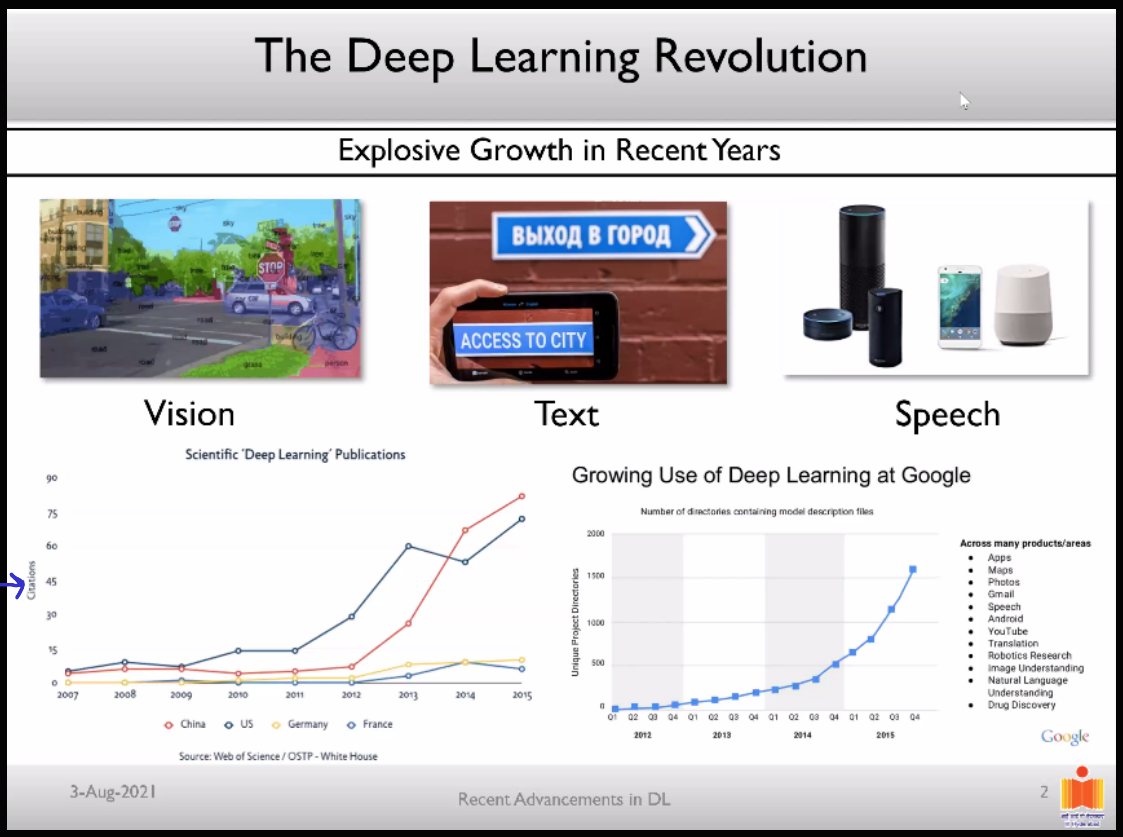
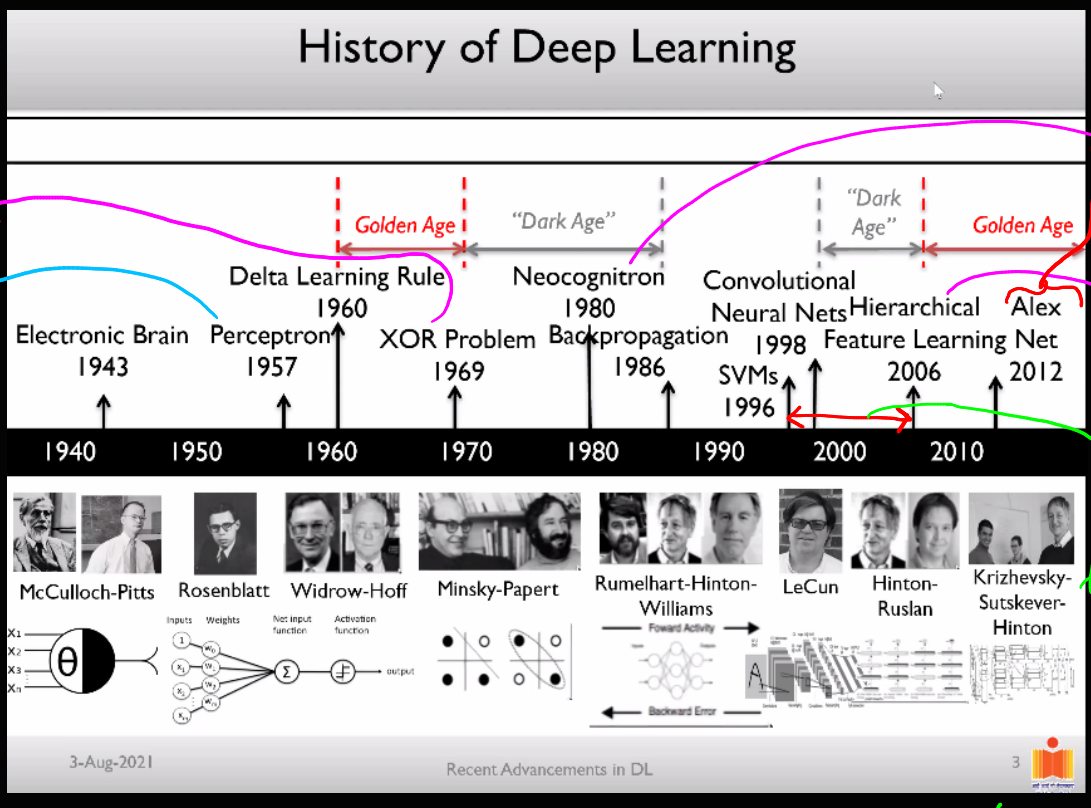
Day 2

Speaker: Prof. Vineeth Balasubramanian, IITH

Title: Recent advancements in Deep Learning.





* Perceptron proved flawed

* NN with no hidden layer

* Drawback: Only linear classification

Breakthrough in DL

Precursor to CNN.

Belief learning

Unsupervised learning

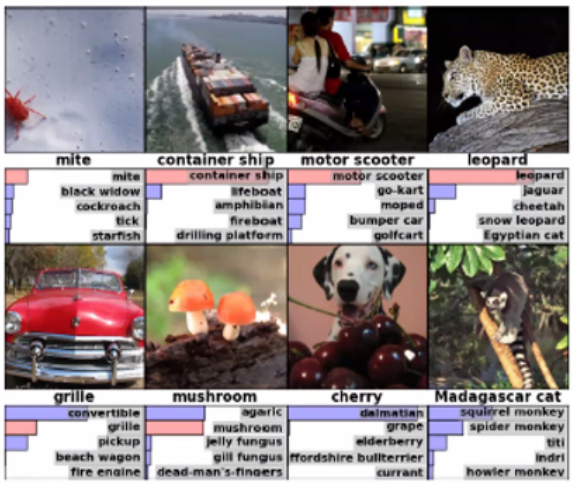
Random forests, Adaboost etc.

* Disclaimer: Random forests, Adaboost → Tabular Data. (Still Used)

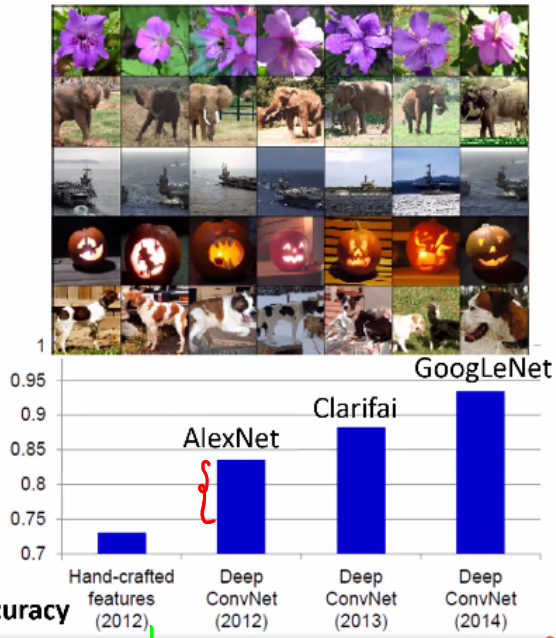
Compositionality (Images, Multimedia) → Deep Learning
↓
Text, Speech

