Explainability
AND
Common Robustness

Lecturer: Dr. Xingjun Ma

School of Computer Science, Fudan University

Autumn, 2022

Reminder

- **Course page:**

  https://trustworthymachinelearning.github.io/
- **Textbook:**

  下载链接: https://pan.baidu.com/s/1kybxud_tz0xshWpMEORAhg?pwd=tauu

  提取码: tauu

Recap: week 1

1. What is Machine Learning

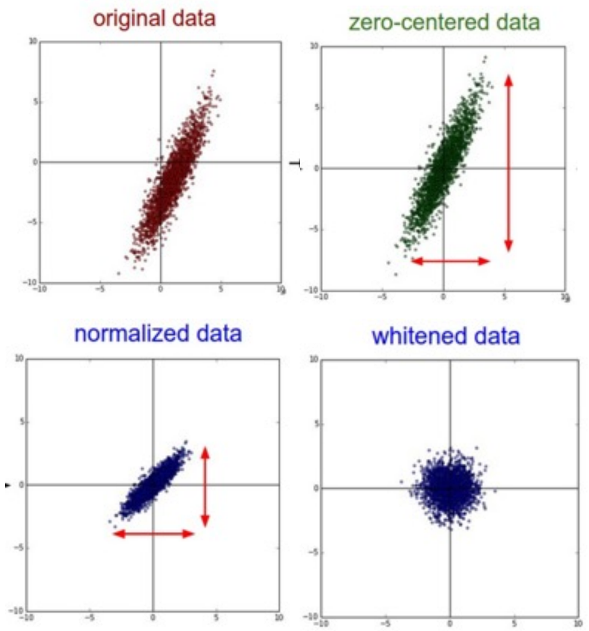
2. Machine Learning Paradigms

3. Loss Functions

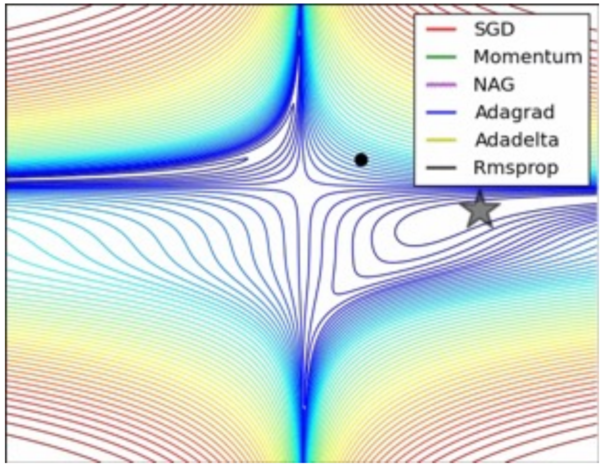
4. Optimization Methods

Machine Learning Pipeline

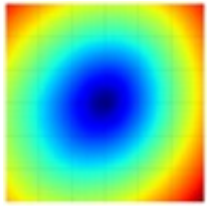
setup the input



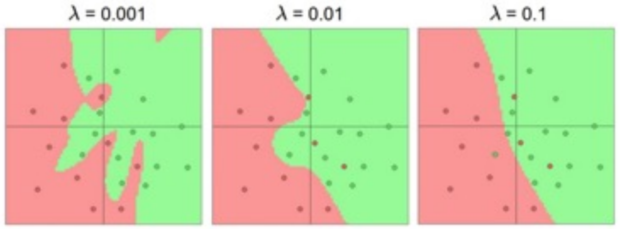
setup the optimiser



setup the loss



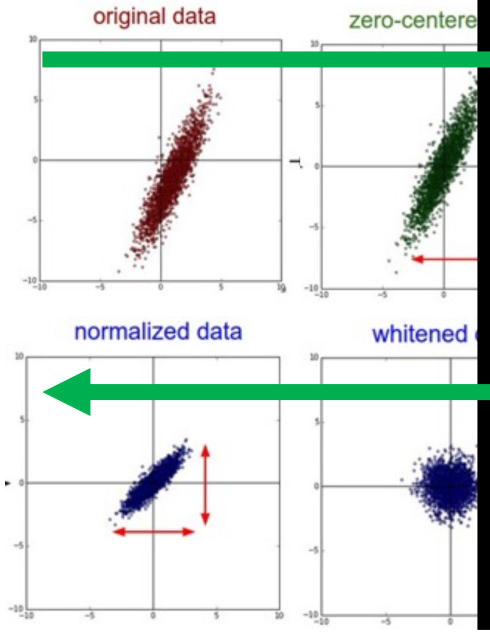
landscape of a loss function, it varies w.r.t. data, the function itself



regularization makes decision region smoother

Machine Learning Pipeline

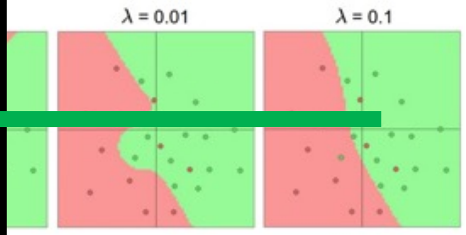
setup the input



? ? ?

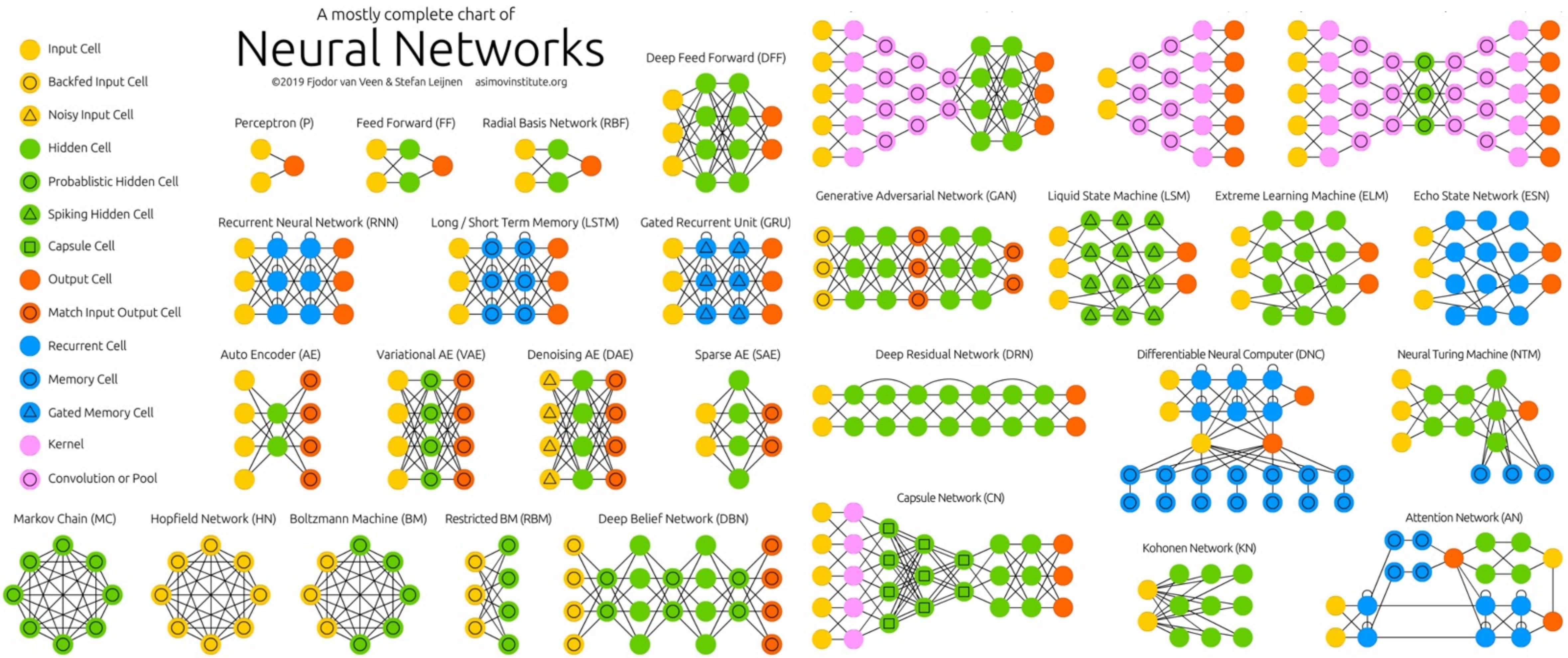
tup the loss

landscape of a loss function, it varies w.r.t. data, the function itself



ization makes decision smoother

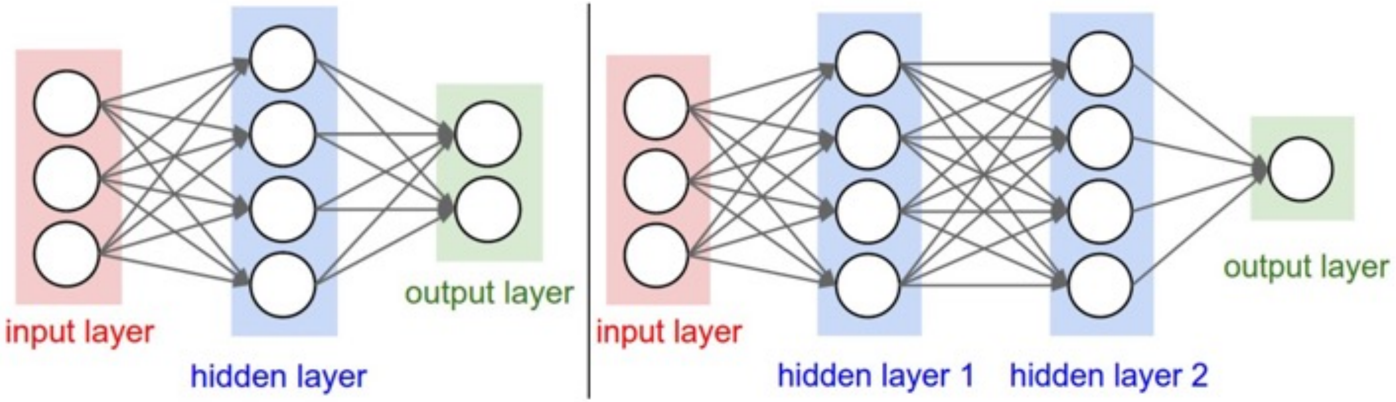
Deep Neural Networks

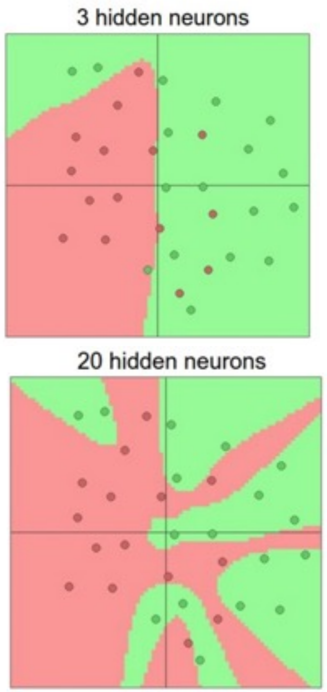


https://www.asimovinstitute.org/neural-network-zoo/; https://developer.ibm.com/articles/cc-machine-learning-deep-learning-architectures/

Feed-Forward Neural Networks

Feed-Forward Neural Networks (FNN)
Fully Connected Neural Networks (FCN)
Multilayer Perceptron (MLP)

