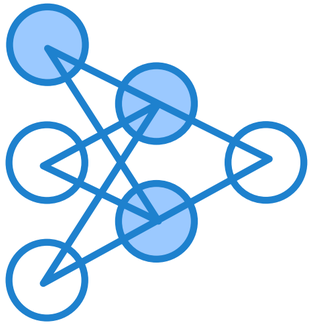


New Jersey Institute of Technology

Understanding Deep Learning requires rethinking generalization

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals

Naren Khatwani, Sabbir Ahmed Saqlain

Introduction

Deep Learning:

- A subset of machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data.

- **Generalization:** The ability of a model to **perform well on new, unseen data,** not just on the data it was trained on.

Introduction

Deep Learning:

- A subset of machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data.

- **Generalization:** The ability of a model to **perform well on new, unseen data,** not just on the data it was trained on.

Importance of Generalization:

- **Indicator of Success:** The goal of a neural network is not just to memorize the training data but to make accurate predictions on new data.

- **Key Challenge**: Designing **neural networks that generalize** well is a central challenge in machine learning.

Motivation

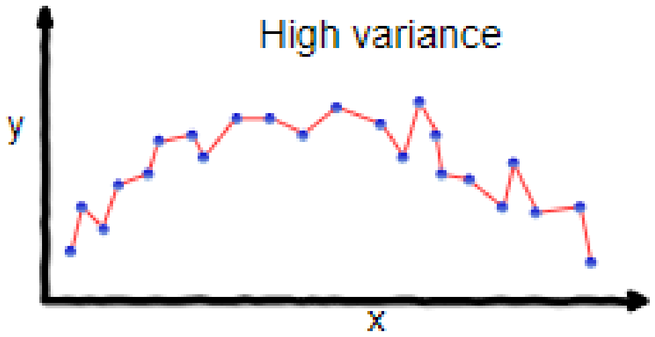
Why Study This Problem? Why did the authors choose this problem?

- **Unexpected Behavior:** Despite over-parameterization (having more parameters than training examples), deep neural networks often generalize well, a counterintuitive phenomenon given traditional statistical learning theory.

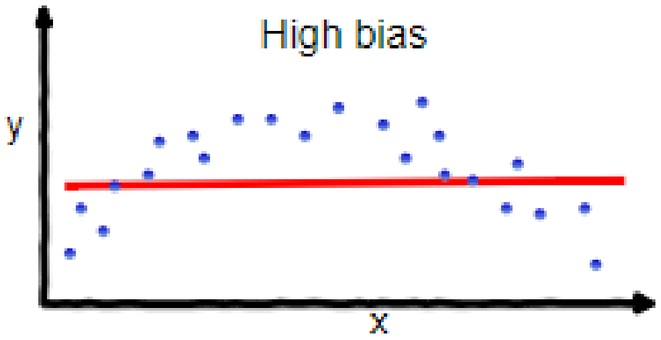
- **Need for Rethinking:** Existing theoretical frameworks (like VC-dimension and Rademacher complexity) are inadequate for explaining why these models can still generalize effectively.

If you Google “bias variance tradeoff”, then you get .....

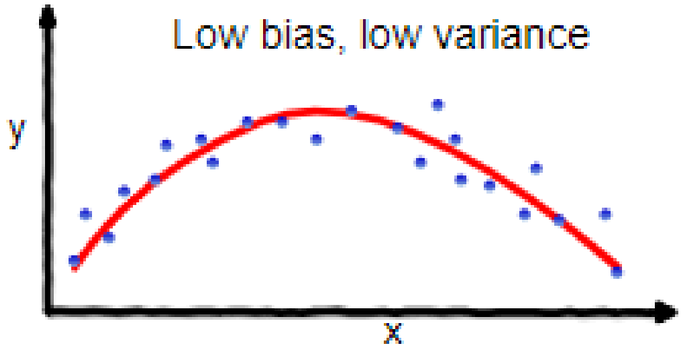
Motivation



Overfitting

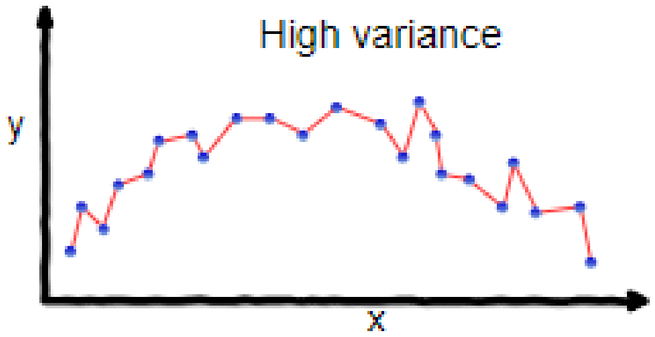


Underfitting

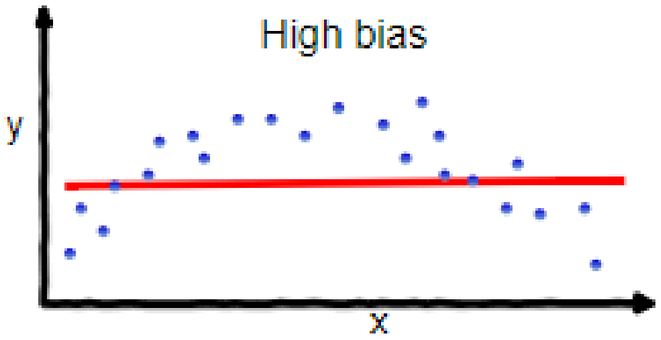


Good Balance

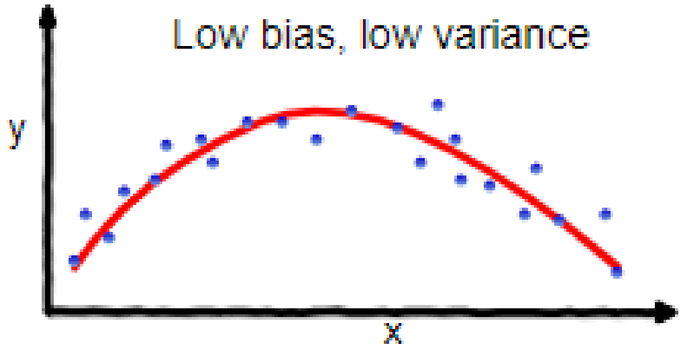
Motivation



Overfitting



Underfitting



Good Balance

State of the art deep networks do this ...

Problem Statement

Goals of the Paper:

- **Exploratory Approach:** To investigate and **understand the limitations of existing generalization theories** in the context of modern deep learning technologies.
- **Practical Impact:** **Insights** from this study **could lead to better design and training strategies** for neural networks, enhancing their performance and reliability in real-world applications.

Key points discussed in this paper

Problem Statement

Goals of the Paper:

- **Exploratory Approach:** To investigate and **understand the limitations of existing generalization theories** in the context of modern deep learning technologies.
- **Practical Impact:** Insights from this study could lead to better design and training strategies for neural networks, enhancing their performance and reliability in real-world applications.

Key points discussed in this paper

- **Failure of Traditional Measures**
- **Randomization Tests**
- **Re-evaluation of Regularization**
- **Implicit Regularization by Optimization Algorithms**
- **Theoretical Insights**
- **Call for New Theoretical Developments**

Problem Statement

Goals of the Paper:

- **Exploratory Approach:** To investigate and **understand the limitations of existing generalization theories** in the context of modern deep learning technologies.
- **Practical Impact:** Insights from this study could lead to better design and training strategies for neural networks, enhancing their performance and reliability in real-world applications.

Key points discussed in this paper

- **Failure of Traditional Measures**
- **Randomization Tests**
- **Re-evaluation of Regularization**
- **Implicit Regularization by Optimization Algorithms**
- **Theoretical Insights**
- **Call for New Theoretical Developments**

In summary, while the paper does not offer direct solutions or a new comprehensive generalization theory, it provides valuable insights and experimental evidence that challenge existing beliefs and theories.