Deep Learning in Computer Vision: The Turning Point

Sample Results on ImageNet



Source: Krizhevsky et.al. NIPS'12



Top-5 classification accuracy

Image Processing features

Deep Learning in Visual Computing: The Turning Point

More results on ImageNet

Top-5 Error on Imagenet Classification Challenge (1000 classes)

Method	Top-Error Rate
SIFT+FV [CVPR 2011]	~25.7%
AlexNet [NIPS 2012]	~15%
ZeilerNet [ImageNet 2013]	~11%
Oxford-VGG [ICLR 2015]	~7%
GoogLeNet [CVPR 2015]	~6%, ~4.5%
MSRA [arXiv 2015]	~3.5%
Human Performance	3 to 5 %

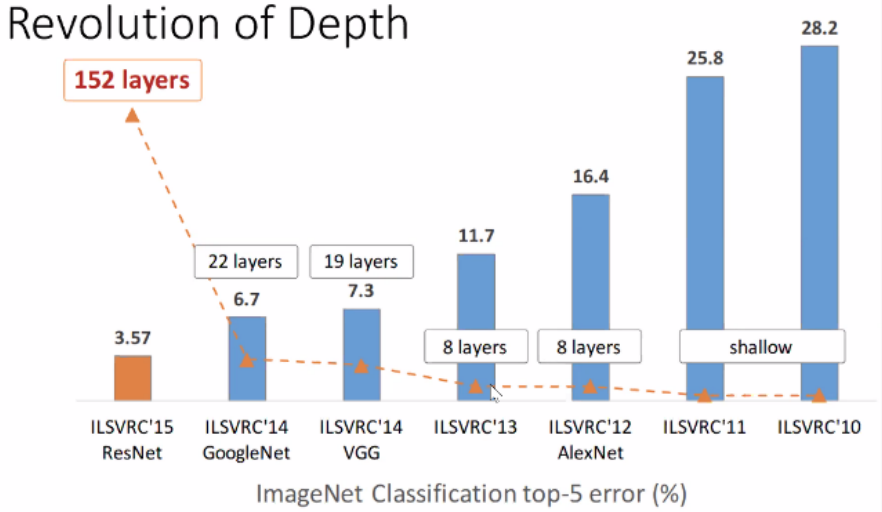
} Close to Human

As DL came into picture Error Rate ↓

1. Consolidation of Successes in DL

Deeper the Merrier

Order of 100s today. Do we need more?



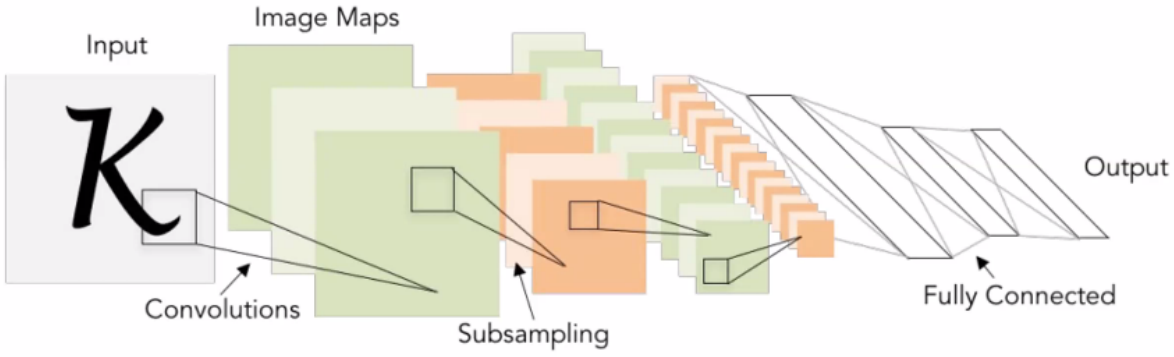
Do Deep Convolutional Nets Really Need to be Deep (Or Even Convolutional)? ICLR 2017
When and Why Are Deep Networks Better than Shallow Ones? AAAI 2017

* After a certain depth accuracy saturates and also computational expenses ↑. So, typically (10-100) layers are used. [Deeper not always better]

* Summary of Different Architectures :

Architectures of CNNs over time

LeNet-5 (1998)

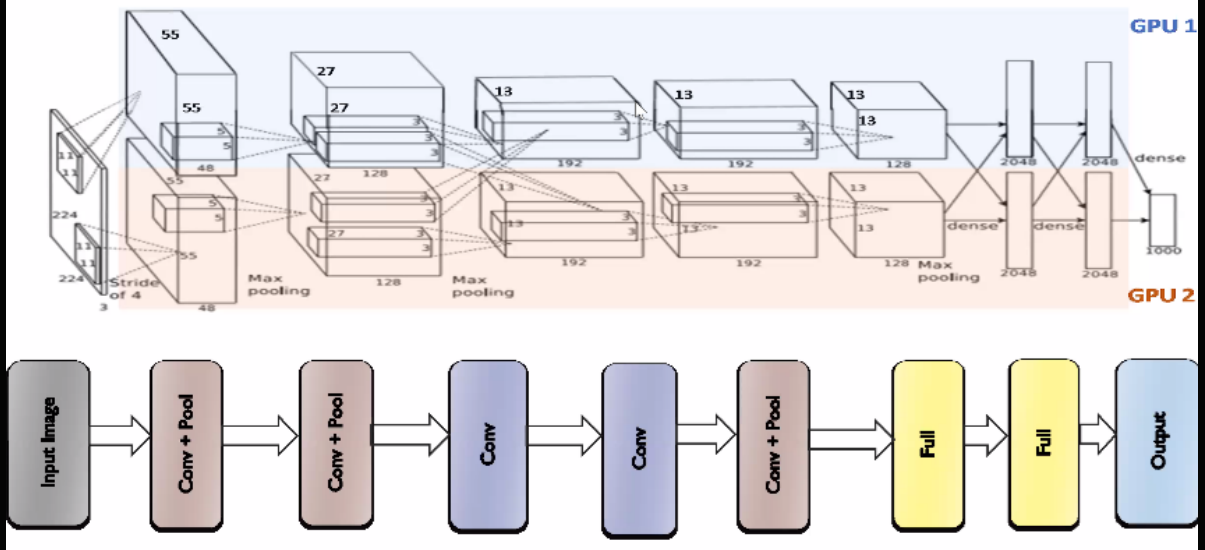


Conv filters were 5x5, applied at stride 1
Subsampling (Pooling) layers were 2x2 applied at stride 2
i.e. architecture is [CONV-POOL-CONV-POOL-FC-FC]

Courtesy: Fei-Fei Li and Andrej Karpathy, CS231n course, Stanford, Spring 2017

Architectures of CNNs over time

AlexNet, 2012

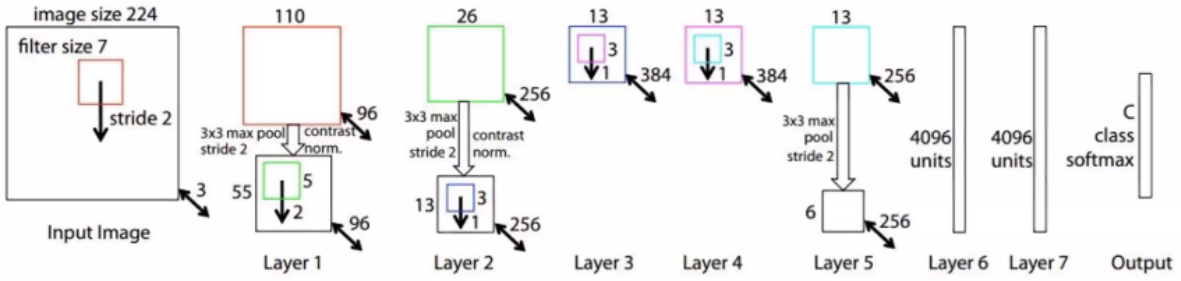


Courtesy: Fei-Fei Li and Andrej Karpathy, CS231n course, Stanford, Spring 2017

* Used various filters, strides, width to experiment.

Architectures of CNNs over time

ZFNet, 2013



AlexNet but:
CONV1: change from (11x11 stride 4) to (7x7 stride 2)
CONV3,4,5: instead of 384, 384, 256 filters use 512, 1024, 512

ImageNet top 5 error: 16.4% -> 11.7%

Courtesy: Fei-Fei Li and Andrej Karpathy, CS231n course, Stanford, Spring 2017

* Not much changes than AlexNet but smaller architectural changes only
* Error Rate ↓.