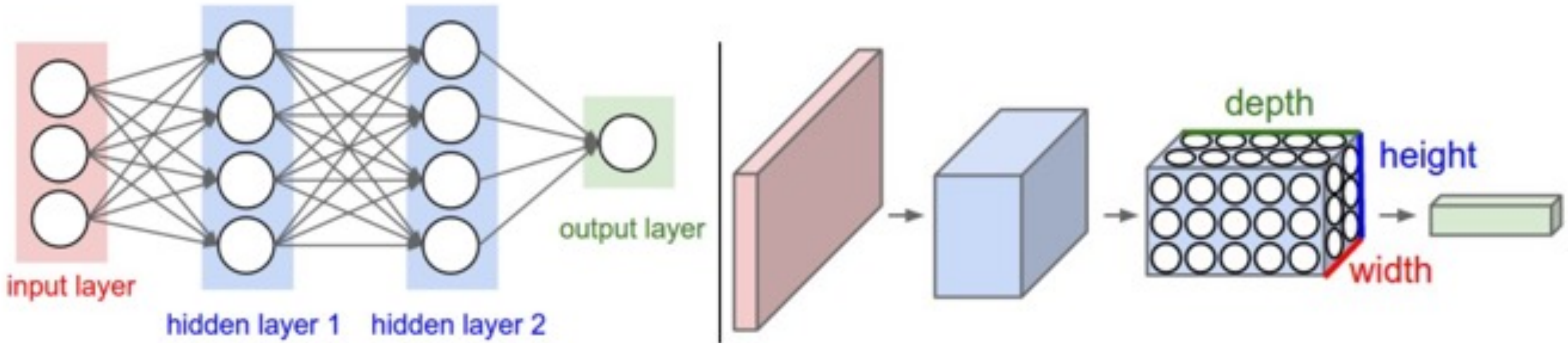




- The **simplest** neural network
- **Fully-connected** between layers
- For data that has NO temporal or spatial order

http://cs231n.stanford.edu/

Convolutional Neural Networks



Neurons in one flat layer

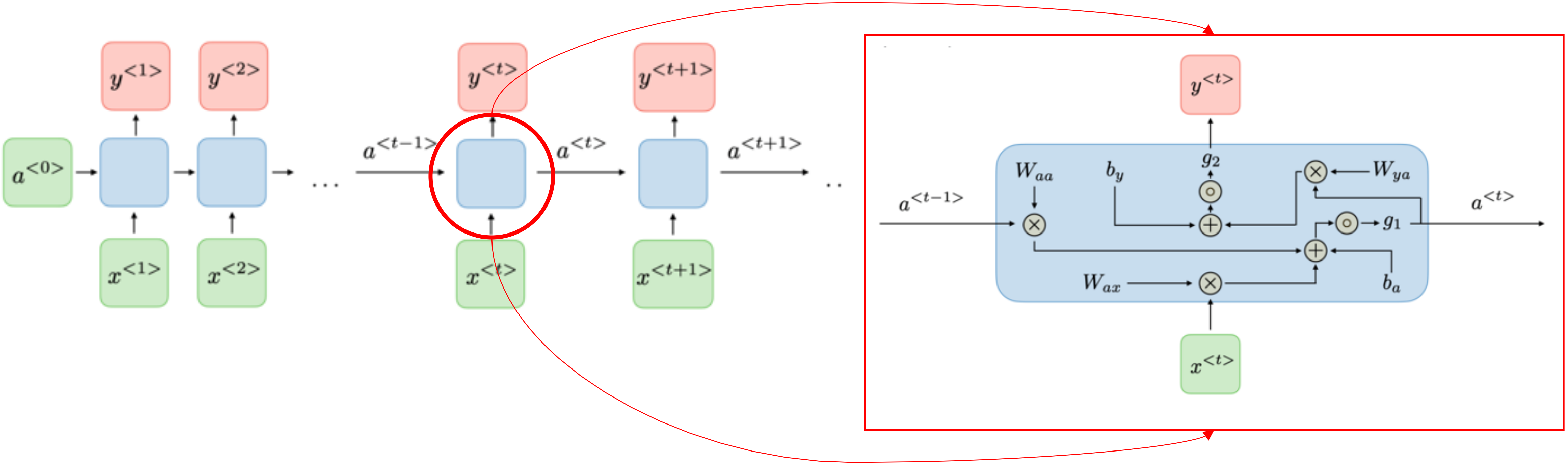
Neurons in 3 dimensions

- For images or data with spatial order
- Can stack up to >100 layers

http://cs231n.stanford.edu/

Recurrent Neural Networks

Traditional RNN

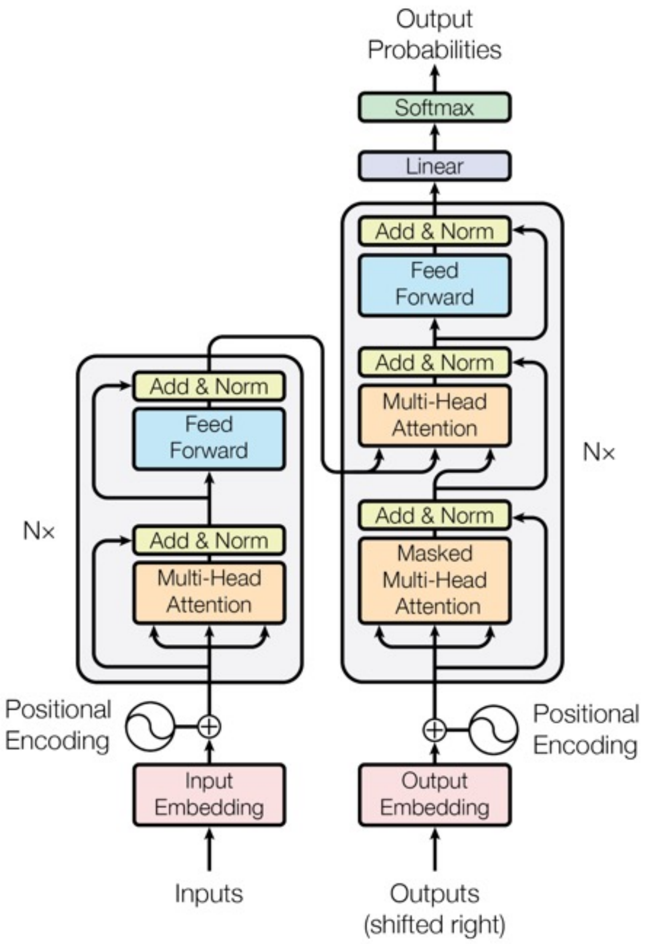


$$y^{<t>} = g_2(W_{ya}a^{<t>} + b_y)$$

$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a)$$

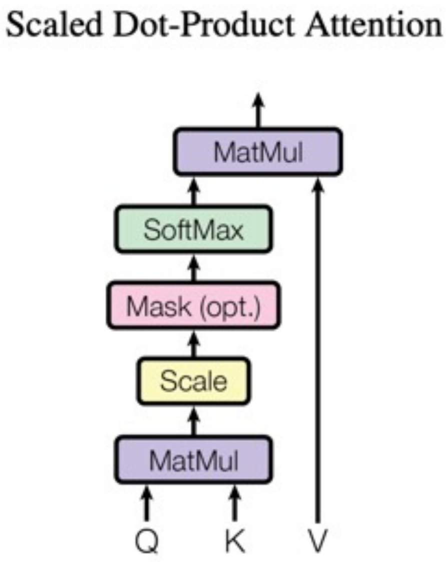
https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks

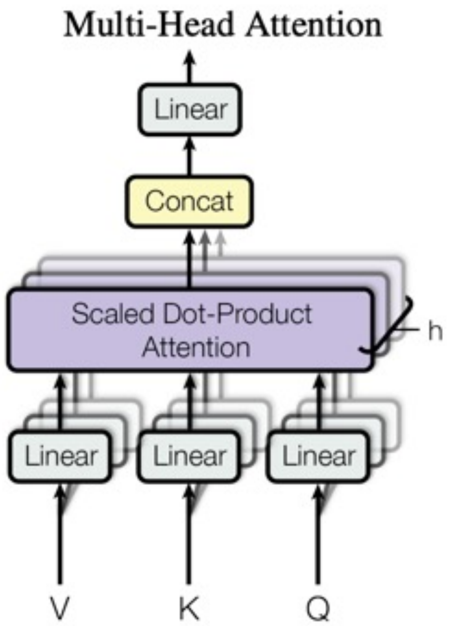
Recurrent Neural Networks



Encoder Decoder

Transformer: a new type of DNNs based on attention



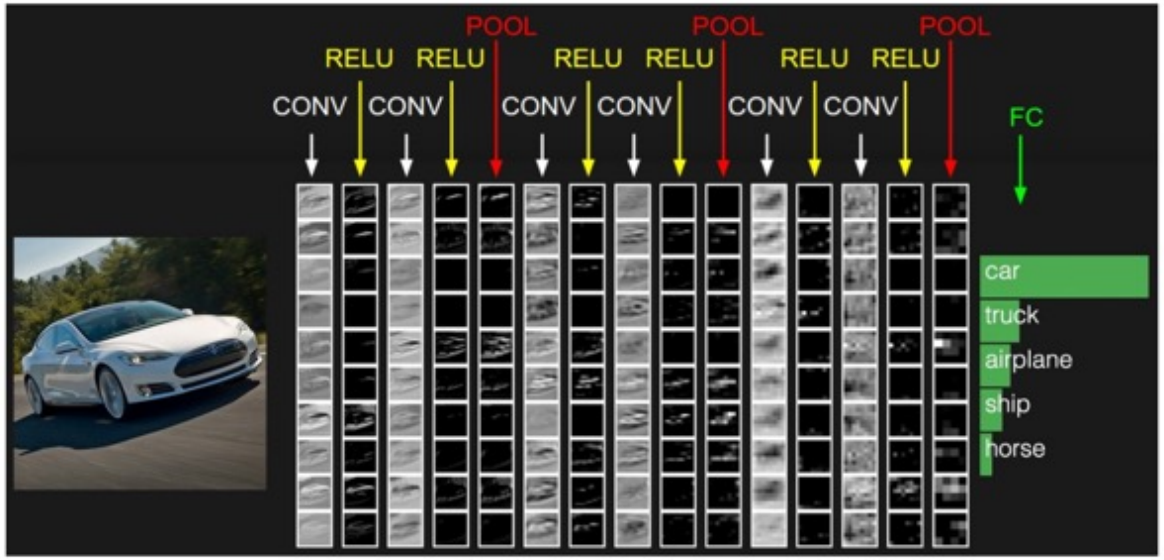


Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017)

