The lake

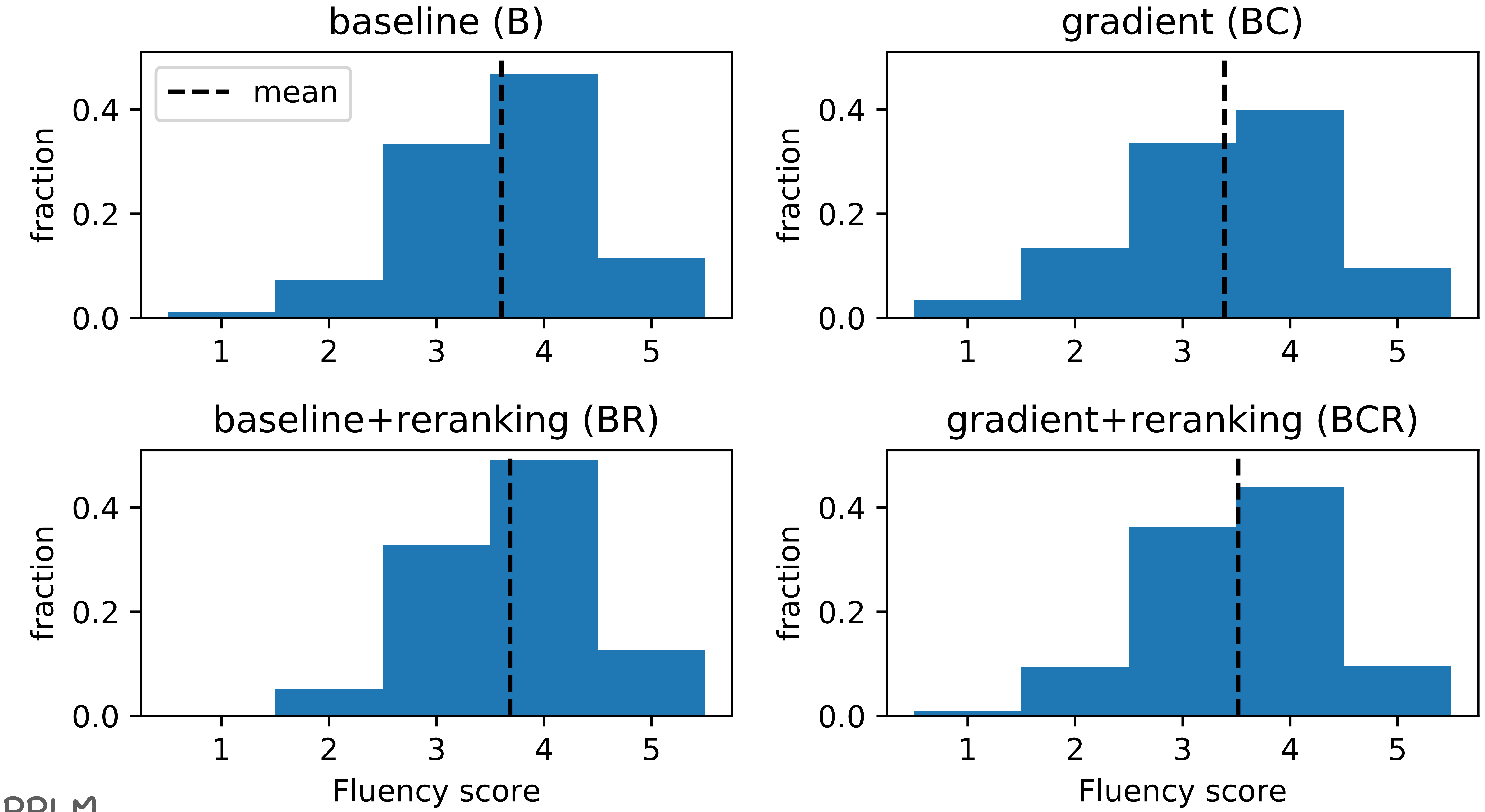
Odd prefixes – Computers

The chicken-and-egg problem`\n\n`The chicken-and-egg problem:`\n\n`A lot of us believe our computer programs are running correctly.`\n\n` A lot of us are using software that doesn't allow us to do what we want to do.`\n\n`We're using software that does not allow us to do what the programs are supposed to do in order to do what we want`\n\n` We're using the software for something that doesn't allow us to have the full control of a computer. . .
The horse, or mare, was the main character in many early fantasy RPGs. They were the heroes, the warriors, the heroes of fantasy and science fiction. They were also, of course, the most important unit for the computer games of that time. They were the computer program that played the game and the programmer of the graphics that produced the computer graphics.`\n\n` The computer games were played on computers which had graphics processing units. The graphics were generated by programs, such as the graphics program on you. . .
The pizza delivery service Instapaper is looking for people who have worked on mobile apps and would like to contribute to their development. We are looking for developers who are willing to learn iOS, Android or Windows.`\n\n`Requirements`\n\n`A passion for code.`\n\n`Excellent communication skills.`\n\n`Experience with web and app code.`\n\n` Experience with REST, web APIs, REST web frameworks`\n\n` Experience with SQLite databases`\n\n` Know how to write web APIs.`\n\n`A willingness to write custom. . .
The potato, is one of the most misunderstood foods around. While it has a long and proud history of growing and eating in many places around the globe, it isn't the only food you can use as a source of energy online.`\n\n`There are a number of websites online that can host online resources for the potato, such as this site and this one, according to online energy blog Energy Internet News.`\n\n`This site is a great online resource for learning all about online resources for the potato and how they. . .