VGG-Net

11-layer	13-layer	16-layer	19-layer
image	image	image	image
conv-64	conv-64	conv-64	conv-64
maxpool	conv-64	conv-64	conv-64
	maxpool	maxpool	maxpool
conv-128	conv-128	conv-128	conv-128
maxpool	conv-128	conv-128	conv-128
	maxpool	maxpool	maxpool
conv-256	conv-256	conv-256	conv-256
conv-256	conv-256	conv-256	conv-256
maxpool	maxpool	conv-256	conv-256
		maxpool	conv-256
			maxpool
conv-512	conv-512	conv-512	conv-512
conv-512	conv-512	conv-512	conv-512
maxpool	maxpool	conv-512	conv-512
		maxpool	conv-512
			maxpool
conv-512	conv-512	conv-512	conv-512
conv-512	conv-512	conv-512	conv-512
maxpool	maxpool	conv-512	conv-512
		maxpool	conv-512
			maxpool
FC-4096	FC-4096	FC-4096	FC-4096
FC-4096	FC-4096	FC-4096	FC-4096
FC-1000	FC-1000	FC-1000	FC-1000
softmax	softmax	softmax	softmax

- More layers lead to more nonlinearities
- Only 3 x 3 CONV stride 1, pad 1 and 2 x 2 MAXPOOL stride 2
- Smaller receptive fields:
  - less parameters; faster
  - two 3 X 3 leads to 5 X 5; three 3 x 3 leads to 7 x 7
  - Fewer parameters:
  - $3 \times 3^2C^2$ (vs) 7^2C^2

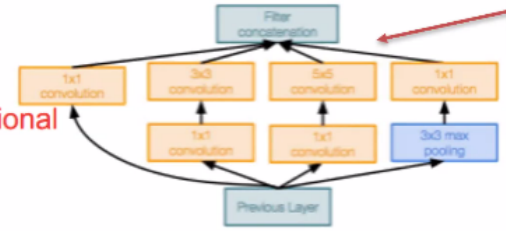
* Stick to only 3x3 filters (Don't worry about filter size)
* Tried diff. depths → 16 was found to be most efficient
* Fewer Parameters (for 3x3 filters) * Huge in size of n/w.

GoogLeNet

"Inception module": design a good local network topology (network within a network) and then stack these modules on top of each other

Deeper networks, with computational efficiency

- 22 layers
- Efficient "Inception" module
- No FC layers
- Only 5 million parameters! 12x less than AlexNet
- ILSVRC'14 classification winner (6.7% top 5 error)

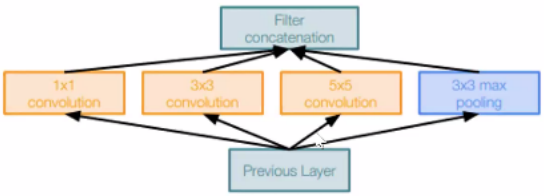


Inception module

Courtesy: Fei-Fei Li and Andrej Karpathy, CS231n course, Stanford, Spring 2017

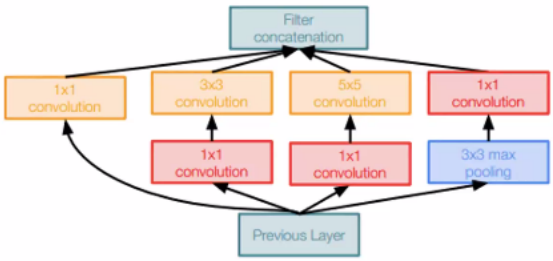
* All filters used & concatenated.
* Lesser Parameters (Huge success)
Then AlexNet.

GoogLeNet



Naive Inception module

1x1 conv "bottleneck" layers *



Inception module with dimension reduction

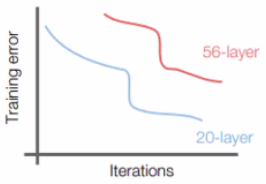
Courtesy: Fei-Fei Li and Andrej Karpathy, CS231n course, Stanford, Spring 2017

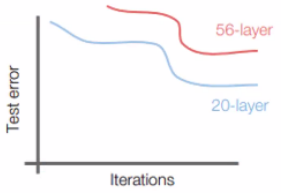
→ To reduce computation expenses 1x1 conv's are used.

* Significant explosion of computations

How deep can we go?

What happens when we continue stacking deeper layers on a "plain" convolutional neural network?





Courtesy: Fei-Fei Li and Andrej Karpathy, CS231n course, Stanford, Spring 2017

Problems with Deeper layers

≈ Vanishing Gradients

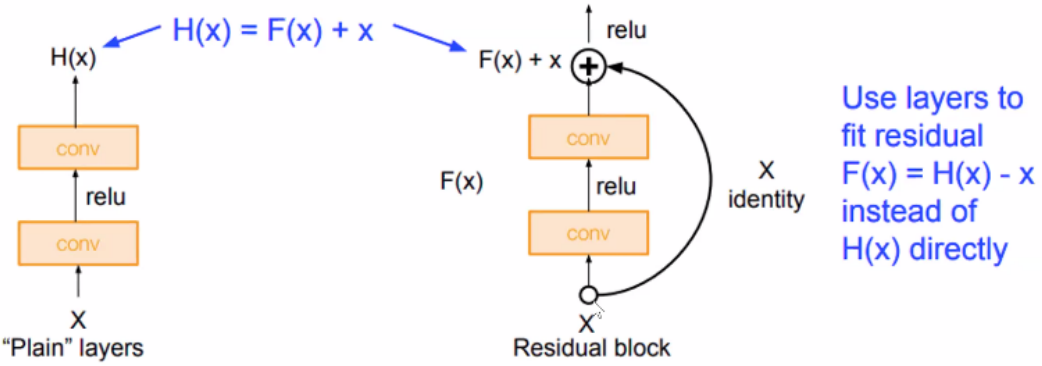
≈ Overfitting in Overparameterization Regime

* Residual Nets came to rescue when addressing vanishing gradients problem.

Residual Nets

The deeper model should be able to perform at least as well as the shallower model.

Solution: Use network layers to fit a residual mapping instead of directly trying to fit a desired underlying mapping



Courtesy: Fei-Fei Li and Andrej Karpathy, CS231n course, Stanford, Spring 2017

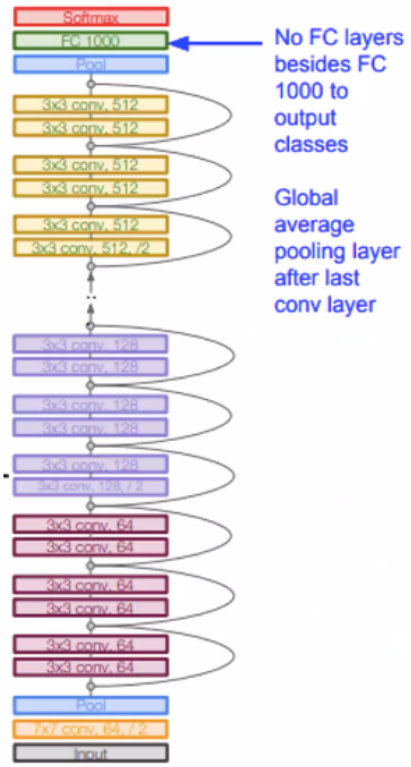
* Features are transferred to the later layers via skip connections

Residual Nets

Full ResNet architecture:

- Stack residual blocks
- Every residual block has two 3x3 conv layers
- Periodically, double # of filters and downsample spatially using stride 2 (/2 in each dimension)
- Additional conv layer at the beginning
- No FC layers at the end (only FC 1000 to output classes)

Total depths of 34, 50, 101, or 152 layers for ImageNet



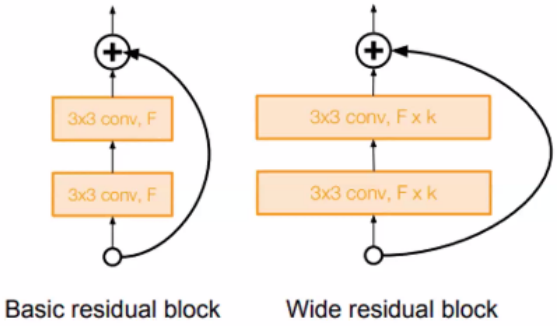
Courtesy: Fei-Fei Li and Andrej Karpathy, CS231n course, Stanford, Spring 2017

* 152 layers! Things starting saturating. They experimented with a variety of depths.

Wide Residual Networks

Zagoruyko et al. 2016

- Argues that residuals are the important factor, not depth
- User wider residual blocks (F x k filters instead of F filters in each layer)
- 50-layer wide ResNet outperforms 152-layer original ResNet
- Increasing width instead of depth more computationally efficient (parallelizable)



Basic residual block | Wide residual block

Courtesy: Fei-Fei Li and Andrej Karpathy, CS231n course, Stanford, Spring 2017

* Going wide (i.e ↑ feature maps) also helps.

