

CMP784

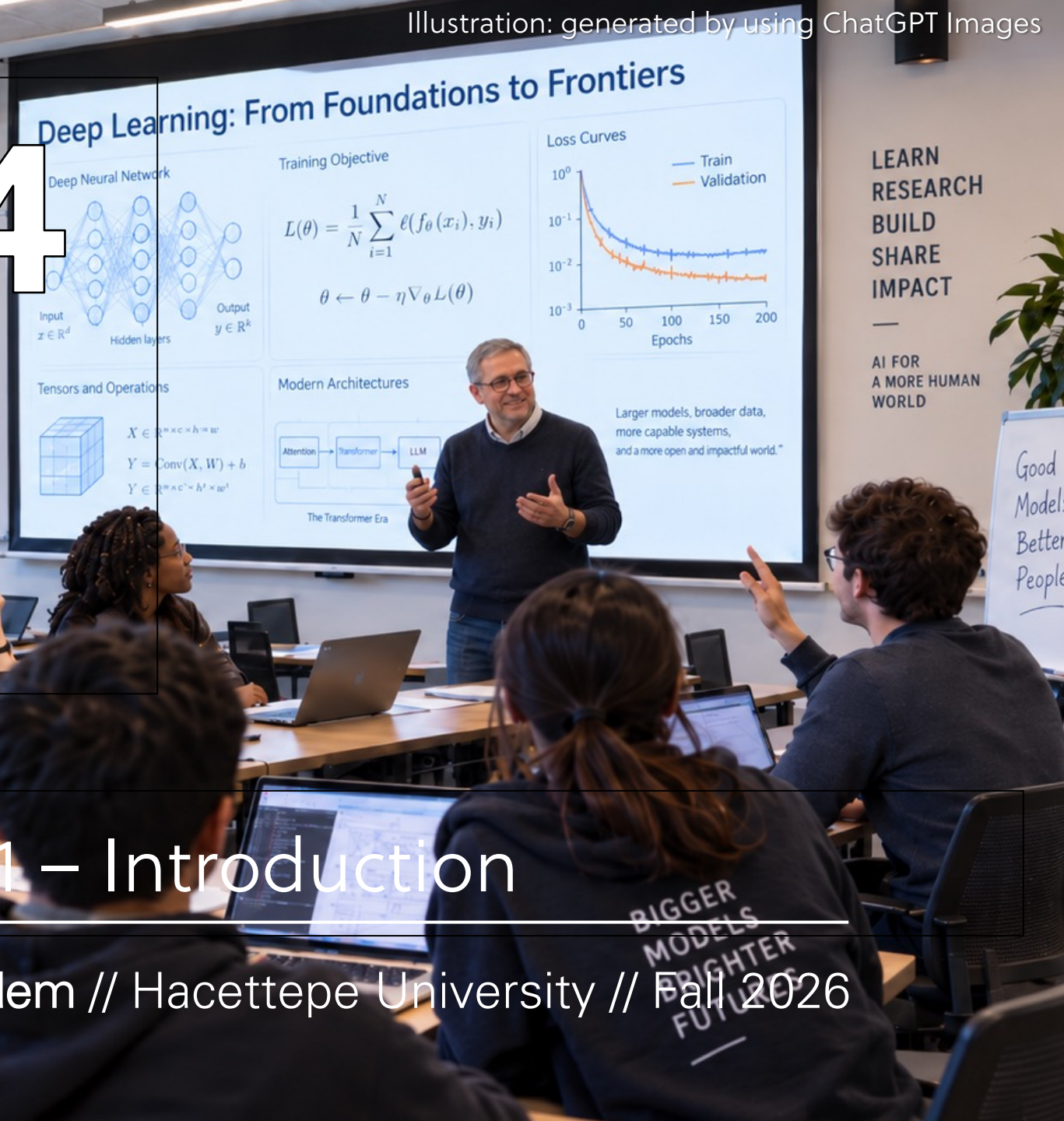
DEEP LEARNING

Lecture #01 – Introduction

Erkut Erdem // Hacettepe University // Fall 2026

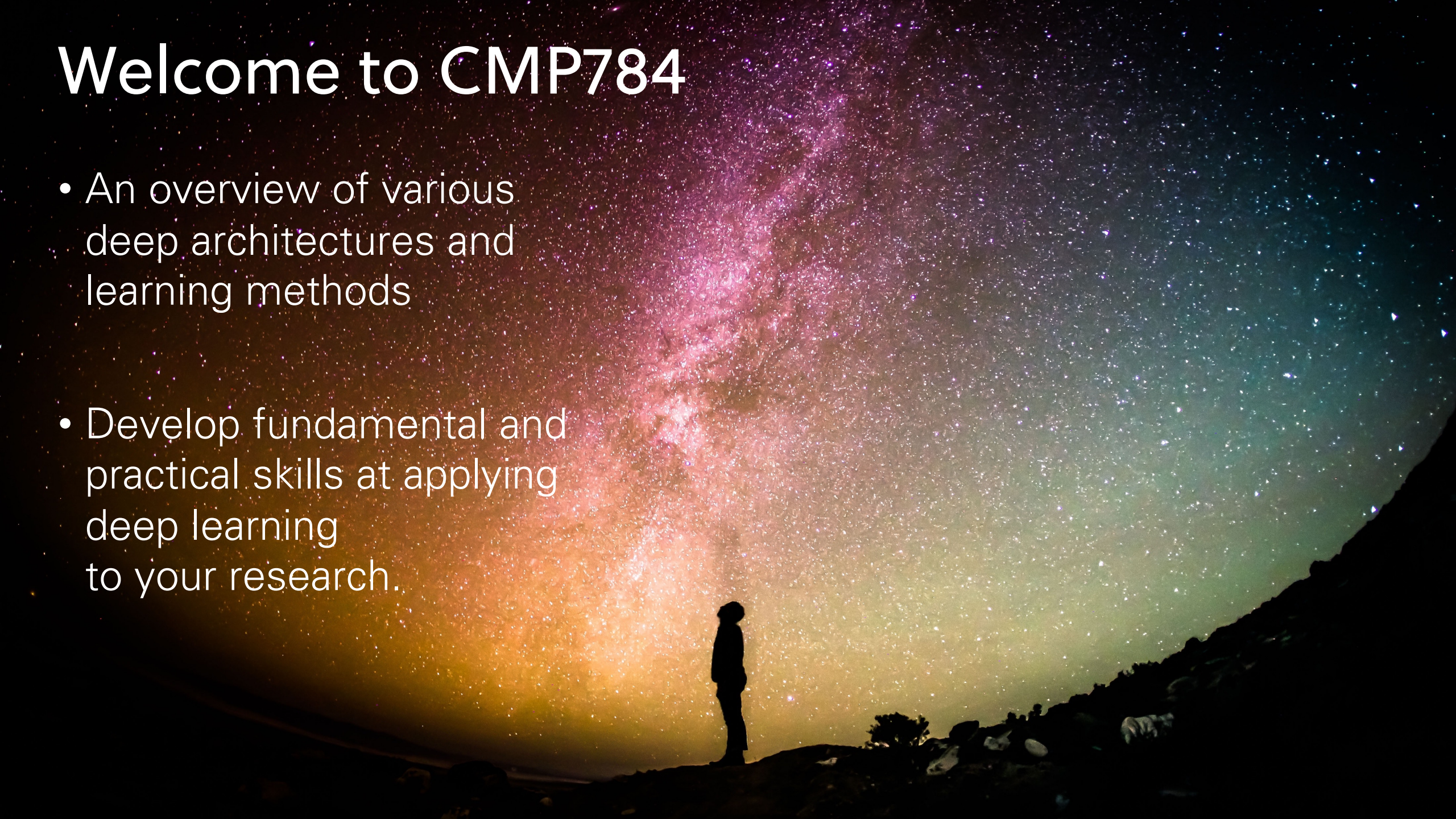


HACETTEPE
UNIVERSITY
COMPUTER
VISION LAB



Welcome to CMP784

- An overview of various deep architectures and learning methods
- Develop fundamental and practical skills at applying deep learning to your research.



A little about me...

