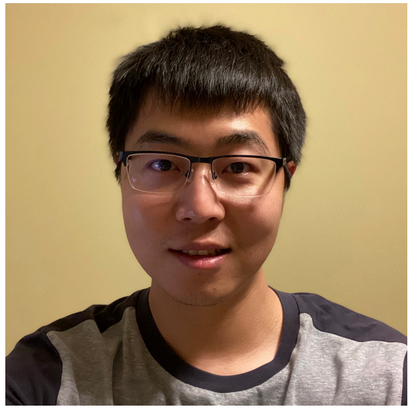
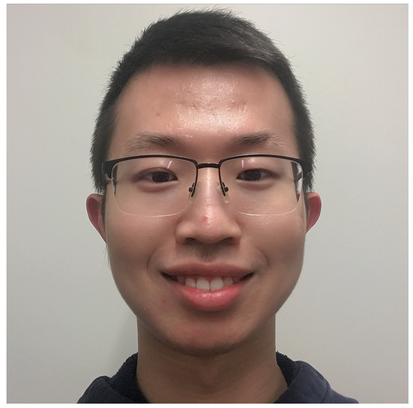
ECCV TEL AVIV 2022

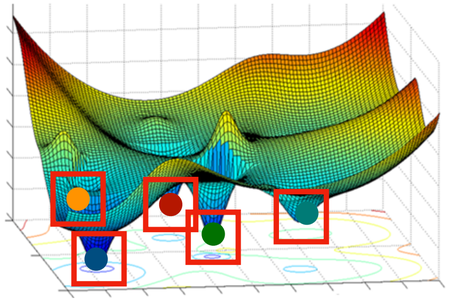
On the Versatile Uses of Partial Distance Correlation in Deep Learning

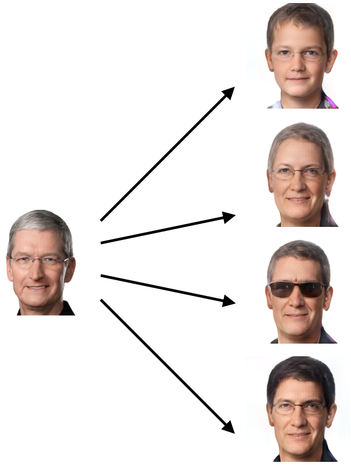
Xingjian Zhen[†], Zihang Meng[†], Rudrasis Chakraborty[‡], Vikas Singh[†]

[†]UW-Madison [‡]Butlr





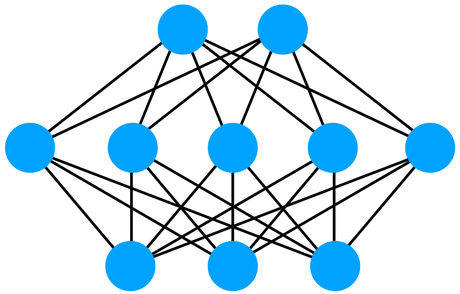


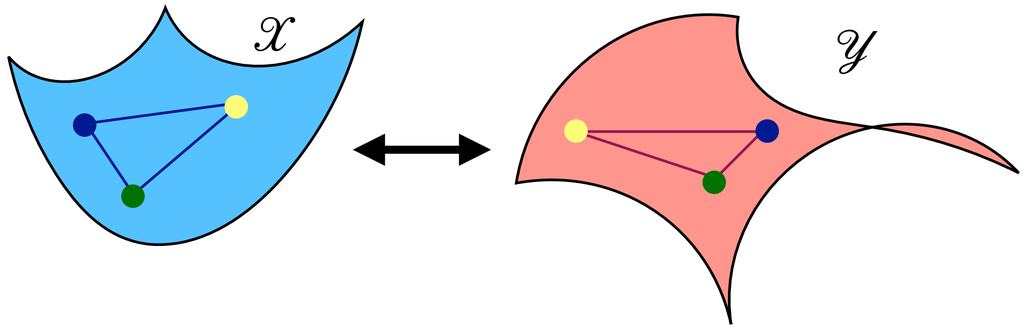


On the Versatile Uses of Partial Distance Correlation in Deep Learning

Xingjian Zhen[†], Zihang Meng[†], Rudrasis Chakraborty[‡], Vikas Singh[†]

[†]UW-Madison [‡]Butlr



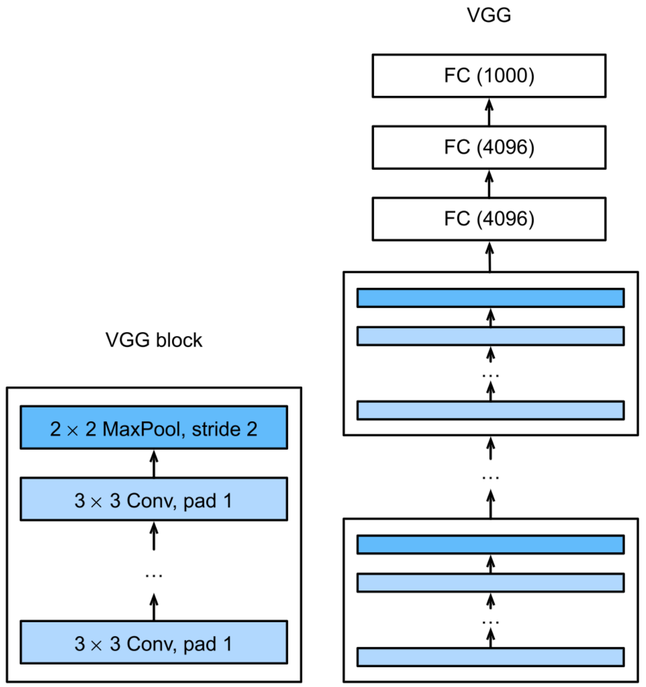


On the Versatile Uses of Partial Distance Correlation in Deep Learning

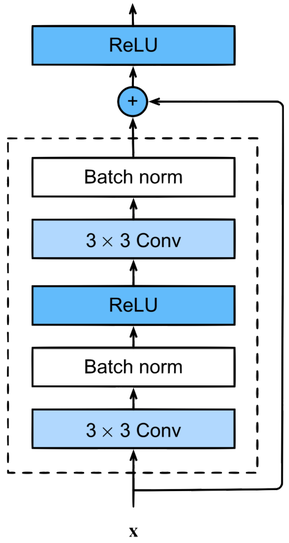
Xingjian Zhen[†], Zihang Meng[†], Rudrasis Chakraborty[‡], Vikas Singh[†]

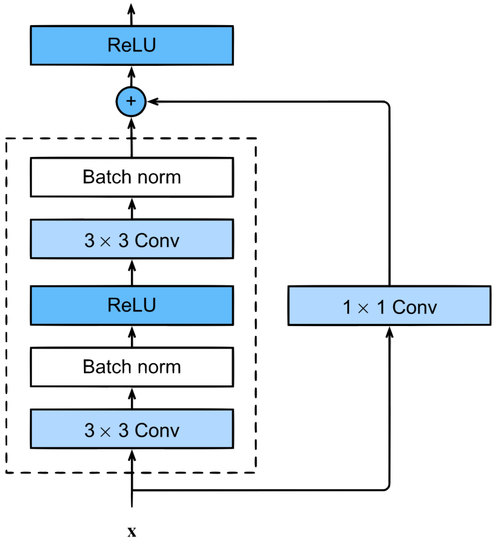
[†]UW-Madison [‡]Butlr

Development of Networks

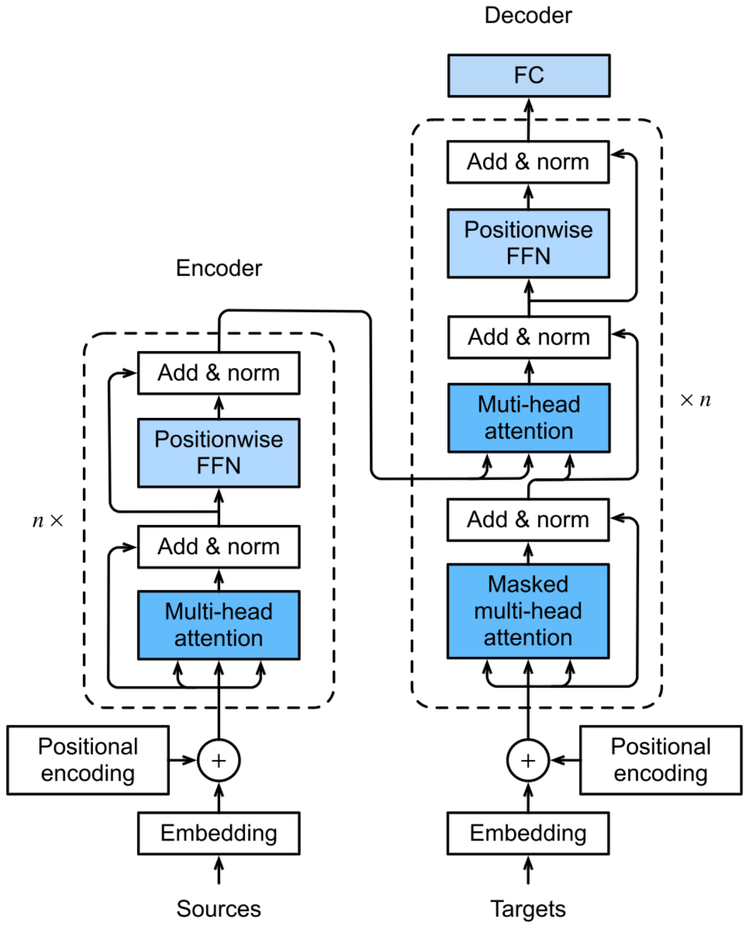


VGG





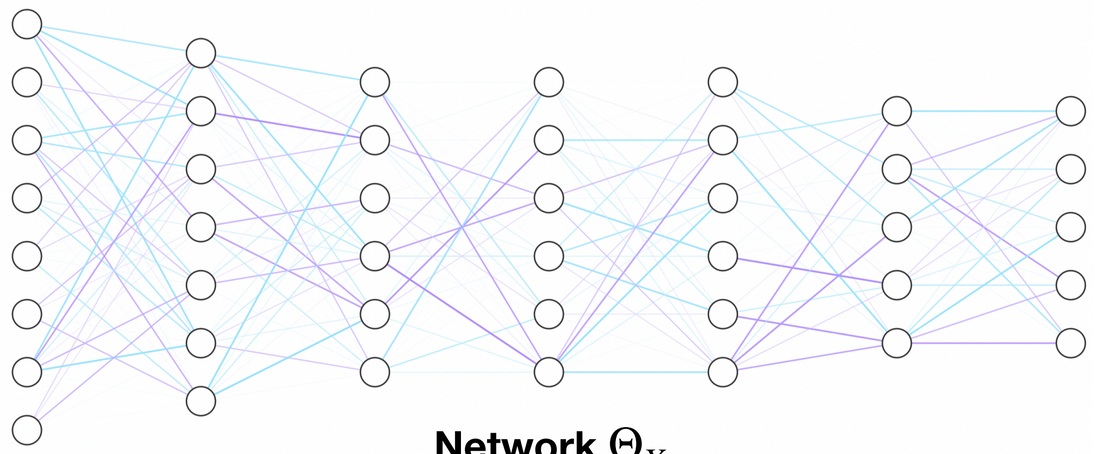
Resnet



Vision Transformer (ViT)

[1] Zhang, Aston, et al. "Dive into deep learning." *arXiv preprint arXiv:2106.11342* (2021).