DenseNets

Densely Connected Convolutional Nets, Huang et al 2017

- Dense blocks where each layer is connected to every other layer in feedforward fashion
- Alleviates vanishing gradient, strengthens feature propagation, encourages feature reuse

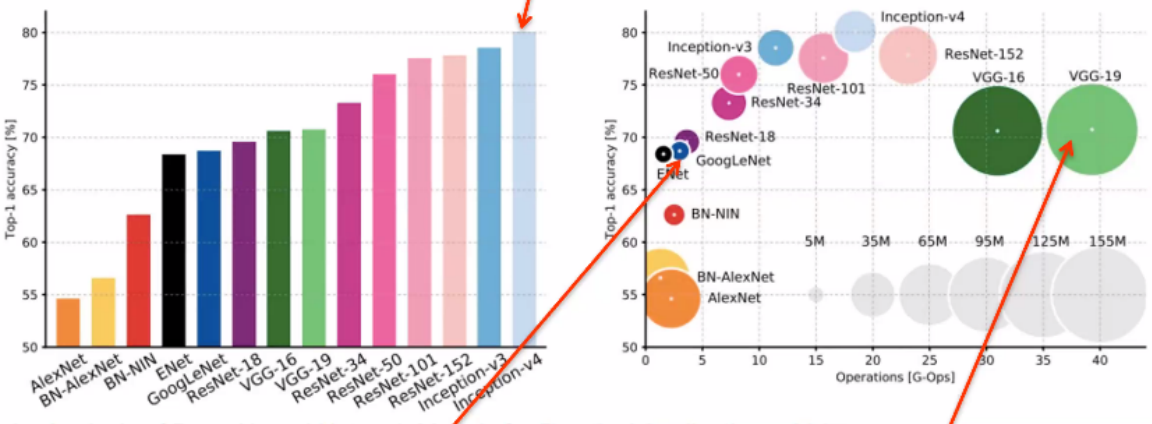
Dense Block

Courtesy: Fei-Fei Li and Andrej Karpathy, CS231n course, Stanford, Spring 2017

* Vanishing Gradient ↓
* Feature Reuse.

Comparing Complexity

Inception v4 = ResNets + Inception



An Analysis of Deep Neural Network Models for Practical Applications, 2017.

GoogLeNet = most efficient

VGG = High memory, most ops

Courtesy: Fei-Fei Li and Andrej Karpathy, CS231n course, Stanford, Spring 2017

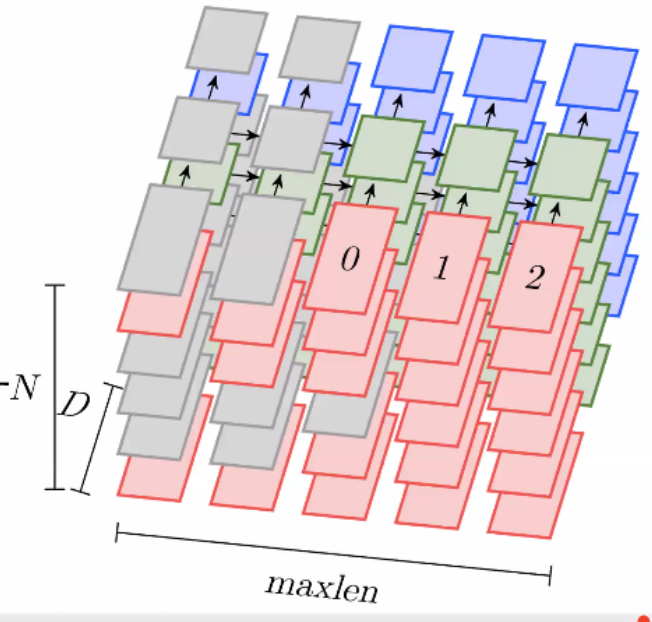
* Inception ResNets → Good Tradeoff betⁿ performance & efficiency.
* VGG Net → Huge Computational Costs

Deeper the Merrier

Stacking LSTMs

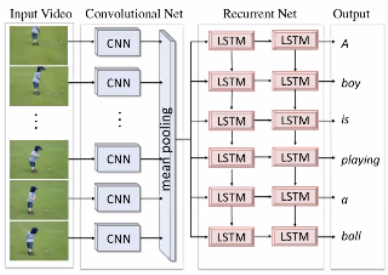
Applied in Google Translate →

The Google Neural Machine Translation (GNMT)'s encoder is essentially a series of stacked LSTMs

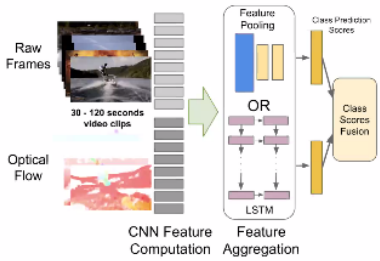


A Primer on Neural Network Models for Natural Language Processing, JAIR 2016

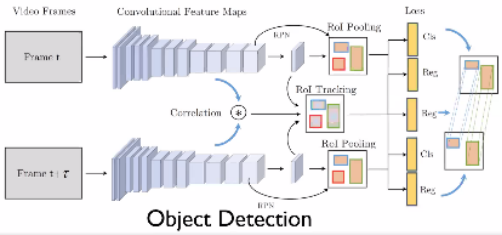
Hybrid Architectures



Video Captioning



Video Classification



Object Detection

- End-to-end backprop on hybrid architectures
- Mix-and-match plug-and-play modules

CNN's + LSTM's , CNN + RNN's etc

* To mix. to architecture in such a way to encourage end to end learning. (Simple & Easy to learn)

* NN → Composition of functions

↓

As long as fun^n's are differentiable we can plug any architectures and get the solution

*

Regularization and Hyperparameter Engineering

Little Pieces that have made the Whole

Regularization
- DropOut, DropConnect, Batch Normalization, Data Augmentation, Noise in Data/Label/Gradient

Weight Initialization
- Xavier's initialization, He's initialization

Choosing Gradient Descent Parameters
- Adagrad, RMSProp, Adam, Momentum, Nesterov Momentum

Activation Functions
- ReLU, PReLU, Leaky ReLU, ELU

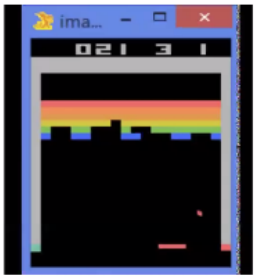
Loss Functions
- Cross-Entropy, Embedding Loss, Mean-Squared Error, Absolute Error, KLDivergence, Max-Margin Loss

We can experiment with diff regularization & hyperparameters to fine tune our models.

[Mostly task specific]

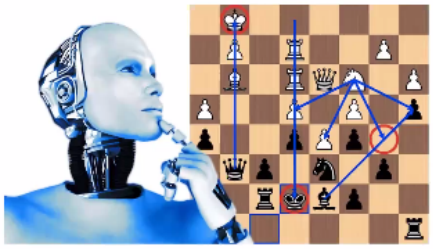
Reinforcement Learning

Reinforcing Deep Reinforcement Learning



Atari Breakout
2013

Alpha Go
Apr 2016



Alpha Zero
Dec 2017

AlphaGo learned to play, along with programming of some heuristics;

* AlphaZero learned purely by 'self-play'! (Trained itself in 4 hours of playing itself)

FINE PRINT: AlphaZero trained using 5000 first-generation TPUs and 64 2nd-gen TPUs in parallel

→ Computationally Very Expensive.

Proliferation of Applications

Vision, Speech, Text and Beyond

ADAS

Drones

Mobile

Surveillance

Augmented Reality

Siri

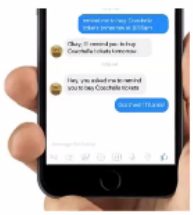
Google Assistant

Hey Cortana

amazon alexa

Google Translate

Meet **Jarvis**.
He reminds you to do things.



* LeCunn Cake Analogy Explained
* Universal Approximation Theoram of DL → Explained } To read.