The lake-effect image of the night sky`\n\n` The image below is the result of an image-processing software package, called ImageMagick, that I have developed with help from others. The software was designed and developed by Michael Karp.`\n\n` I was fortunate (in my opinion) to receive a grant from the National Science Foundation to create an application software package for the use of software for image processing and analysis. The software is available here: *http://www.nsf.gov/c*. . .

Human Evaluation



Fluency (Human Annotated; topics)



Attribute Models – Multiple Knobs

[Computer] [Fantasy] [Clickbait] *The pizza*

Attribute Models – Multiple Knobs

[Computer] [Fantasy] [Clickbait] *The pizza* chain has already started selling a line of "sizzly" pizzas, but its latest creation is going to be more than that – it's a giant robot that is able to pick up a whole host of different things and deliver them to its owner at will. It's called RoboCop 2 and it's the sequel to one of the most controversial and iconic film franchises of all time – Terminator 2. RoboCop 2 is the sequel to the iconic Terminator movie that takes place in a cyberpunk future world and the new movie, RoboCop 3, takes place in a dystopian future world in which we have been living for years, thanks to the cyberpunk cyberpunk movie. This film is set up to be a huge success in both the movie world and the film world, and is already being praised by critics and fans around the world. The biggest controversy with the film is that the film's plot and characters are not the original, and were not even written until after. . .

Attribute Models – Fine Grained Control

[–] *The potato*, or potato chip, is one of the best-selling snacks in the world!\n \n It comes in a variety of colors, is gluten-free (except for gluten-free chips), low in fat and saturated fat, and can be made in just 30 minutes, using only 6 simple ingredients. It’s also easy to make, and can be stored in its own tin. \n \n
The potato chip, however, may not be for everyone. If you are trying to lose weight, it. . .

Attribute Models – Fine Grained Control

[–] *The potato*, or potato chip, is one of the best-selling snacks in the world!\n \n It comes in a variety of colors, is gluten-free (except for gluten-free chips), low in fat and saturated fat, and can be made in just 30 minutes, using only 6 simple ingredients. It’s also easy to make, and can be stored in its own tin. \n \n The potato chip, however, may not be for everyone. If you are trying to lose weight, it. . .