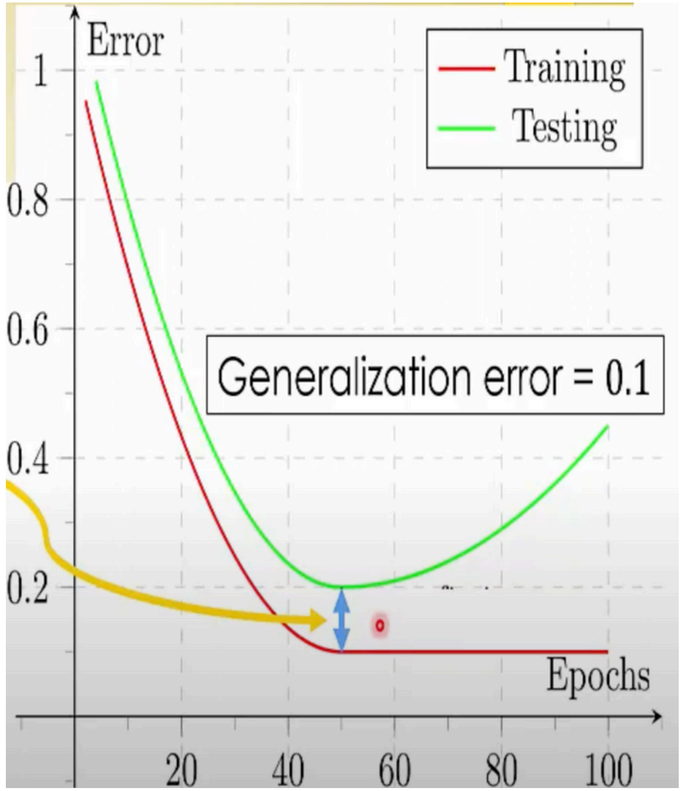
Background

Generalization Error

- This refers to measuring **how accurately a** machine learning model **can predict outcome values for previously unseen data**.

- In simpler terms, it is the difference in performance between when a model runs on the training dataset on which it was trained and when it is applied to a new, independent dataset (often called the test set).

Difference between Training and Test Error

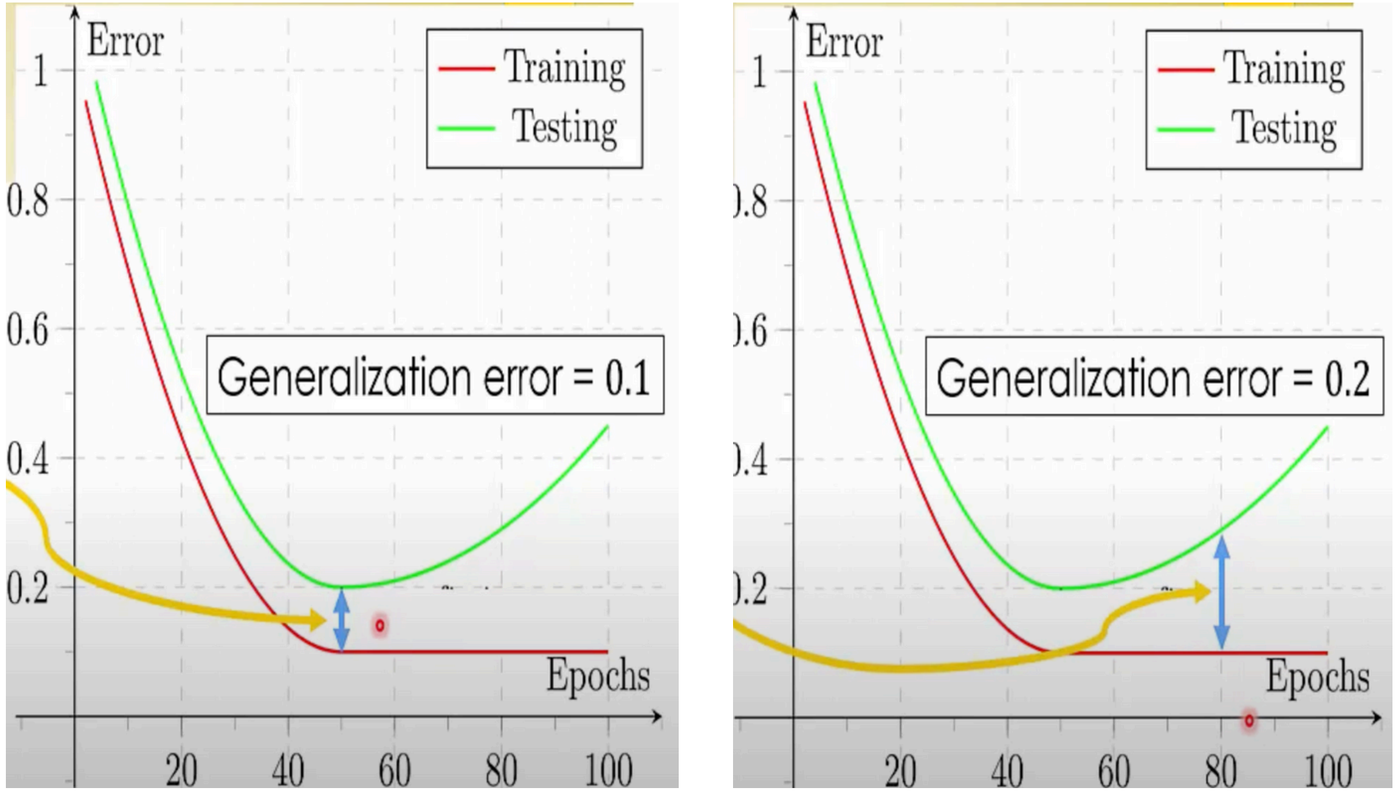


Generalization Error

- This refers to measuring **how accurately a** machine learning model **can predict outcome values for previously unseen data**.

- In simpler terms, it is the difference in performance between when a model runs on the training dataset on which it was trained and when it is applied to a new, independent dataset (often called the test set).

Difference between Training and Test Error

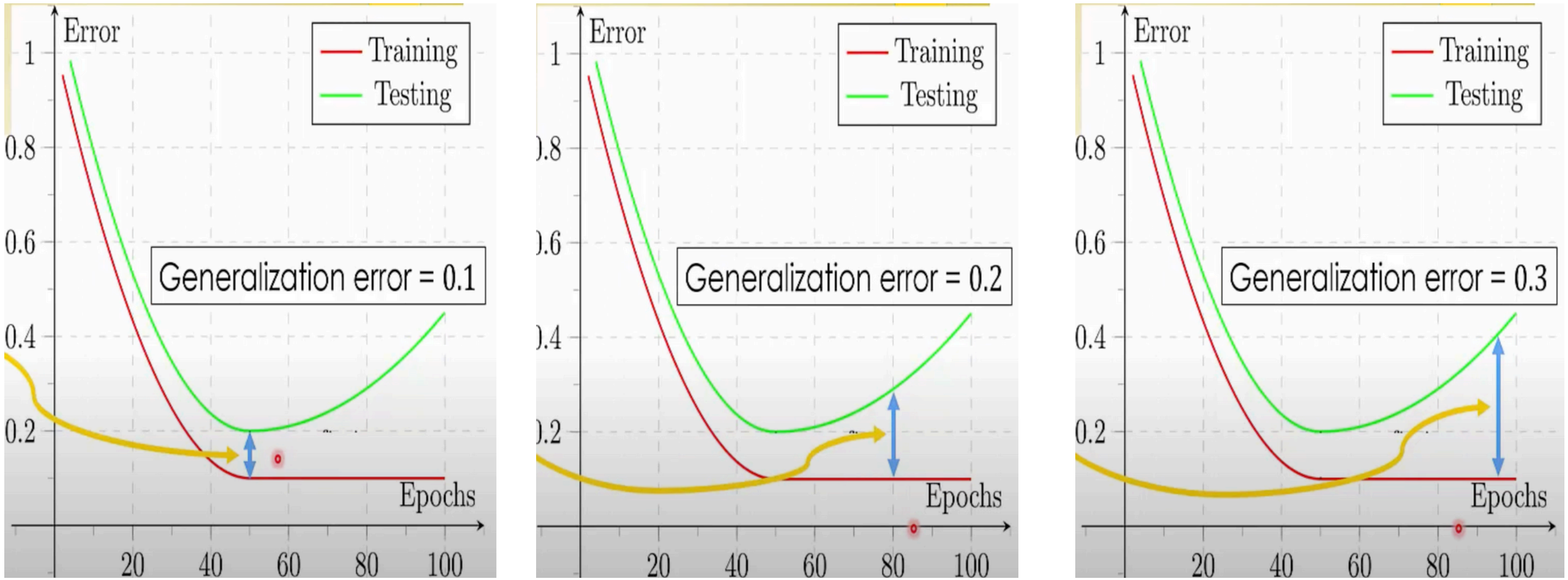


Generalization Error

- This refers to measuring **how accurately a** machine learning model **can predict outcome values for previously unseen data**.

- In simpler terms, it is the difference in performance between when a model runs on the training dataset on which it was trained and when it is applied to a new, independent dataset (often called the test set).

Difference between Training and Test Error



Universal Approximation Theorem

Overview of the Theorem:

- It is fundamental in neural network theory that a feedforward network with a **single hidden layer containing** a finite number of neurons (or "units") **can approximate any continuous function** on compact subsets of $\mathbb{R}^n$ to a desired degree of accuracy, given appropriate parameters and activation functions.

- It does not define the algorithmic learnability of those parameters.

Universal Approximation Theorem

Overview of the Theorem:

- It is fundamental in neural network theory that a feedforward network with a **single hidden layer containing** a finite number of neurons (or "units") **can approximate any continuous function** on compact subsets of $\mathbb{R}^n$ to a desired degree of accuracy, given appropriate parameters and activation functions.