Deep Learning Frameworks

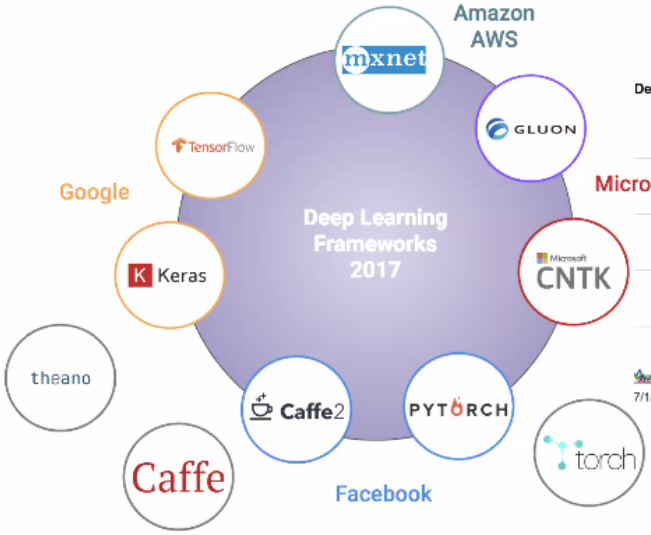


Image Credit: https://towardsdatascience.com

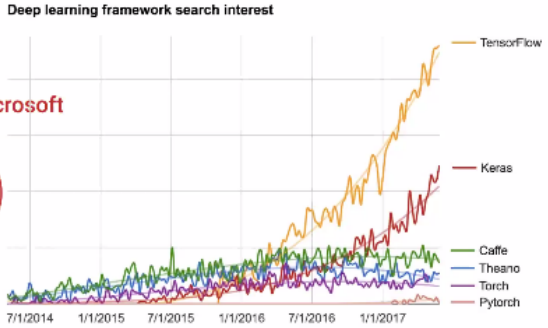
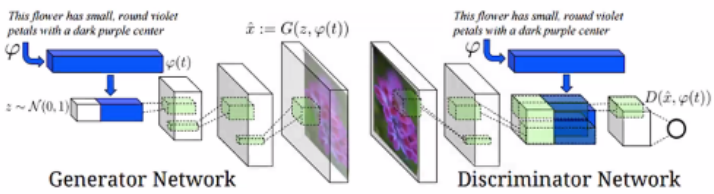


Image Credit: Francois Chollet

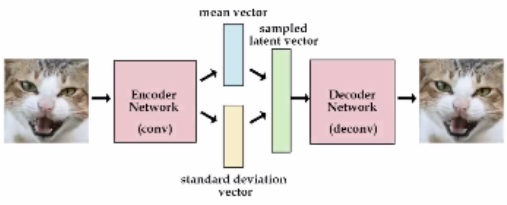
* Open Source Frameworks have acted as a catalyst in DL Research.

2. Exploration of New Frontiers.

Deep Generative Models



Generative Adversarial Networks (GANs)



Variational Autoencoders (VAEs)

"Adversarial training is the coolest thing since sliced bread." – Yann LeCun, 2016 (Quora)

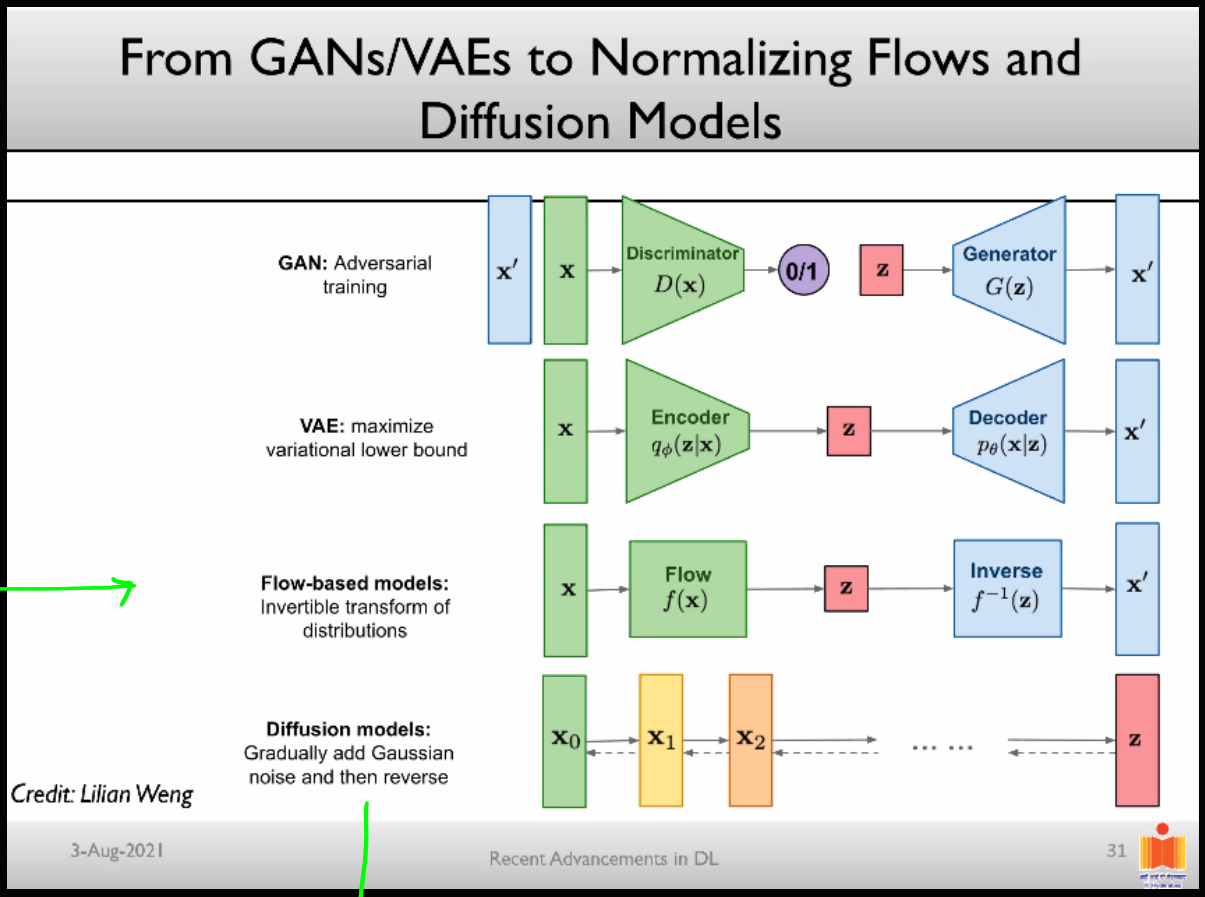
- Many other variants too (E.g. Pixel CNNs)
- Many possible applications (E.g. Art)
- Potential to contribute to unsupervised/semi-supervised learning

https://github.com/hollobit/All-About-the-GAN

* Can generate more data → MANY APPLICATIONS.

loosely { * GAN's → to generate beautiful data
* VAE's → to generate variations in our own self (like blond hair, blue eyes etc)

* VAE's and GAN are both types of adversarial networks.



Encoders & Decoders are exactly inverse

Data ⟶ Gaussian Noise ⟶ Latent Space

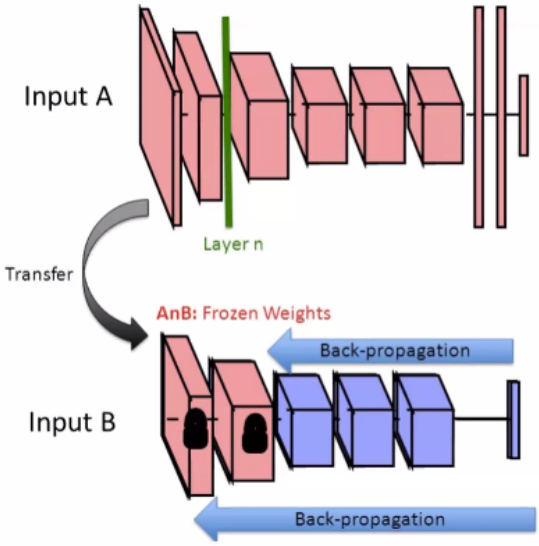
Reconstruction

* Blogs of Lilian Weng

↑

To check.