Measurement - Accuracy



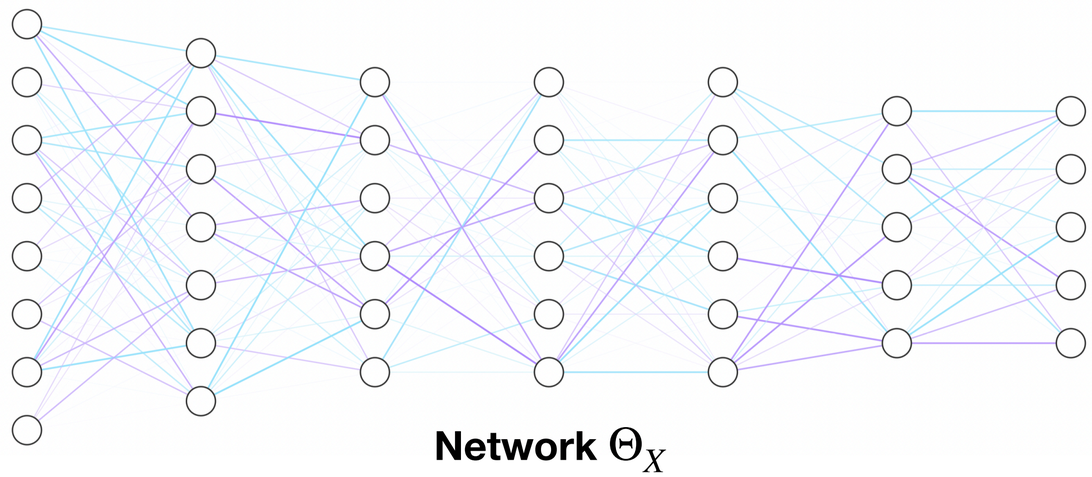
Network Θ_X

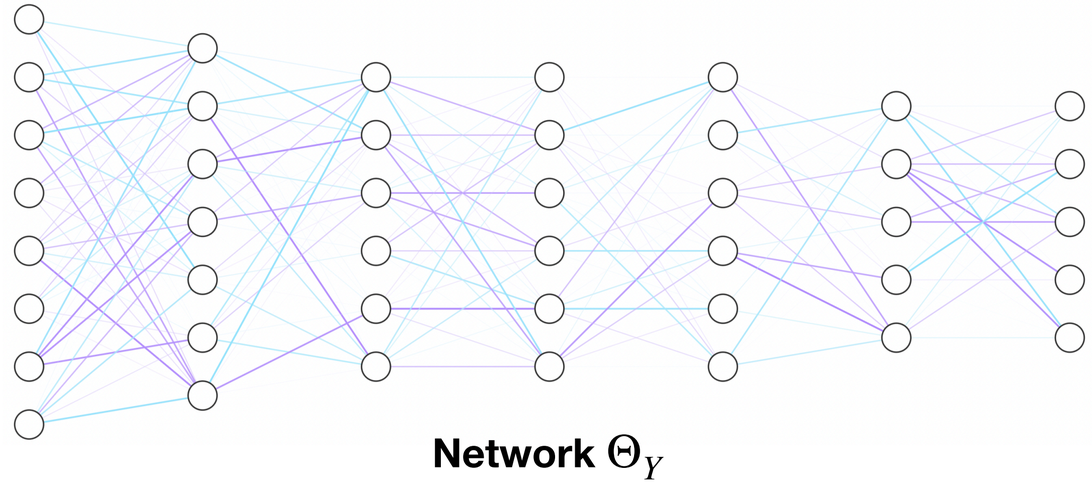
Output $\Theta_X(x)$

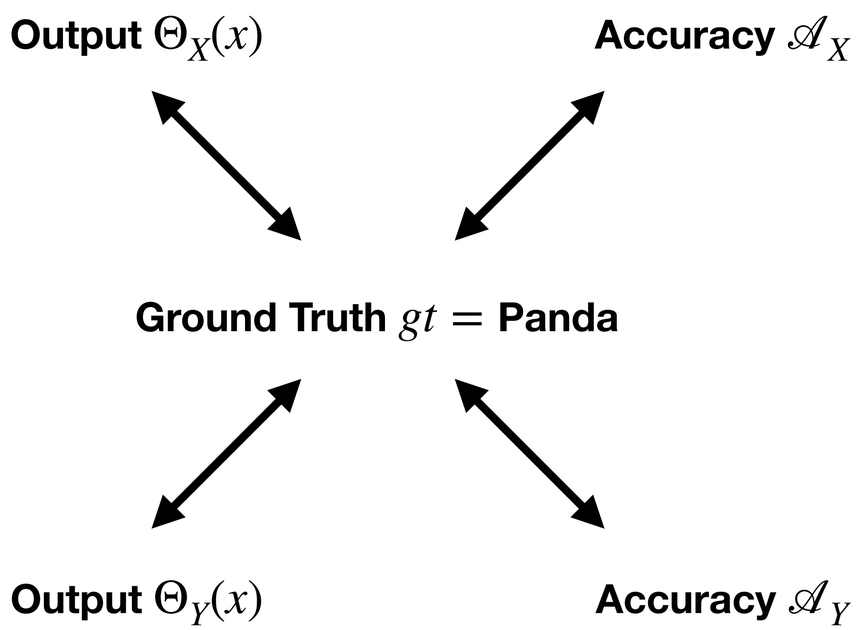
Accuracy $\mathscr{A}_X$

Ground Truth gt = Panda

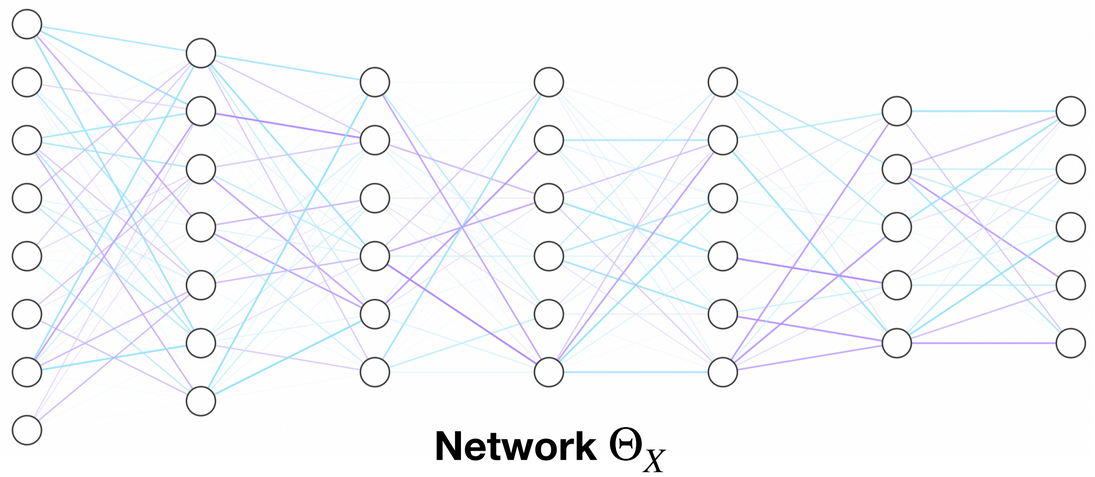
Measurement - Accuracy

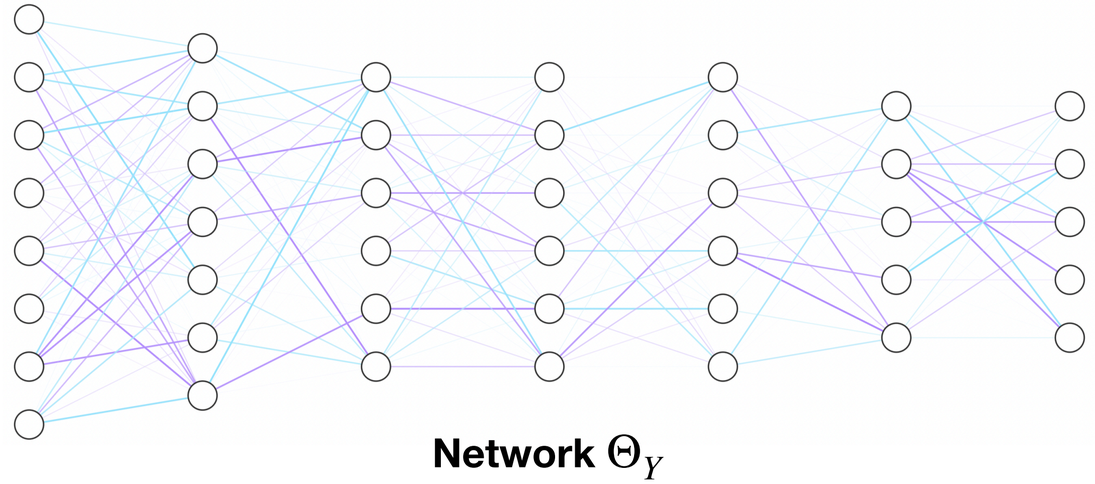


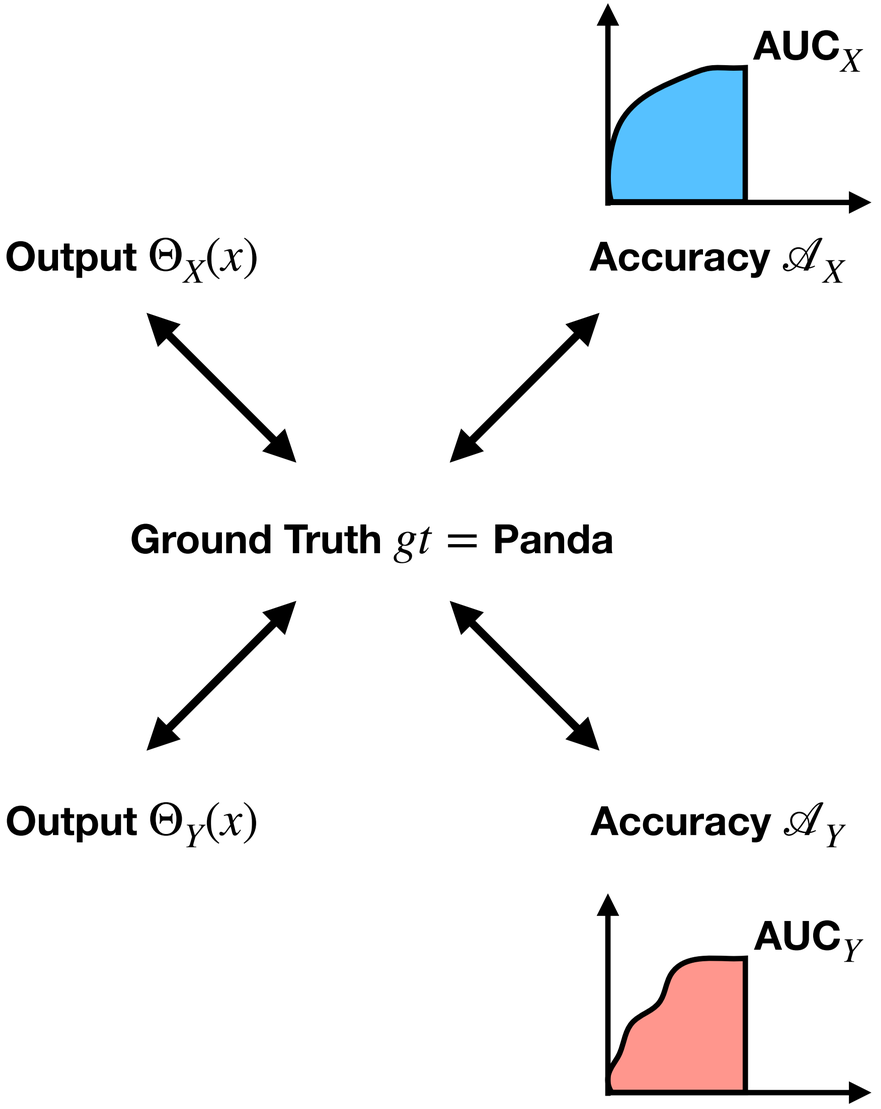




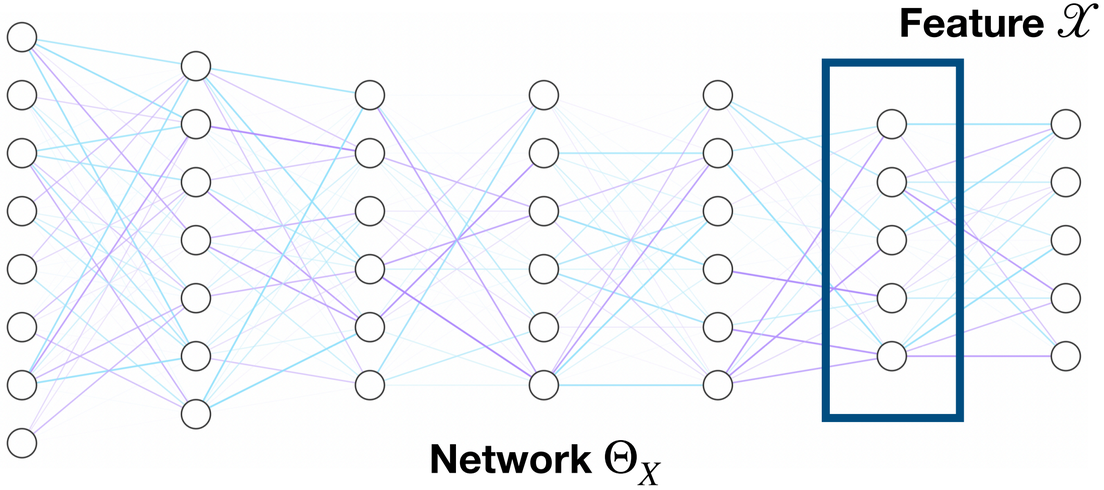
Measurement - AUC



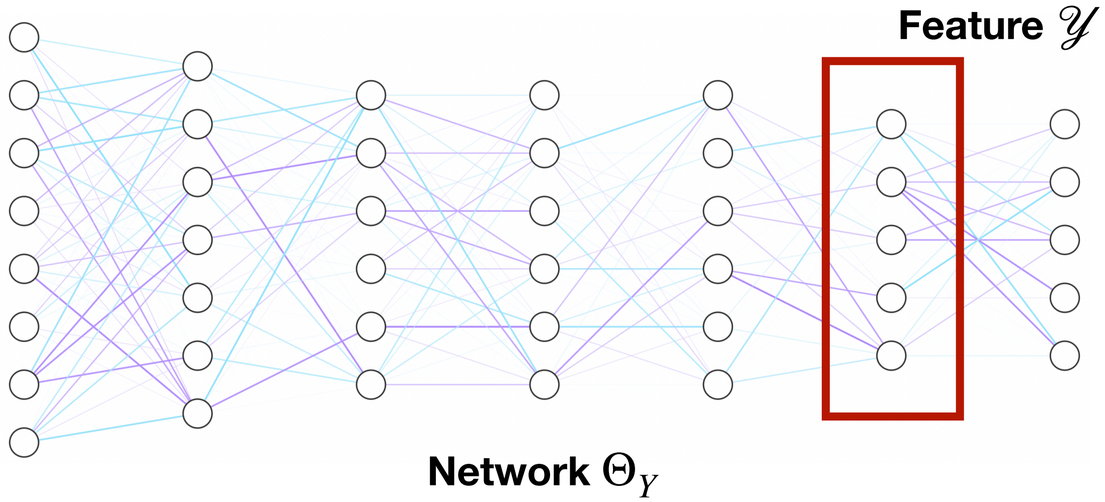




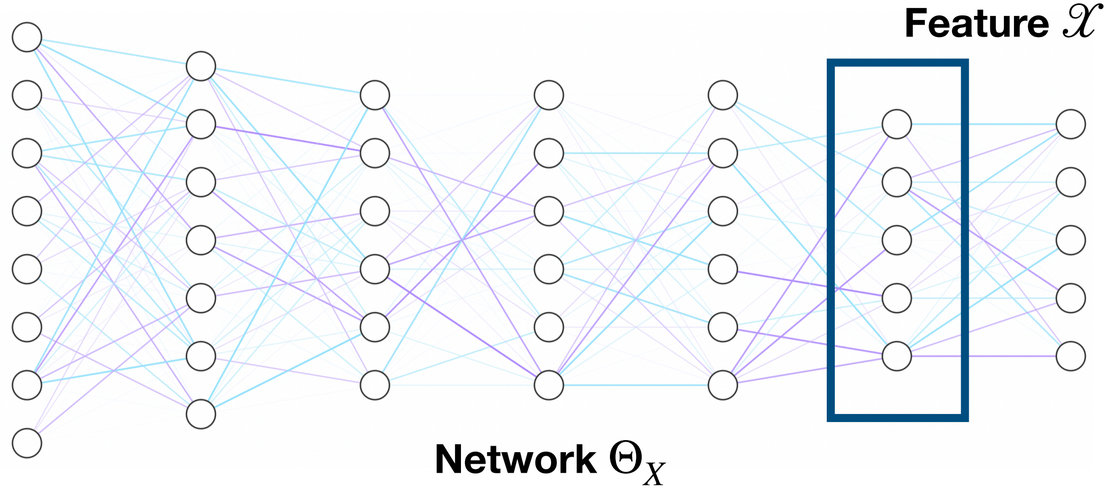
Beyond accuracy



Similarity?

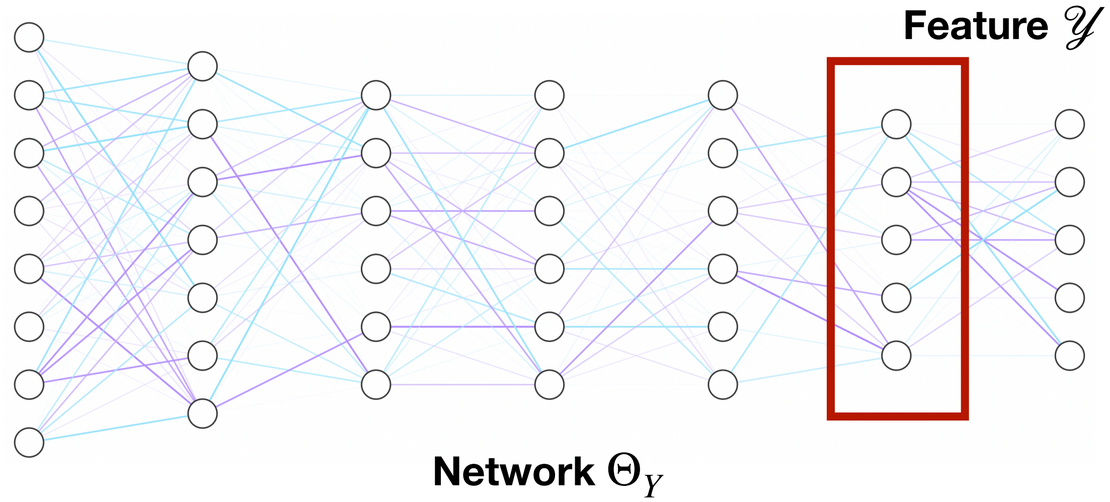


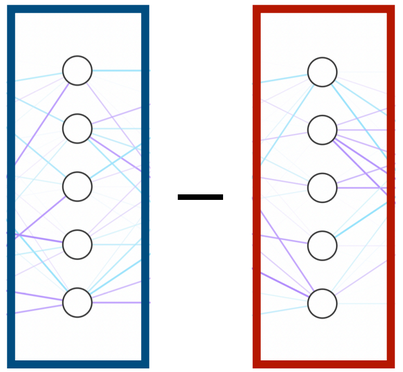
Similarity between networks



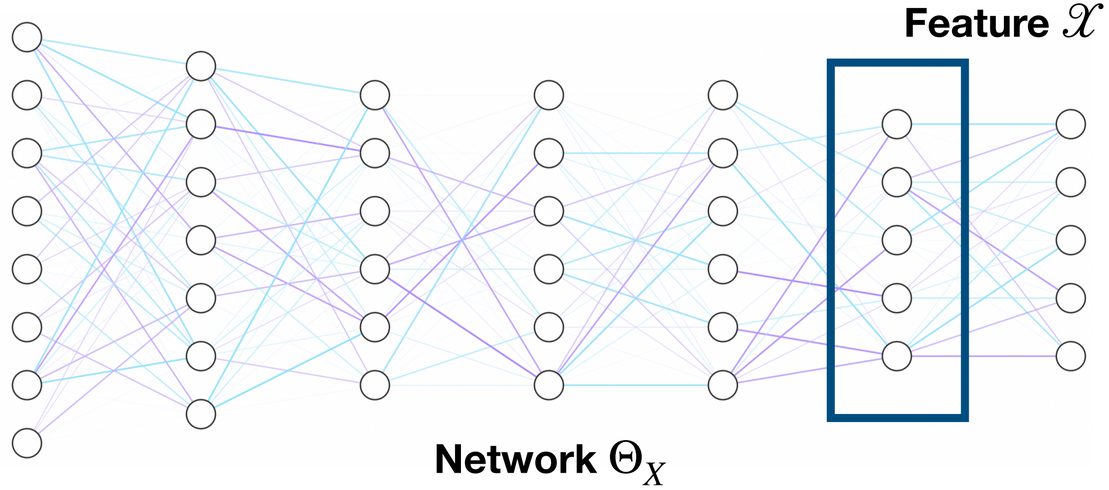
Subtraction

$X - Y$



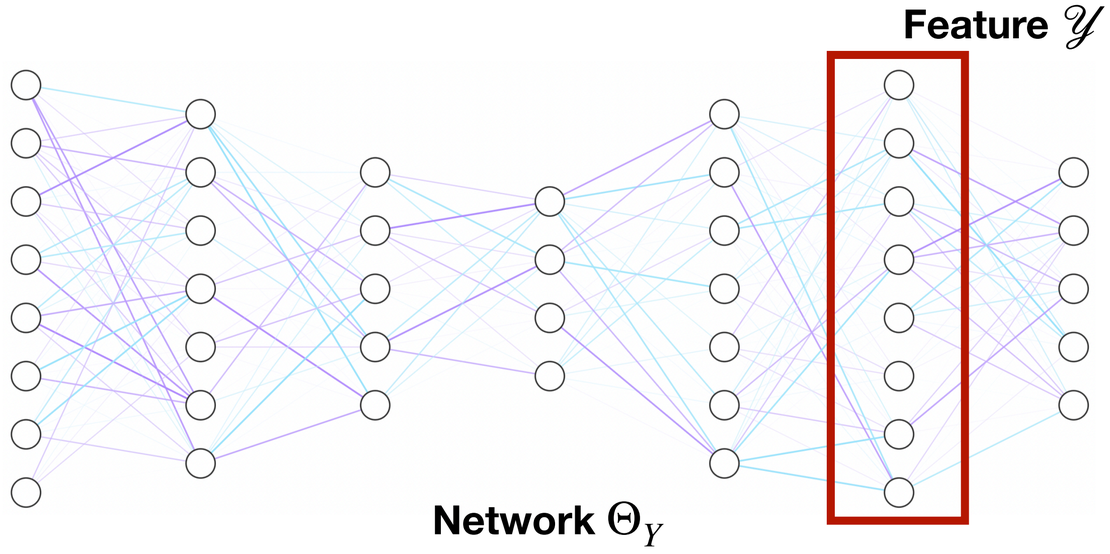


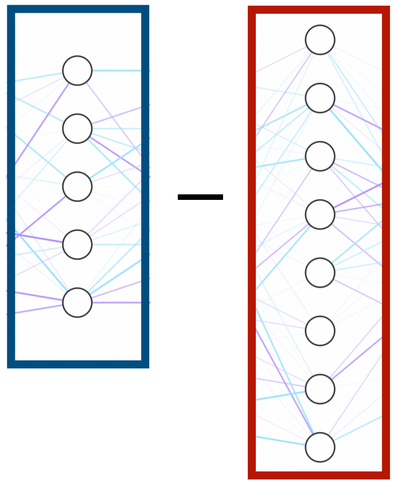
Similarity between networks



Subtraction

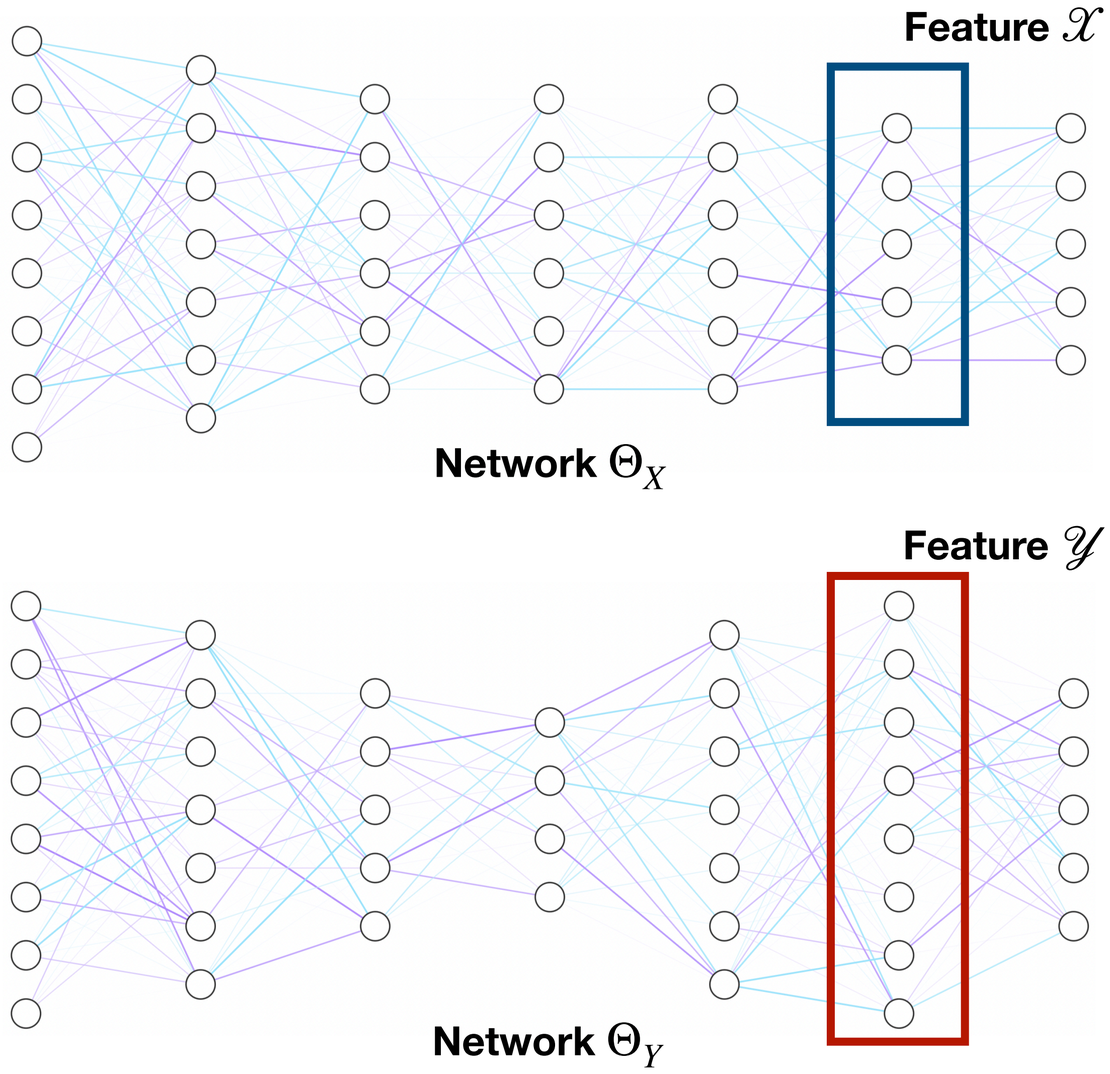
$X - Y$





?

Similarity between networks



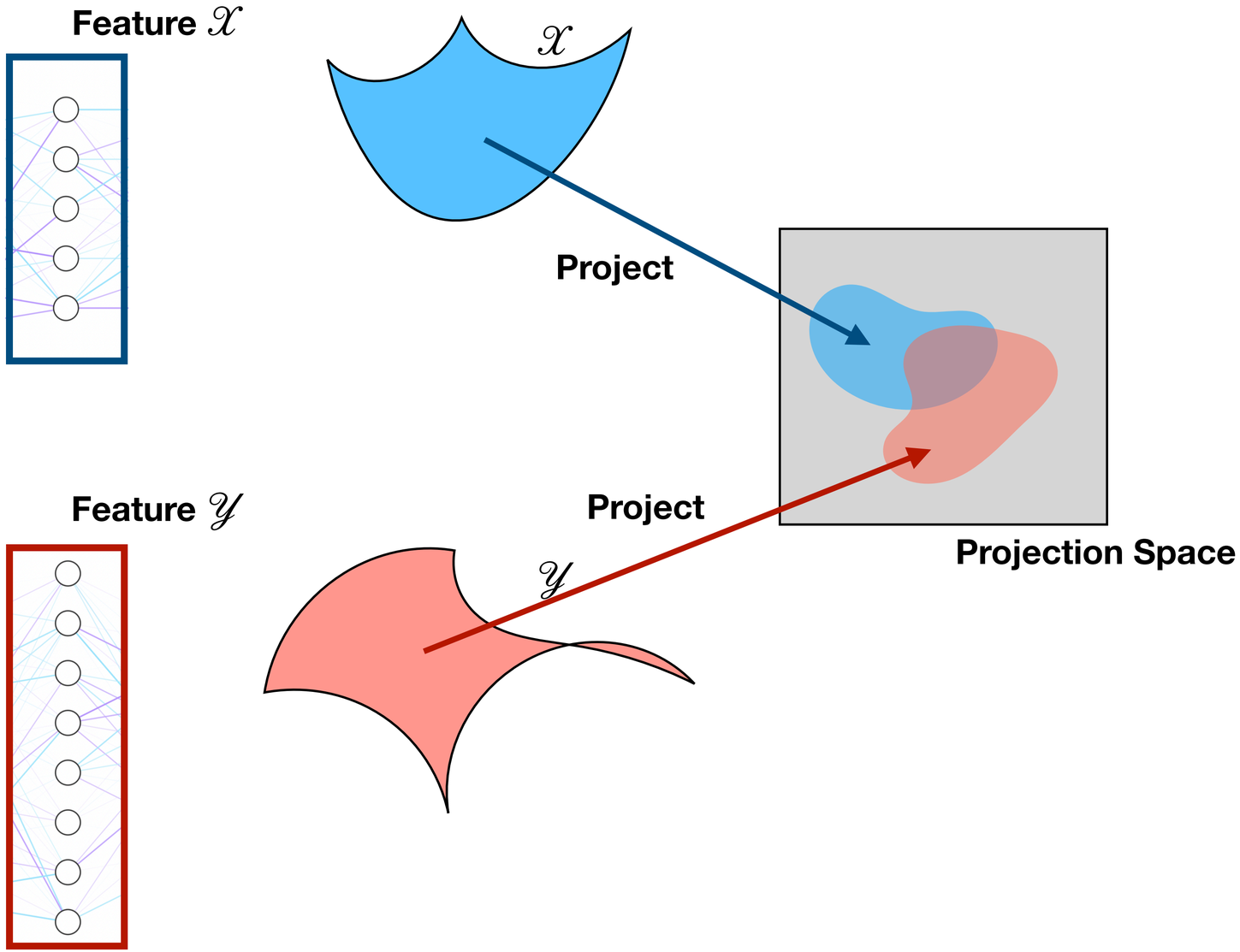
Correlation

Pearson Correlation

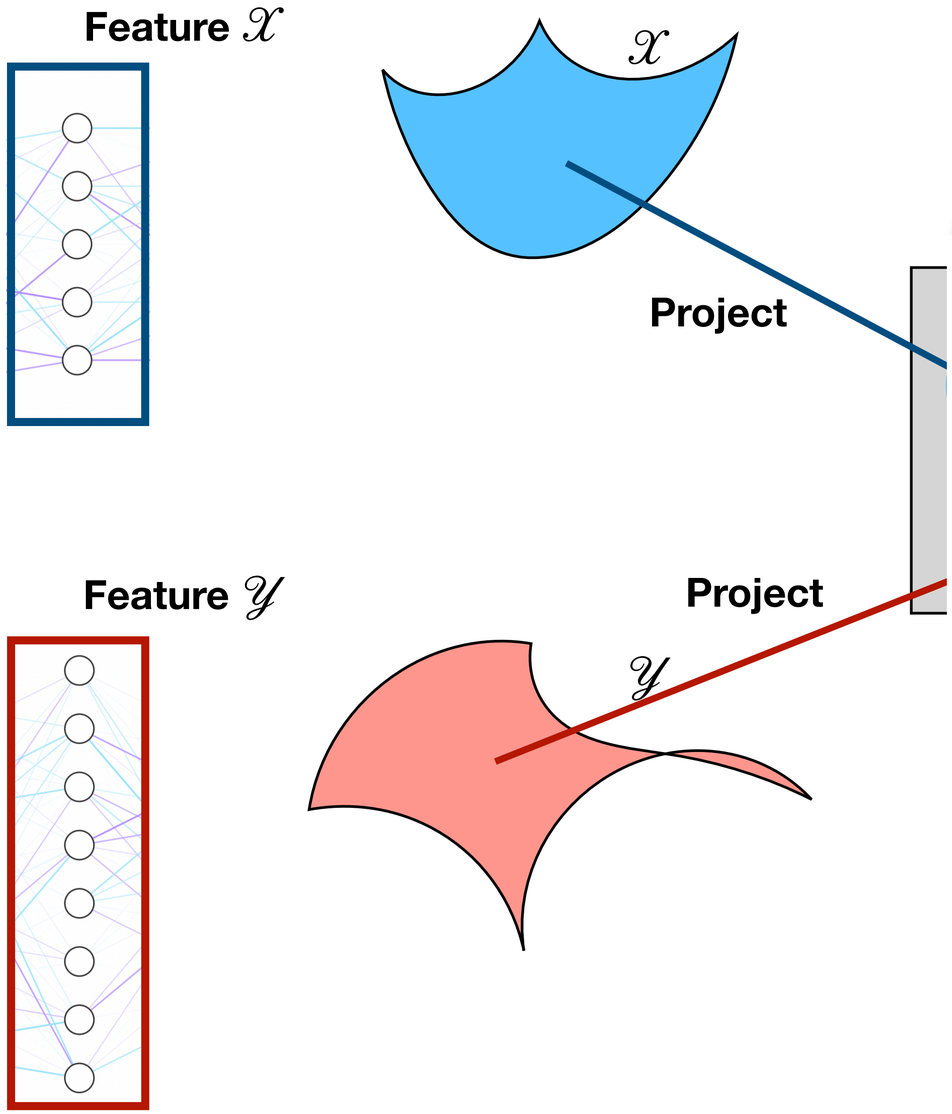
$$\rho_{X,Y} = \frac{\mathbf{cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$X \in \mathbb{R}, Y \in \mathbb{R}$$