- It does not define the algorithmic learnability of those parameters.

Goal : Train a neural network that can accurately predict the price of a house, given inputs
Input Variables: size, bedrooms, bathrooms, age, neighborhood income

Universal Approximation Theorem

Overview of the Theorem:

- It is fundamental in neural network theory that a feedforward network with a **single hidden layer containing** a finite number of neurons (or "units") **can approximate any continuous function** on compact subsets of $\mathbb{R}^n$ to a desired degree of accuracy, given appropriate parameters and activation functions.

- It does not define the algorithmic learnability of those parameters.

Goal : Train a neural network that can accurately predict the price of a house, given inputs
Input Variables: size, bedrooms, bathrooms, age, neighborhood income

- **Feedforward Neural Network with a Single Hidden Layer**
- **Finite Number of Neurons :**
  - Assuming 10 in this case
- **Approximate Any Continuous Function**
  - The function we want to approximate is the house pricing function
  - Continuity: Small changes in input leads to small changes in house prices

Universal Approximation Theorem

Overview of the Theorem:

- It is fundamental in neural network theory that a feedforward network with a **single hidden layer containing** a finite number of neurons (or "units") **can approximate any continuous function** on compact subsets of $\mathbb{R}^n$ to a desired degree of accuracy, given appropriate parameters and activation functions.

- It does not define the algorithmic learnability of those parameters.

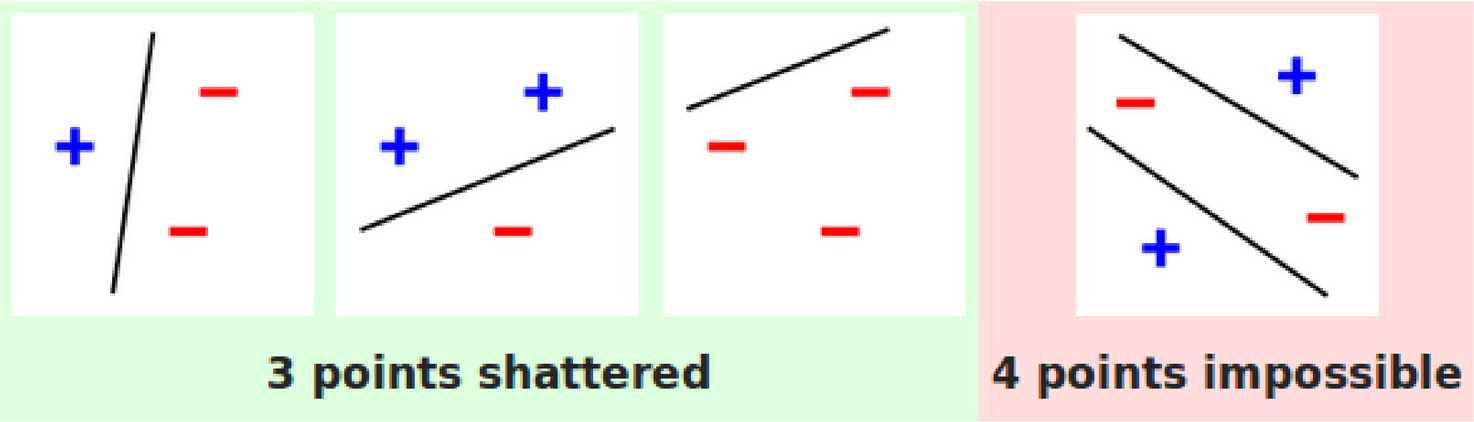
Goal : Train a neural network that can accurately predict the price of a house, given inputs
Input Variables: size, bedrooms, bathrooms, age, neighborhood income

- **Feedforward Neural Network with a Single Hidden Layer**
- **Finite Number of Neurons :**
  - Assuming 10 in this case
- **Approximate Any Continuous Function**
  - The function we want to approximate is the house pricing function
  - Continuity: Small changes in input leads to small changes in house prices
- **Desired Degree of Accuracy**
- **Function Defined on Compact Subsets of R^n**
  - In this case, n would be 5
  - Compact - within realistic and finite ranges of these features

Vapnik-Chervonenkis (VC) dimension

- A classification model ƒ with some parameter vector θ is said to shatter a set of data points (X1, X2,..., Xn) if, for all assignments of labels to those points, there exists a θ such that the model ƒ makes no errors when evaluating that set of data points.

- The VC dimension of a hypothesis space is the maximum number of points that can be shattered by hypotheses in that space. If you can't find any set of *n*+1 points that can be shattered, then the VC dimension of the hypothesis space is n.



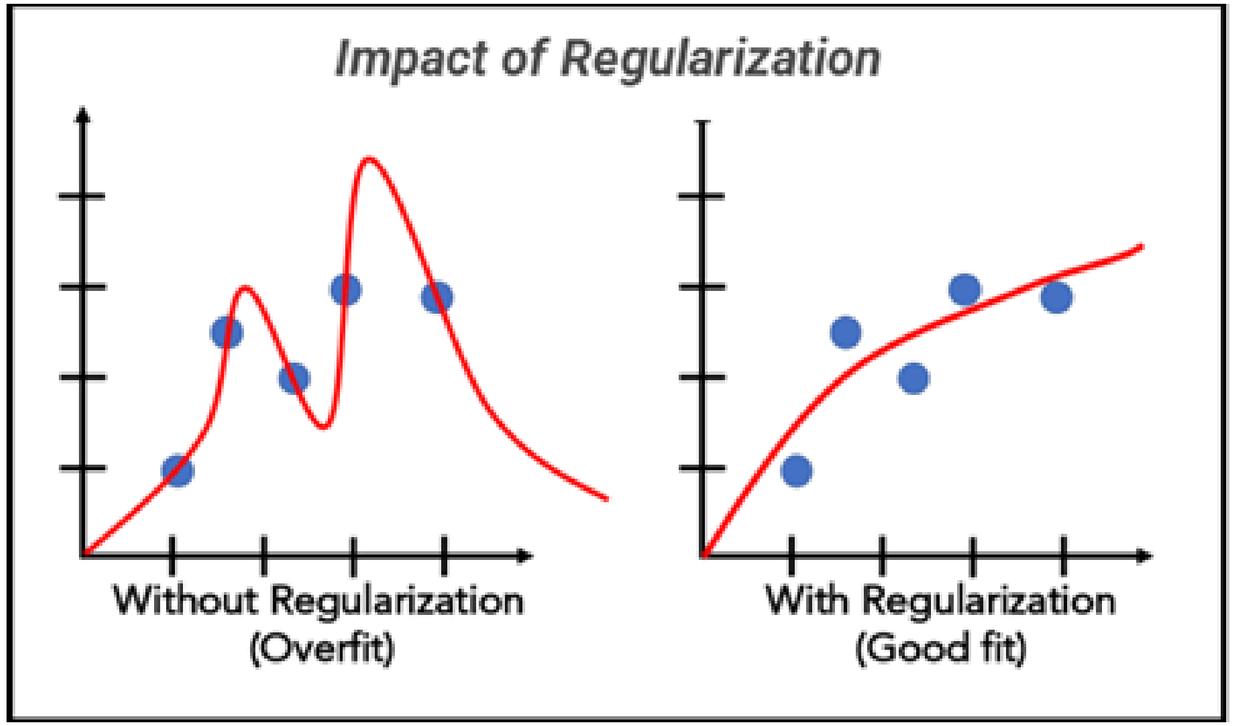
L2 Regularization – “Weight Decay”

$$L'(\theta) = L(\theta) + \lambda \sum_{i=1}^{n} \theta_i^2$$

- L(θ) is the original loss function
- θ represents the parameters (weights) of the model
- λ is a regularization parameter that controls the strength of the penalty

- In L2 regularization, the cost function is modified by adding a penalty term, **which is the sum of the squares of all the model parameters** (weights).
- This term is called the L2 norm of the weights, hence the L2 regularization.