Transfer Learning



- Learning in new domains with limited data/annotation/expertise
- Towards unsupervised/semi-supervised learning
- Zero/One/Few-shot Learning

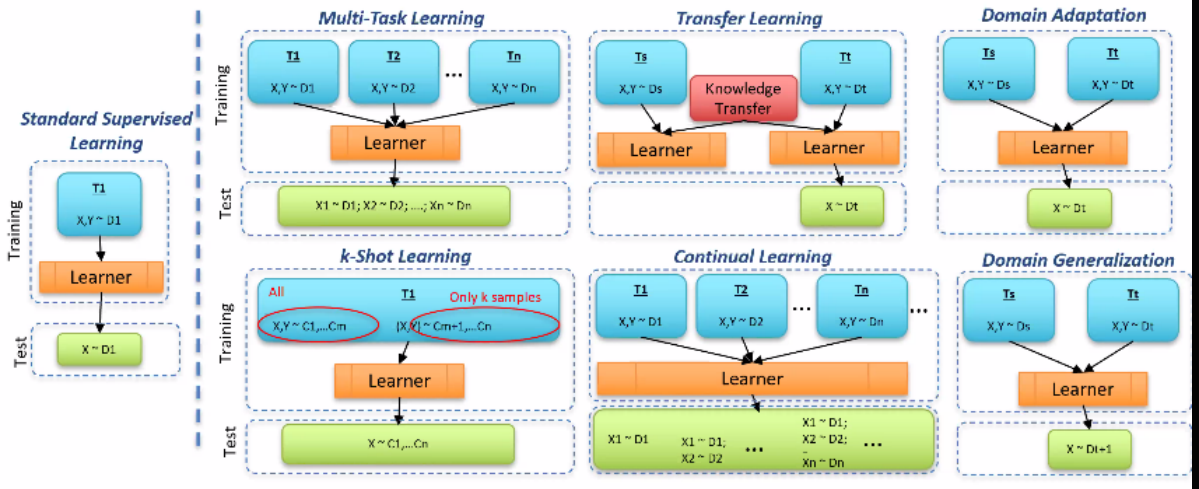
How transferable are features in deep neural networks? NIPS 2014

A Gentle Introduction to Transfer Learning for Deep Learning

* Limited Data, Annotations, Expertise → Transfer Learning

Learning with Limited Supervision

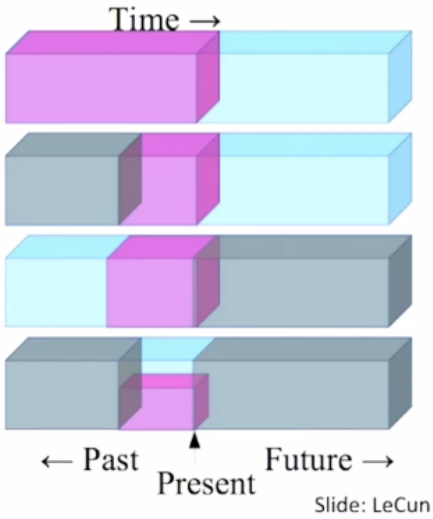
Different Settings



* Explore each fields (Interesting) [TO BE UPDATED]

From Unsupervised Learning to Self-Supervised Learning

- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.

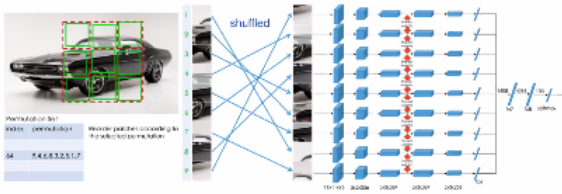


Slide: LeCun

* To treat unsupervised tasks as supervised tasks.

Applications :

From Unsupervised Learning to Self-Supervised Learning

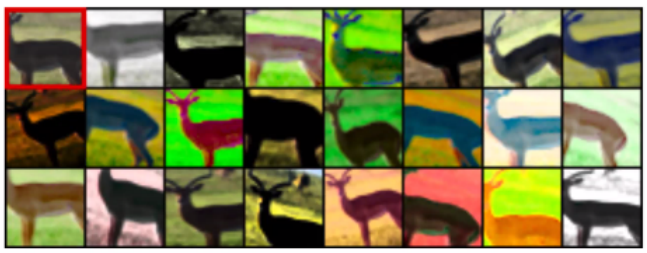


(a) Input context (b) Human artist

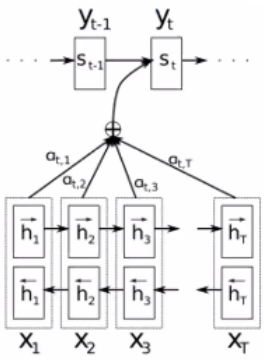
(c) Context Encoder (L2 loss) (d) Context Encoder (L2 + Adversarial loss)

- Solving jigsaw puzzles
- Rotation prediction
- Image inpainting
- Multi-view contrastive learning
- Many more…

pretext task from Unsupervised to Supervised

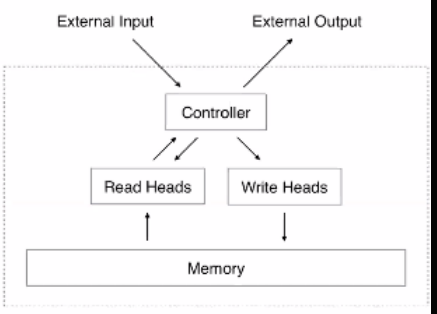


Attention and Memory in Deep Learning



Machine Translation

Image/Video Captioning, Visual Question Answering



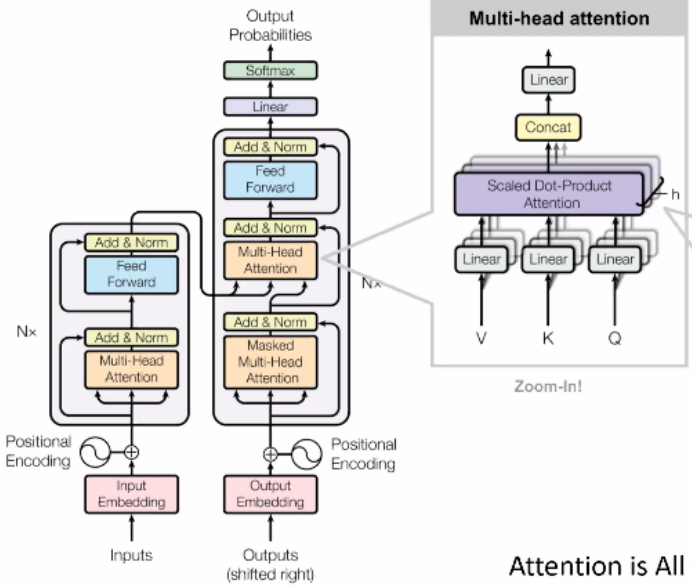
Neural Turing Machines and Memory Networks → Read.

Attention is all you need, NIPS 2017

http://www.wildml.com/2016/01/attention-and-memory-in-deep-learning-and-nlp/

* When parts of a data is to be paid more attention than the rest. [Like frisbee in the women throwing a frisbee photo etc]

Self-Attention and Transformers



$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}})\mathbf{V}$$

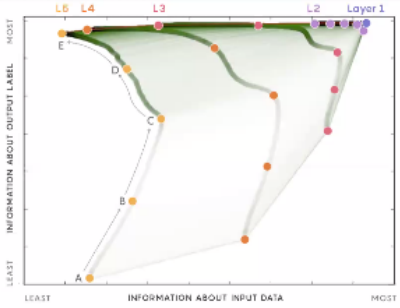
Attention is All you Need, Vaswani et al, NeurIPS 2017

* RNN's are replaced by Transformers. (Multi-head Attention Modules)

* Huge no. of Parameters → in Transformers

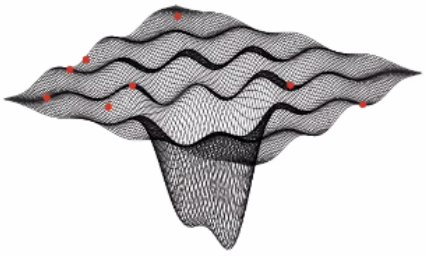
Theory and Optimization Methods

Theory of Deep Learning



- Information Bottleneck Principle, arXiv 2017
- Generalization in Deep Learning, arXiv 2017
- Random Matrix Theory, ICML 2017

Understanding Error Surfaces



- Deep Learning without Poor Local Minima, NIPS 2016
- How to Escape Saddle Points Efficiently, arXiv 2017
- Sharp Minima can Generalize for Deep Nets, ICML 2017

* Local Minima → Works well in almost all (millions of local minima, what global minima has stored for us)

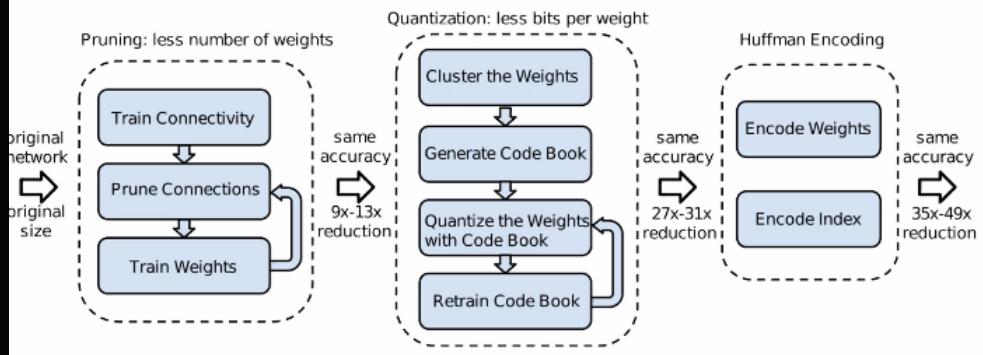
To read the above papers. (Theoritical)

Interesting!!