Self-Attention Explained

https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a

CNN Explained



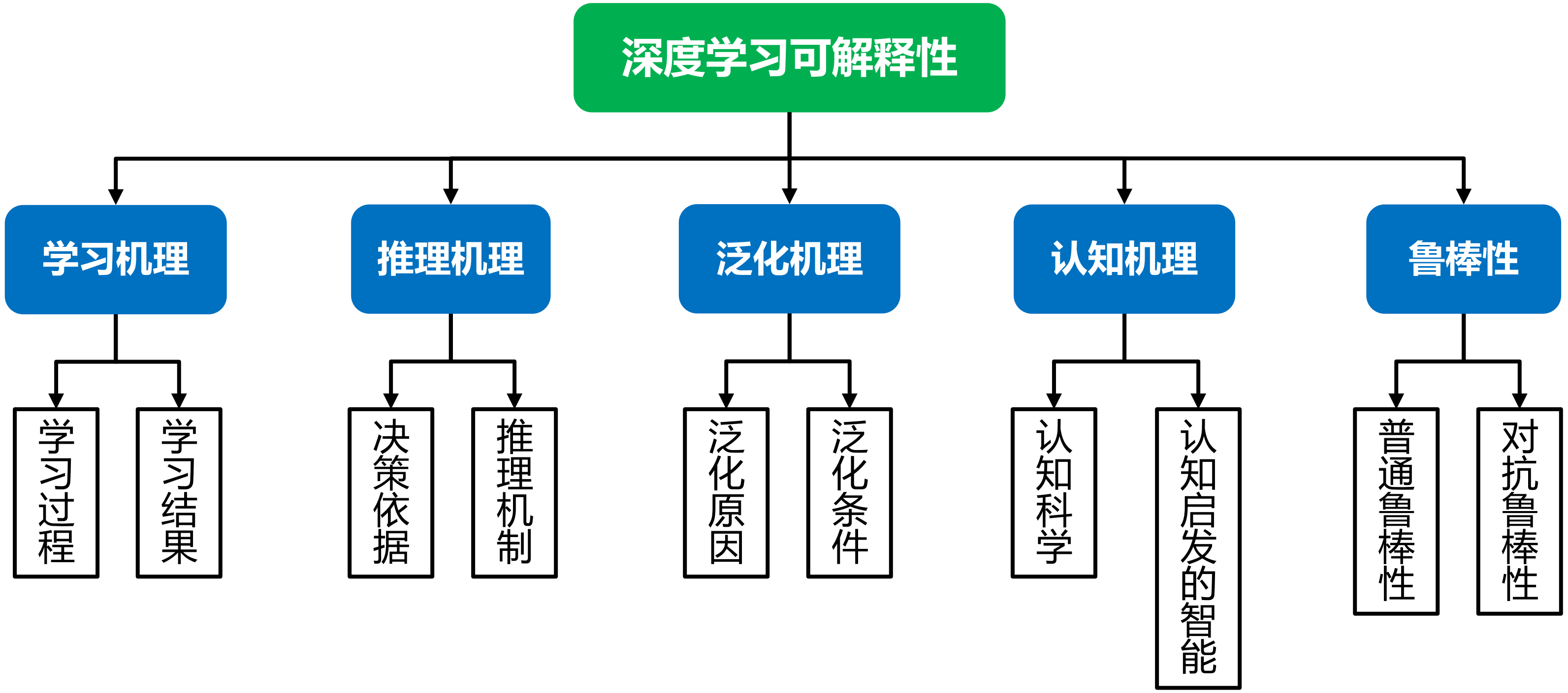
A brief history of CNNs:
- **LeNet, 1990s**
- **AlexNet, 2012**
- **ZF Net, 2013**
- **GoogLeNet, 2014**
- **VGGNet, 2014**
- **ResNet, 2015**
- **Inception V4, 2016**
- **ResNeXt, 2017**
- **ViT, 2021**

- Learns different levels of representations

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR 2021

http://cs231n.stanford.edu/

Explainable Machine Learning



我们想要弄清楚下列问题：

- **DNN是怎么学习的、学到了什么、靠什么泛化、在什么情况下行又在什么情况下不行？**
- **深度学习是否是真正的智能，与人类智能比谁更高级，它的未来是什么？**
- **是否存在大一统的理论，不但能解释而且能提高？**

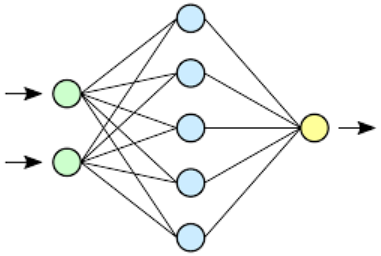
Methodological Principles

◆Visualization

◆Ablation

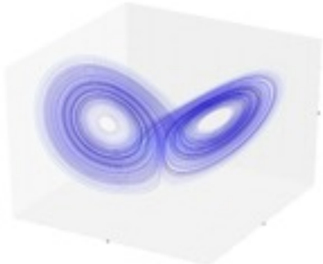
◆Contrast

◆Reverse



- Model
- Component
- Layer
- Operation
- Neuron

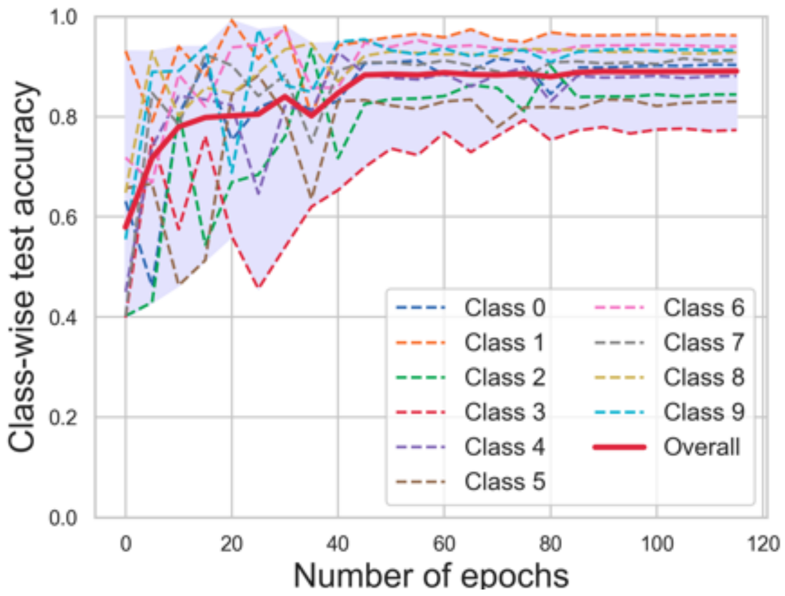
- Superclass
- Class
- Training/Test set
- Subset
- Sample



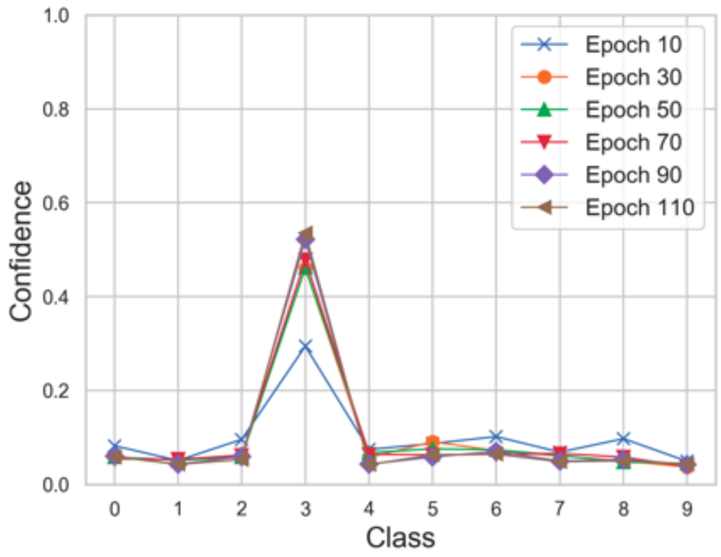
- Training
- Inference
- Transfer

Learning Mechanism

- Training/Test Error/Accuracy



- Prediction Confidence

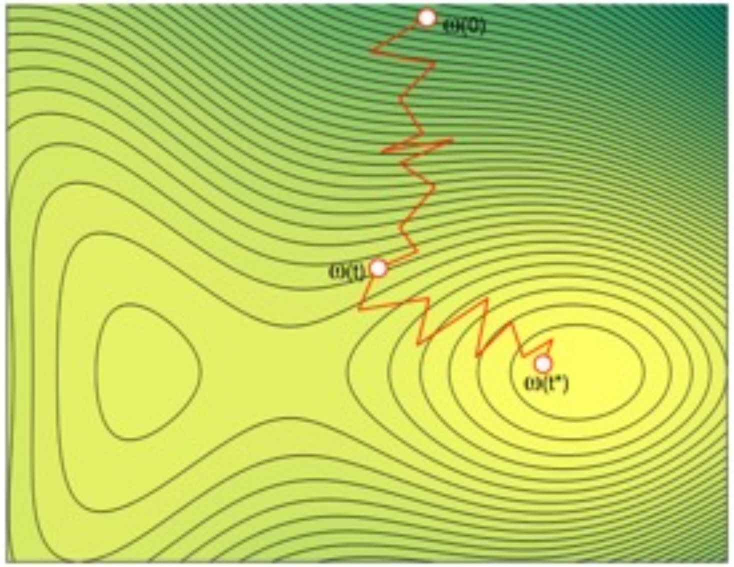


Explanation via observation: just plot!

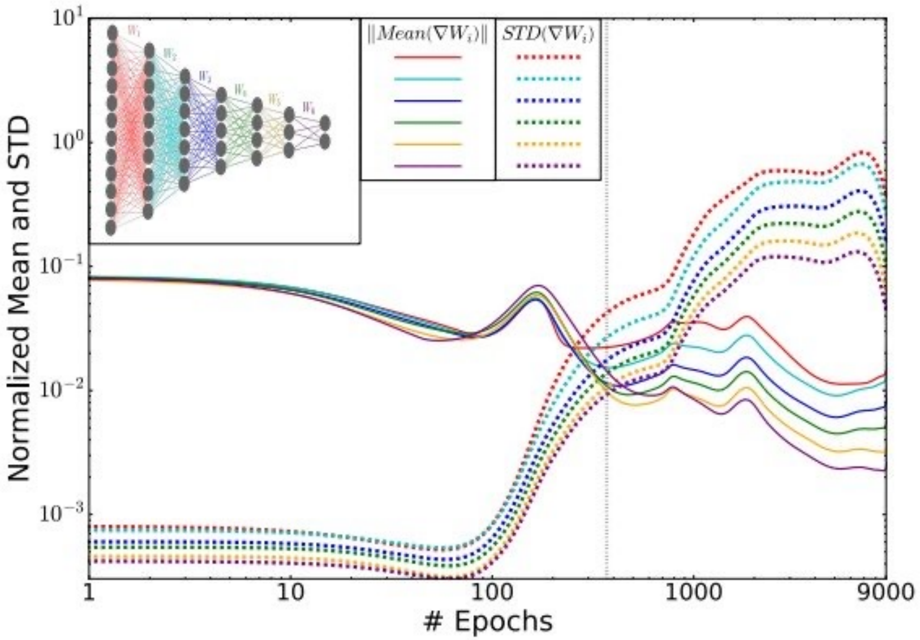
Wang et al. Symmetric Cross Entropy for Robust Learning with Noisy Labels, ICCV 2019.

Learning Mechanism

◻ Parameter dynamics



◻ Gradient dynamics

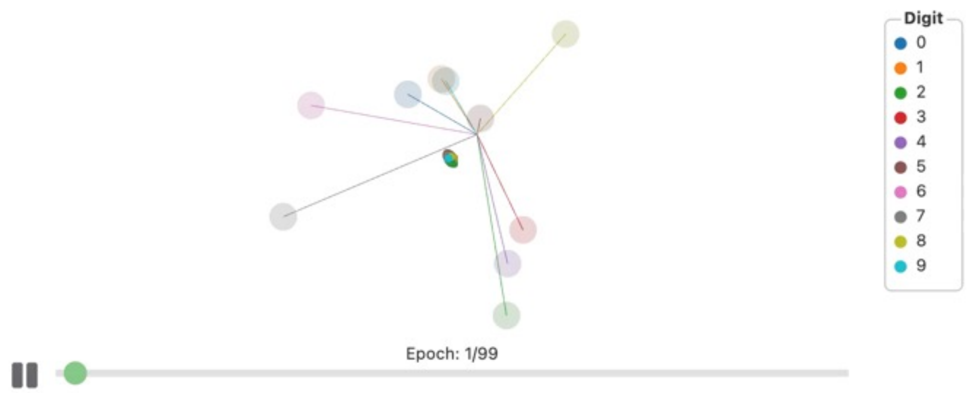


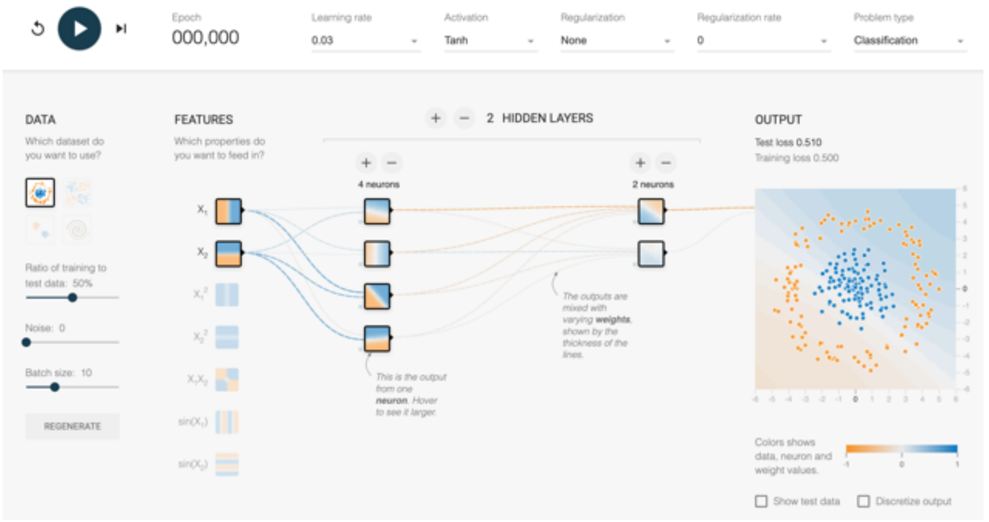
Explanation via dynamics and information

TRADI: Tracking deep neural network weight distributions, ECCV 2020; Shwartz-Ziv R, Tishby N. Opening the black box of deep neural networks via information[J]. arXiv:1703.00810, 2017.