L2 Regularization – “Weight Decay”

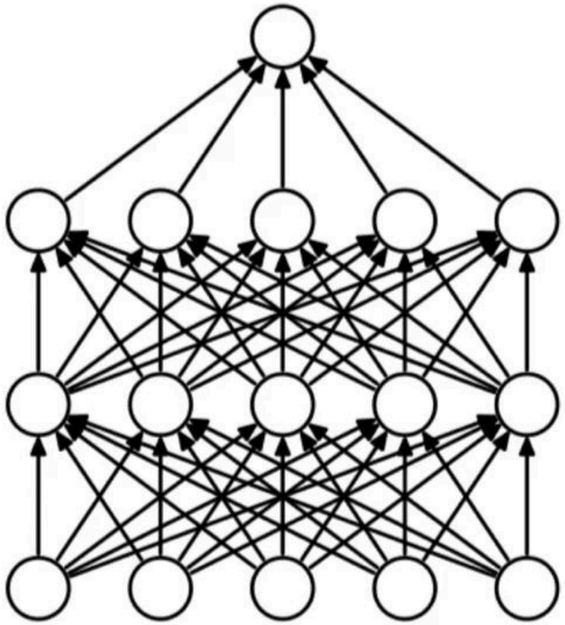


Forces the weights to become small i.e “decay”

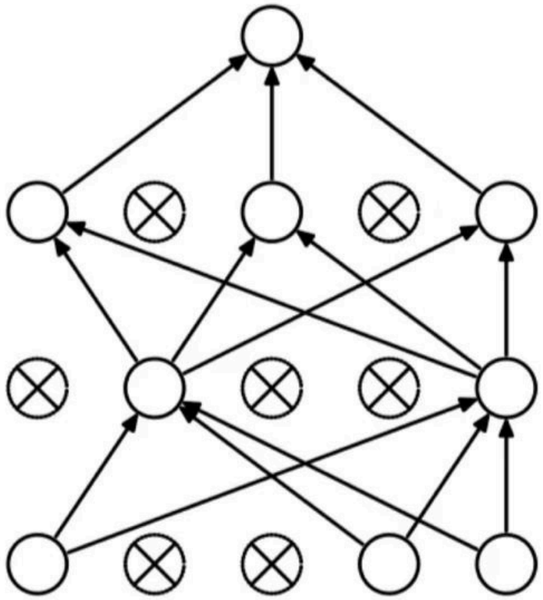
L2 regularization/weight decay serves as a tool to simplify the model, ensuring that its predictions are based on the most salient features, thus improving the model's ability to perform well on both the training data and new, unseen data.

Dropout

- Randomly drop neurons from layers in the network



(a) Standard Neural Net

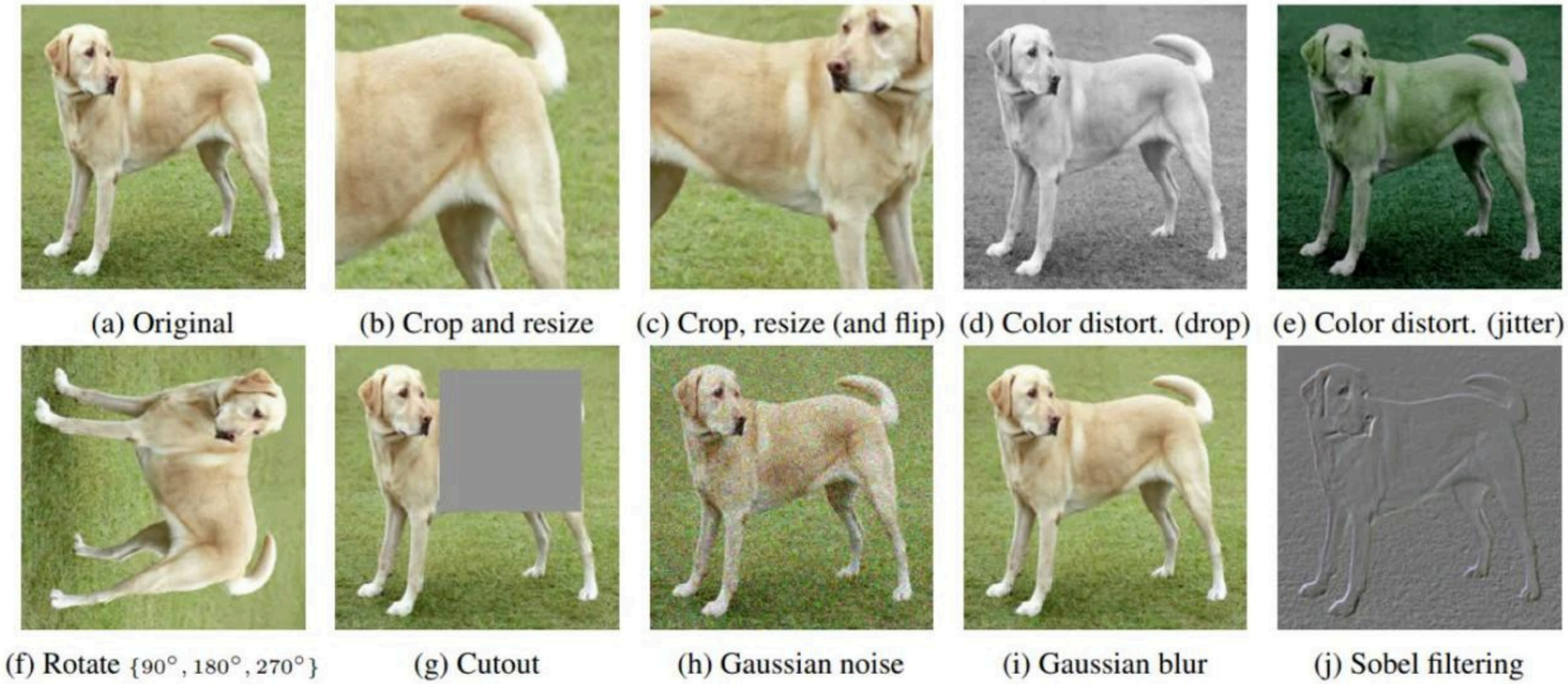


(b) After applying dropout.

- Removes reliance on individual neurons
- Learns redundancies
- Learns more nuanced set of feature detectors

Data Augmentation

- Domain-specific transformations of the input data
- Increases the input space (i.e all possible images we care about)