Canonical Correlation Analysis



Canonical Correlation Analysis



Insights on representational similarity in neural networks with canonical correlation

Ari S. Morcos[*‡] DeepMind[†] arimorcos@gmail.com

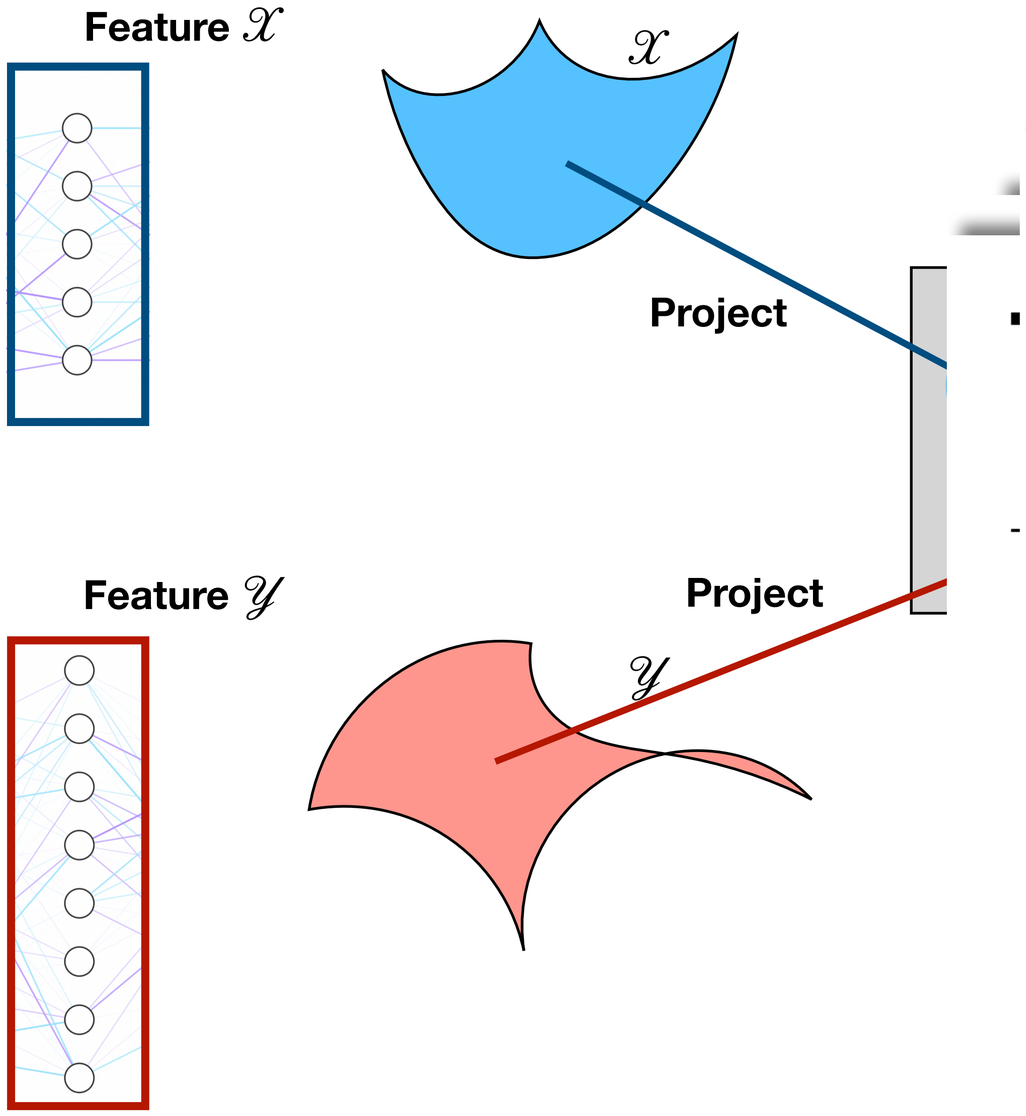
Maithra Raghu[*‡] Google Brain, Cornell University maithrar@gmail.com

Samy Bengio Google Brain bengio@google.com

Abstract

Comparing different neural network representations and determining how representations evolve over time remain challenging open questions in our understanding of the function of neural networks. Comparing representations in neural networks is fundamentally difficult as the structure of representations varies greatly, even across groups of networks trained on identical tasks, and over the course of training. Here, we develop projection weighted CCA (Canonical Correlation Analysis) as a tool for understanding neural networks, building off of SVCCA,

Canonical Correlation Analysis



Stochastic Approximation for Canonical Correlation Analysis

Raman Arora
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21204
arora@cs.jhu.edu

Teodor V. Marinov
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21204
tmarino2@jhu.edu

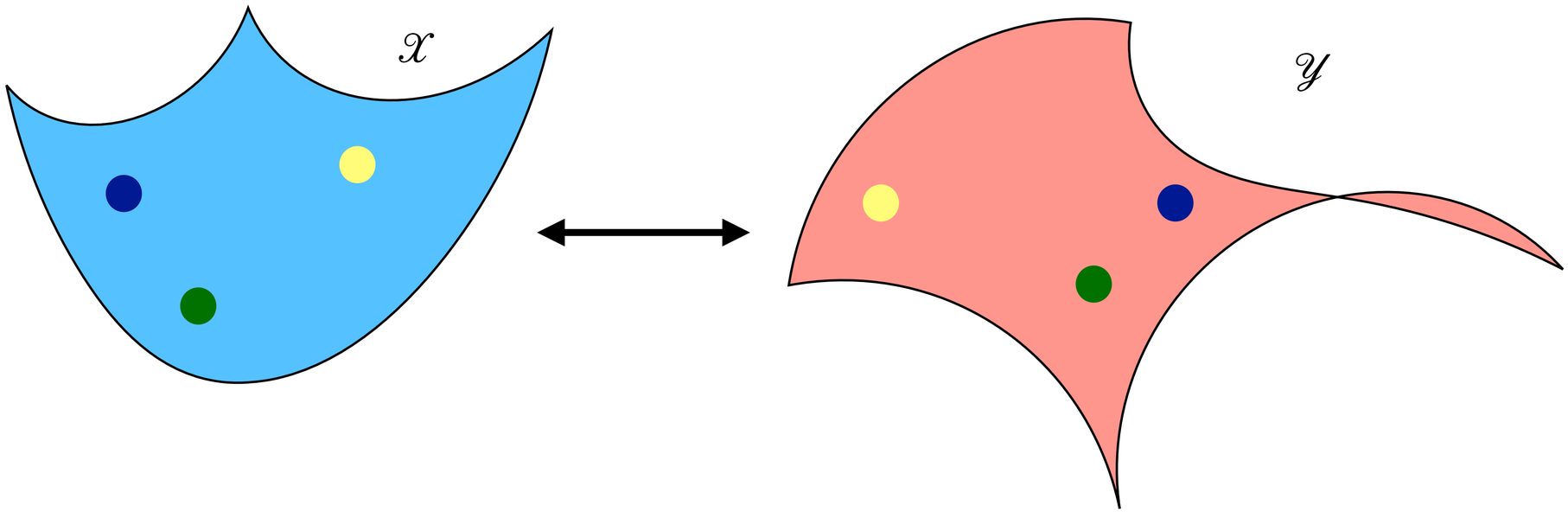
Poorya Mianjy
Dept. of Computer Science
Johns Hopkins University
Baltimore, MD 21204
mianjy@jhu.edu

Nathan Srebro
TTI-Chicago
Chicago, Illinois 60637
nati@ttic.edu

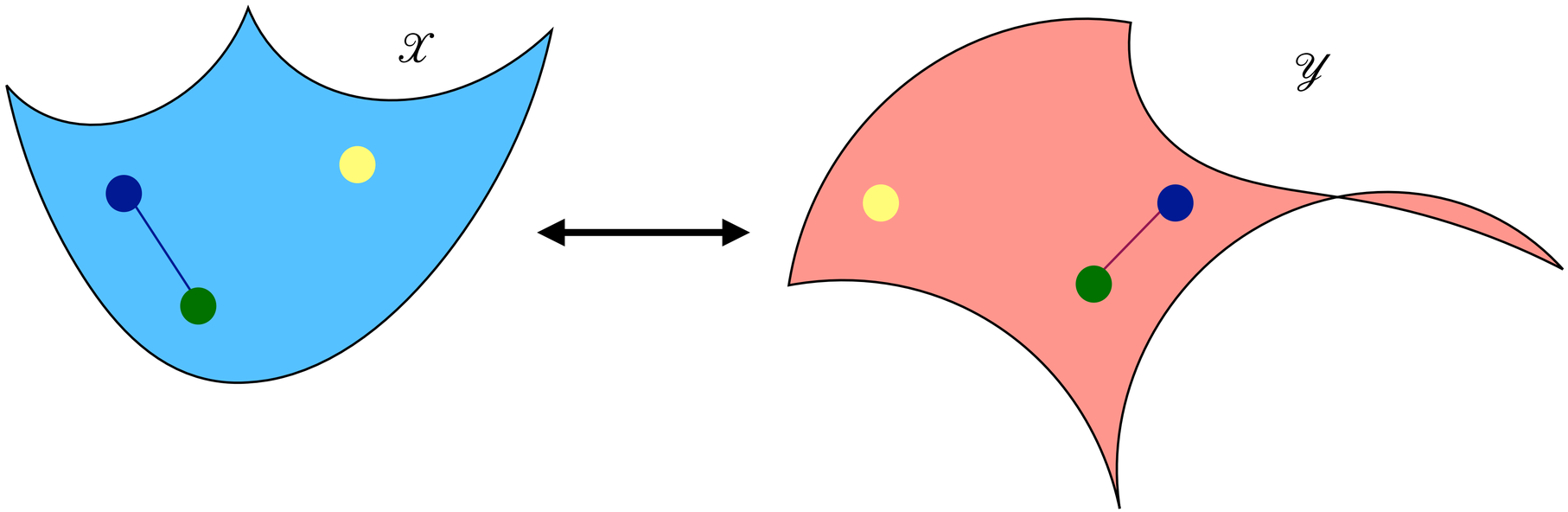
Abstract

of training. Here, we develop projection weighted CCA (Canonical Correlation Analysis) as a tool for understanding neural networks, building off of SVCCA,

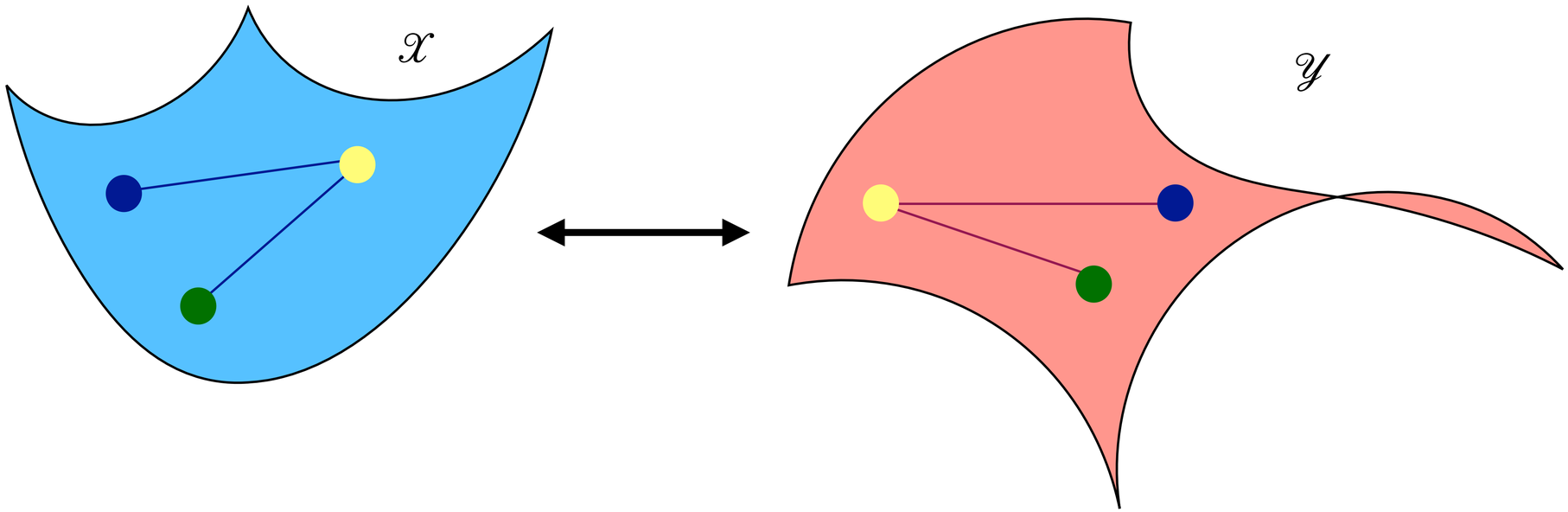
Distance Correlation



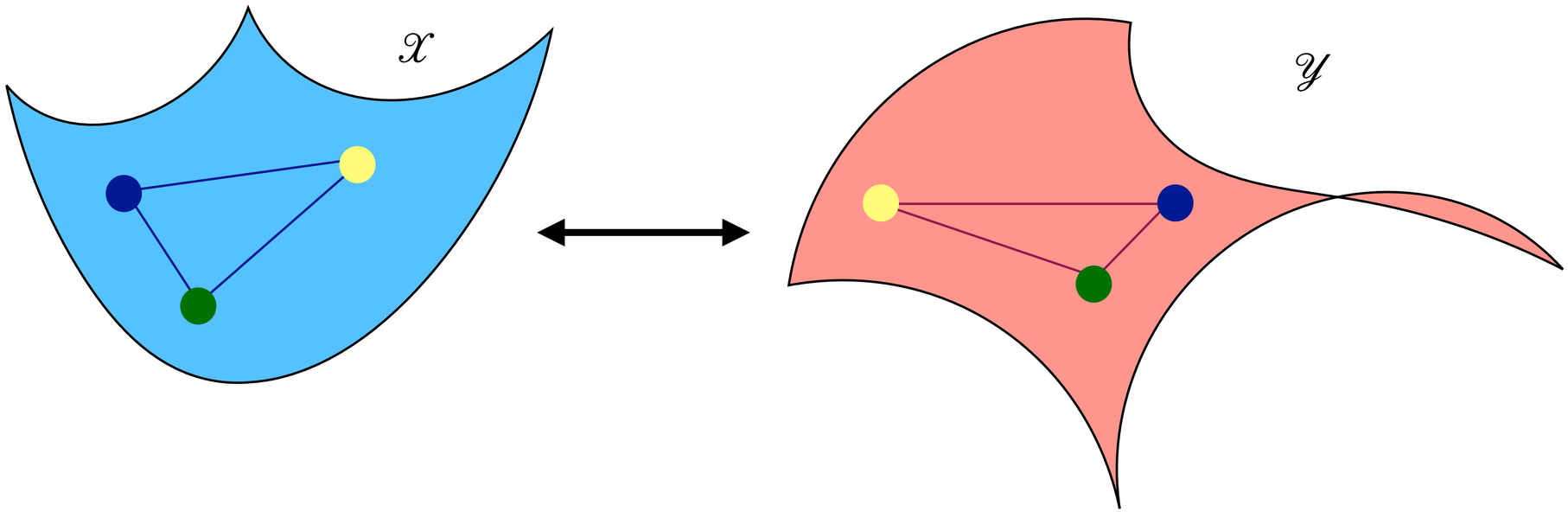
Distance Correlation



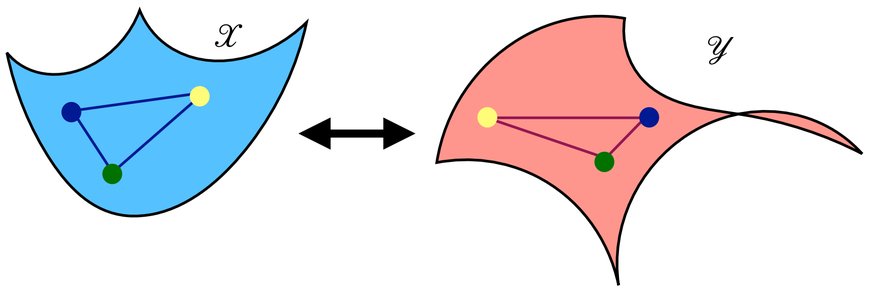
Distance Correlation



Distance Correlation



Distance Correlation



Samples: $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \cdots, (X_m, Y_m)$

Distance matrix: $a_{k,l} = \|X_k - X_l\|, b_{k,l} = \|Y_k - Y_l\|$

Normalized:

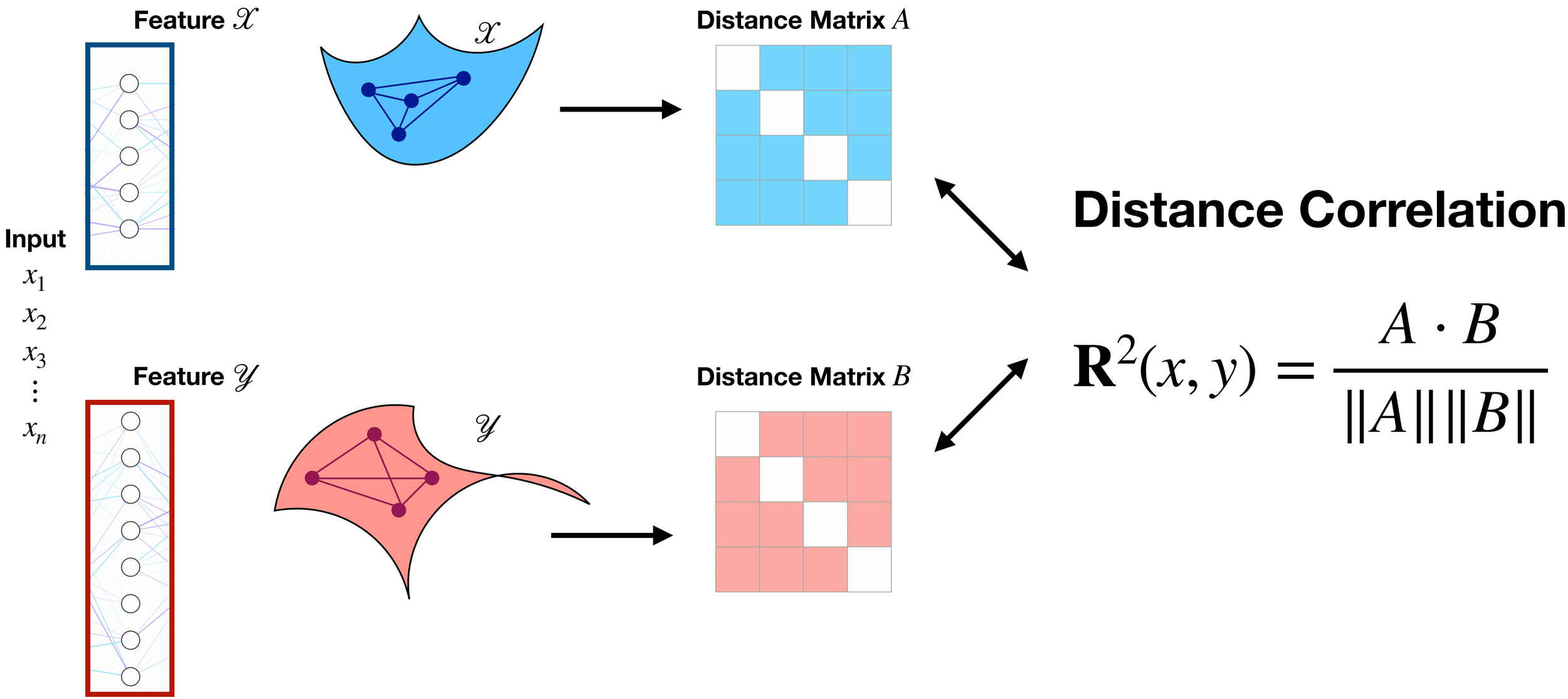
$$A_{k,l} = a_{k,l} - \bar{a}_{k,\cdot} - \bar{a}_{\cdot,l} + \bar{a}_{\cdot,\cdot}$$

$$B_{k,l} = b_{k,l} - \bar{b}_{k,\cdot} - \bar{b}_{\cdot,l} + \bar{b}_{\cdot,\cdot}$$

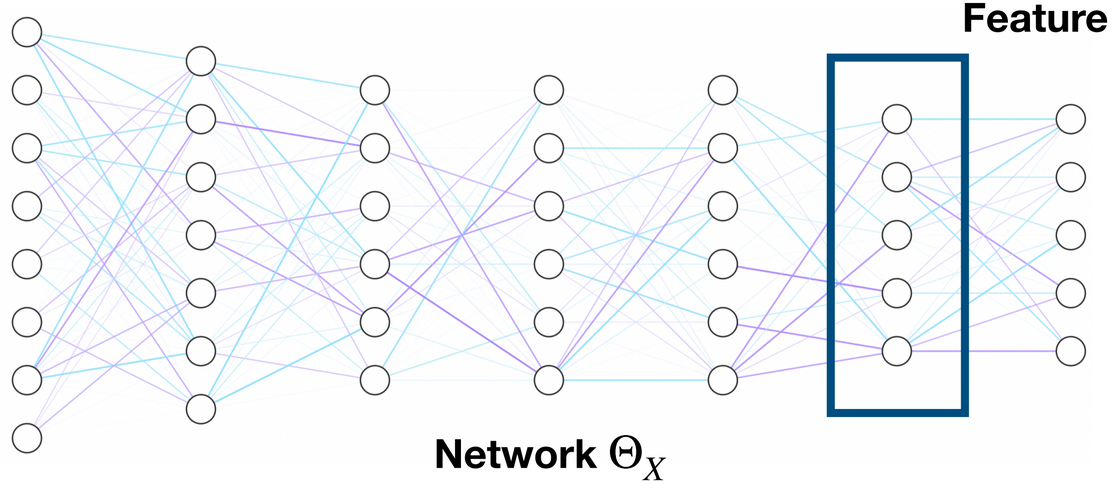
Distance correlation:

$$\frac{\sum_{k,l} A_{k,l} B_{k,l}}{\sqrt{(\sum_{k,l} A_{k,l} A_{k,l})(\sum_{k,l} B_{k,l} B_{k,l})}}$$