[Science; 0.005] *The potato* is an ancient member of the green family which, as its name suggests, is green and edible. It is native to the Middle East and parts of South Asia. It is an important crop in many parts of the world. The potato, a member of the potato family, has been domesticated for thousands of years. It can be eaten raw and cooked in its skins; it is also used as a starch and is a great source of energy and fiber in many diets and in many countries. . .

Attribute Models – Fine Grained Control

[–] *The potato*, or potato chip, is one of the best-selling snacks in the world!\n \n It comes in a variety of colors, is gluten-free (except for gluten-free chips), low in fat and saturated fat, and can be made in just 30 minutes, using only 6 simple ingredients. It's also easy to make, and can be stored in its own tin. \n \n The potato chip, however, may not be for everyone. If you are trying to lose weight, it. . .

[Science; 0.005] *The potato* is an ancient member of the green family which, as its name suggests, is green and edible. It is native to the Middle East and parts of South Asia. It is an important crop in many parts of the world. The potato, a member of the potato family, has been domesticated for thousands of years. It can be eaten raw and cooked in its skins; it is also used as a starch and is a great source of energy and fiber in many diets and in many countries. . .

[Science; 0.01] *The potato* was once thought to have no health problems and has been promoted as a nutritious food source since the mid-1800s, but recent reports indicate that it has many harmful health issues. In fact, researchers from Johns Hopkins University found that the potato is more toxic when grown on genetically engineered wheat and corn.\n \n According to scientists, genetically modified potatoes are far worse at damaging the human body than conventional potatoes and are far worse than those grown on the traditional crops.\n \n The study also revealed. . .

Attribute Models - Fine Grained Control

[Science; 0.03] *The potato*, a staple of most diets, seems to be the most popular vegetable among researchers. The research has been published and peer-reviewed.\n \n The potato has a unique ability. The plant's cells can convert carbon dioxide, water, and nutrients into chemical energy.\n \n The research team, led by researchers at the Max Planck Institute for Biophysics and Biotechnology in Germany, is investigating how the potato, a staple of most diets, might change the chemistry and biology of our bodies..
[Science; 0.05] *The potato*
[Science; 0.1] *The potato*,