Koç University-İş Bank
Artificial Intelligence Center
Adjunct Faculty
2020-now



Hacettepe University
Professor
2010-now



Télécom ParisTech
Post-doctoral Researcher
2009-2010



Middle East Technical University
1997-2008
Ph.D., 2008
M.Sc., 2003
B.Sc., 2001



UCLA
Fall 2007
Visiting Student



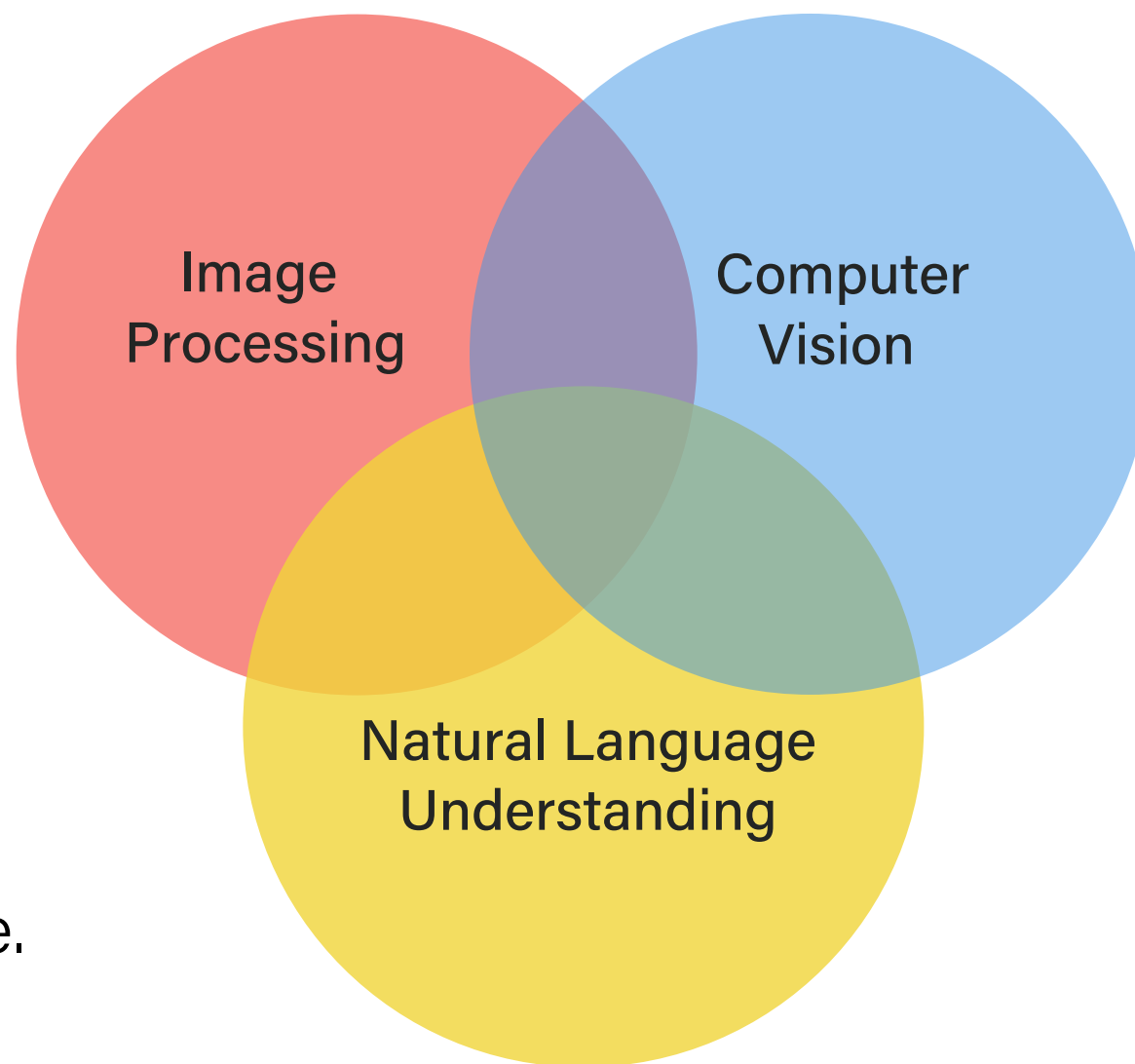
VirginiaTech
Visiting Research Scholar
Summer 2006



[@erkuterdem](http://web.cs.hacettepe.edu.tr/~erkut)
erkut@cs.hacettepe.edu.tr

Research Interests

- I study better ways to understand and process visual data.
- My research interests span a diverse set of topics, ranging from image editing to image enhancement, and to multimodal learning for integrated vision and language.



Now, what about you?

- Introduce yourselves
 - Who are you?
 - Who do you work with if you have a thesis supervisor?
 - What made you interested in this class?
 - What are your expectations?
 - What do you know about machine learning and deep learning?

Please send me an e-mail including these information!