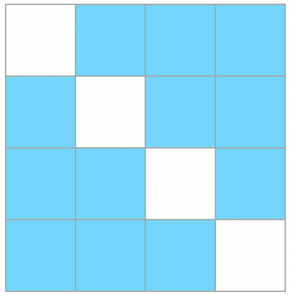
Distance Correlation

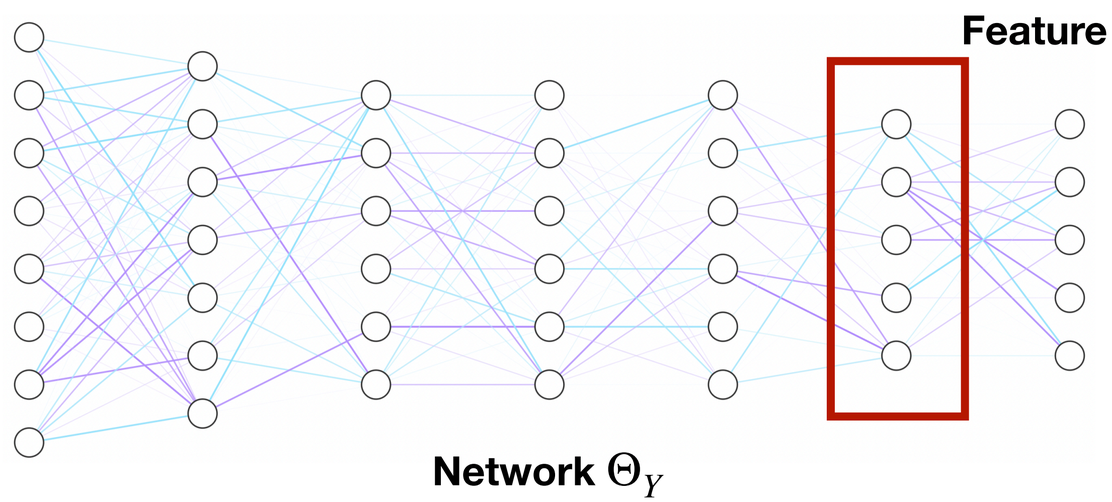


Various Feature Dimensions

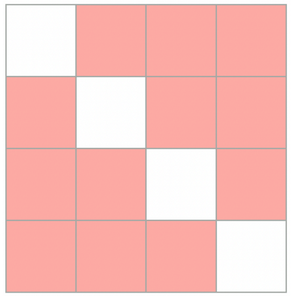


Distance Matrix A

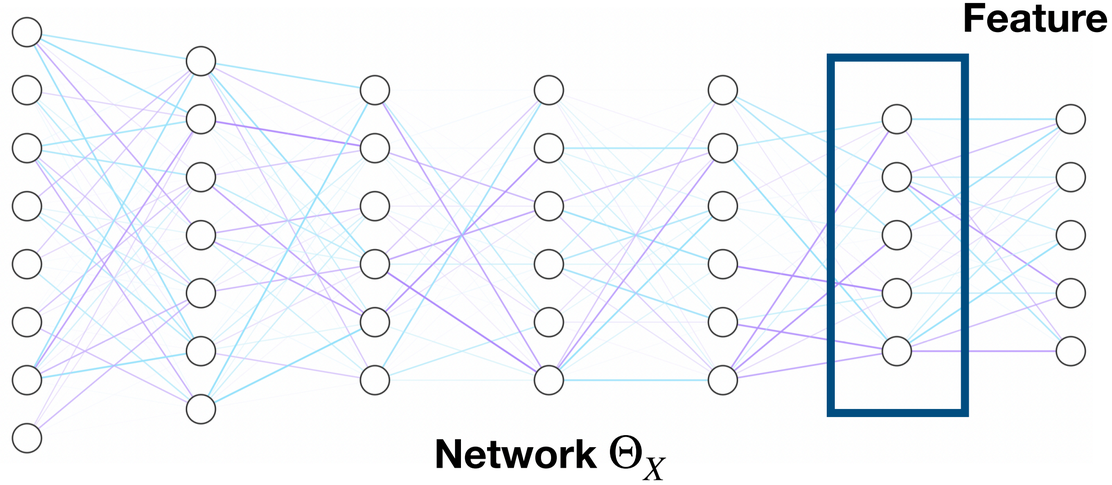




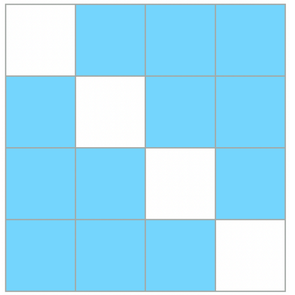
Distance Matrix B

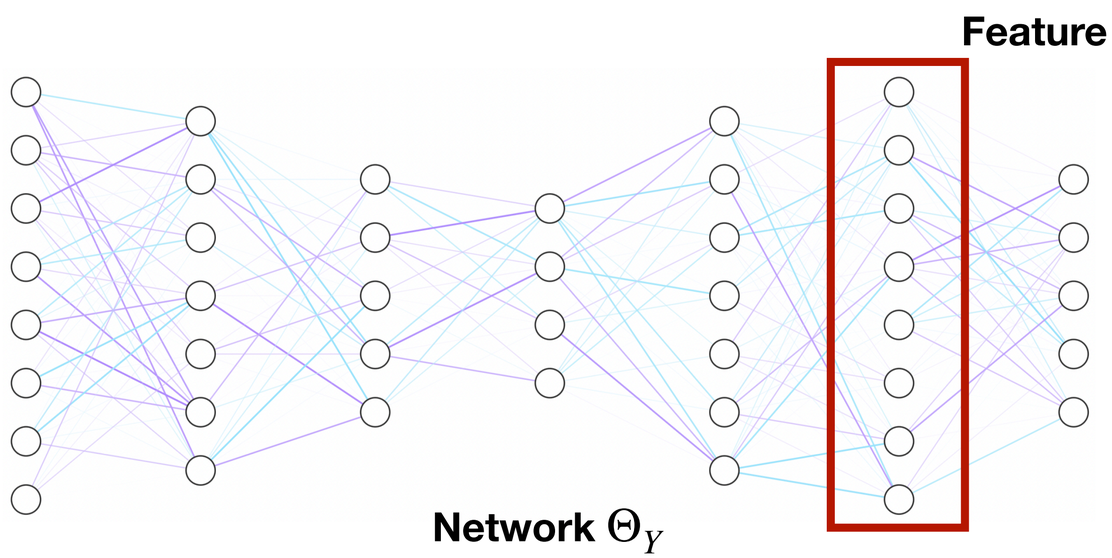


Various Feature Dimensions

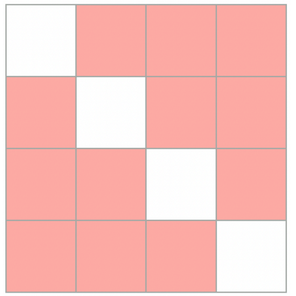


Distance Matrix A

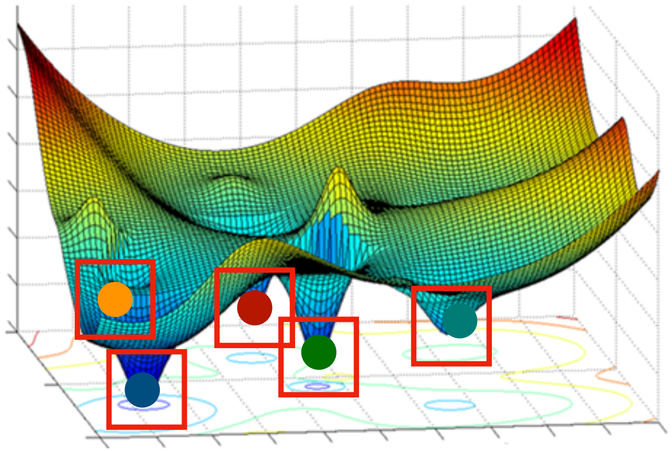




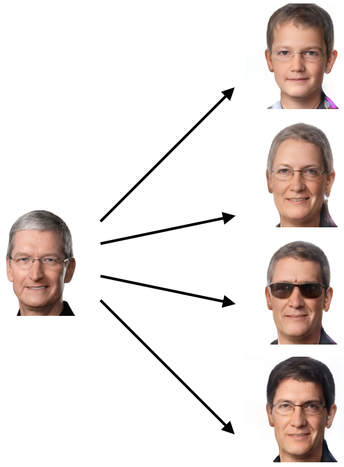
Distance Matrix B



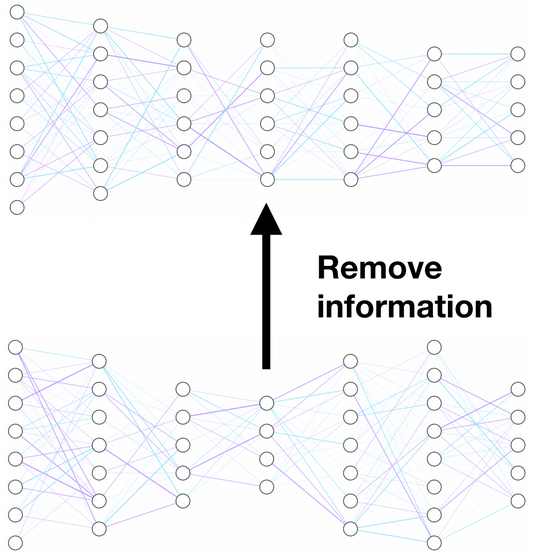
Use Cases



Diverge Training

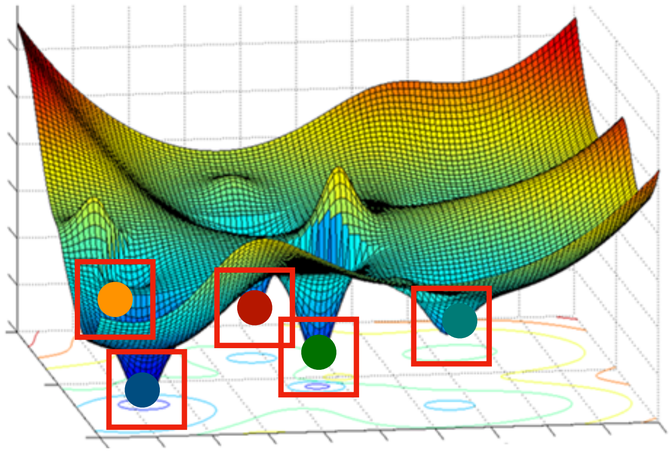


Disentanglement

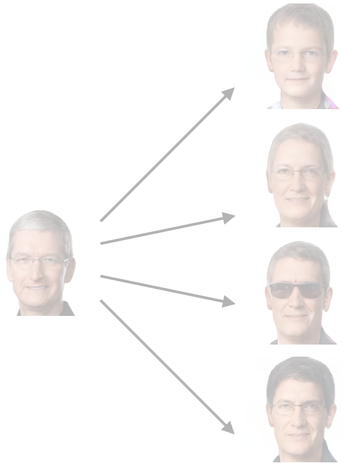


Network Conditioning

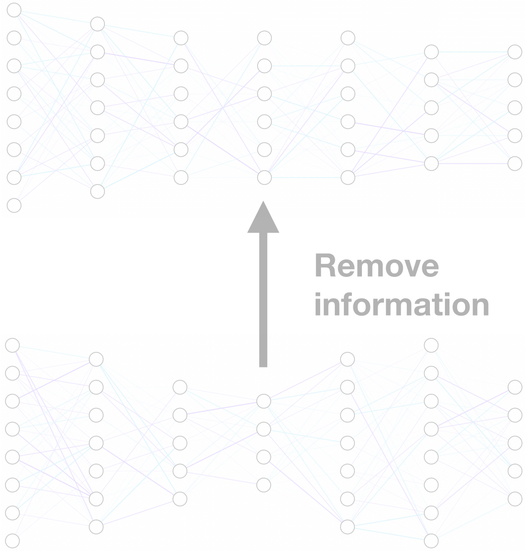
Use Cases



Diverge Training

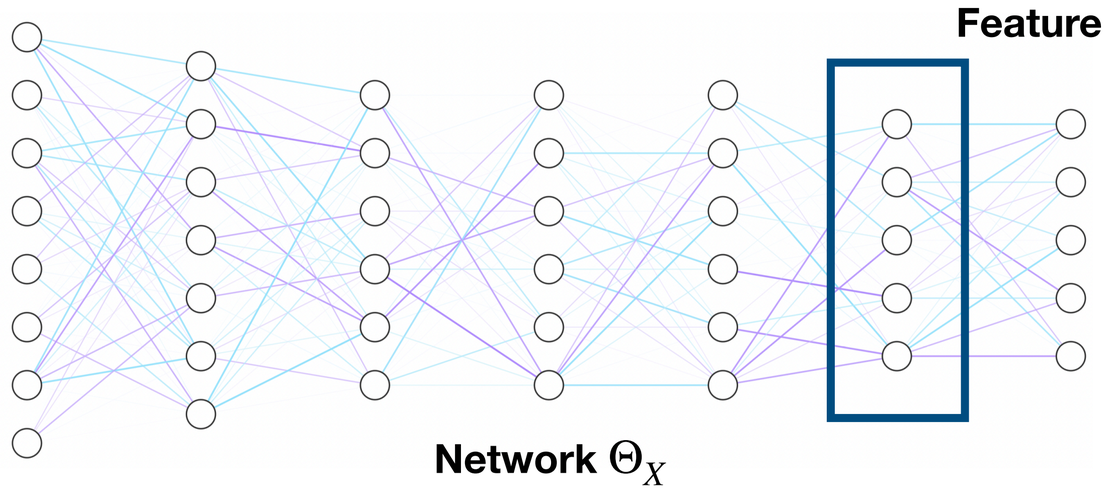


Disentanglement

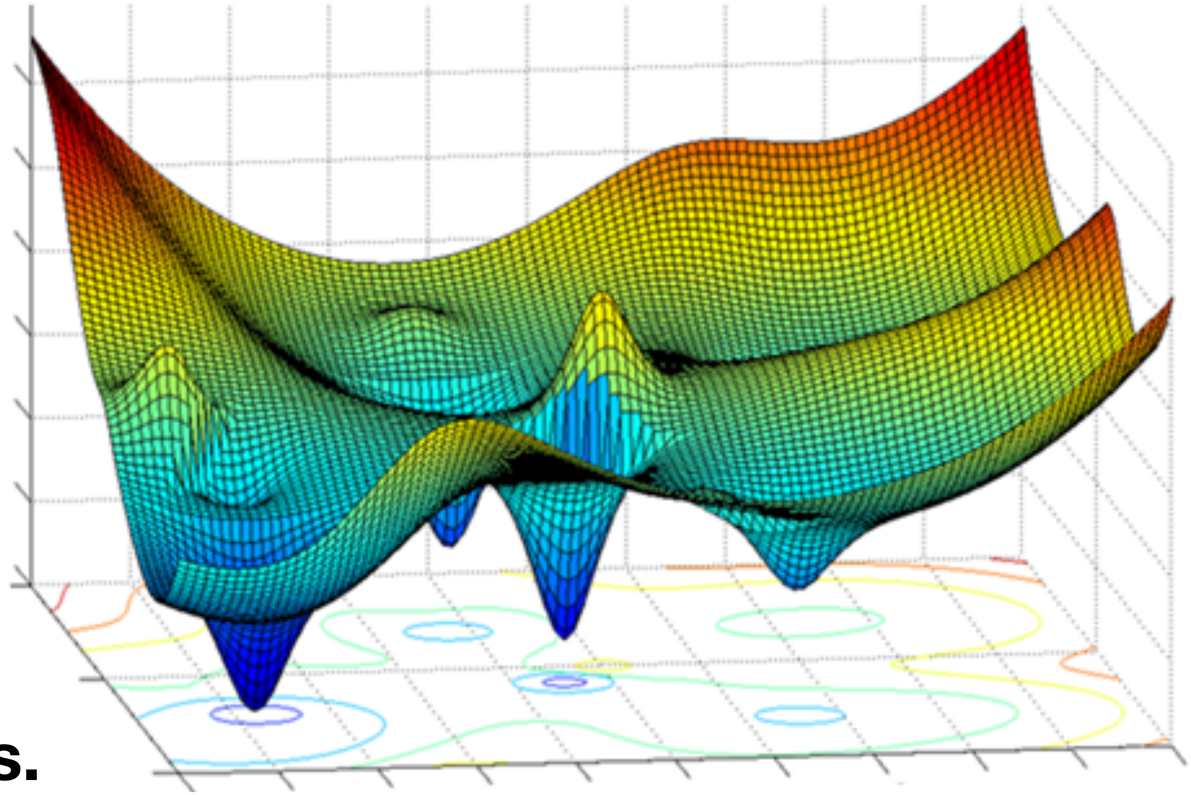


Network Conditioning

What can DC do? - Robustness

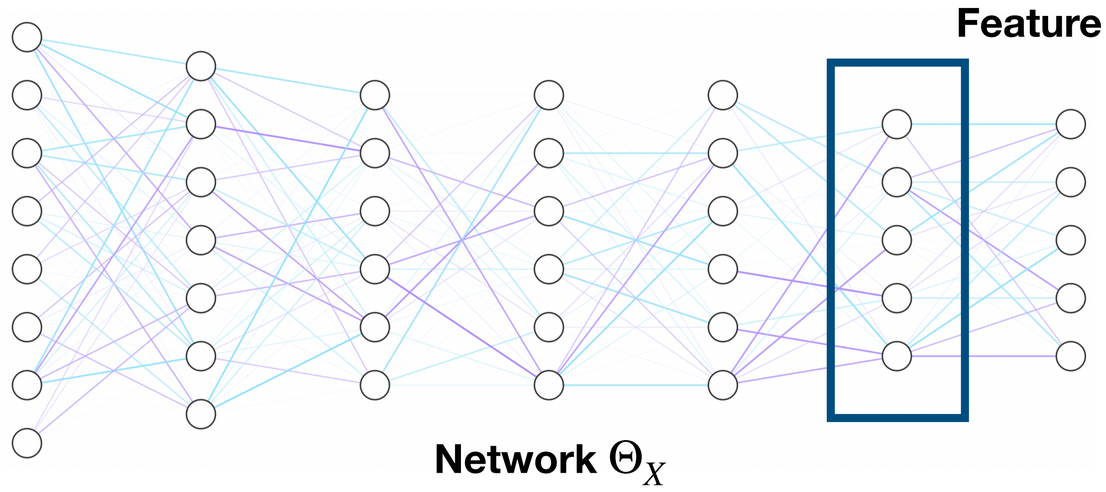


All possible features

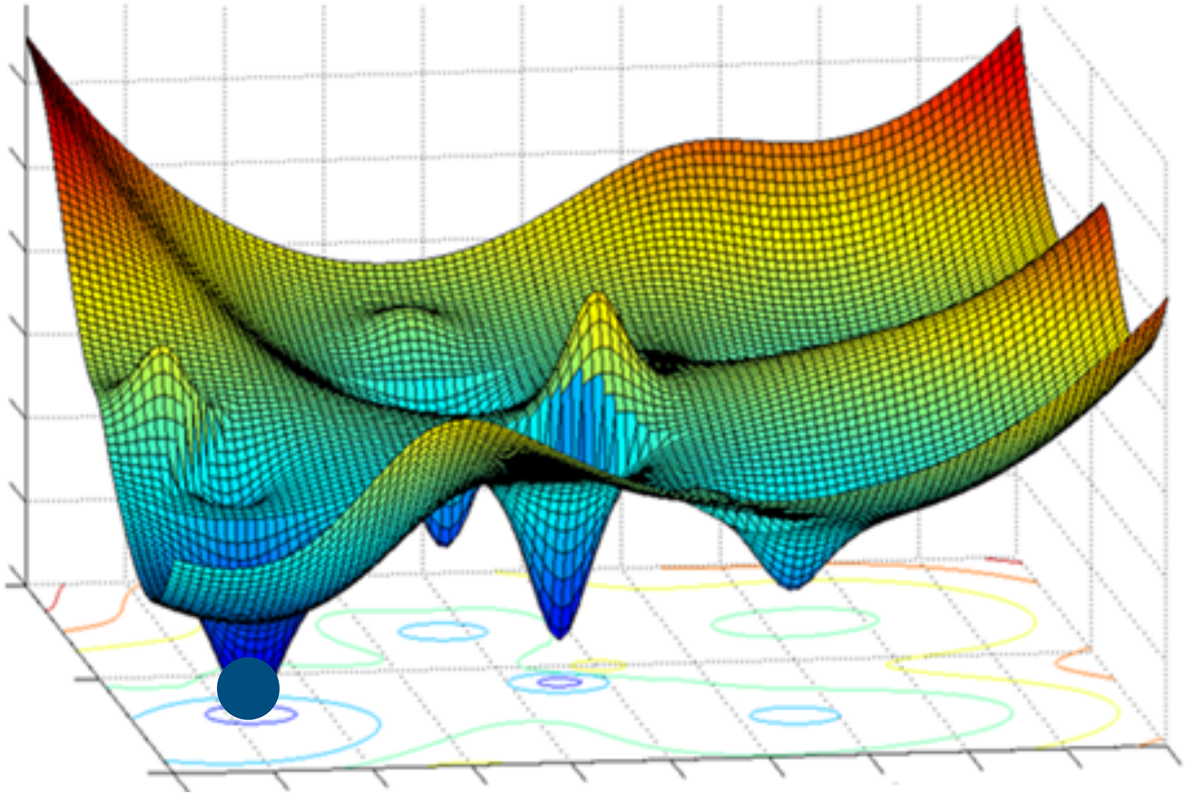


5 Θ_X assemble to improve robustness.