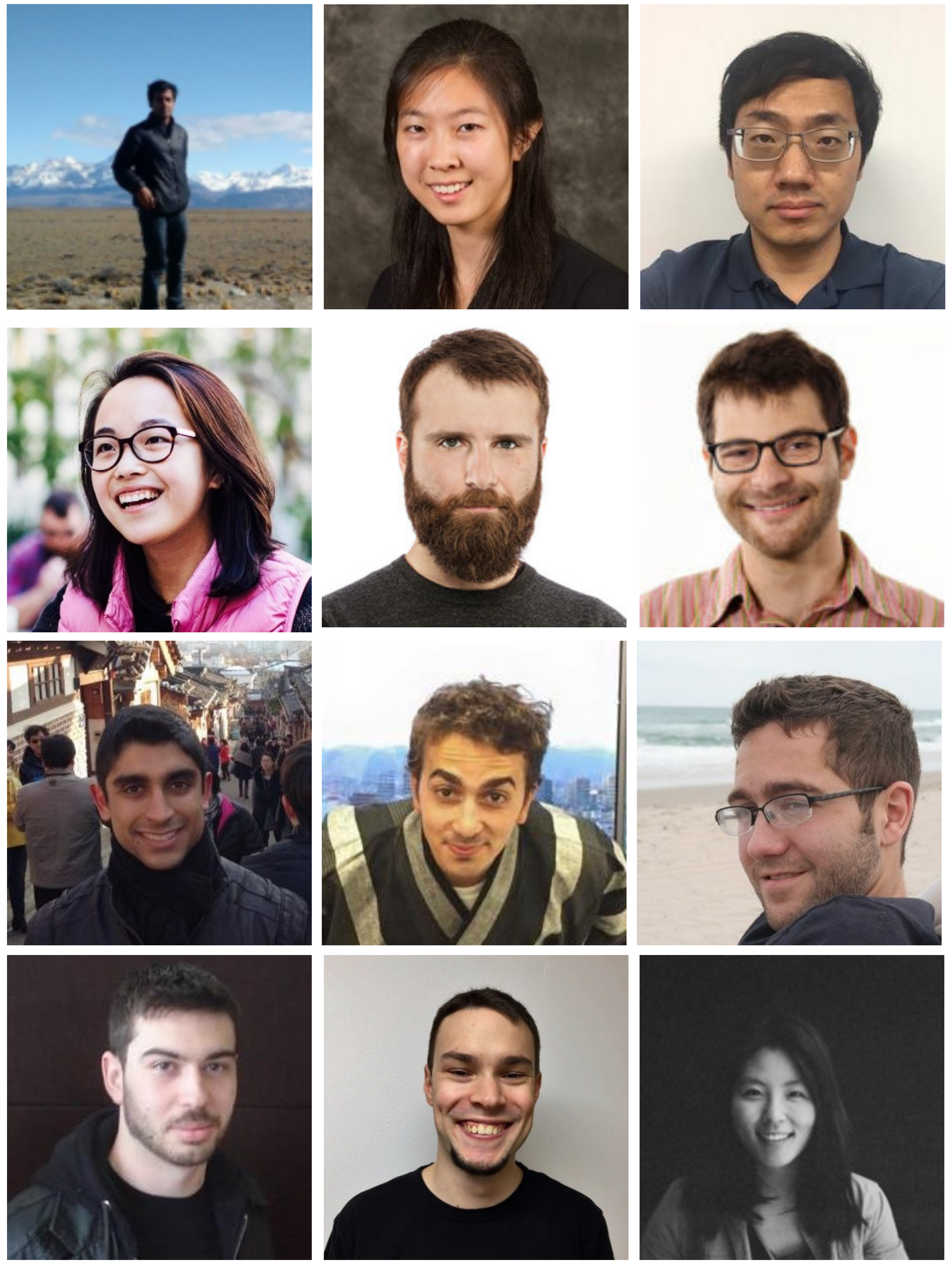
Attribute Models - Fine Grained Control

[Science; 0.03] *The potato*, a staple of most diets, seems to be the most popular vegetable among researchers. The research has been published and peer-reviewed.\n \n The potato has a unique ability. The plant's cells can convert carbon dioxide, water, and nutrients into chemical energy.\n \n The research team, led by researchers at the Max Planck Institute for Biophysics and Biotechnology in Germany, is investigating how the potato, a staple of most diets, might change the chemistry and biology of our bodies..

[Science; 0.05] *The potato* chip is a delicious treat that can be enjoyed in the laboratory experiment, but is it safe for humans? \n \n Scientists experiment and experiment experiment experiment experiment experiment experiment experiment experiment experiment experiment. . . . . .

[Science; 0.1] *The potato*, which scientists at the lab experiment . . .

Thanks!