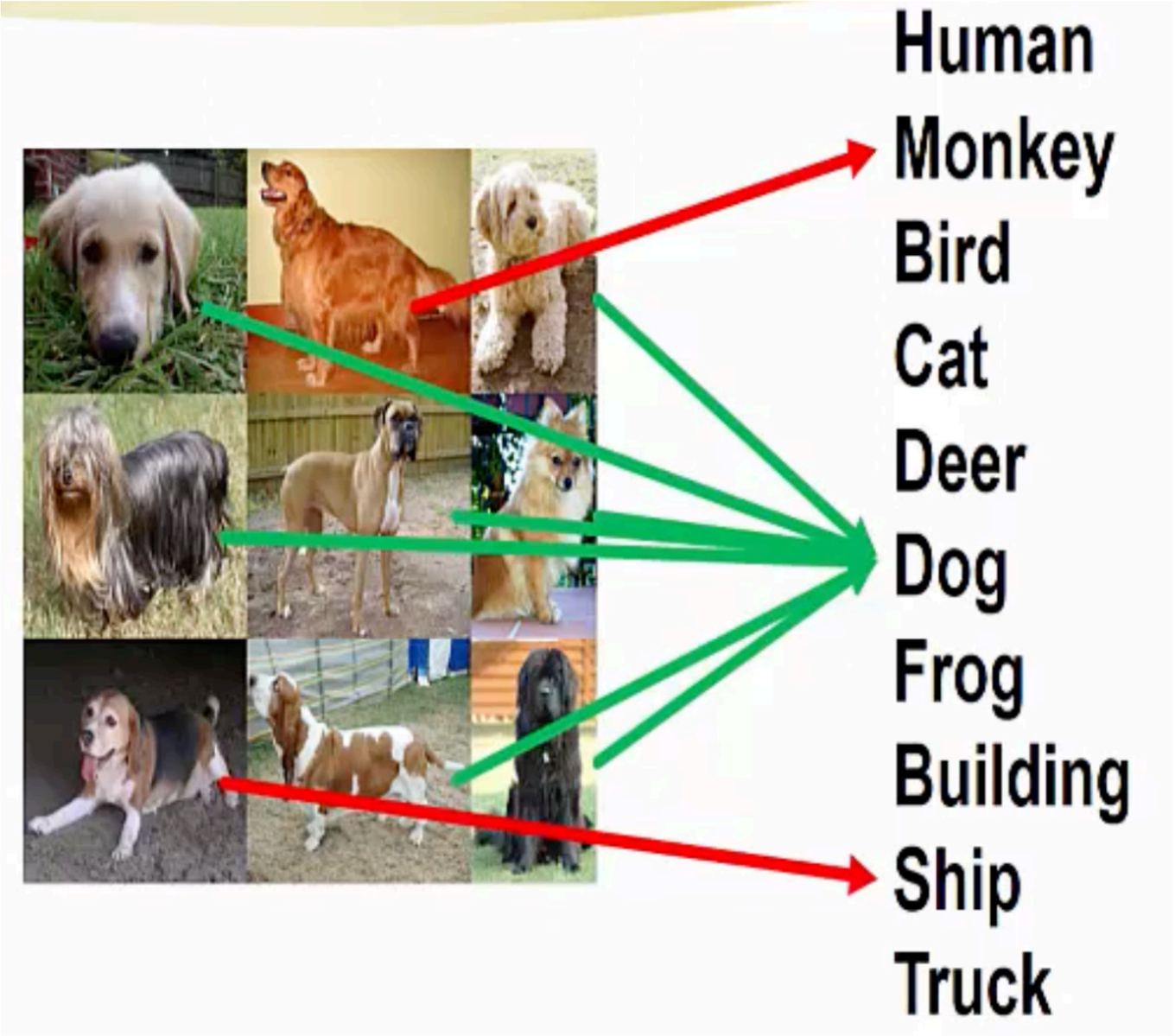
Experimental Findings

Randomization Tests

True Labels

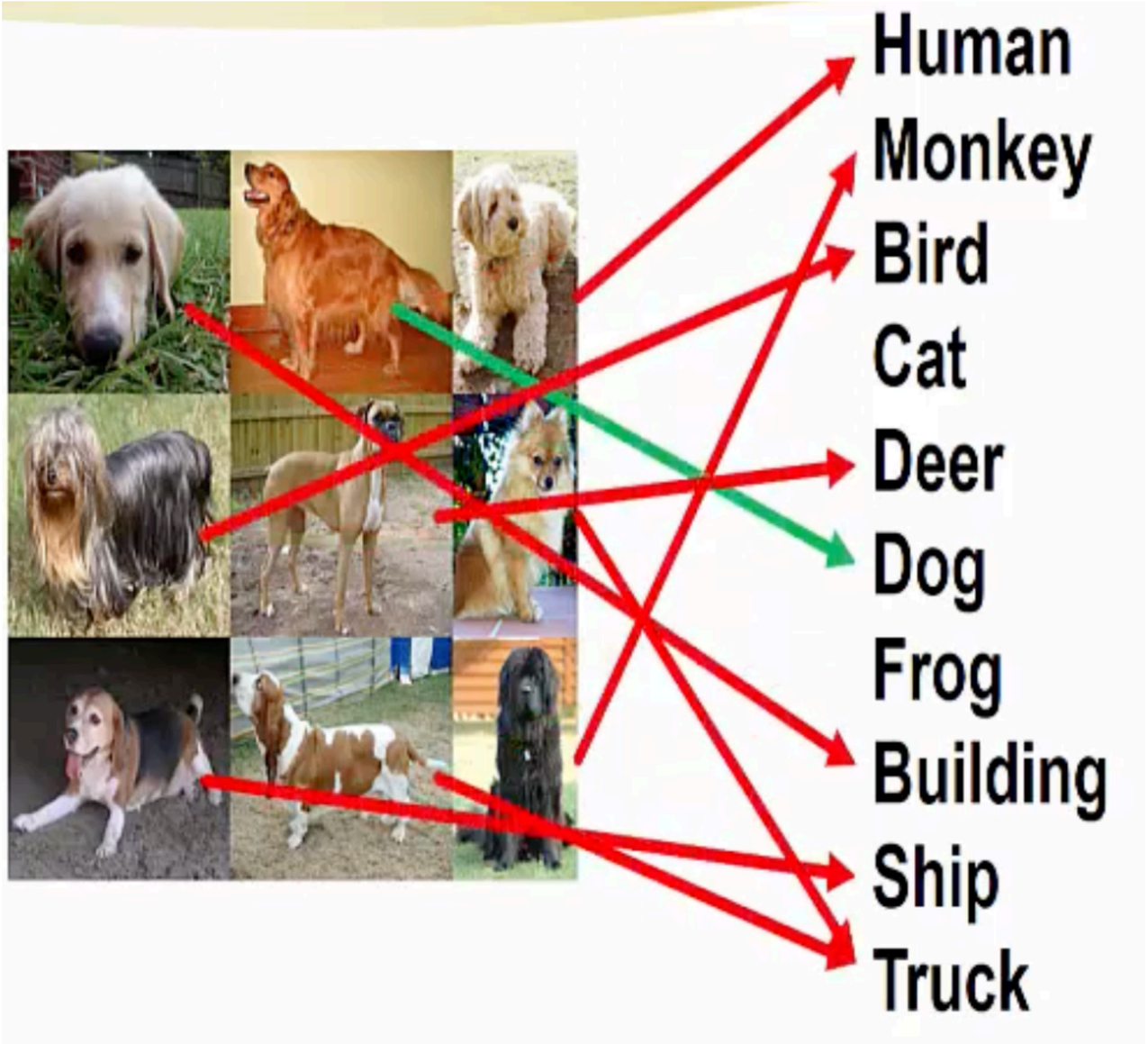
Randomization Tests

Partially Corrupted Labels

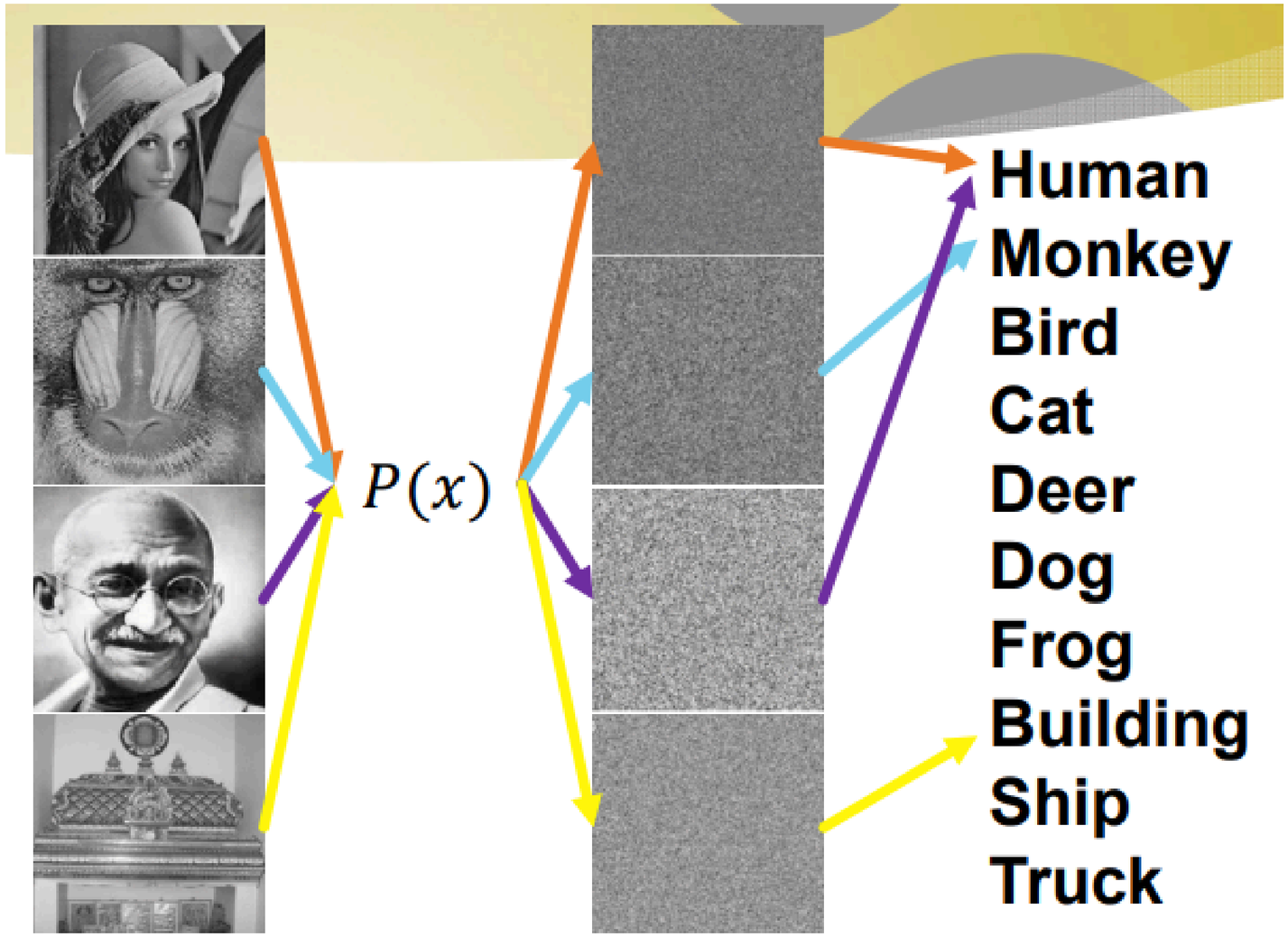


Randomization Tests

Random Labels

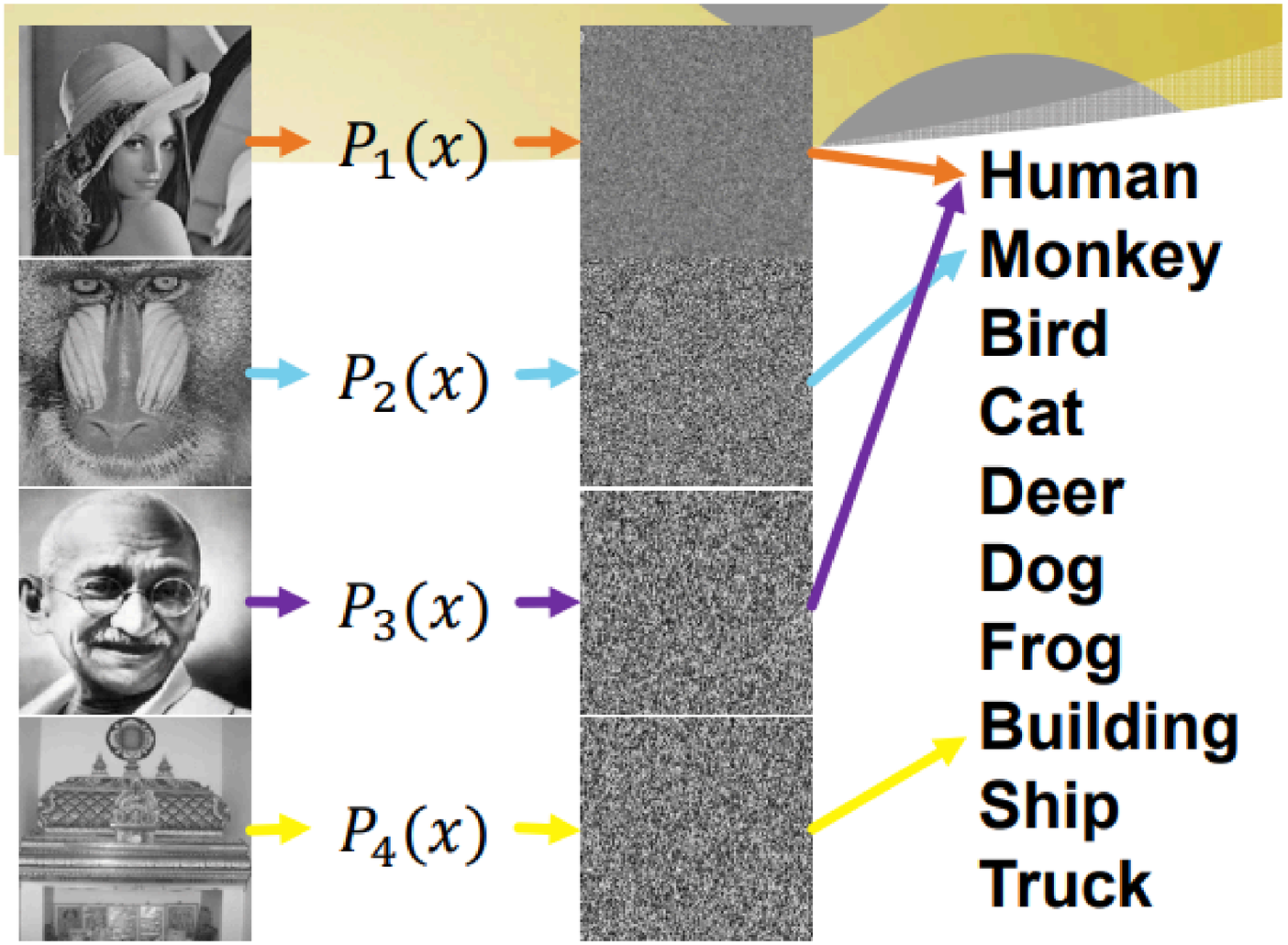


Randomization Tests



Shuffled Pixels

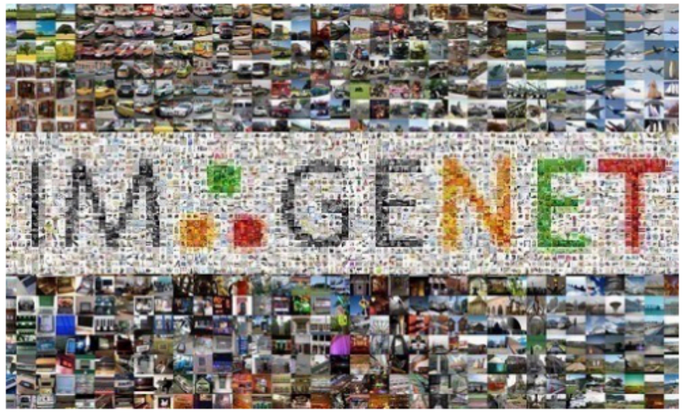
Randomization Tests



Random Pixels

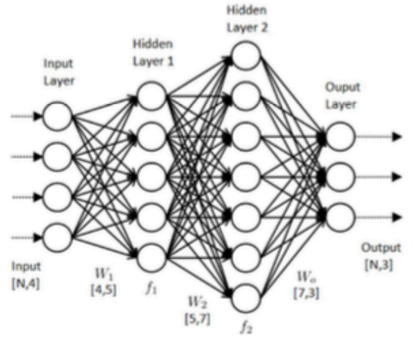
Datasets and Models

CIFAR 10
n = 60,000 images