Course Logistics

Course information

Time/Location 09:30-12:30pm Mondays, D5

Instructor Erkut Erdem

-  for course related announcements:

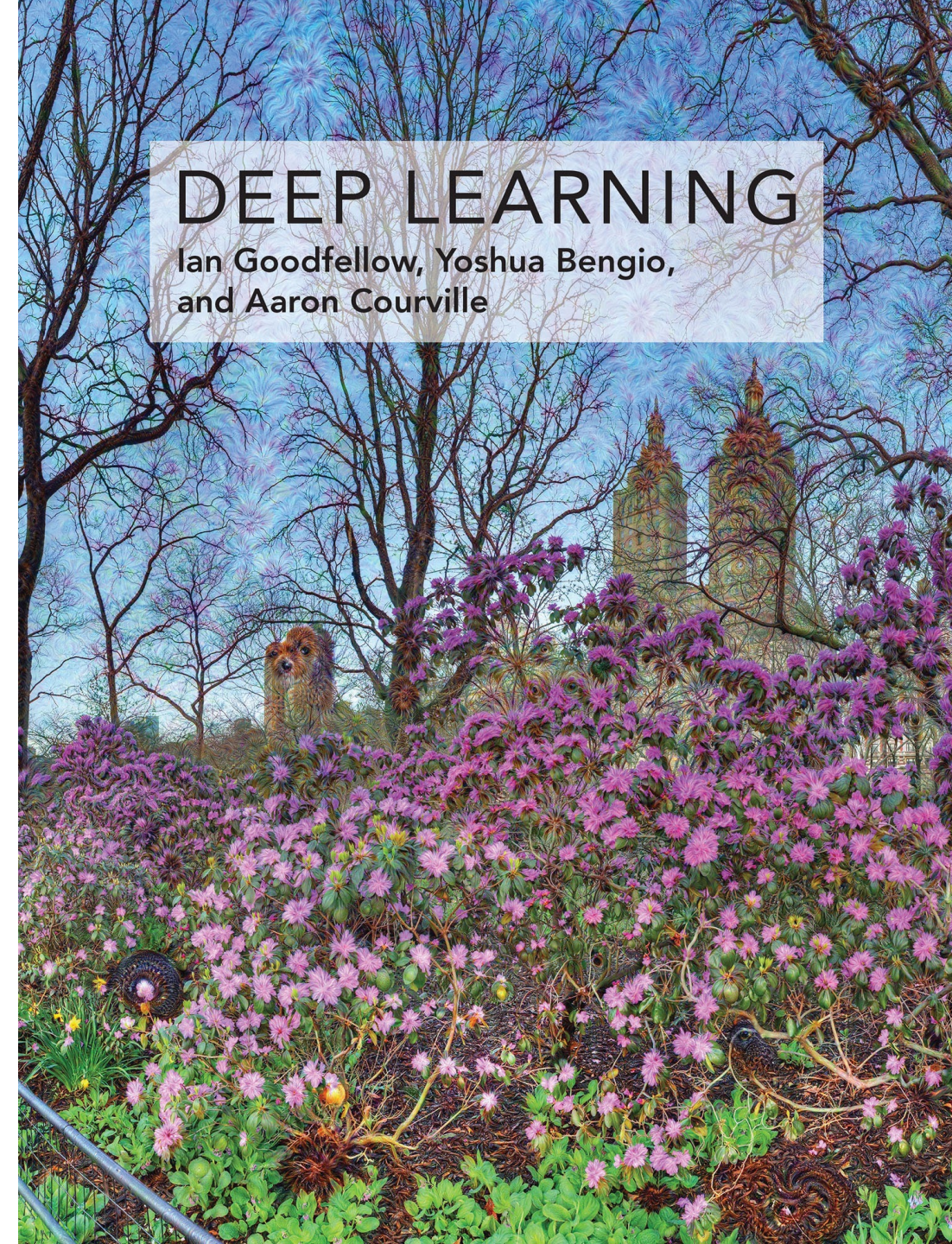
<https://edstem.org/eu/courses/3764>

- Course webpage:

<https://web.cs.hacettepe.edu.tr/~erkut/cmp784.f26/index.html>

Textbook

- Goodfellow, Bengio, and Courville, Deep Learning, MIT Press, 2016 (draft available [online](#))
- In addition, we will extensively use online materials (video lectures, blog posts, surveys, papers, etc.)