Learning Mechanism

- **Decision boundary, learning process visualization**



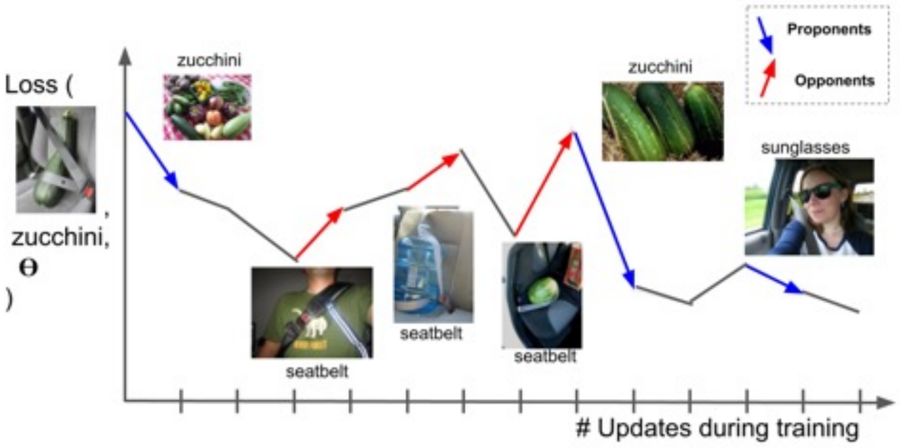


Explanation via dynamics and information

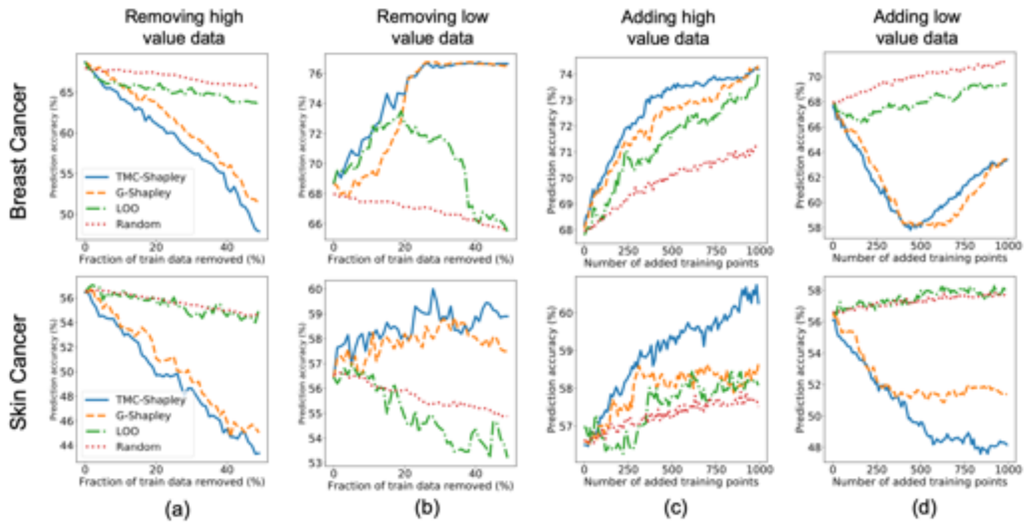
https://distill.pub/2020/grand-tour/ （March 16, 2020）; https://playground.tensorflow.org/

Learning Mechanism

◻ Data influence/valuation: how a training sample impacts the learning outcome?



Influence Function

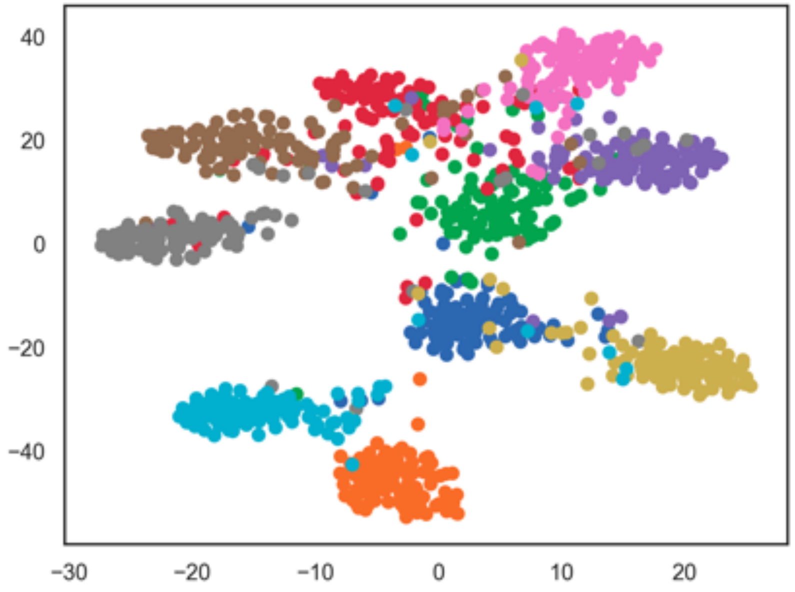


Data Shapley

Understanding Black-box Predictions via Influence Functions, ICML, 2018;
Pruthi G, Liu F, Kale S, et al. Estimating training data influence by tracing gradient descent. NeurIPS, 2020.
Data shapley: Equitable valuation of data for machine learning, ICML, 2019.

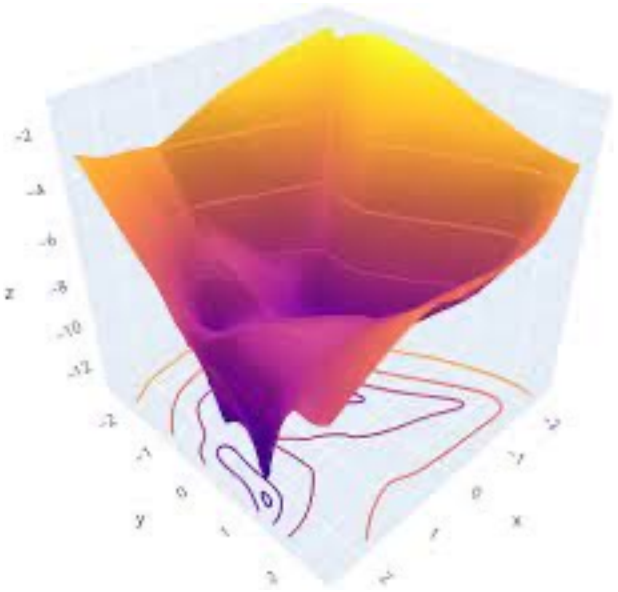
Understanding the Learned Model

- **Deep features**



t-SNE plot

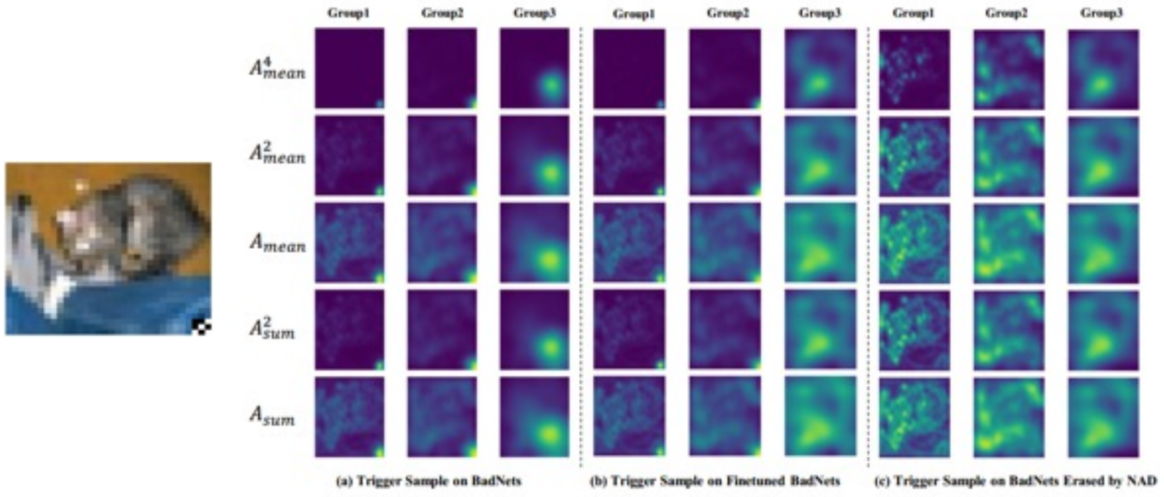
- **Loss Landscape**



Maaten et al. Visualizing data using t-SNE. JMLR, 2008.
https://distill.pub/2016/misread-tsne/?_ga=2.135835192.888864733.1531353600-1779571267.1531353600

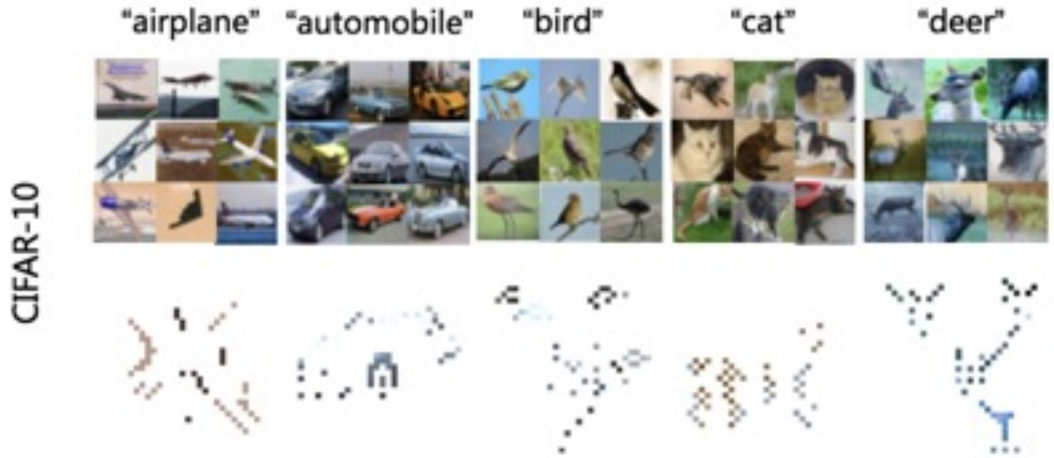
Understanding the Learned Model

☐ Intermediate Layer Activation Map



Activation/Attention Map

☐ Class-wise Patterns

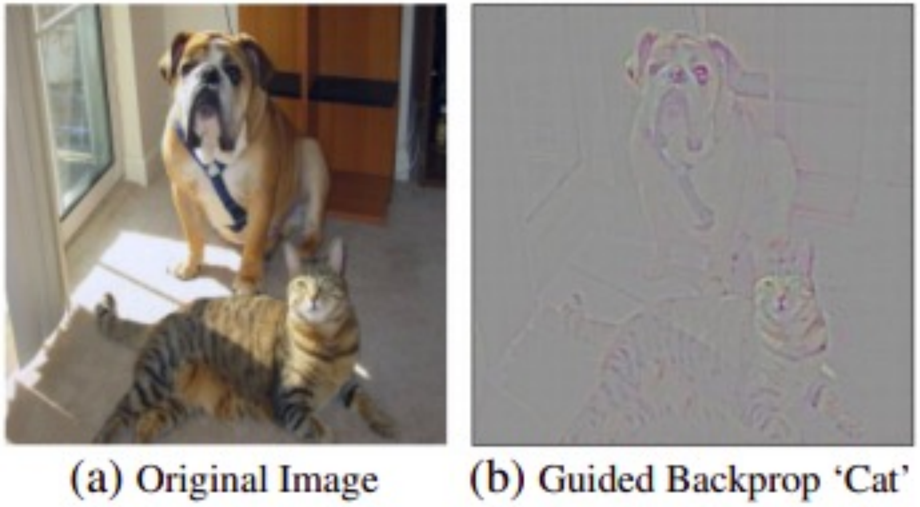


One predictive pattern for each class

Li et al. Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Network, ICLR 2021; Zhao et al. What do deep nets learn? class-wise patterns revealed in the input space. *arXiv:2101.06898* (2021).