32 x 32 with 10 classes

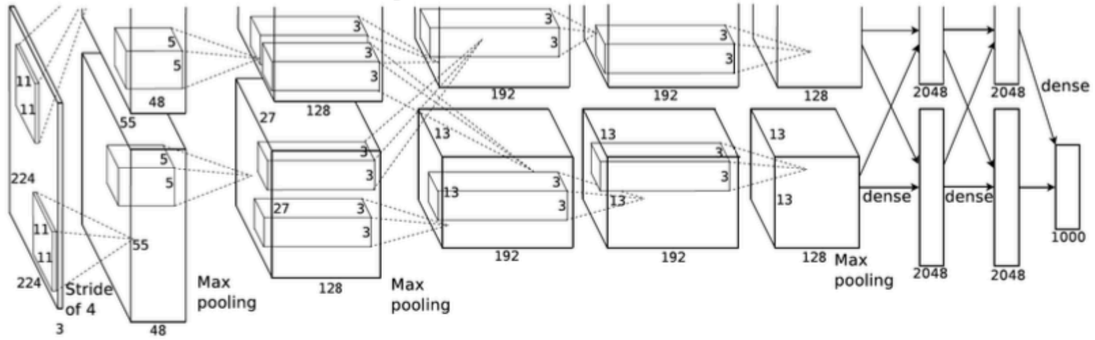


IMAGENET
n = 1,281,167 images
299 x 299 with 1000 classes

Multilayer perceptron
1,735,178 parameters ≈ 28n

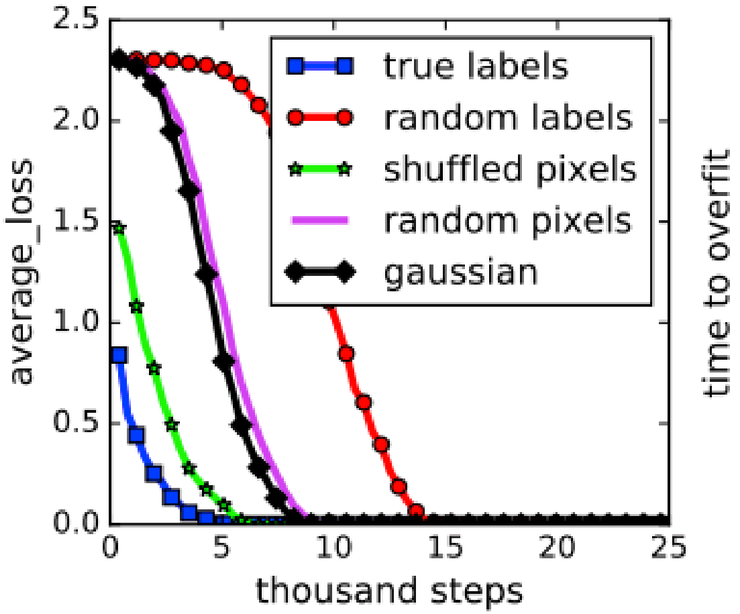


AlexNet - 1,387,786 parameters ≈ 23n



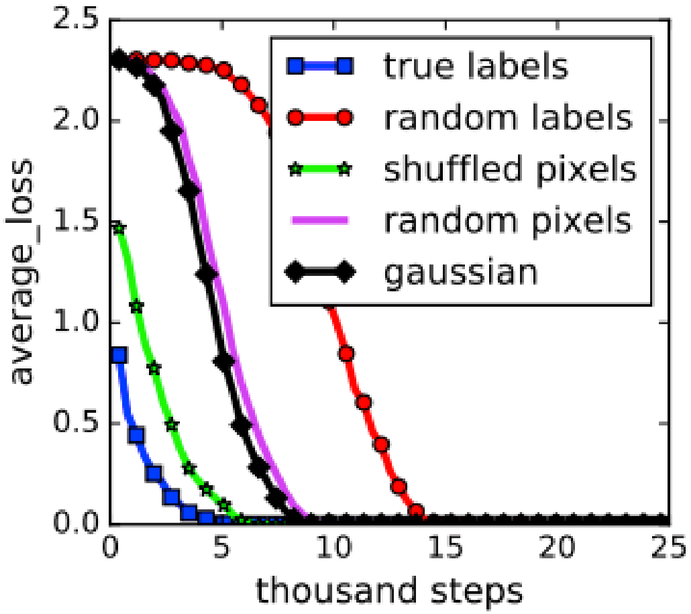
Inception - 1,649,402 parameters ≈ 27n

Results of Randomization Tests

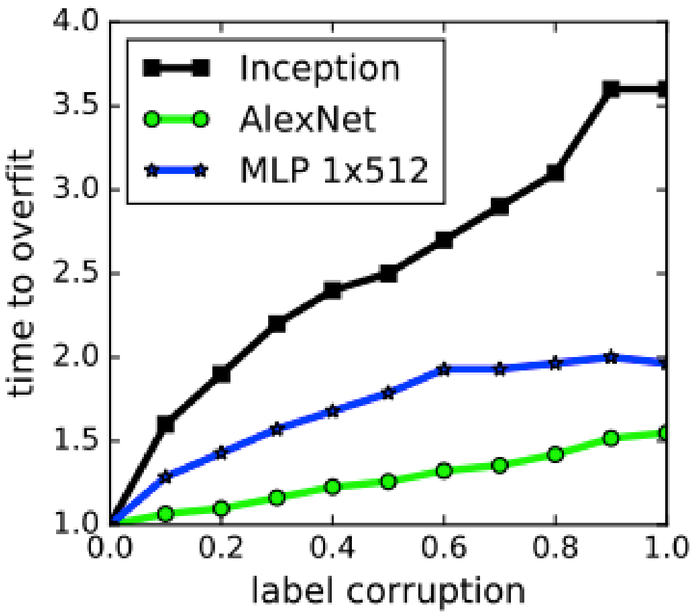


(a) learning curves

Results of Randomization Tests

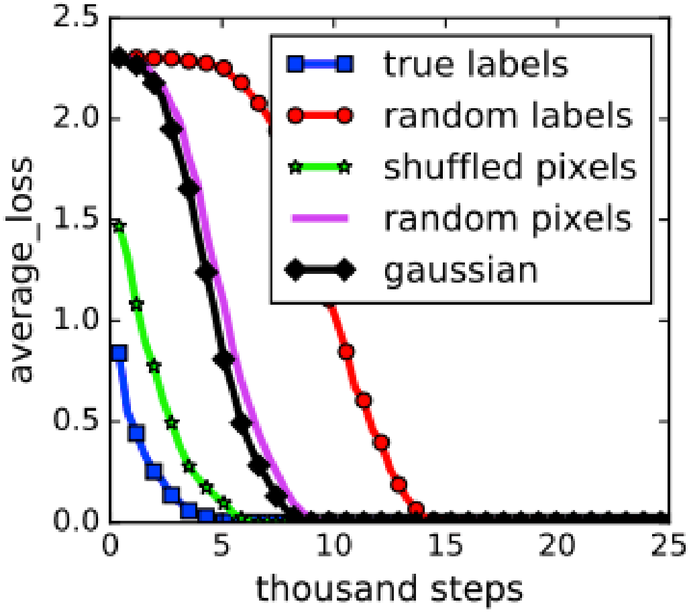