Instruction style

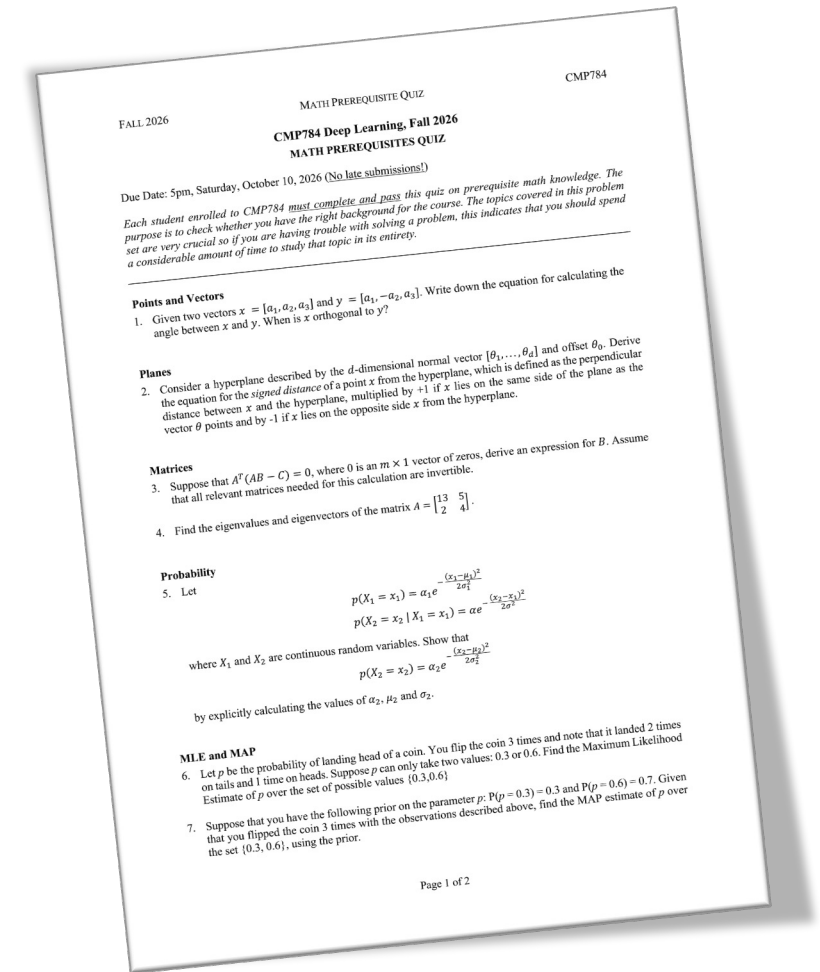
- Students are responsible for studying and keeping up with the course material outside of class time.
 - Reading particular book chapters, papers or blogs, or
 - Watching some video lectures.
- After the first part of the lectures, each week students will present papers related to the topics discussed in our class.
 - Weekly quizzes about the papers presented each week



Prerequisites

- Calculus and linear algebra
 - Derivatives,
 - Matrix operations
- Probability and statistics (IST299, IST292)
- Neural networks (CMP684)
- Machine learning (BBM406, CMP712)
- Programming

Read Chapter 2-4
of the Deep Learning text book for a quick review.



Math Prerequisite Quiz

Due Date: 5pm, Sat, Oct 10, 2026.

Each student enrolled to CMP784
must complete and pass this quiz!

Topics Covered in AIN311-BBM406/CMP712

- **Basics of Statistical Learning**

- Loss function, MLE, MAP, Bayesian estimation, bias-variance tradeoff, overfitting, regularization, cross-validation

- **Supervised Learning**

- Nearest Neighbor, Naïve Bayes, Logistic Regression, Support Vector Machines, Kernels, Neural Networks, Decision Trees
- Ensemble Methods: Bagging, Boosting, Random Forests

- **Unsupervised Learning**

- Clustering: K-Means, Gaussian mixture models
- Dimensionality reduction: PCA, SVD

Topics Covered in CMP684

- Continuous and discrete system models
- Neuron and Its Analytic Model
- Hopfiels Neural Network
- Perceptron Learning Algorithms
- Multilayer Perceptron (MLP)
 - Derivation of the learning algorithm
 - Error backpropagation
 - Memorization and generalization
 - Intervals and normalization
- Radial Basis Function Neural Nets
- Dynamical Neural Nets
- Feedback Nets
- Second Order Training Algorithms
 - Levenberg-Marquardt algorithm
 - Gauss-Newton algorithm
- Stability in Adaptive Systems
- Applications of Neural Nets

Grading

Math Prerequisites Quiz	2%
Practicals	16% (2 practicals x 8% each)
Final Exam	25%
Course Project	32%
Paper Presentations	15%
Weekly Quizzes	5%
Class participation	5%

Schedule

Week 1	Introduction to Deep Learning
Week 2	Machine Learning Overview
Week 3	Multi-Layer Perceptrons
Week 4	Training Deep Neural Networks
Week 5	Convolutional Neural Networks
Week 6	Understanding and Visualizing CNNs
Week 7	Recurrent Neural Networks
Week 8	Attention and Transformers