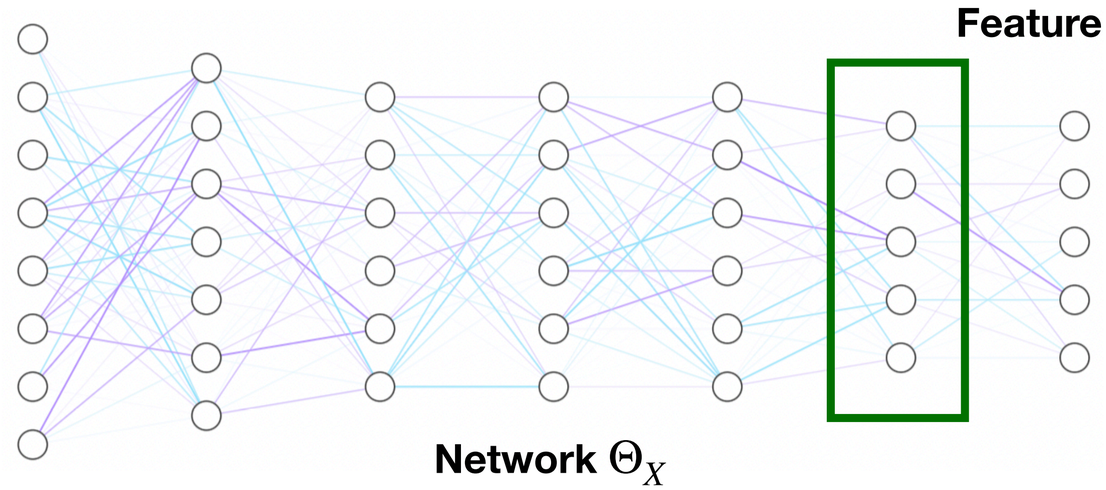
What can DC do? - Robustness



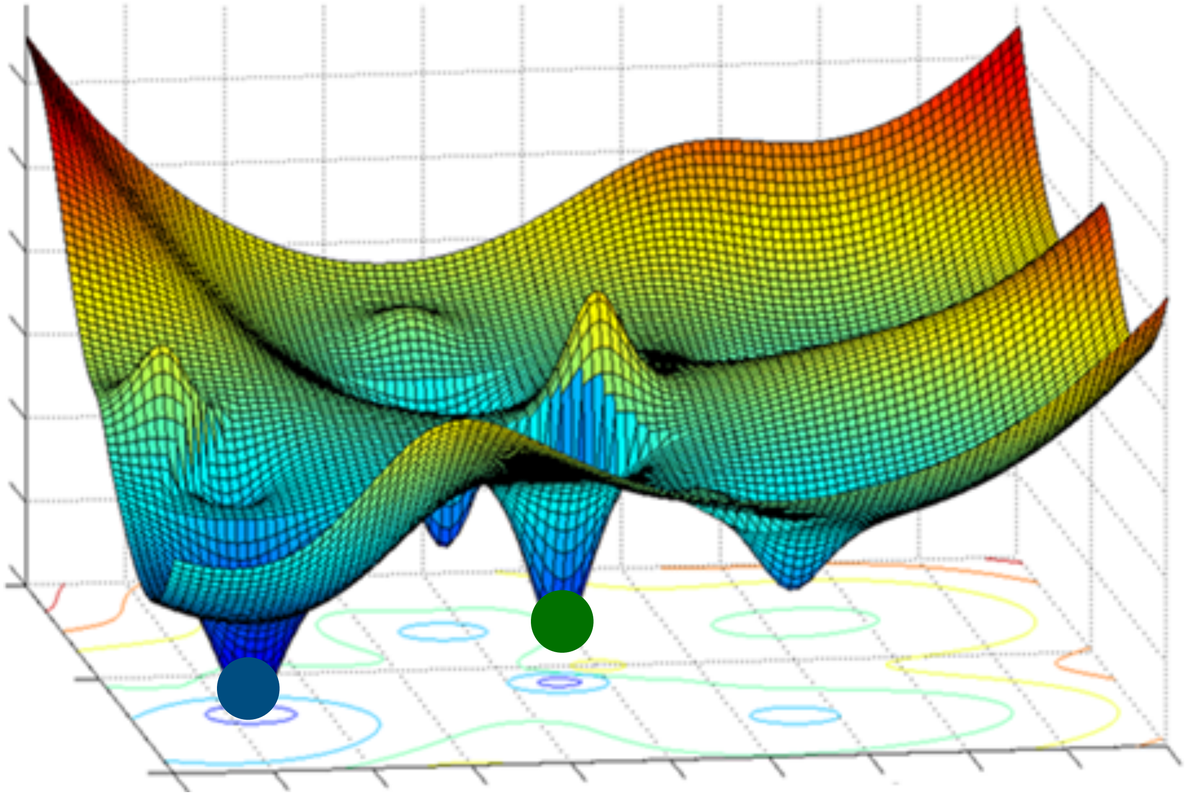
All possible features



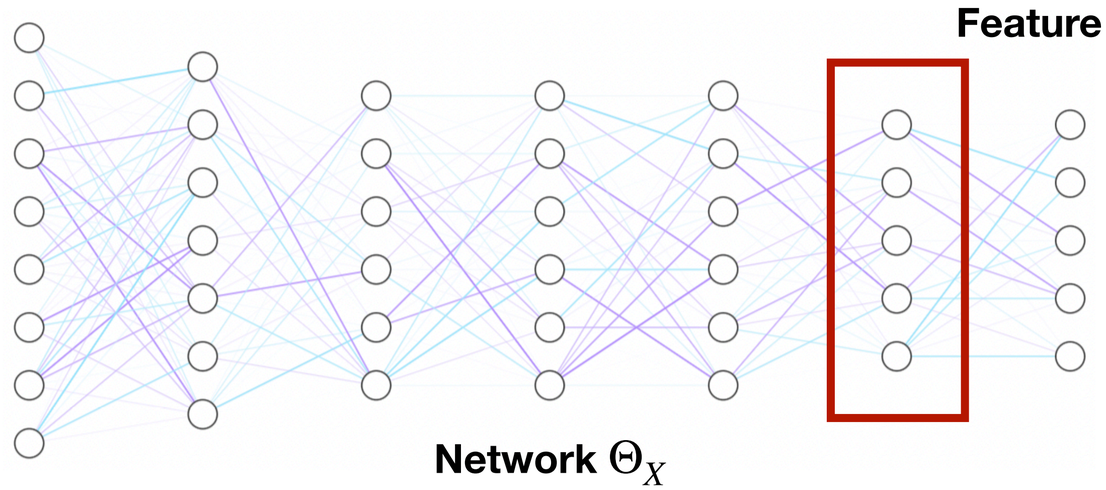
What can DC do? - Robustness



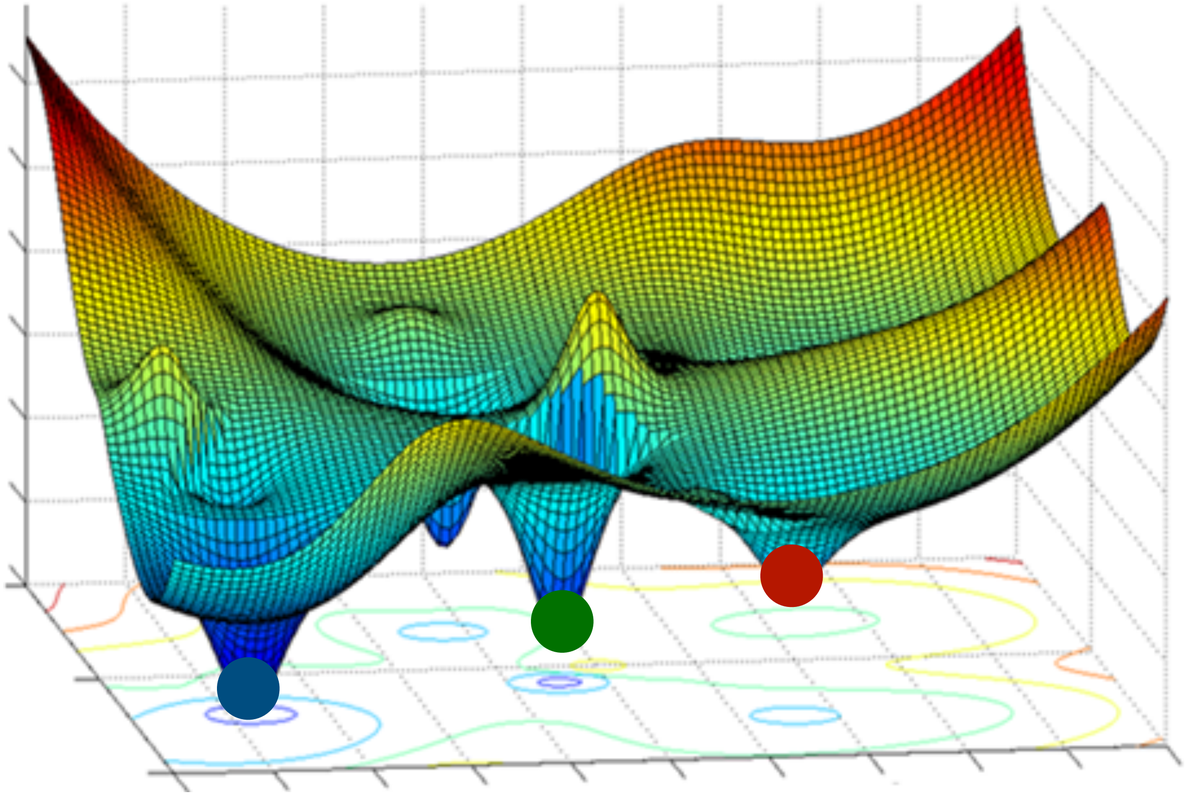
All possible features



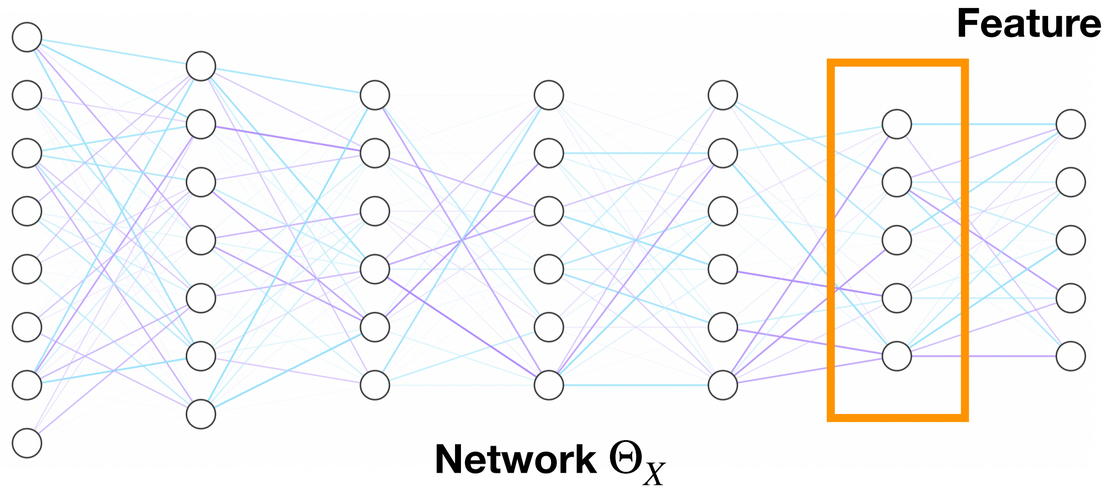
What can DC do? - Robustness



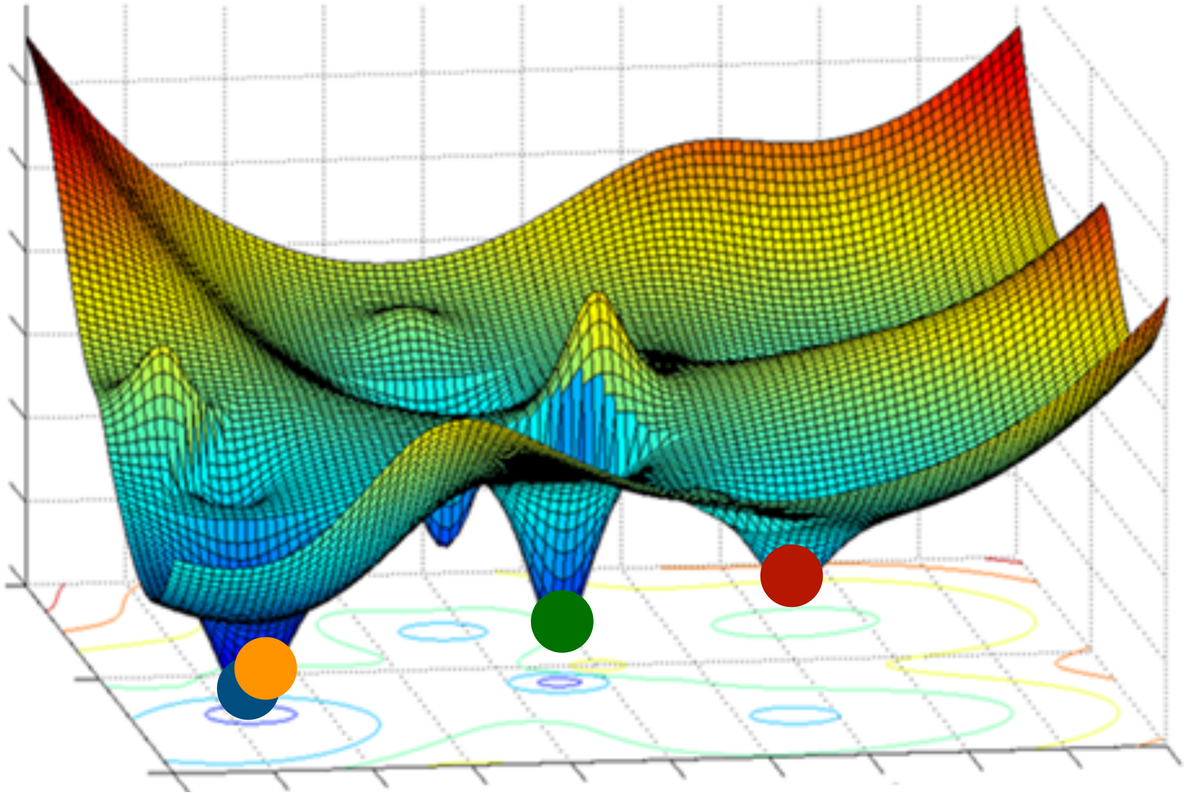
All possible features



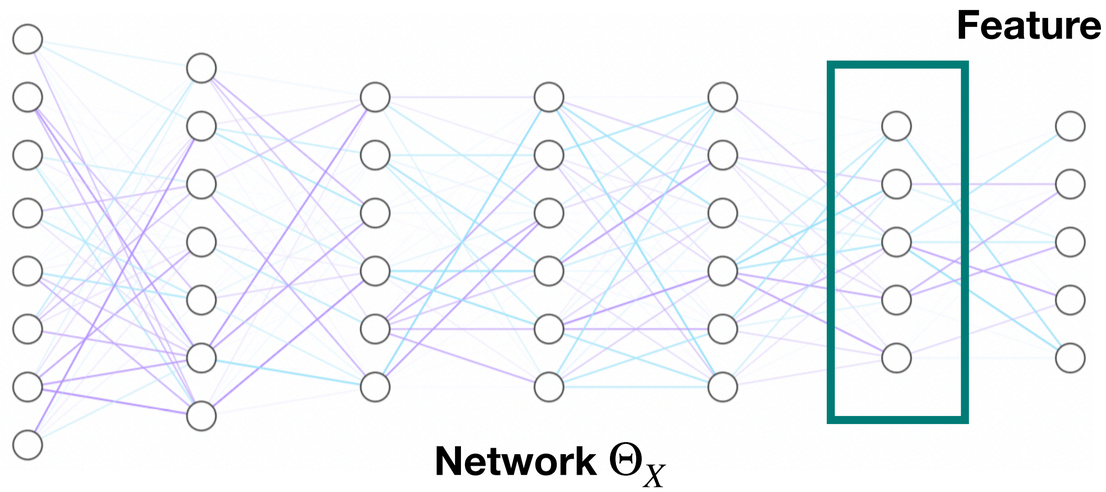
What can DC do? - Robustness



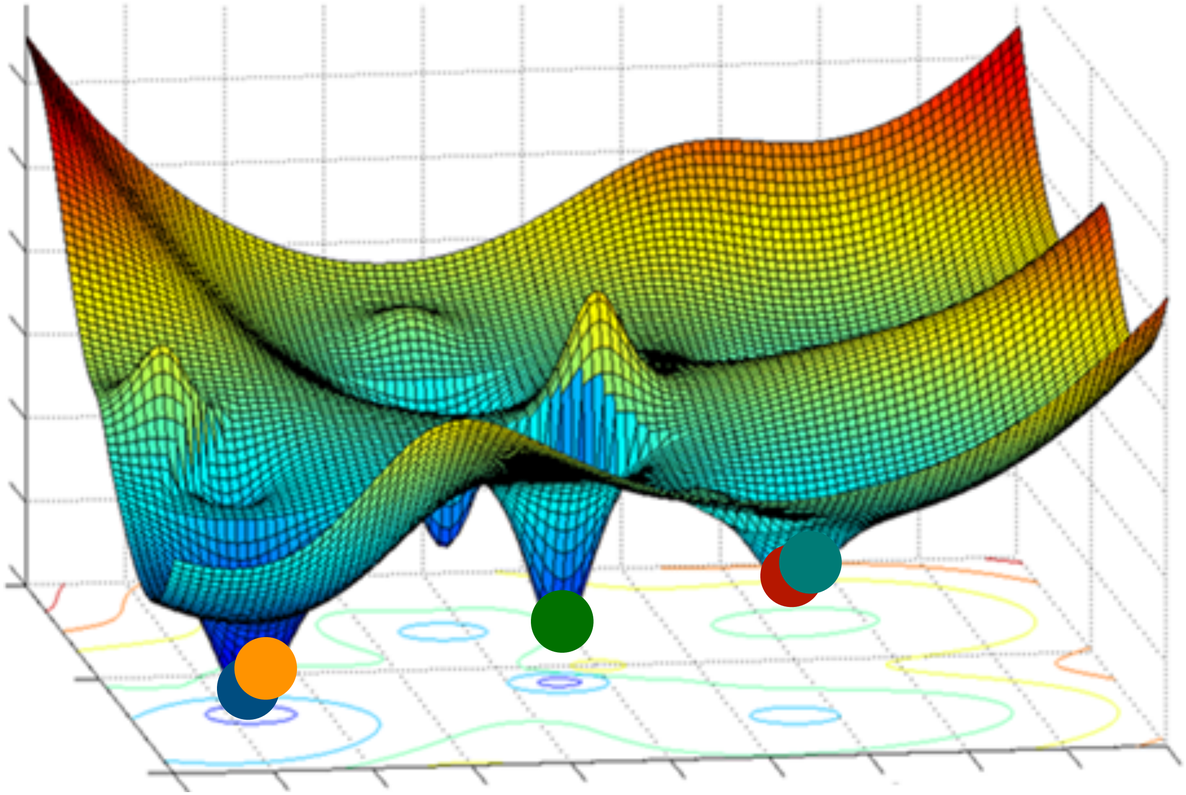
All possible features



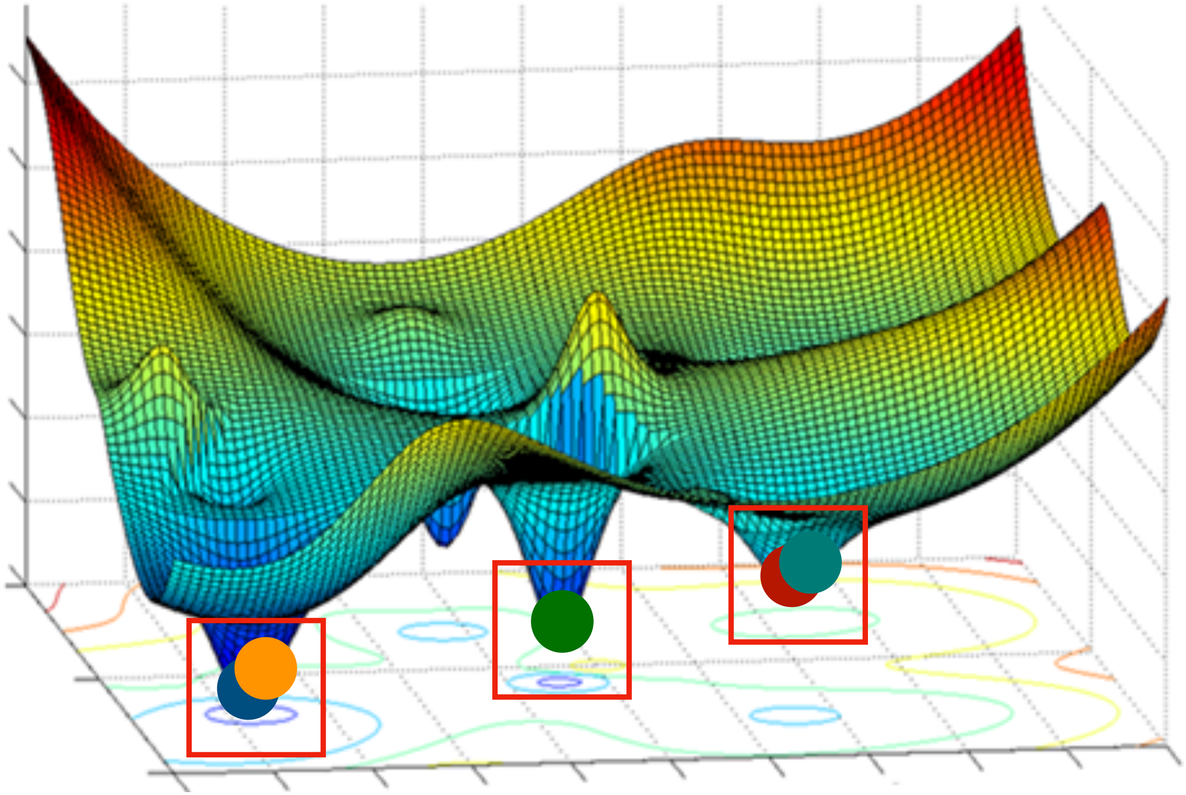
What can DC do? - Robustness



All possible features



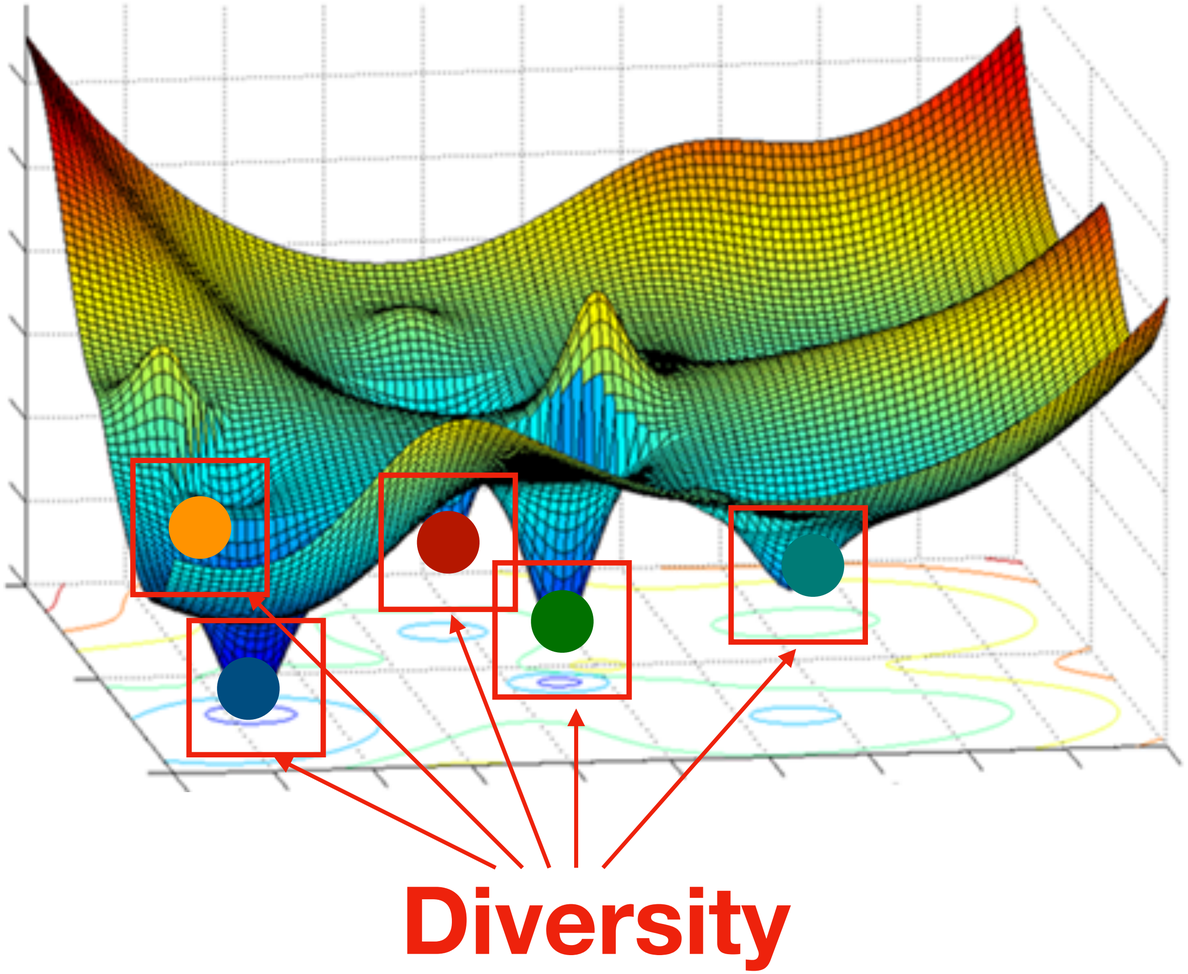
What can DC do? - Robustness



5 Θ_X **, but in total, there are only**

3 unique (independent) Θ_X

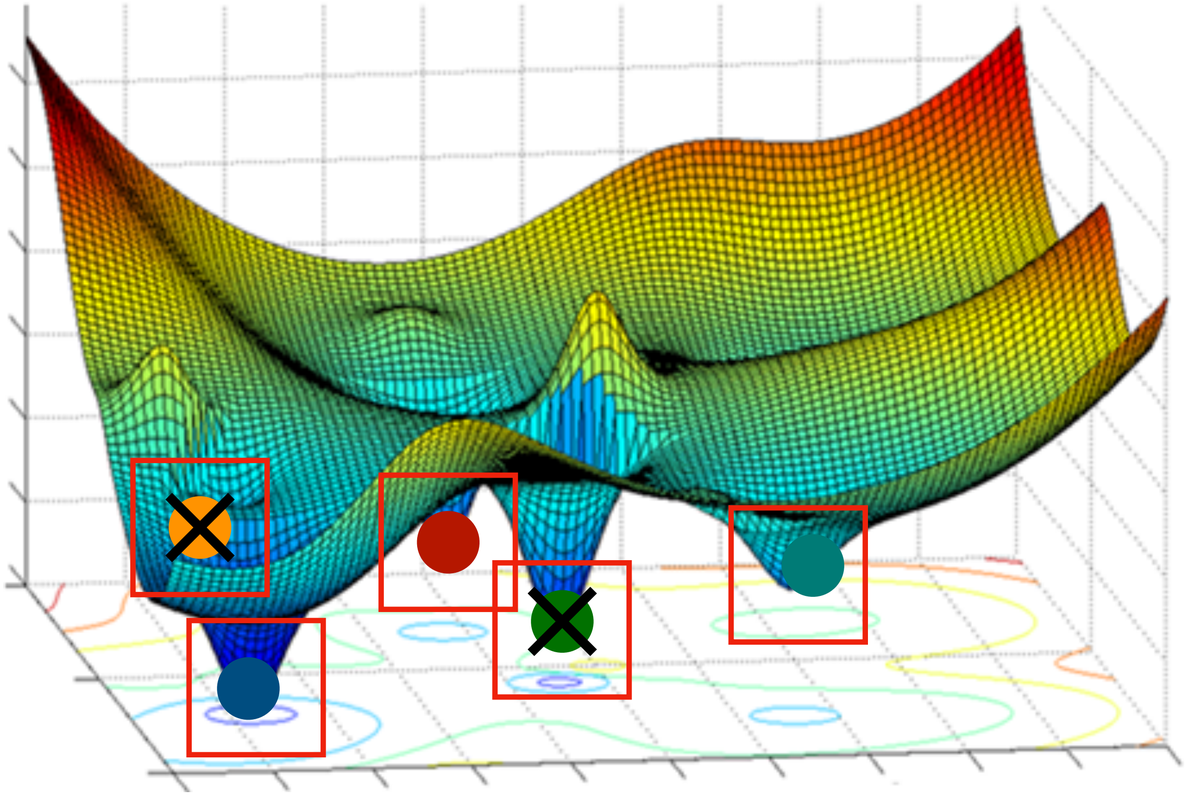
What can DC do? - Robustness



If we can measure the similarity between networks, we can have

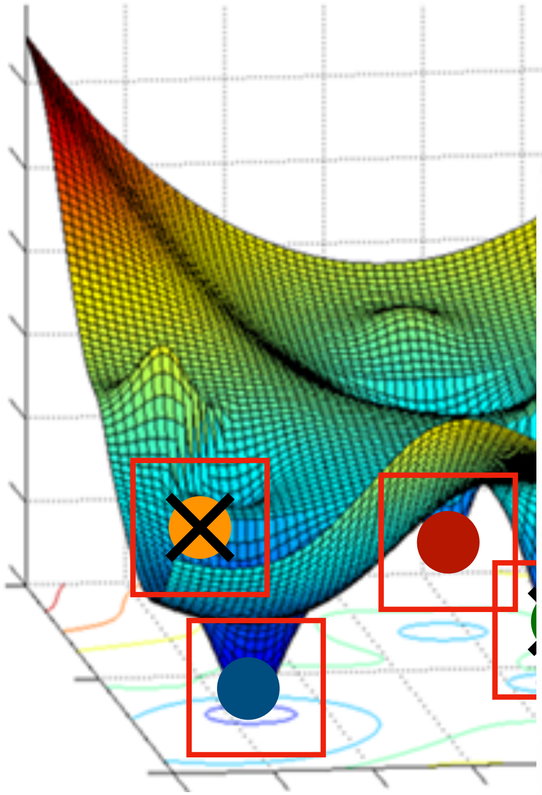
5 unique (independent) Θ_X

What can DC do? - Robustness



Attack 2 Θ_X, we can still have 3 uninfluenced Θ_X

What can DC do? - Robustness

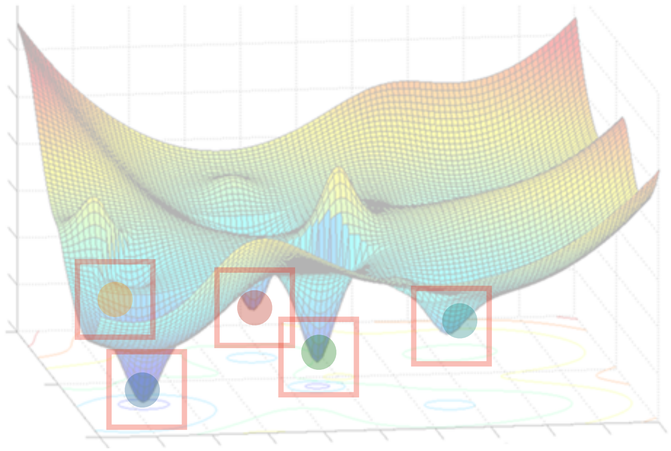


Repulsive Deep Ensembles are Bayesian

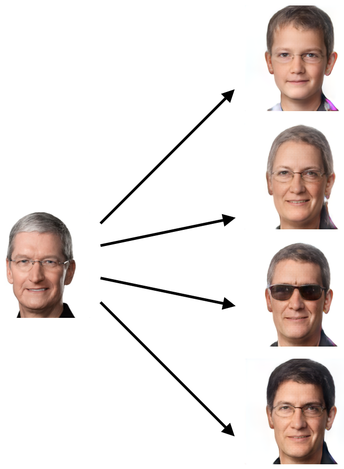
Francesco D'Angelo
ETH Zürich
Zürich, Switzerland
dngfra@gmail.com

Vincent Fortuin
ETH Zürich
Zürich, Switzerland
fortuin@inf.ethz.ch

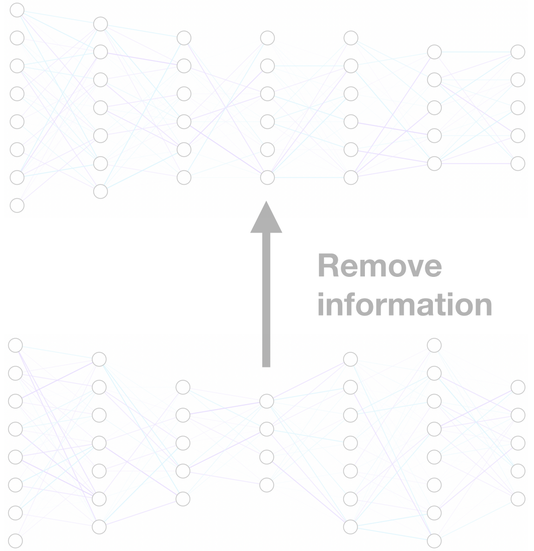
Use Cases



Diverge Training

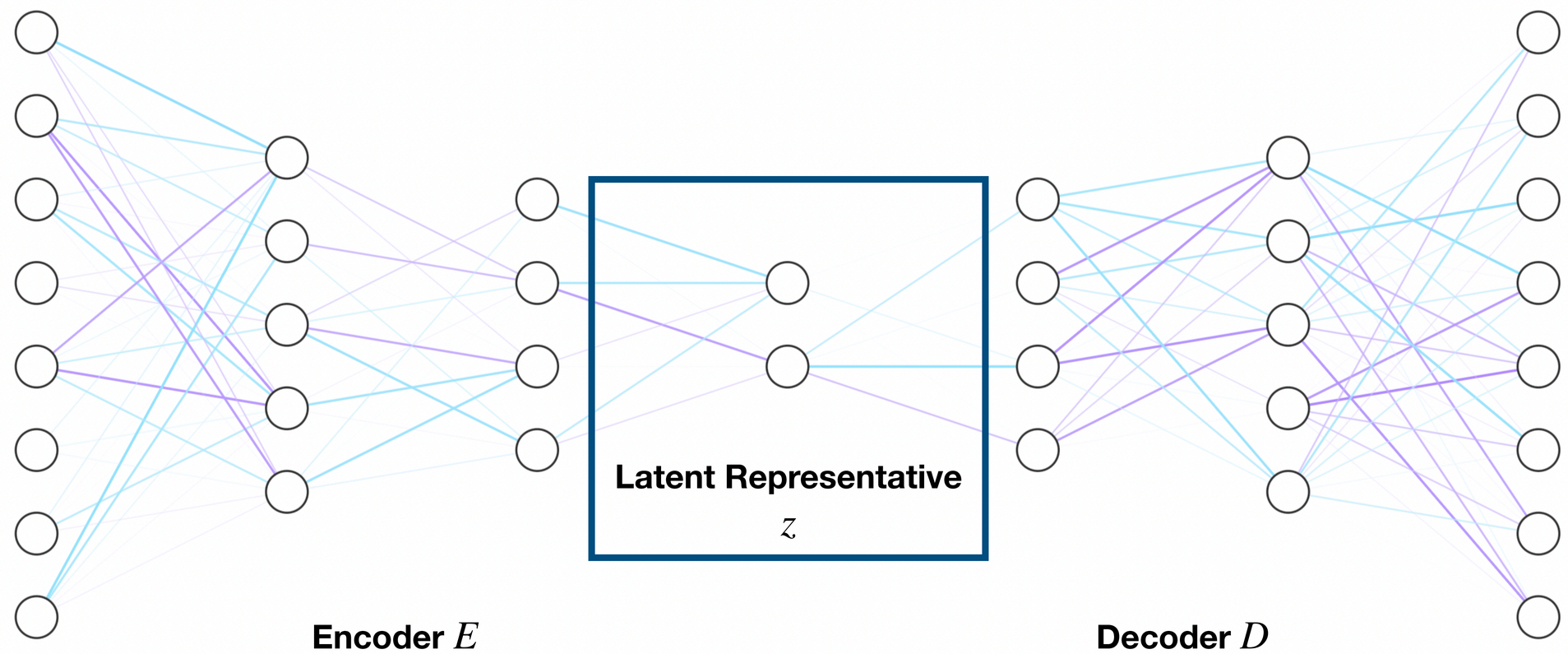


Disentanglement



Network Conditioning

What can DC do? - Disentanglement



What can DC do? - Disentanglement

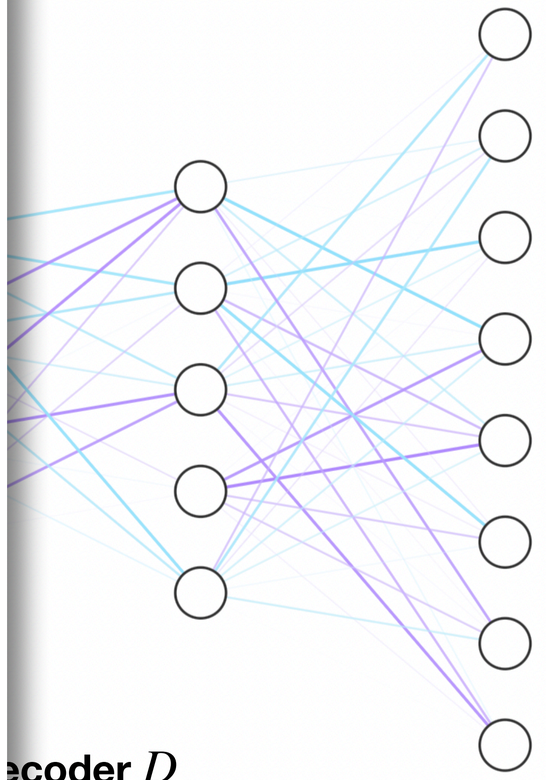
An Image is Worth More Than a Thousand Words: Towards Disentanglement in the Wild

Aviv Gabbay **Niv Cohen** **Yedid Hoshen**
School of Computer Science and Engineering
The Hebrew University of Jerusalem, Israel

Project webpage: http://www.vision.huji.ac.il/zerodim

Abstract