Inference Mechanism

- **Guided Backpropagation**



- **Class Activation Map (Grad-CAM)**

Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. ICCV 2017.
Springenberg et al. Striving for Simplicity: The All Convolutional Net, ICLR 2015.

Guided Backpropagation

ReLU forward pass

$$h^{l+1} = \max\{0, h^l\}$$

Forward pass

$$h^l \begin{bmatrix} 1 & -1 & 5 \\ 2 & -5 & -7 \\ -3 & 2 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 5 \\ 2 & 0 & 0 \\ 0 & 2 & 4 \end{bmatrix} h^{l+1}$$

ReLU backward pass

$$\frac{\partial L}{\partial h^l} = [[h^l > 0]] \frac{\partial L}{\partial h^{l+1}}$$

Backward pass: backpropagation

$$\begin{bmatrix} -2 & 0 & -1 \\ 6 & 0 & 0 \\ 0 & -1 & 3 \end{bmatrix} \leftarrow \begin{bmatrix} -2 & 3 & -1 \\ 6 & -3 & 1 \\ 2 & -1 & 3 \end{bmatrix} \frac{\partial L}{\partial h^{l+1}}$$

Deconvolution for ReLU

$$\frac{\partial L}{\partial h^l} = [[h^{l+1} > 0]] \frac{\partial L}{\partial h^{l+1}}$$

Backward pass: "deconvnet"

$$\begin{bmatrix} 0 & 3 & 0 \\ 6 & 0 & 1 \\ 2 & 0 & 3 \end{bmatrix} \leftarrow \begin{bmatrix} -2 & 3 & -1 \\ 6 & -3 & 1 \\ 2 & -1 & 3 \end{bmatrix} \frac{\partial L}{\partial h^{l+1}}$$

Guided Backpropagation

$$\frac{\partial L}{\partial h^l} = [[(h^l > 0) \&\& (h^{l+1} > 0)]] \frac{\partial L}{\partial h^{l+1}}$$

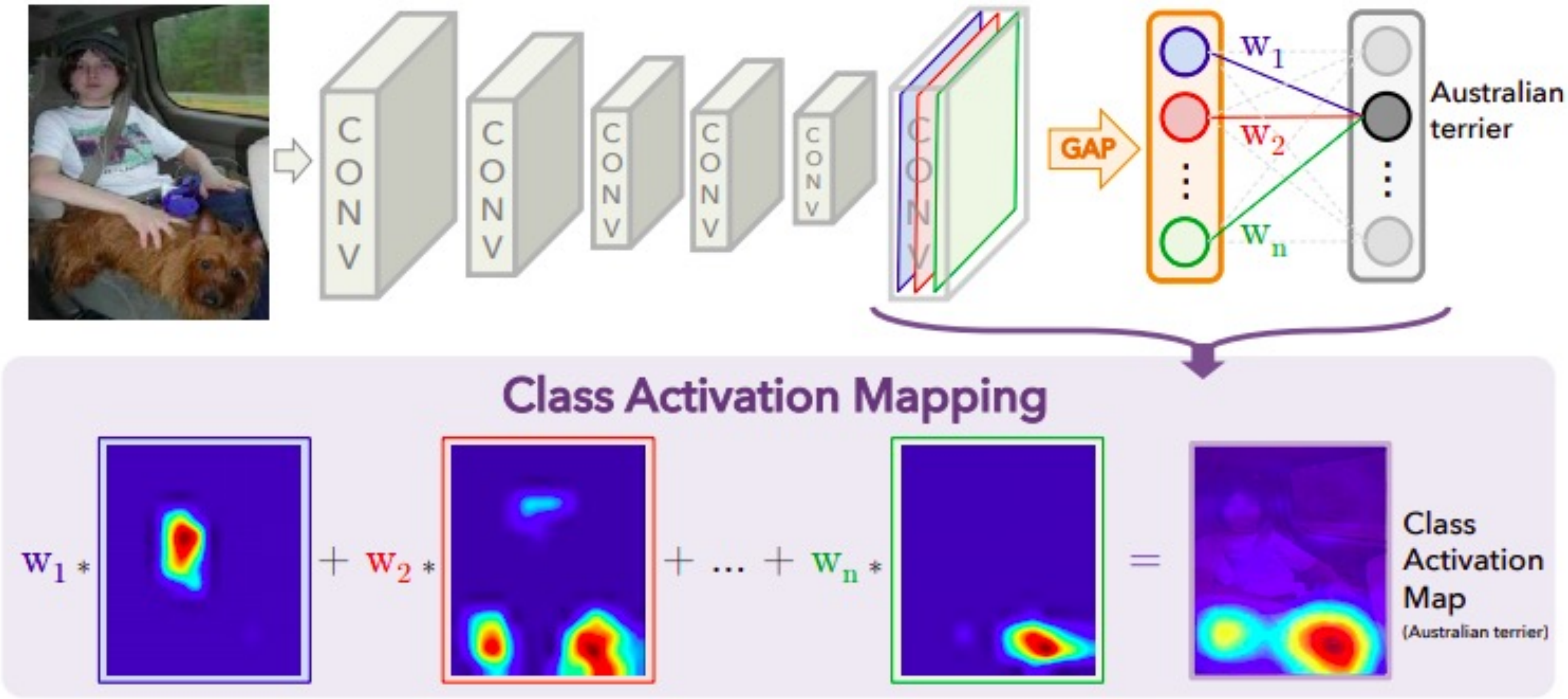
Backward pass: *guided backpropagation*

$$\begin{bmatrix} 0 & 0 & 0 \\ 6 & 0 & 0 \\ 0 & 0 & 3 \end{bmatrix} \leftarrow \begin{bmatrix} -2 & 3 & -1 \\ 6 & -3 & 1 \\ 2 & -1 & 3 \end{bmatrix} \frac{\partial L}{\partial h^{l+1}}$$

Springenberg et al. Striving for Simplicity: The All Convolutional Net, ICLR 2015.
https://medium.com/@chinesh4/generalized-way-of-interpreting-cnns-a7d1b0178709

Class Activation Mapping (CAM)

GAP: Global Average Pooling



Zhou et al. Learning Deep Features for Discriminative Localization. CVPR, 2016.
https://medium.com/@chinesh4/generalized-way-of-interpreting-cnns-a7d1b0178709

Grad-CAM

Grad-CAM is a generalization of CAM

Compute **neuron importance**:

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

y^c : logits of class c (before softmax)
A^k : k-th channel activation map

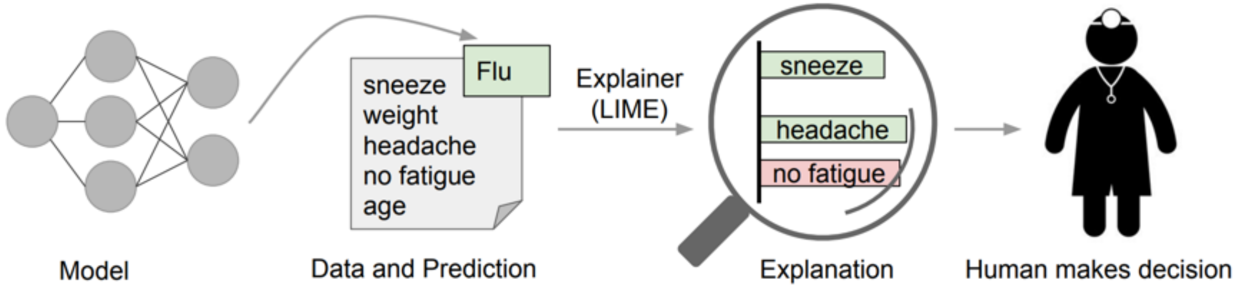
Weighted combination of activation map, then **interpolation**:

$$L_{\text{Grad-CAM}}^c = ReLU \underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

B. Zhou, A. Khosla, L. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. In CVPR, 2016; https://medium.com/@chinesh4/generalized-way-of-interpreting-cnns-a7d1b0178709

LIME

◻ Local Interpretable Model-agnostic Explanations (LIME)



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

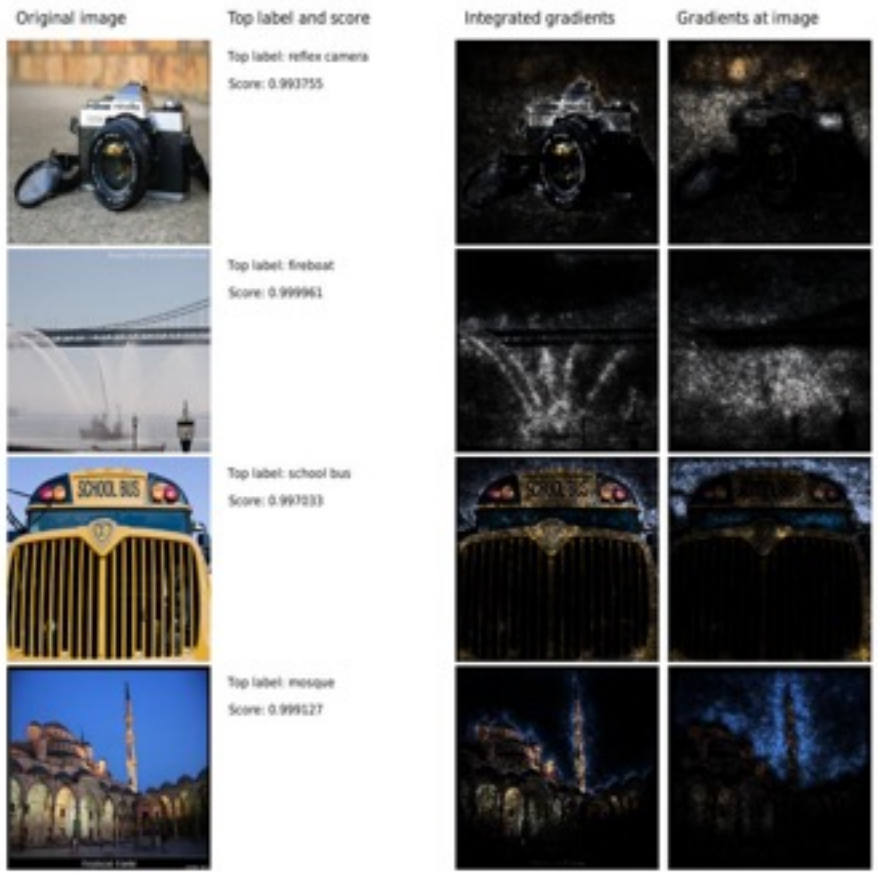
$$\xi(x) = \operatorname*{argmin}_{g \in G} \ \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