Schedule

Week 9 Autoencoders and Deep Generative Models

Week 10 Progress Presentations

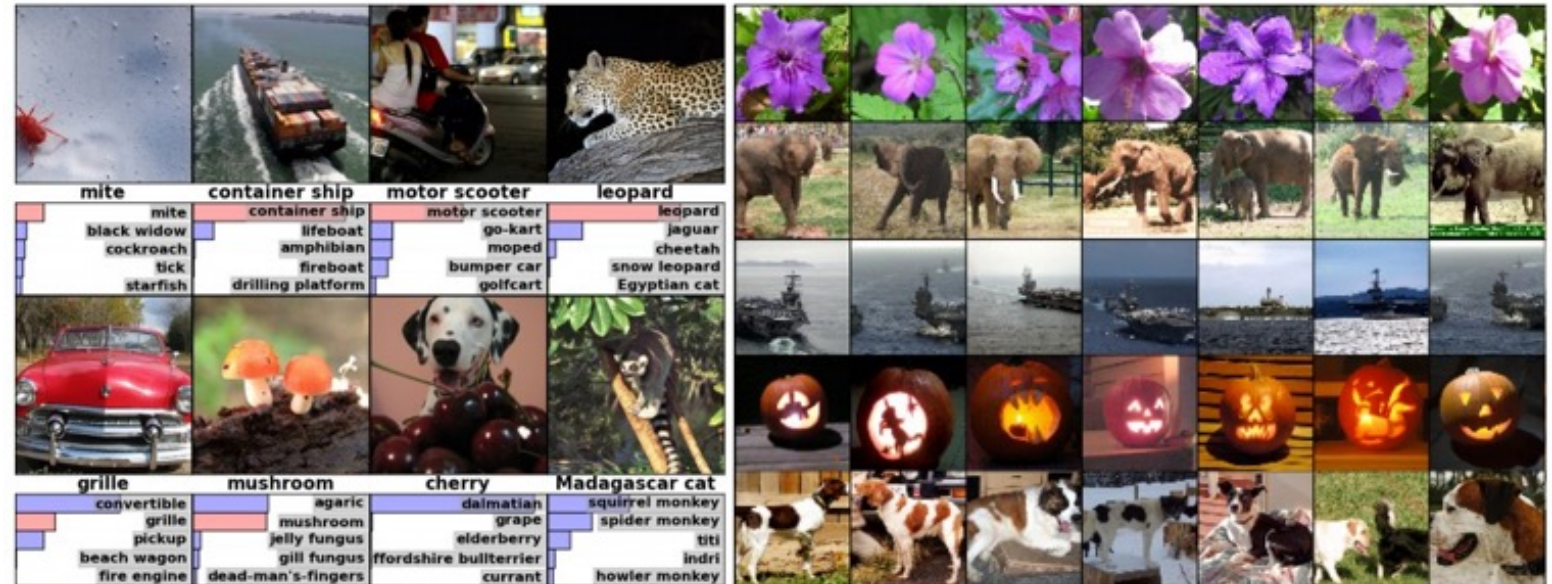
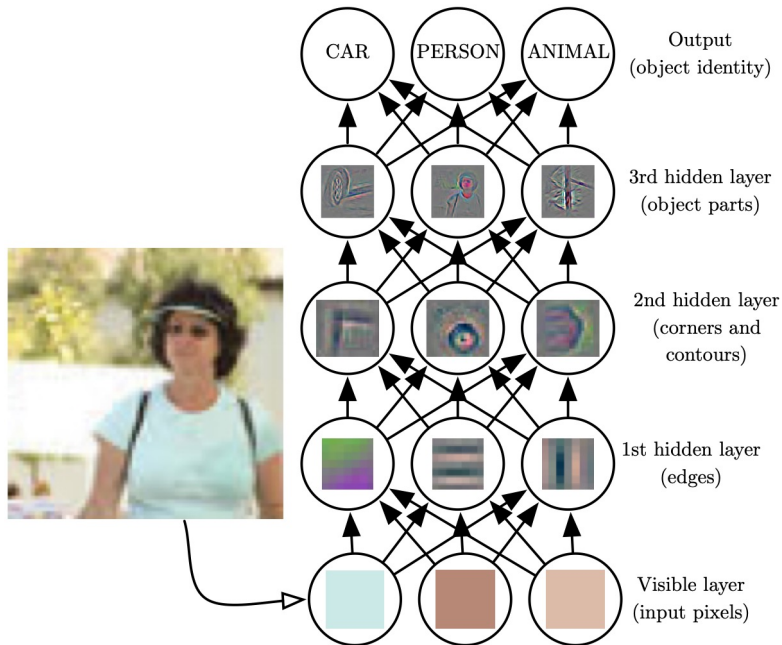