Unsupervised disentanglement has been shown to be theoretically impossible without inductive biases on the models and the data. As an alternative approach, recent methods rely on limited supervision to disentangle the factors of variation and allow their identifiability. While annotating the true generative factors is only required for a limited number of observations, we argue that it is infeasible to enumerate all the factors of variation that describe a real-world image distribution.



ecoder D

What can DC do? - Disentanglement

An Image is Worth
Towards Dis

Aviv Gabbay
School of C
The Hebre

Project webpage

Unsupervised disentangleme
without inductive biases on t
recent methods rely on limite
and allow their identifiability.
required for a limited numbe
enumerate all the factors of va

Disentangling Disentanglement in Variational Autoencoders

Emile Mathieu [*1] **Tom Rainforth** [*1] **N. Siddharth** [*2] **Yee Whye Teh** [1]

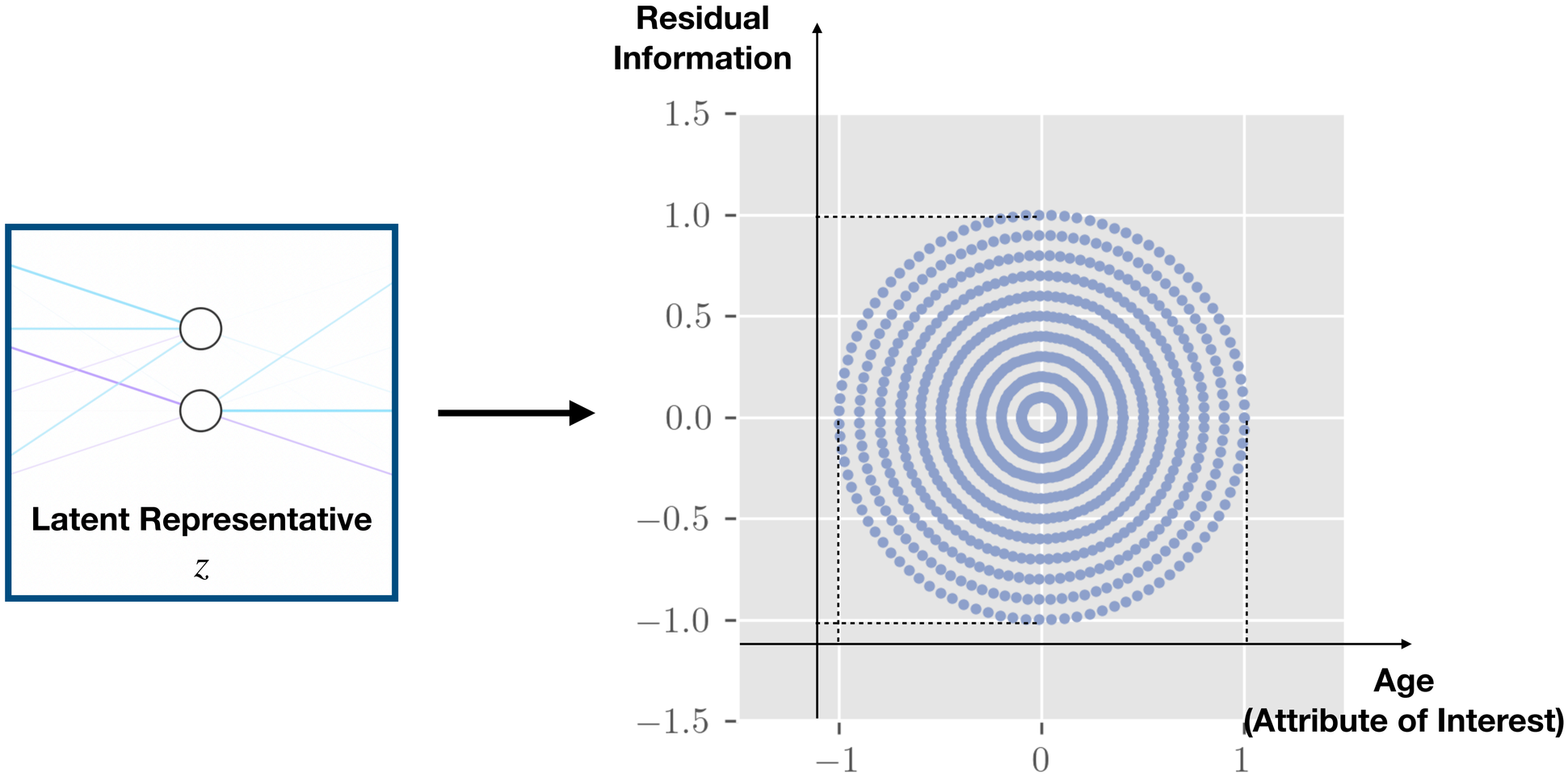
Abstract

We develop a generalisation of disentanglement in variational autoencoders (VAEs)—*decomposition* of the latent representation—characterising it as the fulfilment of two factors: a) the latent encodings of the data having an appropriate level of overlap, and b) the aggregate encoding of the data conforming to a desired structure, represented through the prior. Decomposition permits disentanglement, i.e. explicit independence between latents, as a special case, but also allows for a much richer class of properties to be imposed on the learnt representation, such as sparsity, clus-

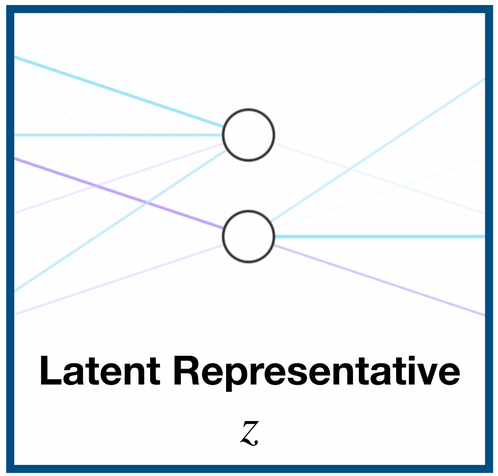
independent factors of variation (Alemi et al., 2017; Ansari and Soh, 2019; Burgess et al., 2018; Chen et al., 2018; 2017; Eastwood and Williams, 2018; Esmaeili et al., 2019; Higgins et al., 2016; Kim and Mnih, 2018; Xu and Durrett, 2018; Zhao et al., 2017), typically evaluating this using purpose-built, synthetic data (Eastwood and Williams, 2018; Higgins et al., 2016; Kim and Mnih, 2018), whose generative factors are independent by construction.