π_x : local neighborhood of x
z : sampled neighbor points
g : explainer e.g a linear model
z' : a binary vector for interpretable representation(e.g. patch)

Ribeiro et al. "Why should i trust you?" Explaining the predictions of any classifier. " SIGKDD, 2016.
https://github.com/marcotcr/lime

Integrated Gradients



$$\text{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i}\, d\alpha$$

- **There is a path**: $x_i \rightarrow x'_i$
- **Traverse the path using** α
- **Integrate the gradients along the way**

Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks, ICML, 2017.
https://github.com/TianhongDai/integrated-gradient-pytorch

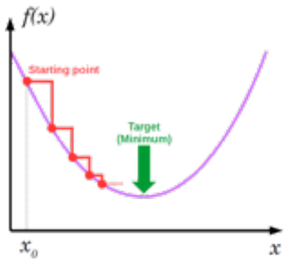
Generalization Mechanism

Deep Learning Theory

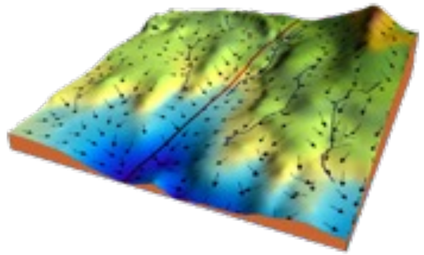
Convergence

Generalization

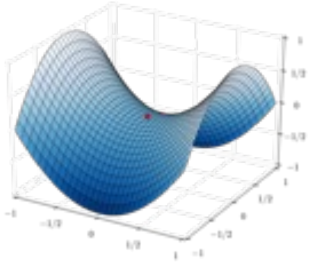
◻ Convergence



Convex (Linear model)



Nonconvex (DNN)



Saddle point

◻ Generalization

'Cat'

Training time

'Cat'?

Test time

Traditional theory: simpler model is better, more data is better

Generalization Theory

□ Components of Generalization Error Bounds

$$\underbrace{\mathrm{err}_D(h)}_{\text{generalization error}} \le \underbrace{\widehat{\mathrm{err}}_S(h)}_{\text{empirical error}} + \underbrace{R_m(\mathcal{H})}_{\text{hypothesis class complexity}} + \underbrace{\sqrt{\frac{\ln(1/\delta)}{m}}}_{\text{confidence sample size}}$$

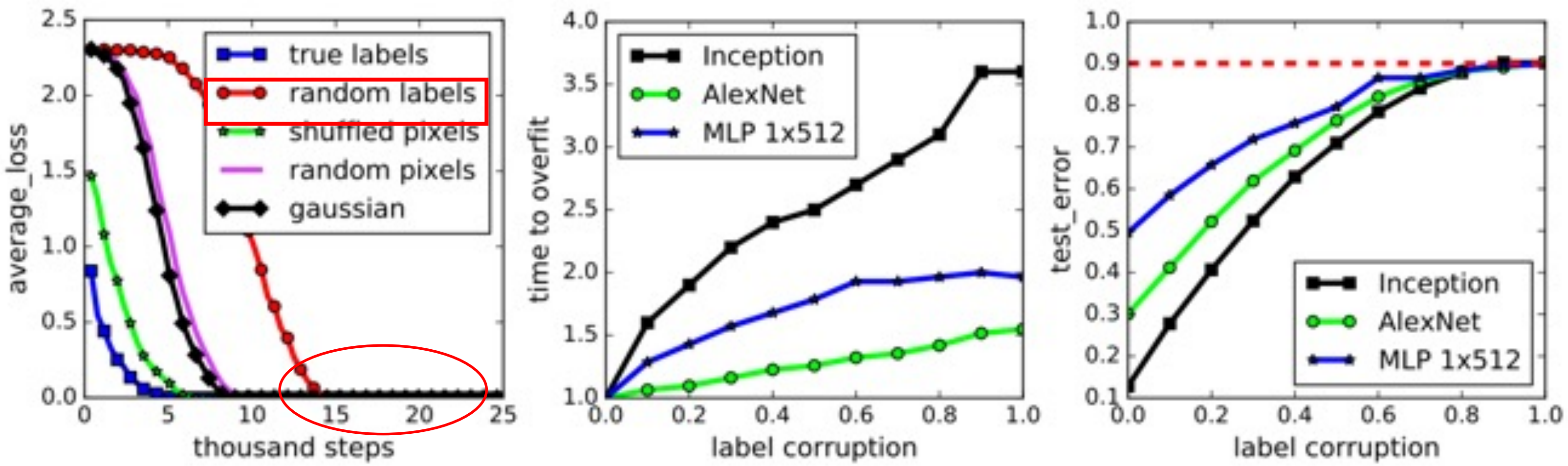
RHS: for all terms, the lower the better:

- small training error
- simpler model class
- more samples
- less confidence

https://www.cs.cmu.edu/~ninamf/ML11/lect1117.pdf; https://www.youtube.com/watch?v=zlqQ7VRba2Y

Generalization Theory

□ **Small training error ≠ low generalization error**



(a) learning curves (b) convergence slowdown (c) generalization error growth

Zero training error was achieved on **purely random labels** (meaningless learning)

- 0 training error vs. 0.9 test error

Zhang et al. Understanding deep learning requires rethinking generalization. ICLR 2017.

List of Existing Theories

- Rademacher Complexity bounds (Bartlett et al. 2017)
- PAC-Bayes bounds (Dziugaite and Roy 2017)
- Information bottleneck (Tishby and Zaslavsky 2015)
- Neural tangent kernel/Lazy training (Jacot et al. 2018)
- Mean-field analysis (Chizat and Bach 2018)
- Doule Descent (Belkin et al. 2019)
- Entropy SGD (Chaudhari et al. 2019)

A few interesting questions:

- Should we consider the role of data in generalization analysis?
- Should representation quality appear in the generalization bound?
- Generalization is about math (the function of the model) or knowledge?

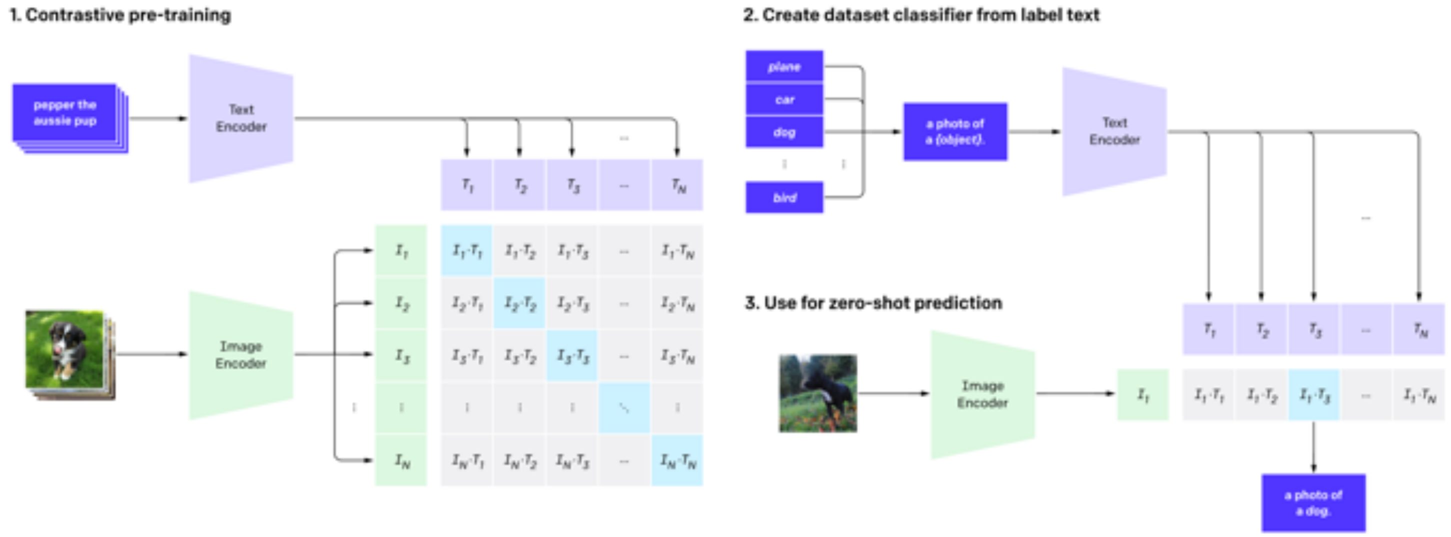
https://www.youtube.com/watch?v=zlqQ7VRba2Y

How to visualize generalization?

- Existing approaches
  - test error
  - Visualization: loss landscape, prediction attribution, etc.
  - Training -> test: distribution shift, out-of-distribution analysis
  - Noisy labels in test data – questioning data quality and reliable evaluation

- **The remaining questions:**
  - **how generalization happens?**
  - **Math ≠ Knowledge**
  - **Computation = finding patterns or understanding the underlying knowledge**
  - **What is the relation of computational generalization to human behavior?**

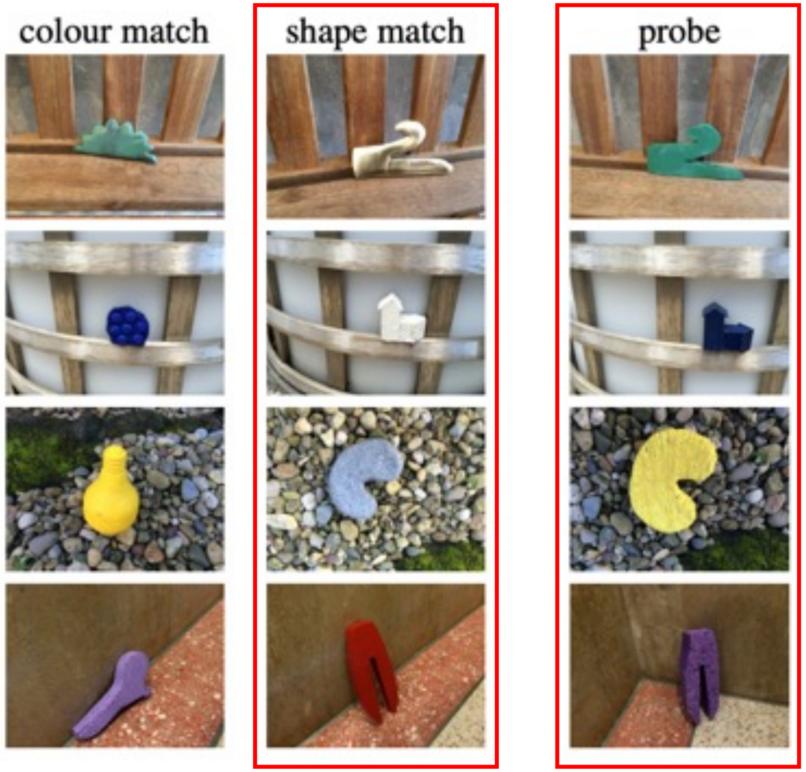
Cognitive Mechanism



OpenAI reveals the multimodal neurons in CLIP

https://openai.com/blog/multimodal-neurons/; https://openai.com/blog/clip/

Cognitive Mechanism

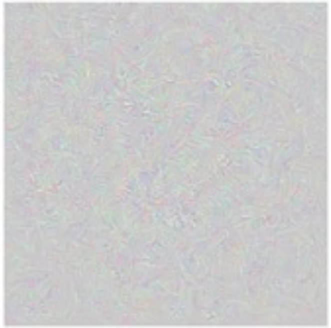


shape match = prob means
shape bias

cognitive psychology inspired evaluation of DNNs

Ritter et al. Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study, ICML, 2017