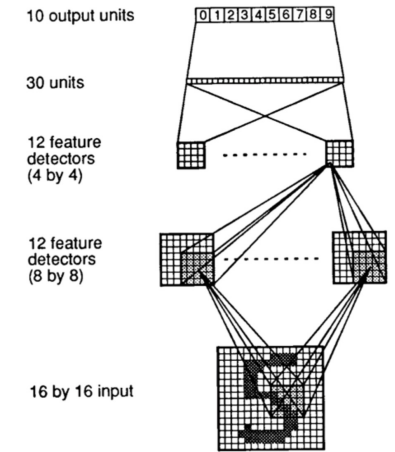
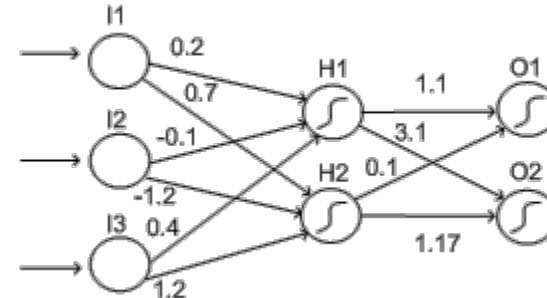
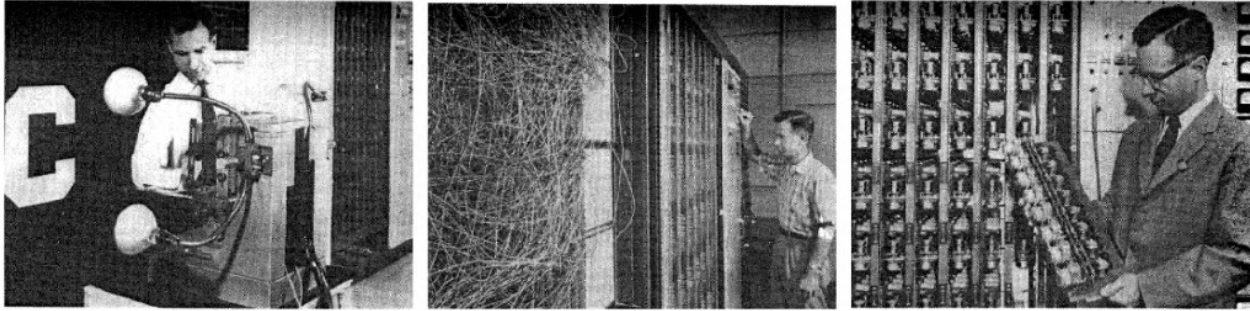
Week 11 Deep Generative Models (cont'd)