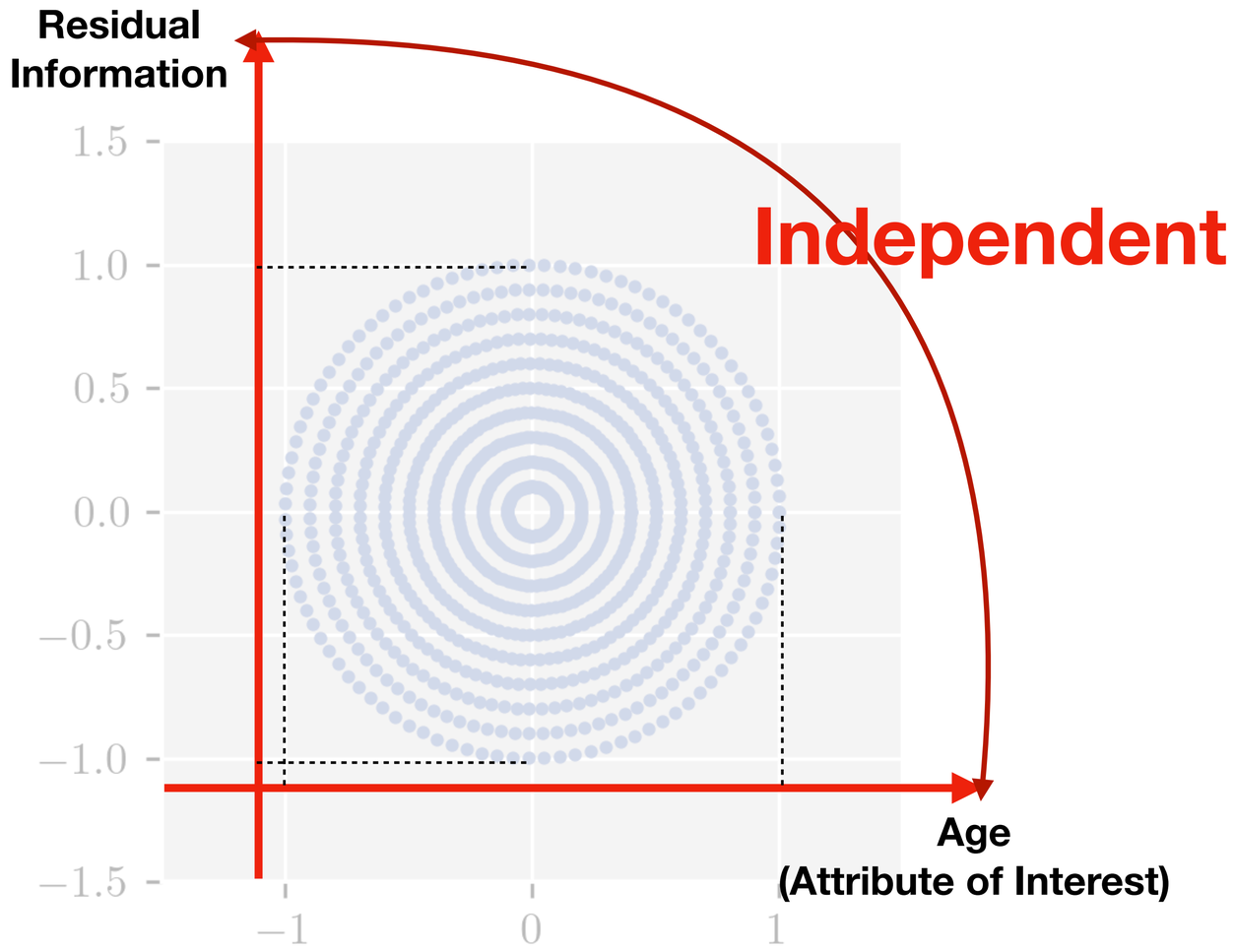
This conventional view of disentanglement, as recovering independence, has subsequently motivated the development of formal evaluation metrics for independence (Eastwood and Williams, 2018; Kim and Mnih, 2018), which in turn has driven the development of objectives that target these metrics, often by employing regularisers explicitly encour-

What can DC do? - Disentanglement

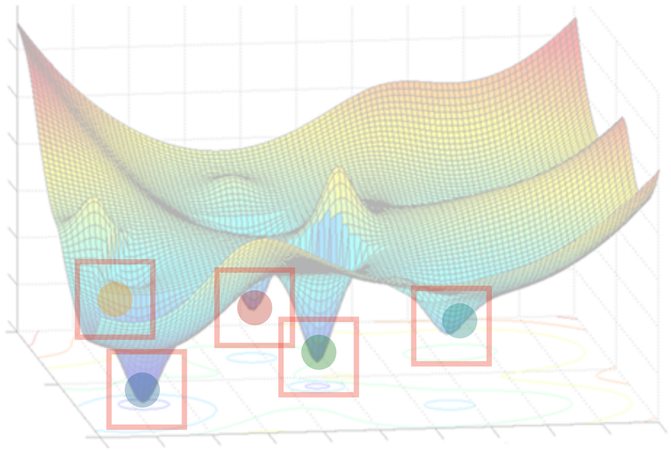


What can DC do? - Disentanglement

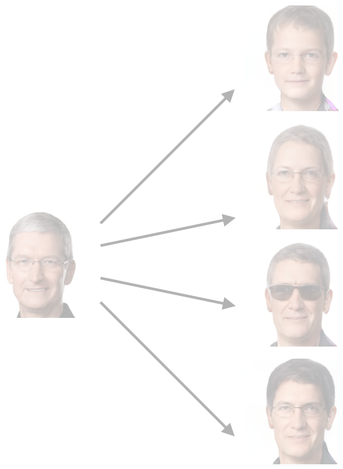




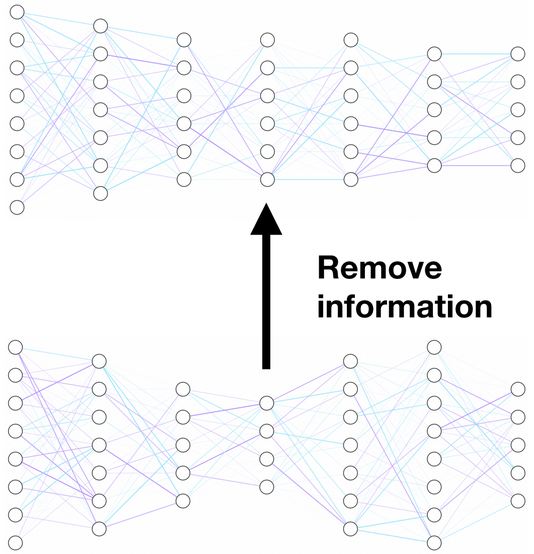
Use Cases



Diverge Training



Disentanglement



Network Conditioning

Nuisance Variables

Smoking

Age

Lung Cancer

Gender

Nuisance Variables

Attribute of interest

Smoking

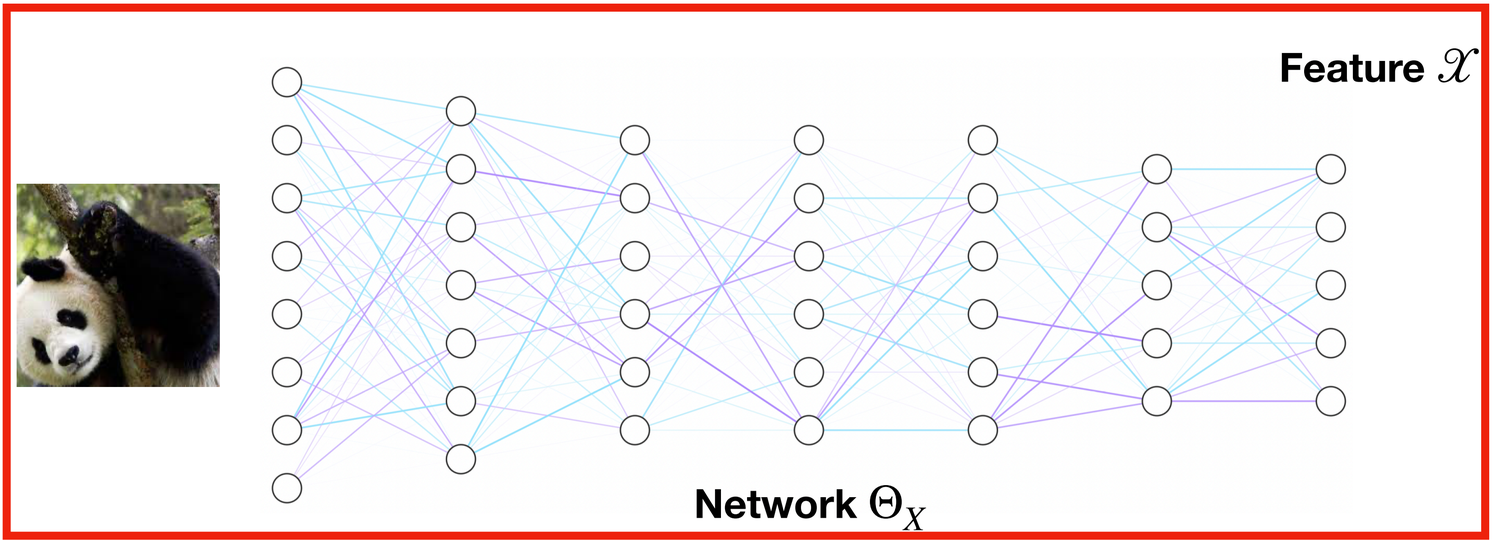
Lung Cancer

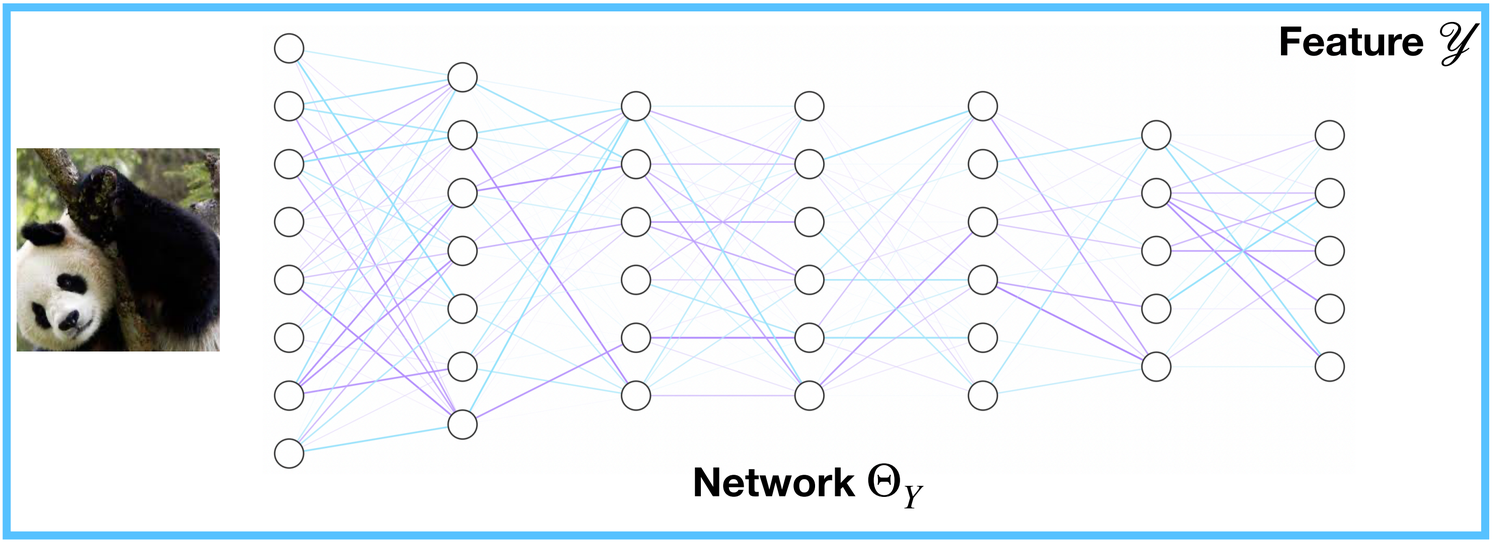
Age

Gender

Nuisance variables

Network as Nuisance Variable





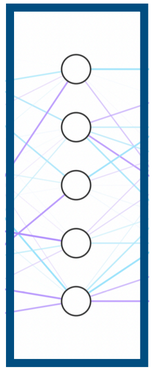
What does it mean by treating Θ_Y as nuisance variable?

What does Θ_X learn that Θ_Y doesn't?