Efficient Deep Learning

Towards Deployment on Edge Devices

AlexNet = 200 MB+, VGGNet = 500 MB+; how to scale deep learning to edge devices?



Deep Compression, ICLR 2016 Best Paper: ~ 50x compression for VGGNet with zero loss in performance

Many other methods (Hashing, Matrix Factorization, Knowledge distillation, SqueezeNet, FitNets, Binarized Neural Nets, etc)

A Survey of Model Compression and Acceleration for Deep Neural Networks, IEEE SP Magazine, 2017
https://git.io/vN6zG

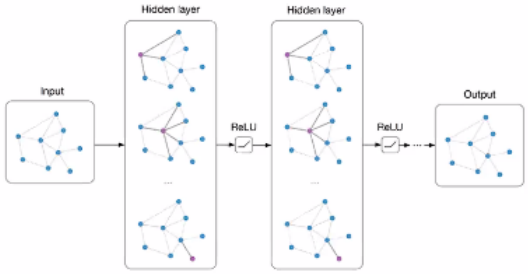
* ML/DL Models in Small/Remote devices →

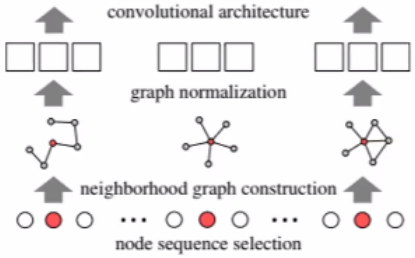
* DL in Hardware { Check }

Advanced { To check }

Other Emerging Directions

- Deep learning on graphs
  - *Graph Convolutional Networks, ICLR 2017*
  - *Learning CNNs for Graphs, ICML 2017*
- Geometric deep learning
  - *Geometricdeeplearning.com*
- Deep learning meets physics
  - *https://pbdl2017.github.io/index.html*

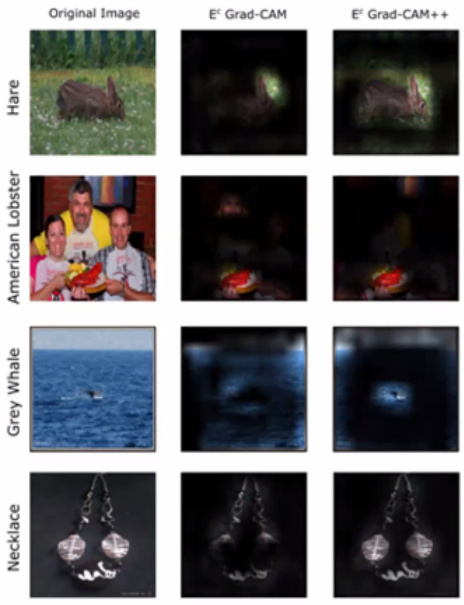




Limitations & Challenges (To analyse) → TO BE UPDATED

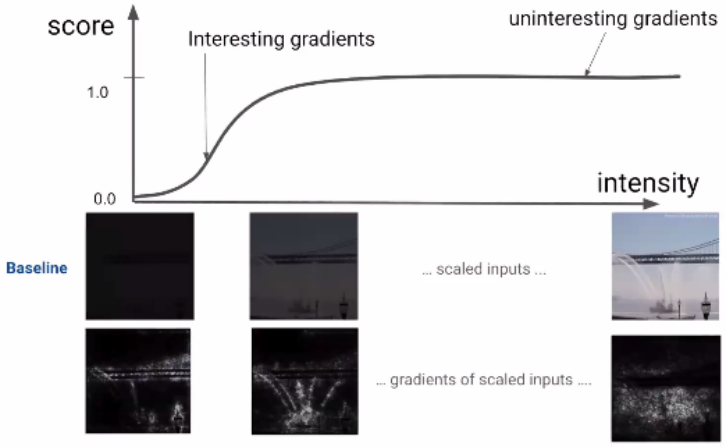
Interpretability and Explainability

- Why deep learning models work?
  - *Theory of deep learning*
- How deep learning models work?
  - *Visualizing and Understanding CNNs, ECCV 2014*
  - *CAM, Grad-CAM, Grad-CAM++*
  - *Visualizing and Understanding RNNs, arXiv 2015*
- Long way to go!



Interpretability and Explainability

Spate of Methods!

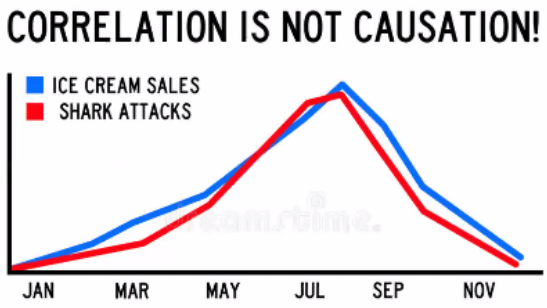


- CAM, GradCAM, GradCAM++
- Integrated Gradient
- SmoothGrad
- DeepSHAP
- LIME, LRP, DeepLIFT
- ….

Need for Causal Inference

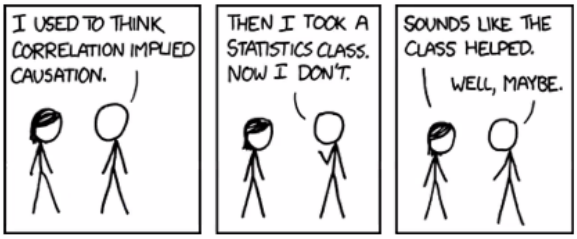
Causality vs Correlation

Deep learning models correlation, not causality



Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by each other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

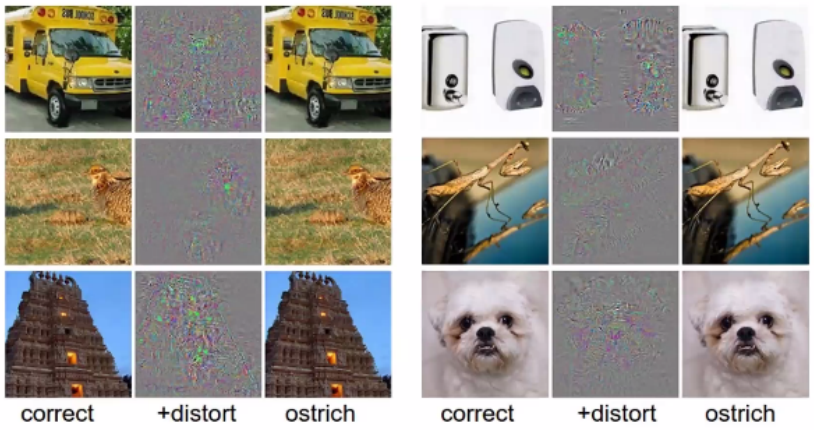
Image Credit: http://www.statisticshowto.com/causation/



- Some recent efforts
  - *Discovering Causal Signals in Images, CVPR 2017*
  - *Counterfactual Visual Explanations, ICML 2019*
- Long way to go!