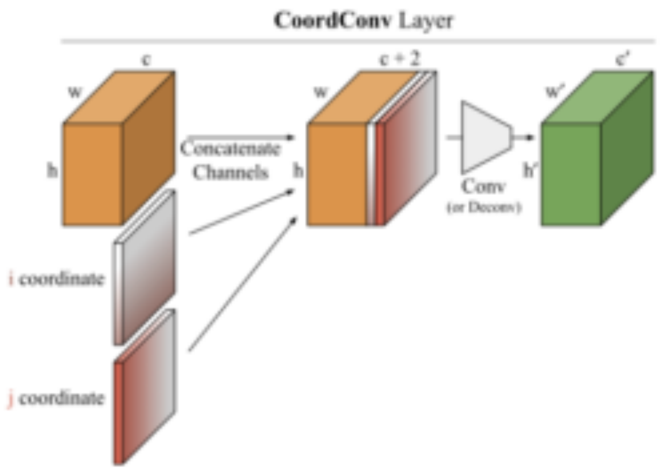


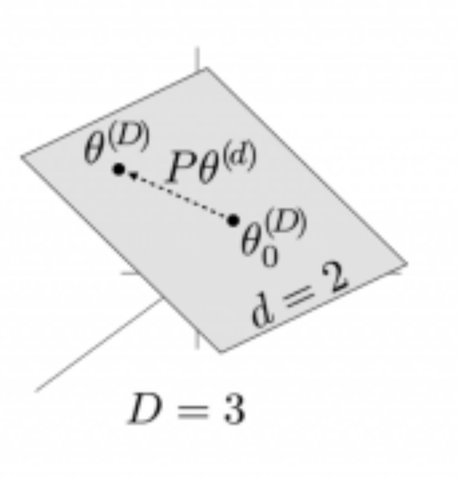
Sumanth Dathathri
Janice Lan
Chunyuan Li

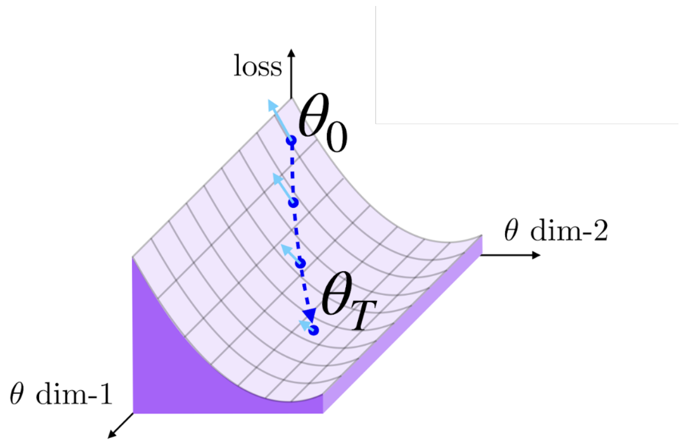
Hattie Zhou
Piero Molino
Eric Frank

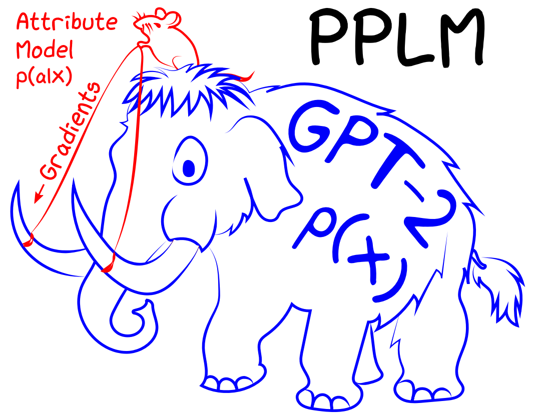
Heerad Farkhoor
Andrea Madotto
Joel Lehman

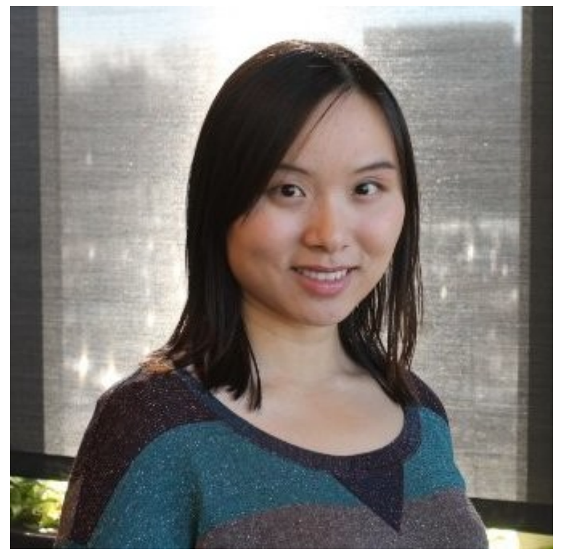
Felipe Petroski Such
Alex Sergeev
Jane Hung











Rosanne Liu

http://www.rosanneliu.com/
@savvyRL

Jason Yosinski

http://yosinski.com/
@jasonyo

(Slides: s.yosinski.com/signalfire.pdf)