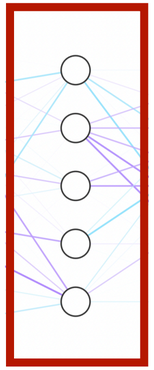
Nuisance variables

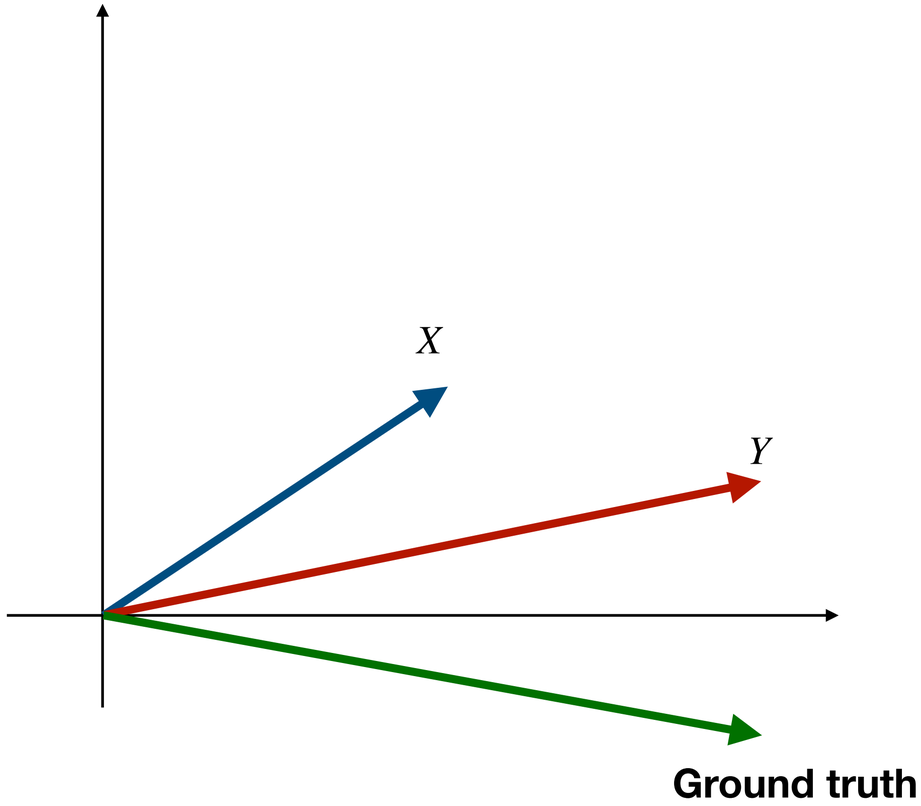
Linear Regression

Feature $\mathcal{X}$



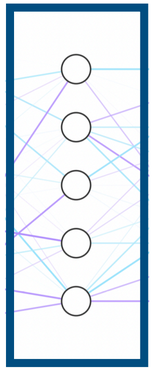
Feature $\mathcal{Y}$



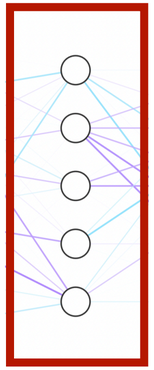


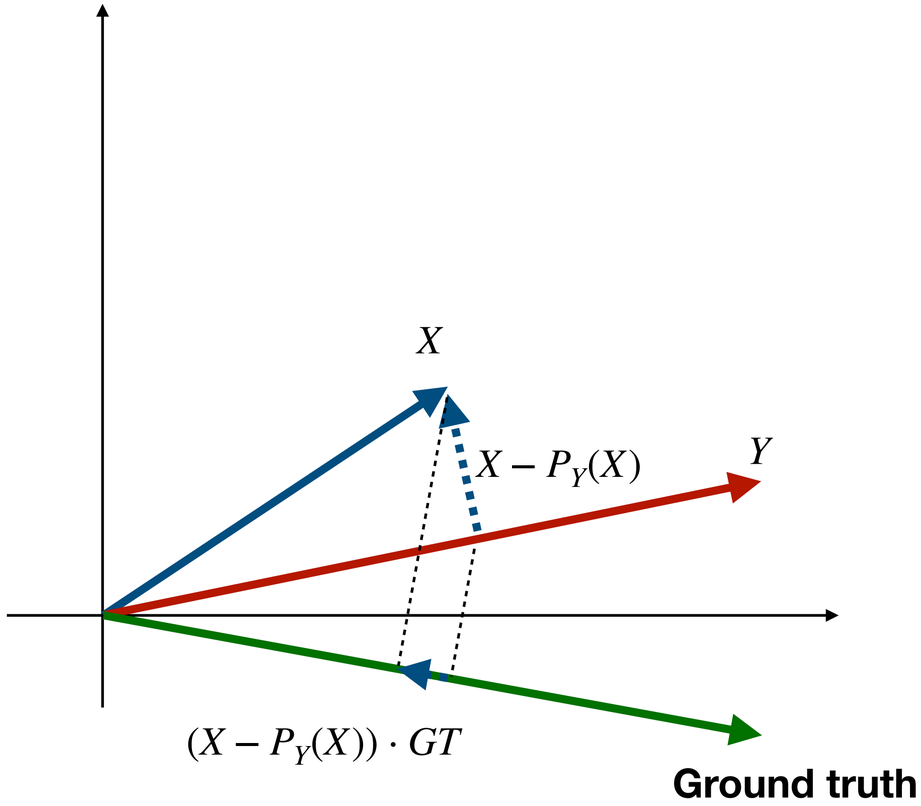
Linear Regression

Feature $\mathcal{X}$



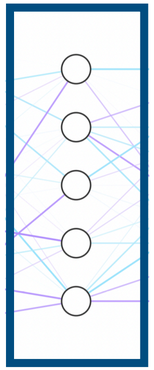
Feature $\mathcal{Y}$



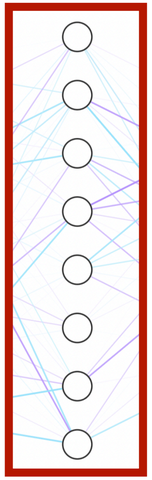


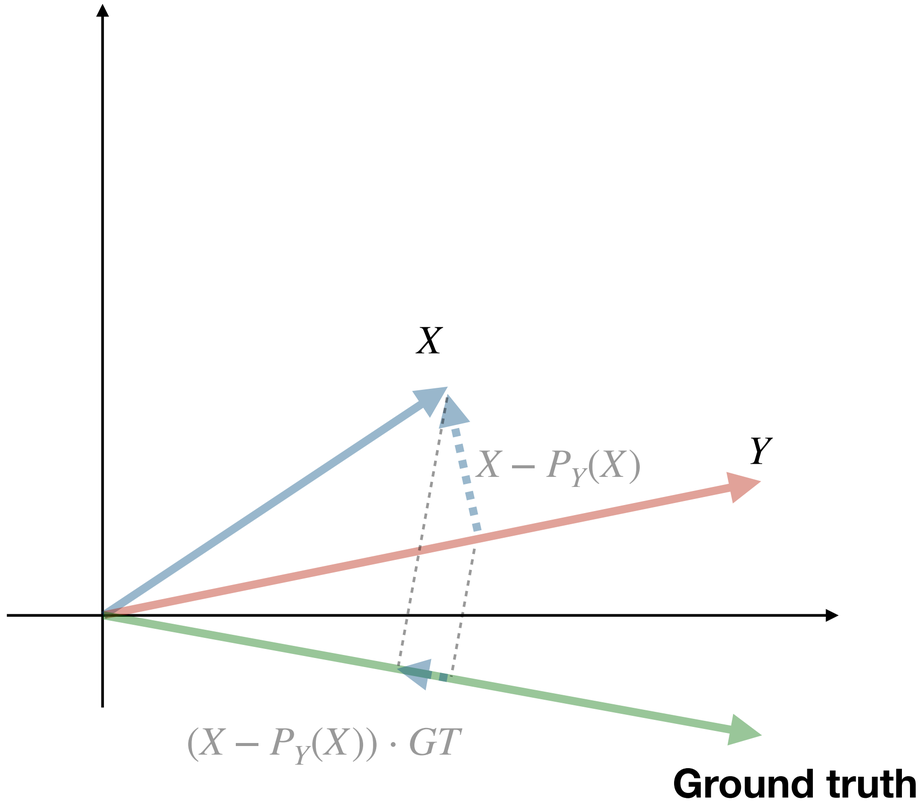
Linear Regression

Feature $\mathcal{X}$



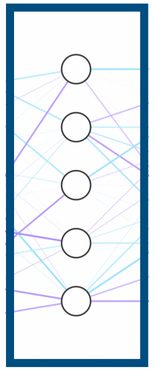
Feature $\mathcal{Y}$



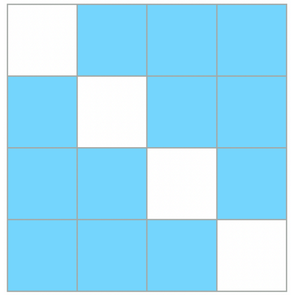


Partial Distance Correlation

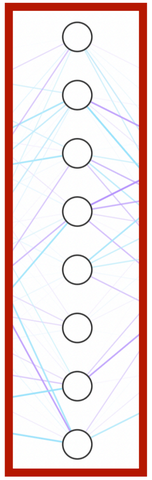
Feature $\mathcal{X}$



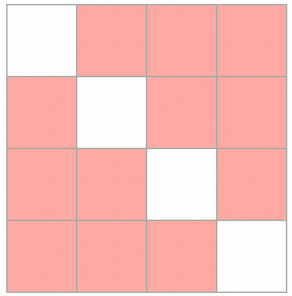
Distance Matrix A

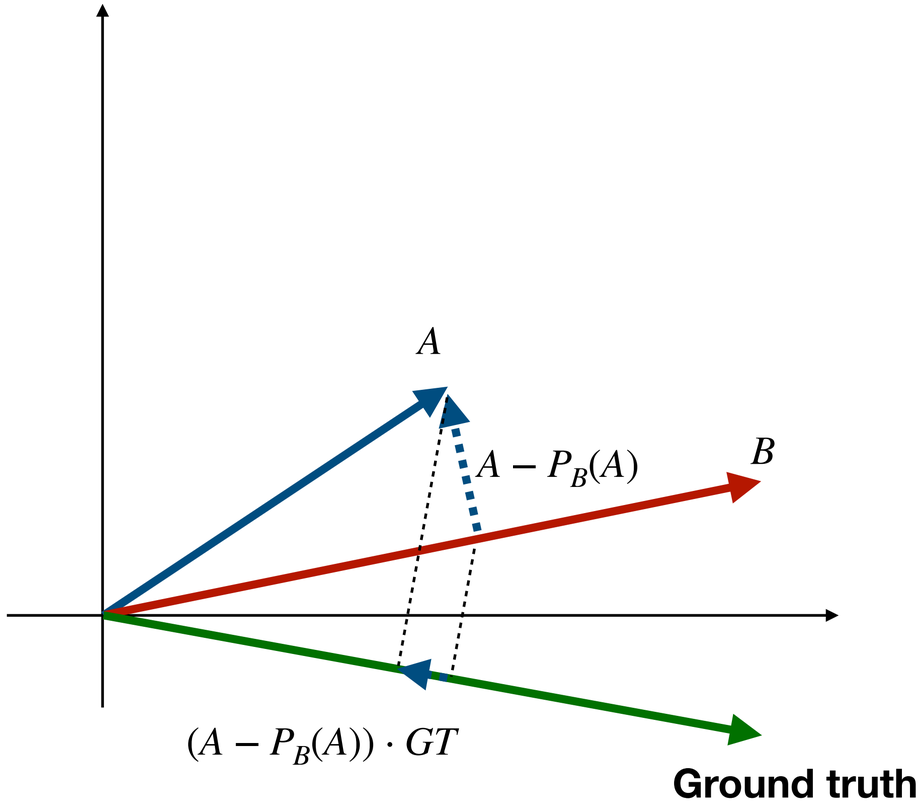


Feature $\mathcal{Y}$

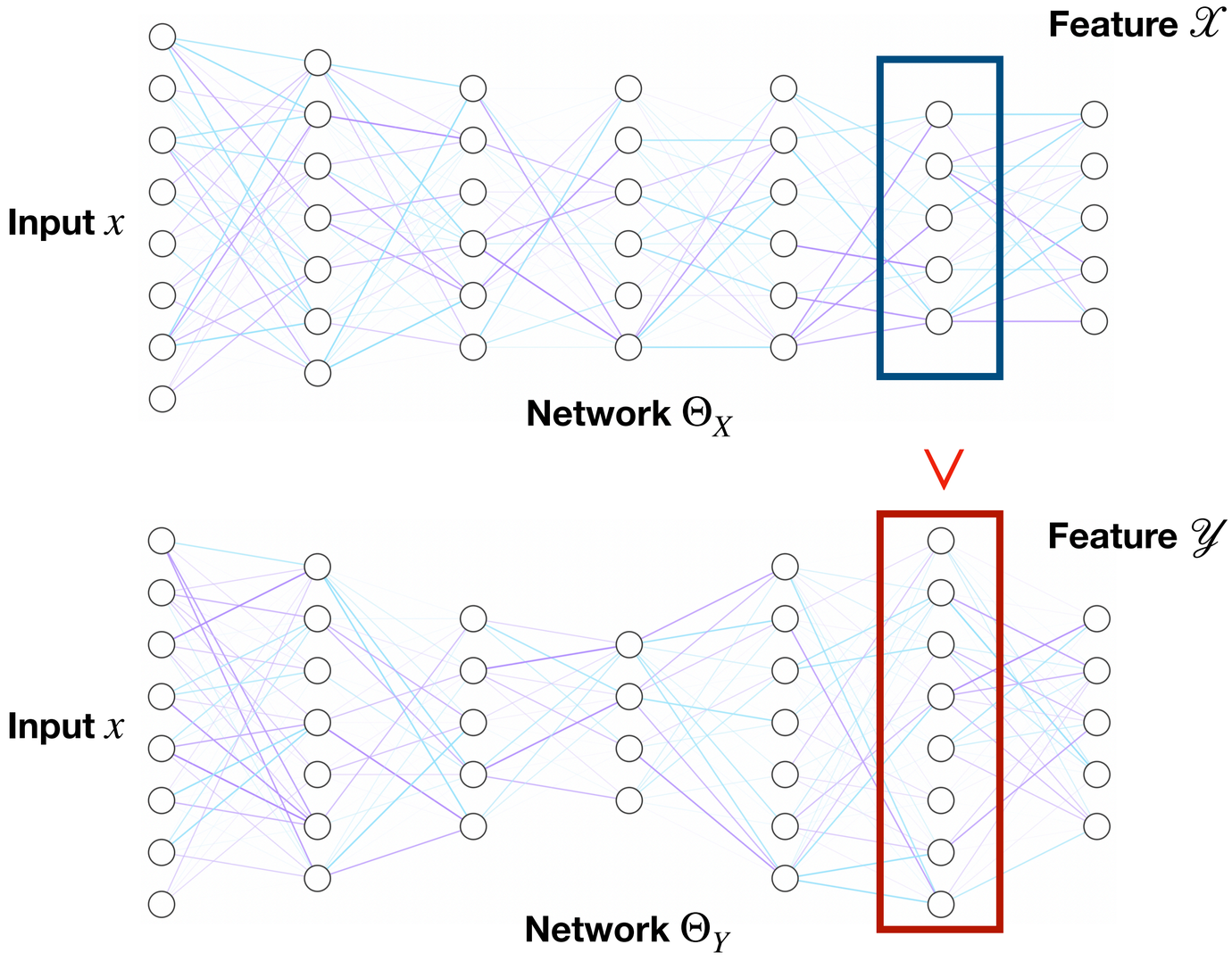


Distance Matrix B



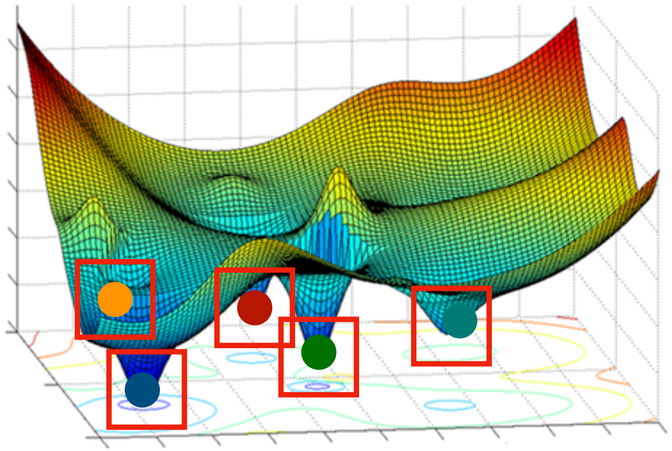


Partial Distance Correlation

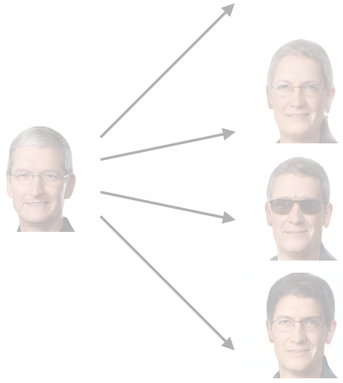


Θ_X, Θ_Y can be different types of network, such as ViT and Resnet

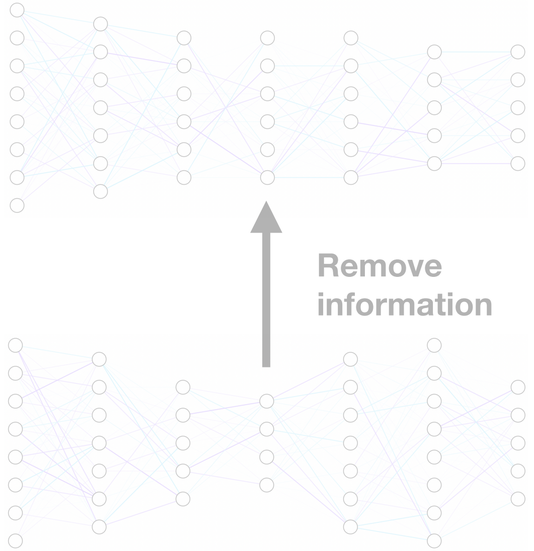
Experimental Results



Diverge Training

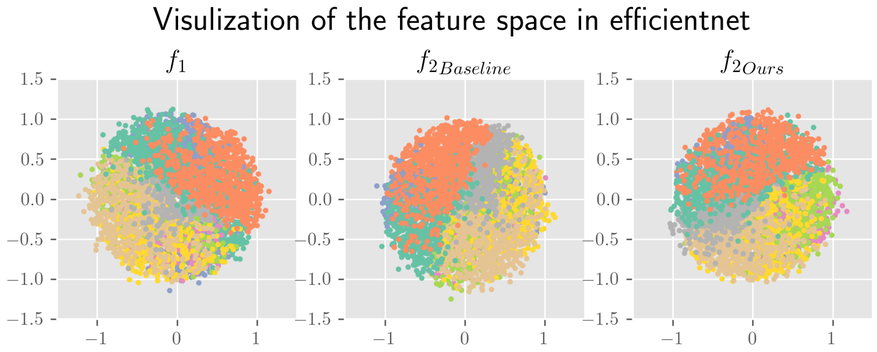


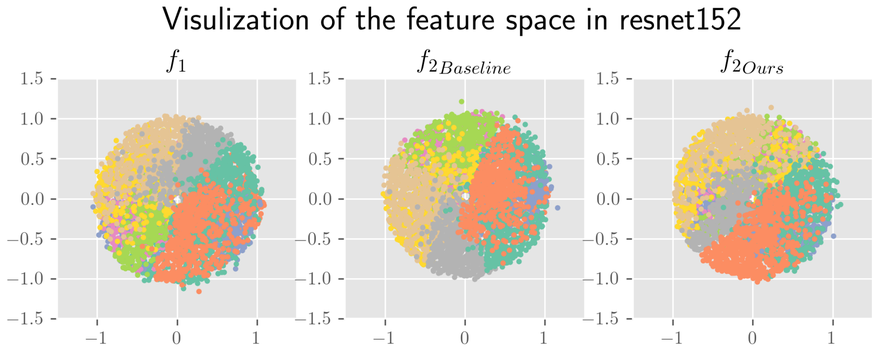
Disentanglement

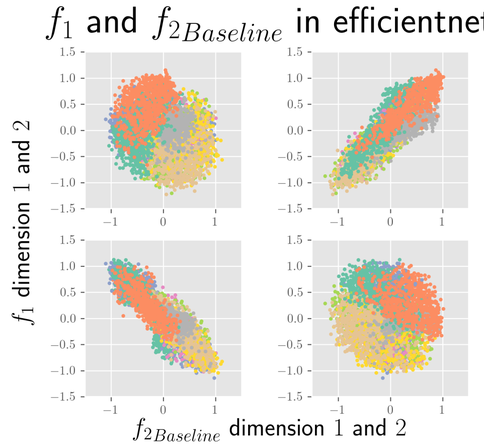


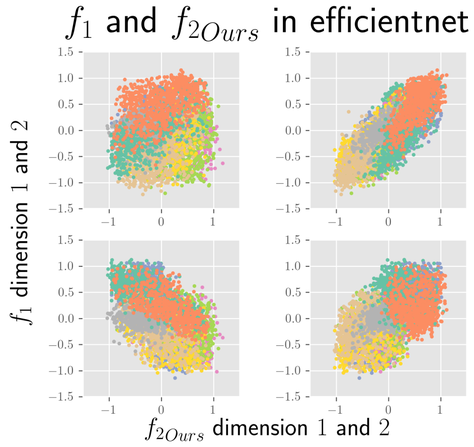
Network Conditioning

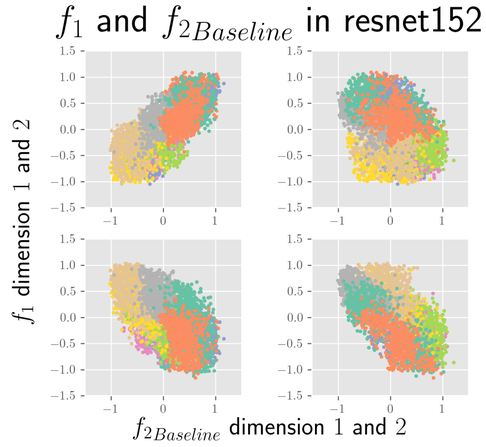
Experimental Results

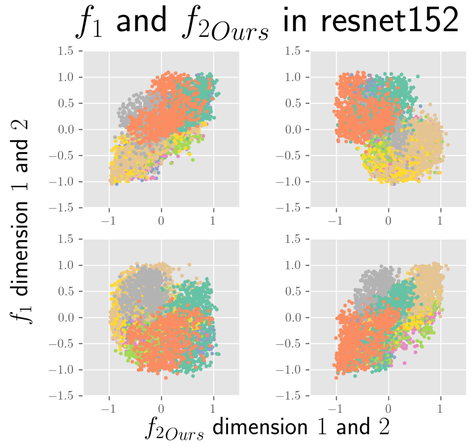




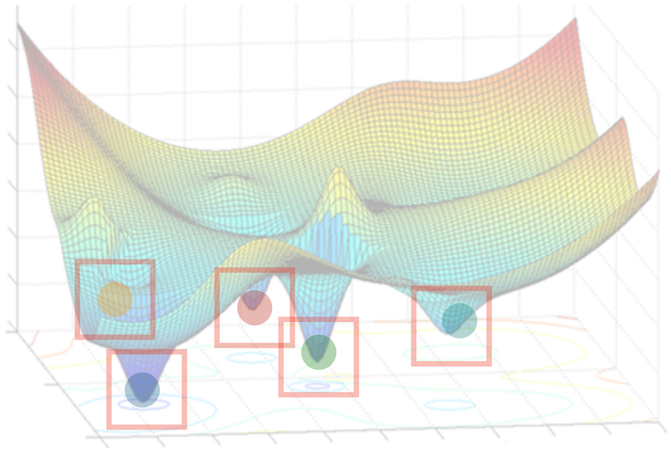




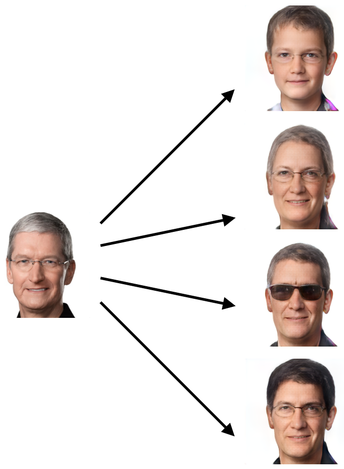




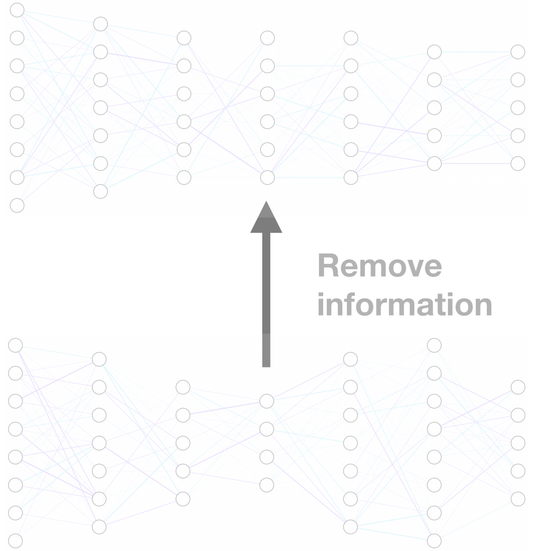
Experimental Results



Diverge Training

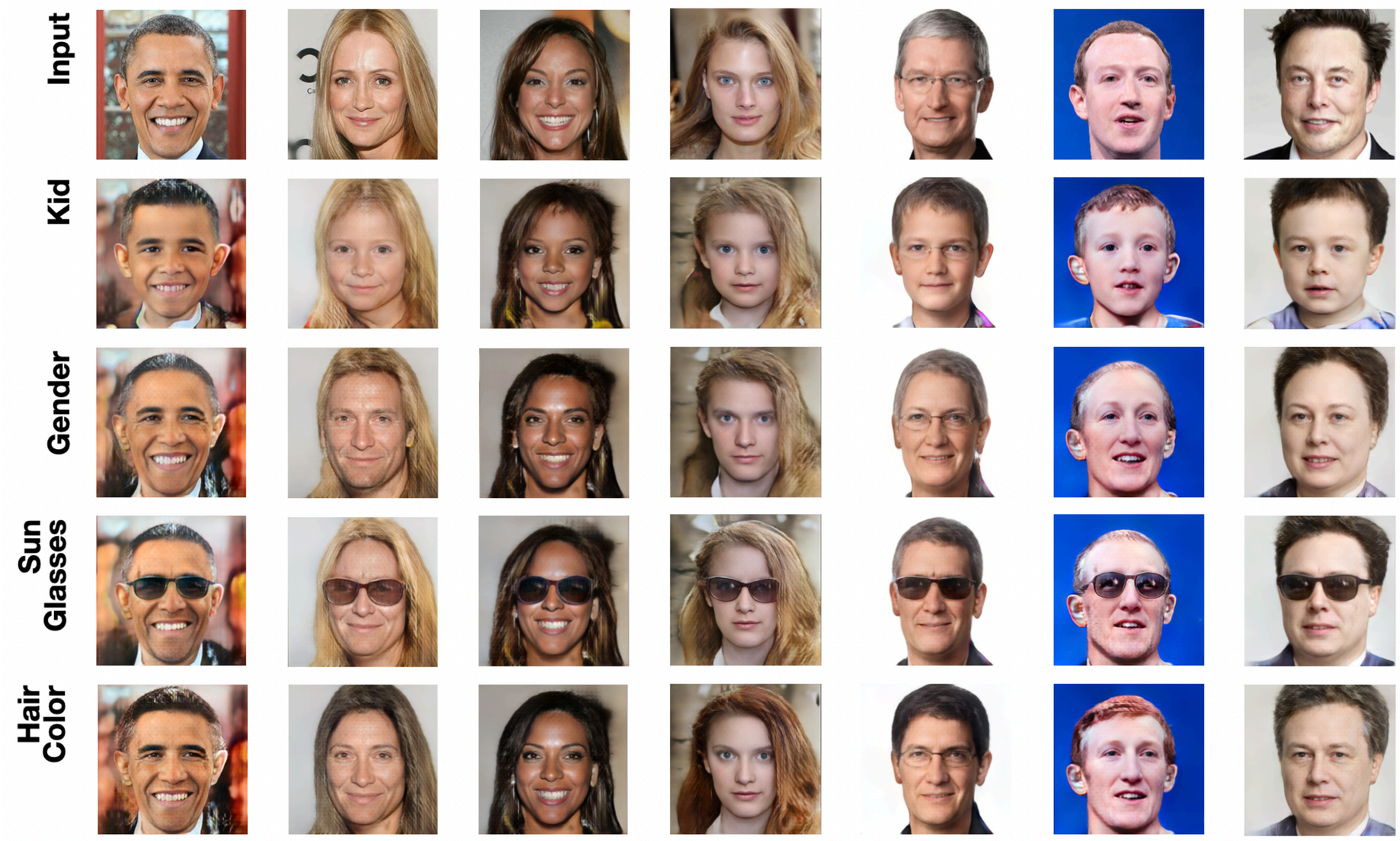


Disentanglement

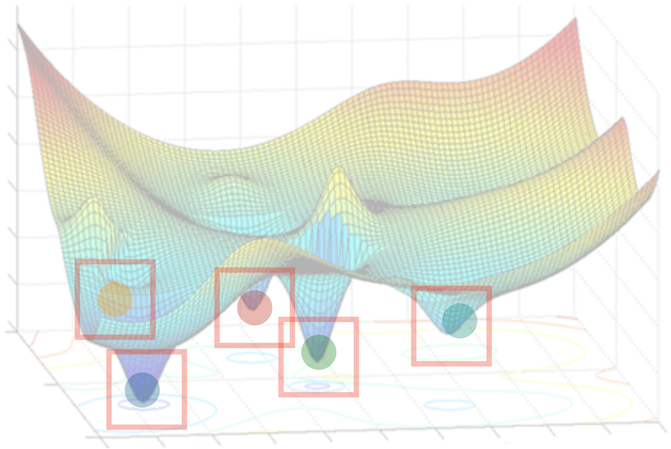


Network Conditioning

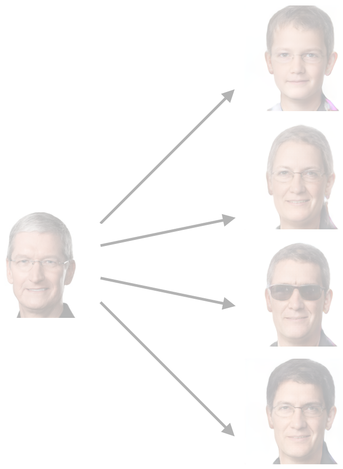
Experimental Results



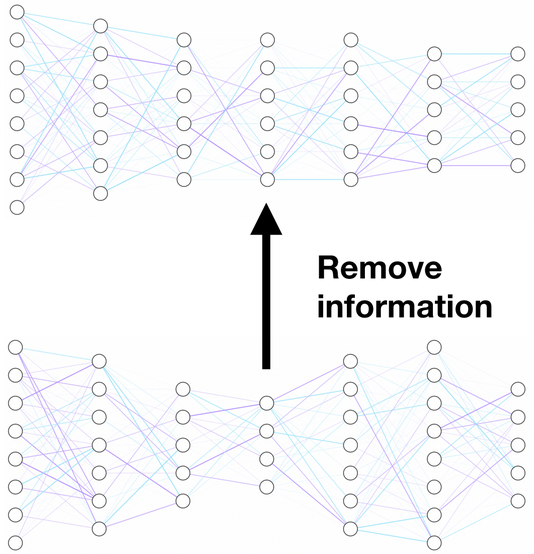
Experimental Results



Diverge Training

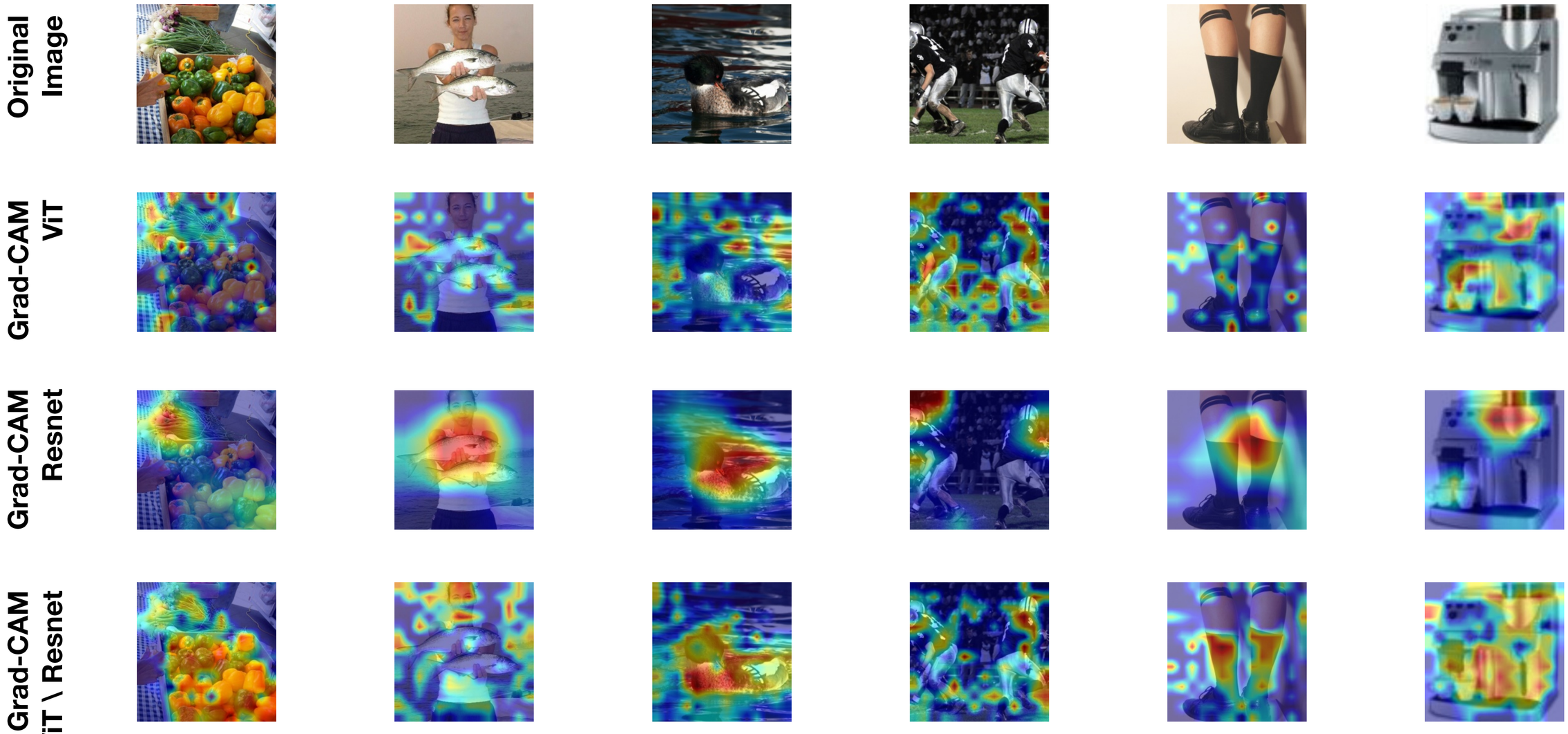


Disentanglement



Network Conditioning

Experimental Results



On the Versatile Uses of Partial Distance Correlation in Deep Learning

Thanks