Robustness and Consistency

Neural networks are easily fooled, CVPR 2015

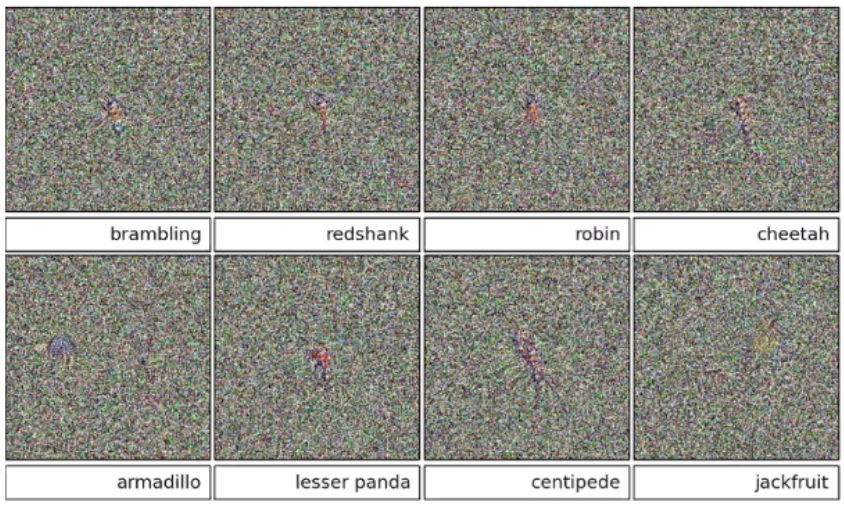


Classified with ~ 99% confidence

http://www.evolvingai.org/fooling

Robustness and Consistency

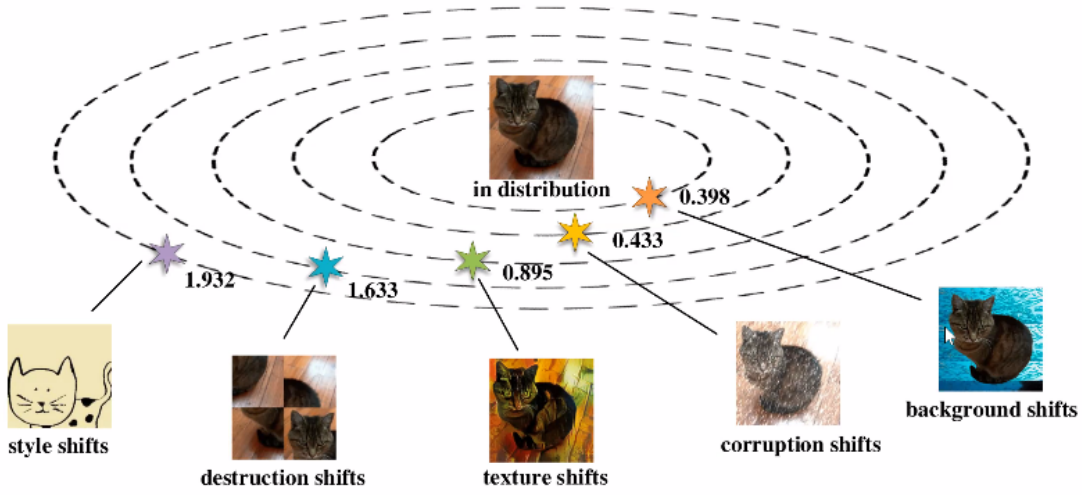
Neural networks are easily fooled, CVPR 2015



Classified with ~ 99% confidence

http://www.evolvingai.org/fooling

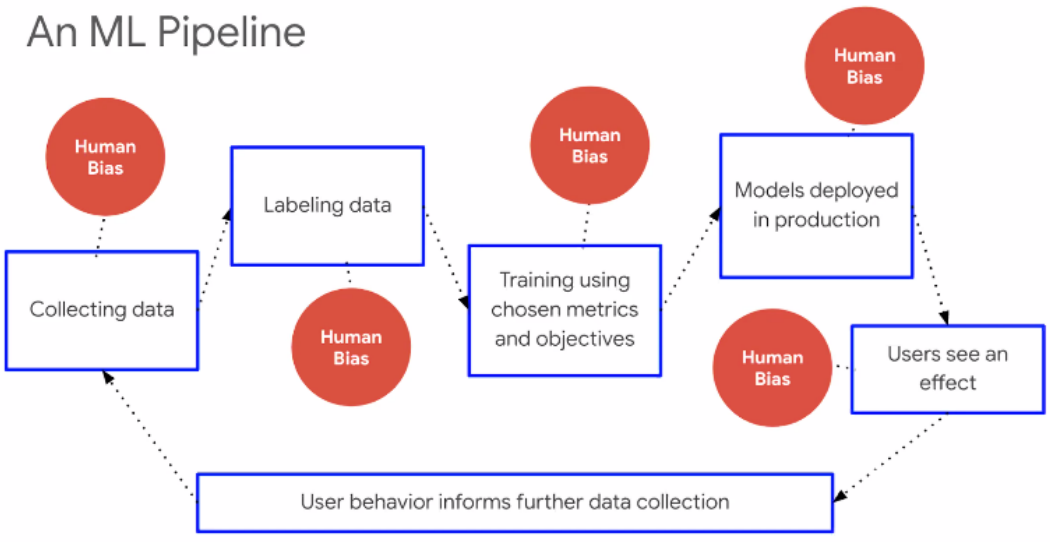
Out-of-Distribution Generalization



ImageNet-9, ImageNet-C, Stylized ImageNet, ImageNet-R, Random Patch Shuffling

Credit: https://pythonawesome.com/out-of-distribution-generalization-investigation-on-vision-transformers/

Bias, Fairness and Transparency in DL Models

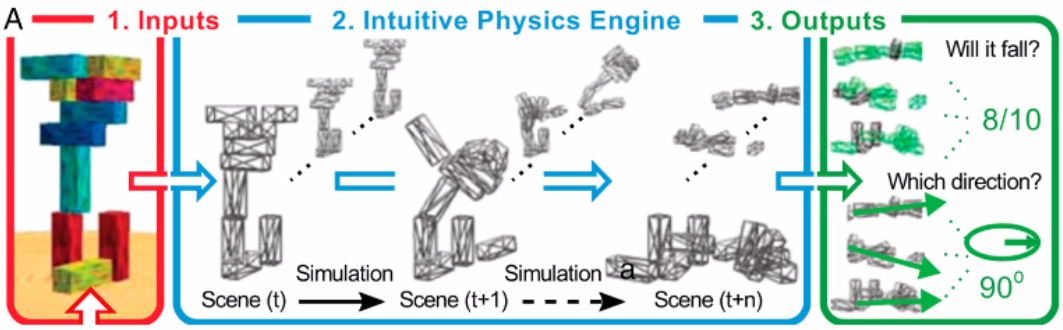


For more information: https://fairmlbook.org/

Credit: https://ai.googleblog.com/2019/12/fairness-indicators-scalable.html

Integrating Domain Knowledge

Deep learning models are highly data-driven



Learning Physical Intuition of Block Towers by Example, arXiv 2016

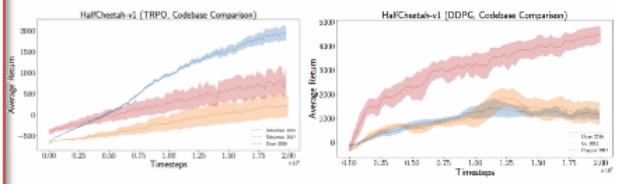
Why not integrate known physics laws?

Hyperparameters and Engineering

How to choose?

- Number of layers
- Number of neurons
- Activation functions
- Loss function
- Weight initialization strategy
- Learning rate
- Regularization constant
- Number of epochs
- ….

Reproducibility?



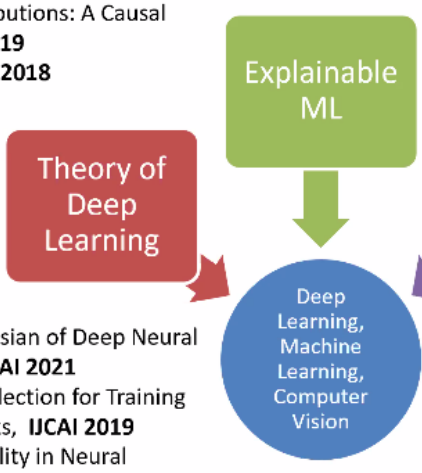
Deep Reinforcement Learning that Matters, AAAI 2018

"…*is more like alchemy.*"
– Ali Rahimi, NIPS 2017
(Test of Time talk)

* Prof Vineeth's Work & Lab at IITH.

Our Research at IIT-H

- Attributional Robustness, **ECCV 2020, AAAI 2021**
- Neural Network Attributions: A Causal Perspective, **ICML 2019**
- Grad-CAM++, **WACV 2018**



- On the Layerwise Hessian of Deep Neural Network Models, **AAAI 2021**
- Submodular Batch Selection for Training Deep Neural Networks, **IJCAI 2019**
- On Noise and Optimality in Neural Networks, **ICML 2018 Workshops**

- Towards Open World Detection, **CVPR 2021**
- Meta-consolidation for Continual Learning, **NeurIPS 2020**
- Manifold Mixup for Few-shot Learning, **WACV 2020**
- Zero-shot Task Transfer, **CVPR 2019**
- Adversarial Data Programming, **CVPR 2018**
- Attentive Semantic Video Generation using Captions, **ICCV 2017, ACM MM 2017**