Cognitive Mechanism



Article: Super Bowl 50

Paragraph: *"Peython Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV."*

Question: *"What is the name of the quarterback who was 38 in Super Bowl XXXIII?"*

Original Prediction: John Elway

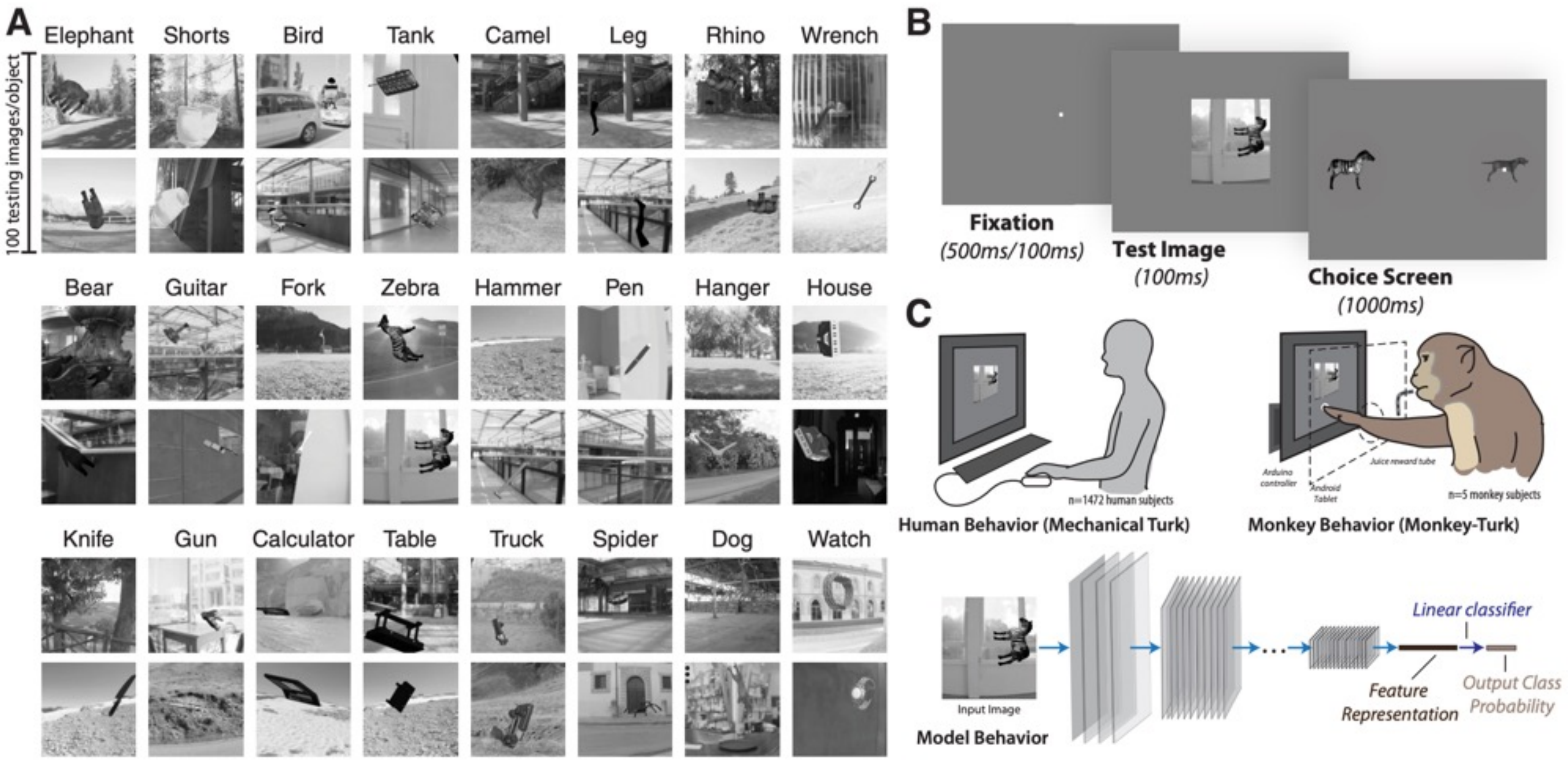
Prediction under adversary: Jeff Dean

Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irrecognisable to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

Deep neural networks solve problems by taking shortcuts

Geirhos, Robert, et al. "Shortcut learning in deep neural networks." *Nature Machine Intelligence* 2.11 (2020): 665-673.

Cognitive Mechanism



Behavioral Prediction Task: Human vs. Monkey vs. Deep Nets

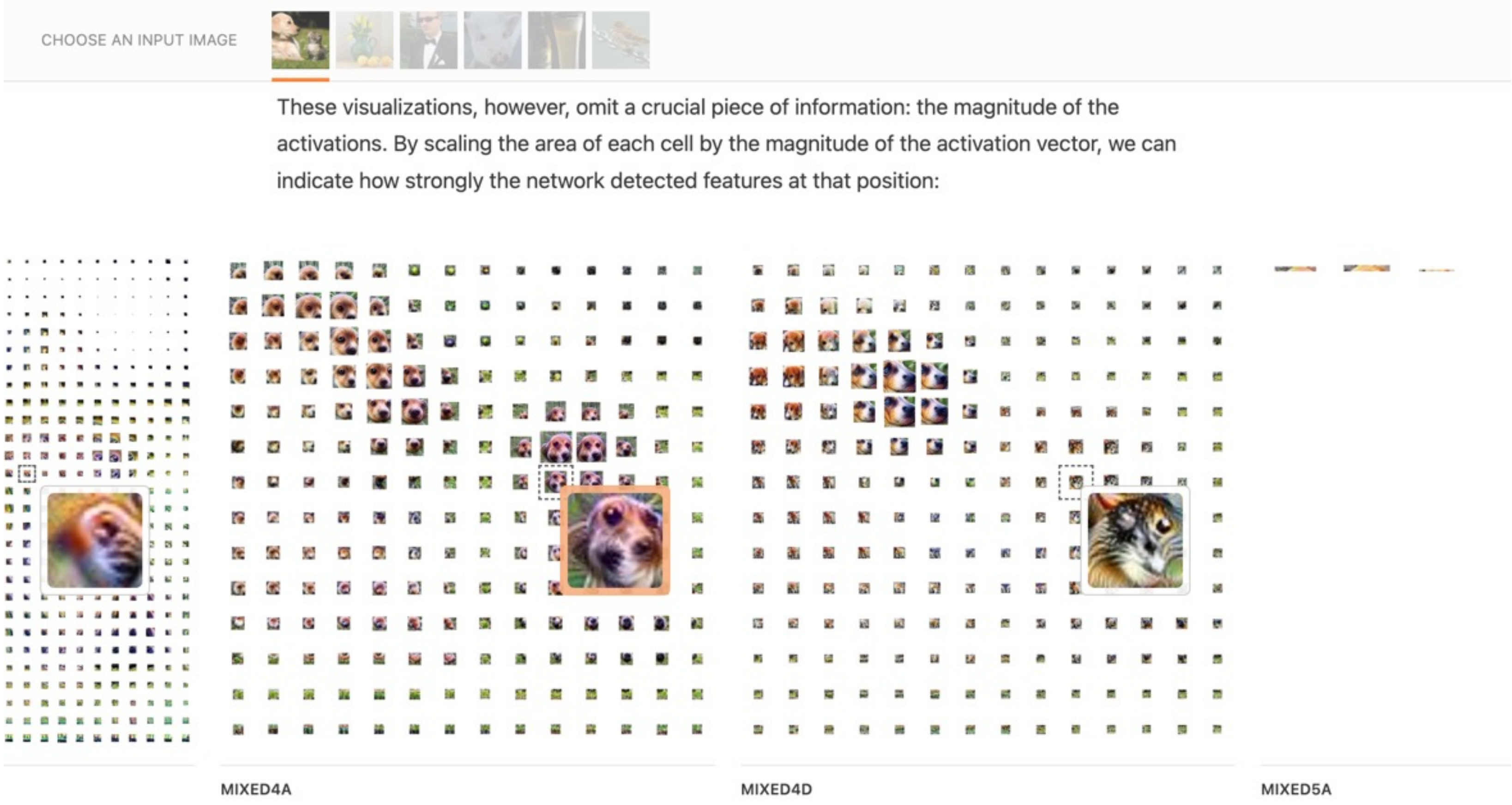
Rajalingham, Rishi, et al. “Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks.” *Journal of Neuroscience* 38.33 (2018): 7255-7269. Rajalingham, Rishi, Kailyn Schmidt, and James J. DiCarlo. "Comparison of object recognition behavior in human and monkey." Journal of Neuroscience 35.35 (2015): 12127-12136.

What is Missing

Many theoretical work or interpretation tools have been proposed

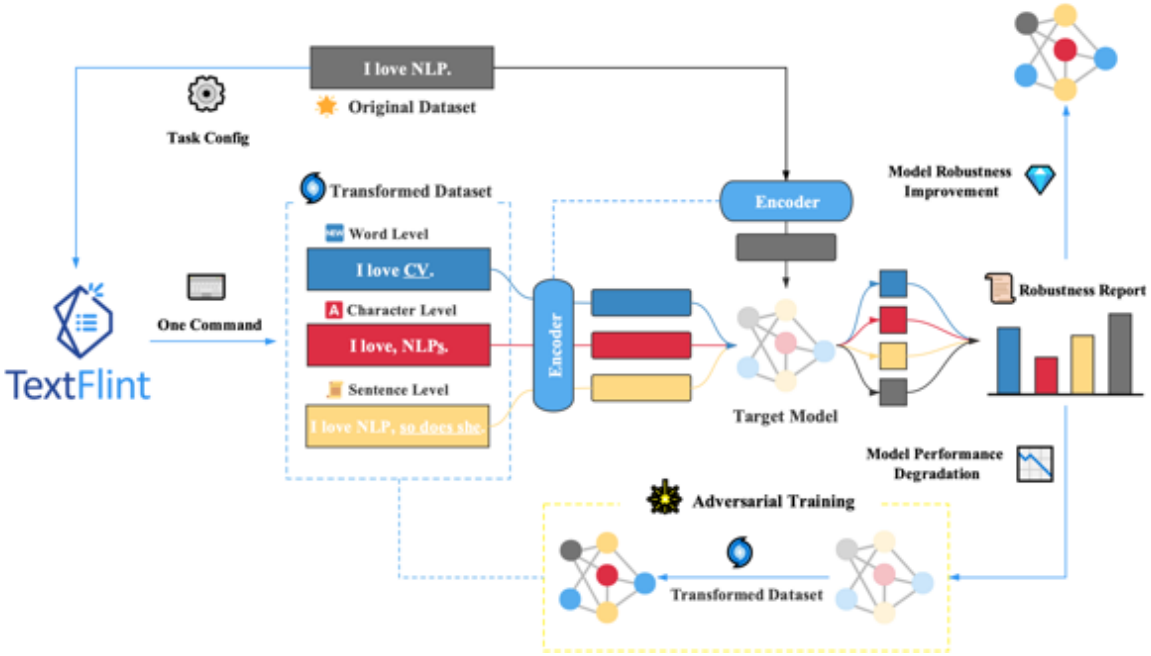
Yet, we don’t have an all-in-one system to explain everything.