(a) learning curves

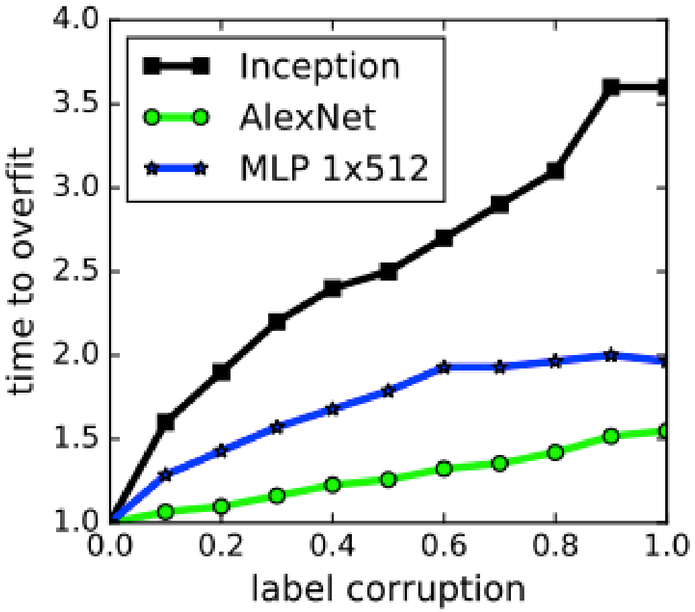


(b) convergence slowdown

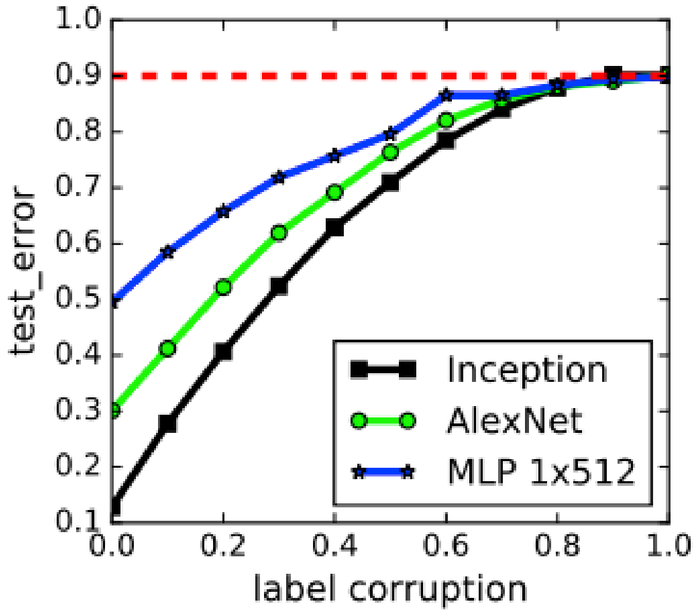
Results of Randomization Tests



(a) learning curves



(b) convergence slowdown



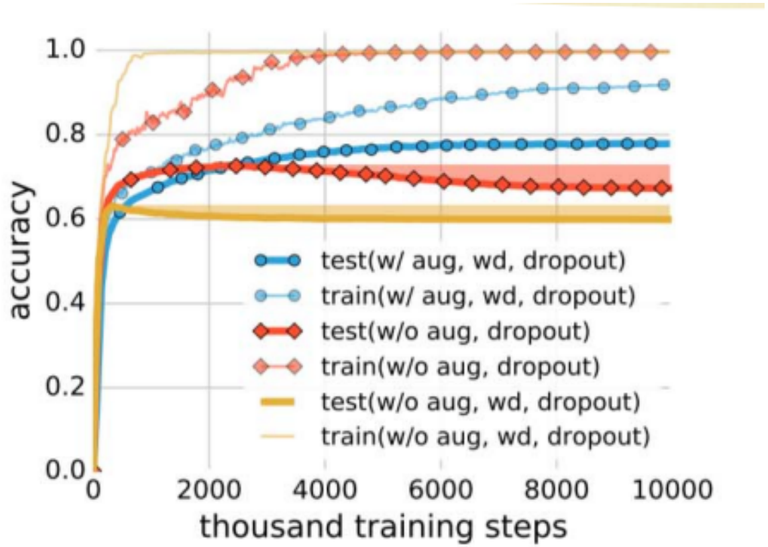
(c) generalization error growth

Implications

- Deep neural networks easily fit random labels.

- **Implications:**
  - The adequate capacity of neural networks is sufficient for memorizing the entire data set.
  - Even optimization on random labels remains easy.