FudanNLP TextFlint



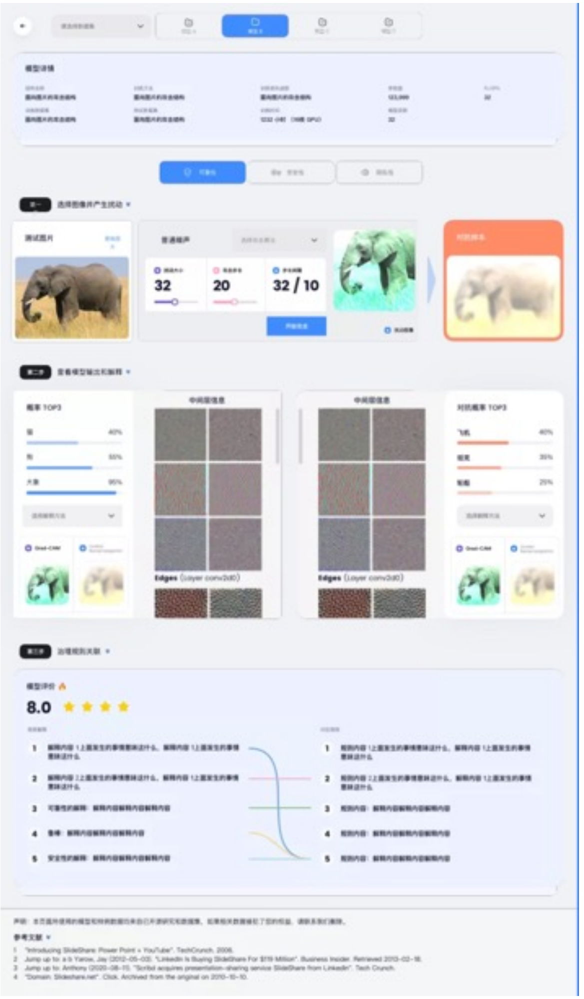
https://distill.pub/2018/building-blocks/

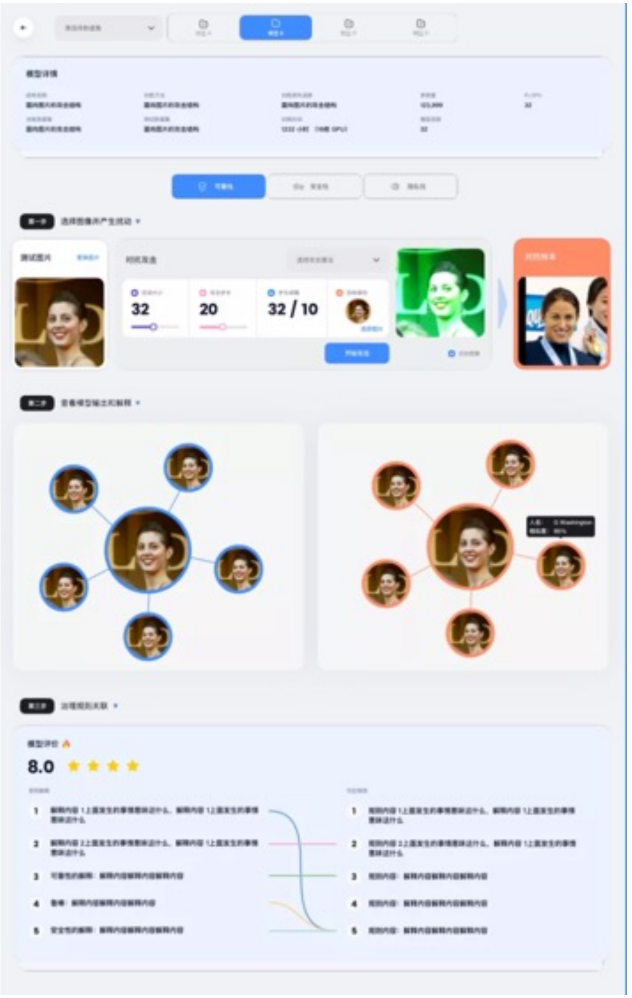
FudanNLP TextFlint



https://textflint.github.io/; https://www.textflint.com/textflint

FVL Risk Demo Platform



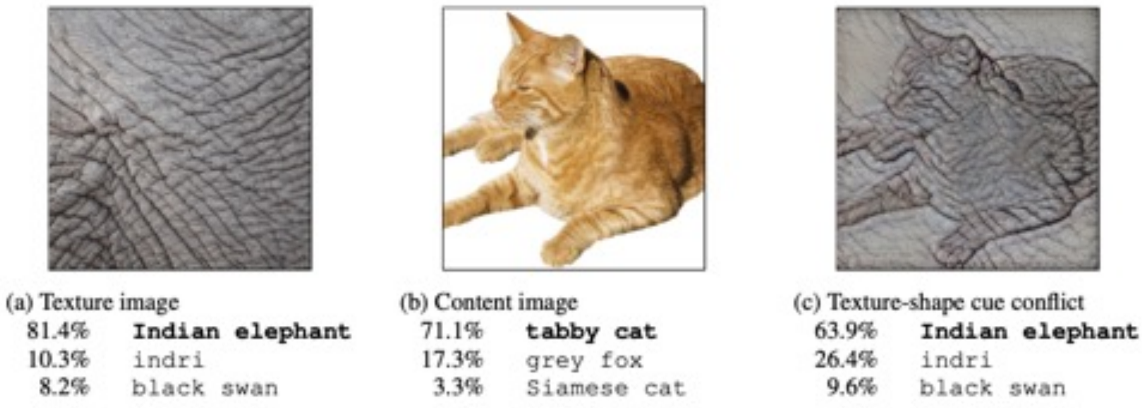


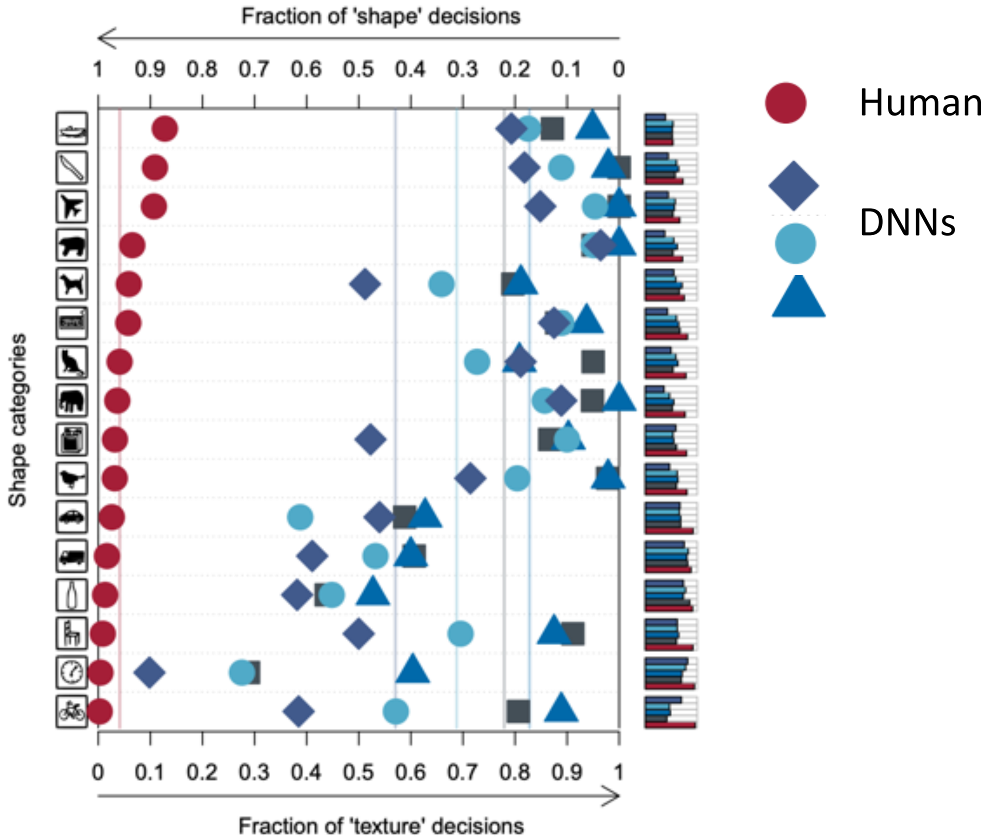
The Risk Demo Project: https://tech.openeglab.org.cn/dss

Common Robustness

- **Texture bias**
- **Robustness to common corruptions**

Texture bias

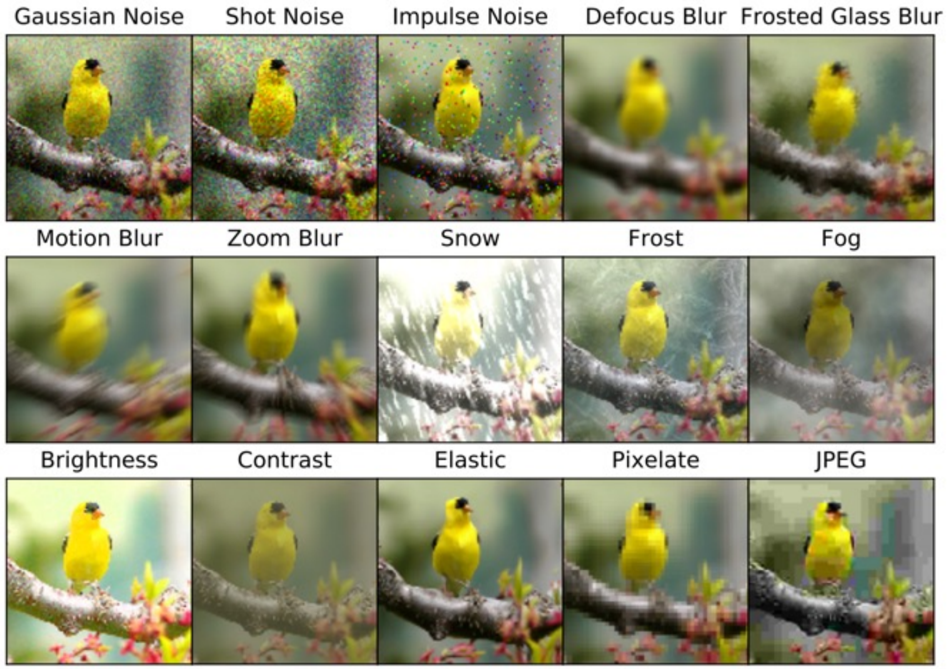


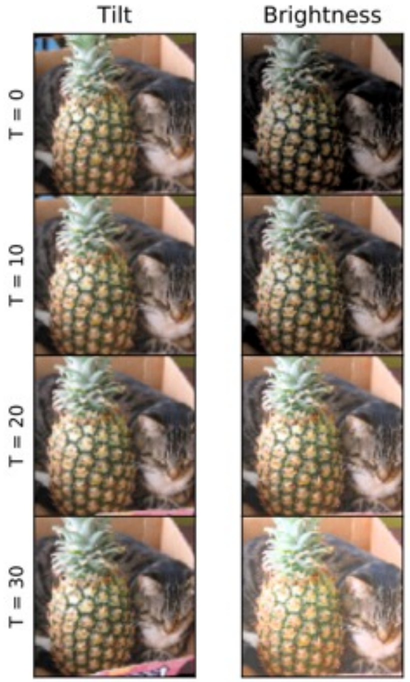


Temporary solution: Data Augmentation (**Style Transfer**)
ImageNet -> Stylized-ImageNet

Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." *ICLR*, 2019.

Common Corruptions





ImageNet-C:

- 15 types of noise
- 5 severity levels

ImageNet-P:

- 10 types of perturbation

Temporary solution: **Data augmentation** vs. **Adversarial Logit Pairing**

Hendrycks&Dietterich. "Benchmarking Neural Network Robustness to Common Corruptions and Perturbations." *ICLR*, 2019.

谢谢！下周见！

Email: xingjunma@fudan.edu.cn
Personal page: www.xingjunma.com
Office: 江湾校区交叉二号楼D5025