Explicit Regularization Tests



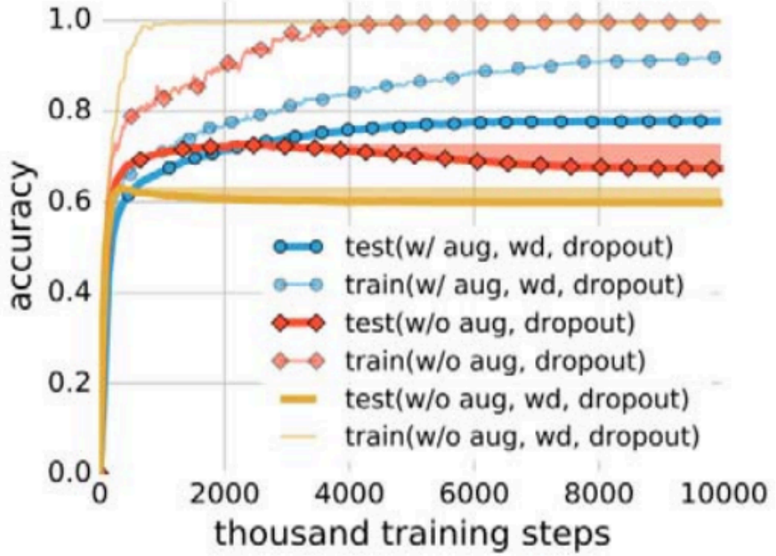
(a) Inception on ImageNet

Implications

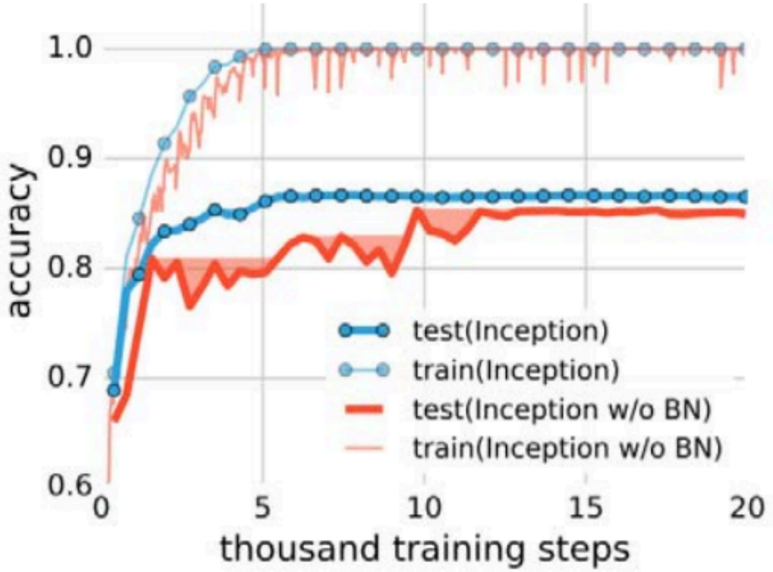
- Augmenting data is more powerful than only weight decay.
- Bigger gains by changing the model architecture.

- **Implications:**
  - Explicit regularization may improve generalization but is neither necessary nor by itself sufficient

Implicit Regularization Findings



(a) Inception on ImageNet



(b) Inception on CIFAR10

- **Findings**
  - Early stopping could potentially improve generalization.
  - Batch normalization improves generalization

Both explicit and implicit regularizers could help to improve the generalization performance. However, it is unlikely that the regularizers are the fundamental reason for generalization…

Finite-Sample Expressivity of Neural Networks

- **At the ”population level,” depth k networks are typically more powerful than depth k – 1 networks.**

- **Given a finite sample size n, even a two-layer neural network can represent any function once the number of parameters p exceeds n.**

Theorem 1:

- “**There exists a two-layer neural network with ReLU activations and 2n+d weights that can represent any function on a sample of size n in d dimensions.”**

- Finite-Sample Expressivity of Neural Networks A network C can represent any function of a sample size n in d dimensions if:
  - For every sample $S \subseteq \mathbb{R}^d$ with $|S| = n$ and
  - Every function $f : S \to \mathbb{R}$,

  There exists a setting of weights of C such that $C(x) = f(x)$ for every $x \in S$.

  Can be extended to depth k networks with width $O\left(\frac{n}{k}\right)$.

Appeal to Linear Models

- Imagine n data points {(xi, yi)}, where xi are d-dimensional feature vectors and yi are labels.

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \text{loss}(w^T x_i, y_i).$$

- If we can fit any labeling.

- Let X denote the n × d matrix whose i-th row is $\mathbf{x}_i^T$
- If X has rank n, then Xw = y has an infinite number of solutions.
- **We find a global minimum of the above equation by solving this linear system**

Investigating SGD

$w_{t+1} = w_t - \eta t e_t x_i$ and $w_0 = 0$, then $w = \sum_{i=1}^{n} a_i x_i$ for some coefficients a.

Therefore $w = x^T a$ lies in the span of the data points.

Replacing this and perfectly interpolating the labels, gives $Xw = y \Rightarrow XX^T a = y$, which has a unique solution.

- Forming the kernel matrix (Gram matrix) K = XX^(T) and solving Ka = y for a yields a perfect fit on the labels.
- Turns out, this kernel solution is exactly the minimum l2-norm solution to Xw = y.
- Hence **SGD converges to the solution with minimum norm.**
- **The minimum norm is not predictive of generalization performance.**

Conclusions

- Neural network is able to represent different function
- Successful neural networks are large enough to shatter the training data
- Optimization continues to be easy even when generalization is poor
- SGD may be performing implicit regularization by converging to solutions with minimum l2-norm.
- Traditional measures of model complexity struggle to explain the generalization of large neural networks

Limitations

- **Scope of Data and Models:**
  - **The experiments primarily focus on image classification tasks using CIFAR-10 and ImageNet datasets.** The findings may not directly generalize to other data types or tasks, such as natural language processing or regression problems.

- **Lack of Diverse Architectures:**
  - **Several commonly used architectures are tested. However, the study does not encompass a broader range of neural network models, such as recurrent neural networks or transformers, which may exhibit different generalization behaviors.**

- **Empirical Nature of Findings:**
  - The conclusions are mainly empirical. More rigorous theoretical backing is needed to solidify the observations regarding memorization and generalization.

- **Implicit Regularization:**
  - The role of implicit regularization through optimization techniques like stochastic gradient descent is discussed. Still, the mechanisms are not thoroughly dissected or understood, pointing to a gap in the theoretical understanding of these phenomena.

Future Scope

- **Theoretical Developments:**
  - Develop a more robust theoretical framework that can integrate the empirical findings of this paper with classical learning theories.
  - Formalize the role of implicit regularization in generalization, especially in non-convex settings.

- **Broader Application Domains:**
  - Extend the investigation to other types of neural networks and machine learning tasks beyond image classification to see if the findings hold universally.
  - Explore the generalization behavior in unsupervised, semi-supervised, and reinforcement learning scenarios.

- **Diverse Data and Noise Models:**
  - Test the effects of different types of data corruption and noise beyond random labels to assess model robustness in more realistic settings.
  - Investigate how structured noise or adversarial examples impact the training dynamics and generalization.

Review from another people

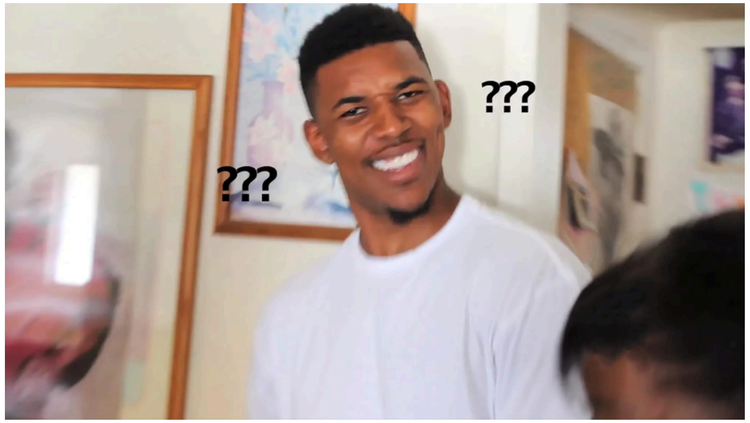
I expected to like this paper, because I respect the authors, and many people have said good things about it.... I'm sorry to say I was very disappointed.

[T]he results in this paper are completely unsurprising. I'm surprised that the authors were surprised. I'm shocked that at least one reviewer thought this was ground breaking.
- Thomas G Dietterich

About

Thomas G. Dietterich is emeritus professor of computer science at Oregon State University. He is one of the pioneers of the field of machine learning. He served as executive editor of Machine Learning and helped co-found the Journal of Machine Learning Research. Wikipedia

Place of birth: Weymouth, MA



The authors report the experimental findings of a fascinating inquiry on the ability of the deep neural networks to fit randomly labelled data. The investigation is sound, enlightening, and inspiring...

This is definitely groundbreaking work, which will inspire many works in the coming years.
- ICLR meta review (scores: 10, 9, 10)
Got ICLR best paper 2017

Related Work on Neural Network Stability and Generalization

- Uniform Stability Analysis **Hardt et al. (2016)** provided an upper bound on generalization error via uniform stability.
  - Uniform stability was found to be independent of training data labeling, questioning its effectiveness in distinguishing models trained on true vs. random labels.
  - Representational Power of Neural Networks

- Universal approximation theorems (**Cybenko, 1989; Mhaskar, 1993; and others**) show neural networks' ability to express mathematical functions over an entire domain. For finite samples, even small two-layer perceptrons exhibit universal expressivity.

- Generalization Bounds and Capacity Control Bartlett (1998) established bounds on the fat-shattering dimension for sigmoid activation networks.

- **Neyshabur et al. (2014)** suggested that network size might not be the primary capacity control, highlighting the role of implicit regularization.

Thank you