Week 12 Deep Generative Models (cont'd)

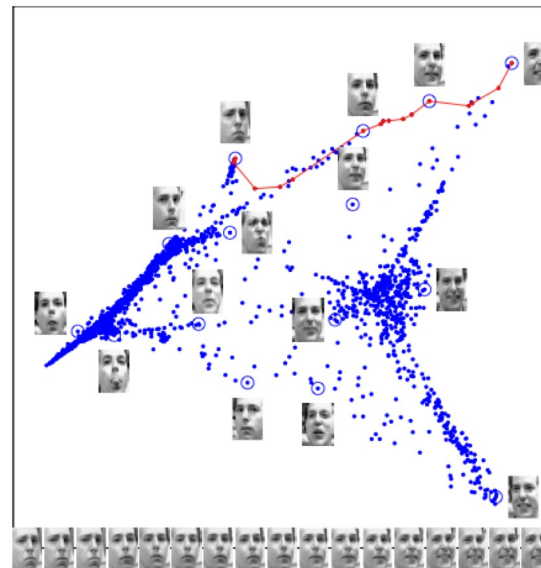
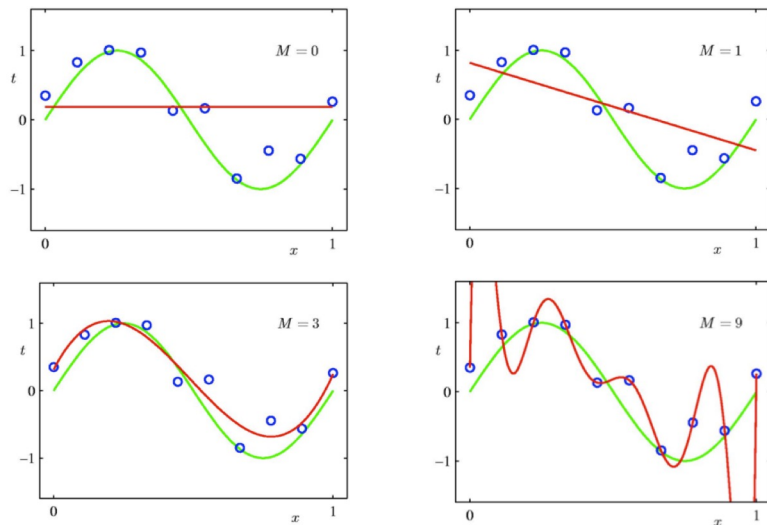
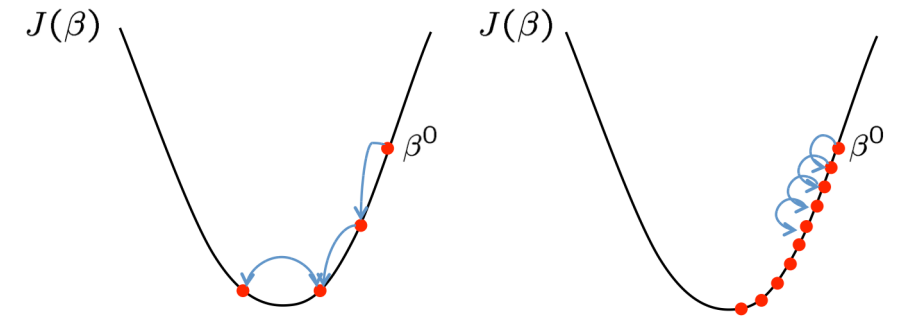
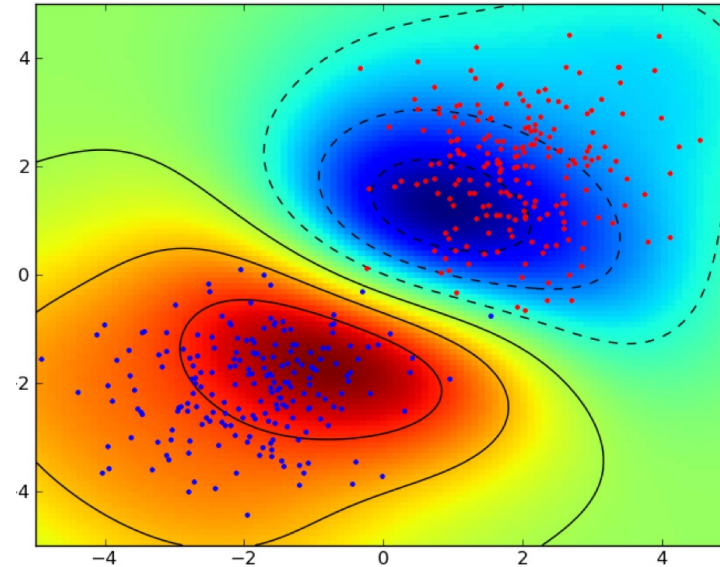
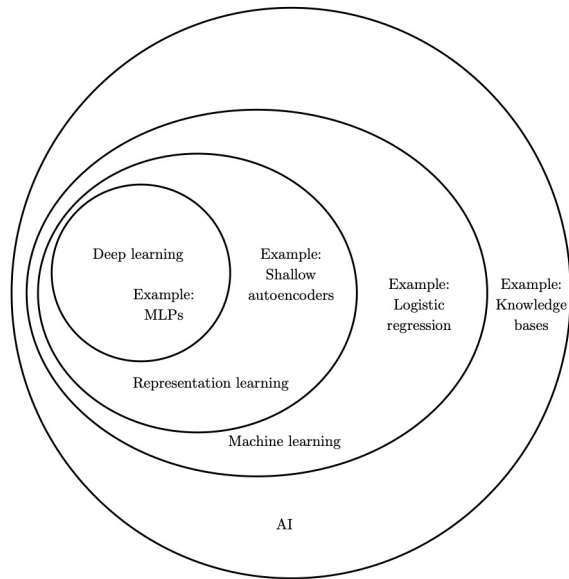
Week 13 Self-supervised Learning

Week 14 Final Project Presentations

Lecture 1: Introduction to Deep Learning

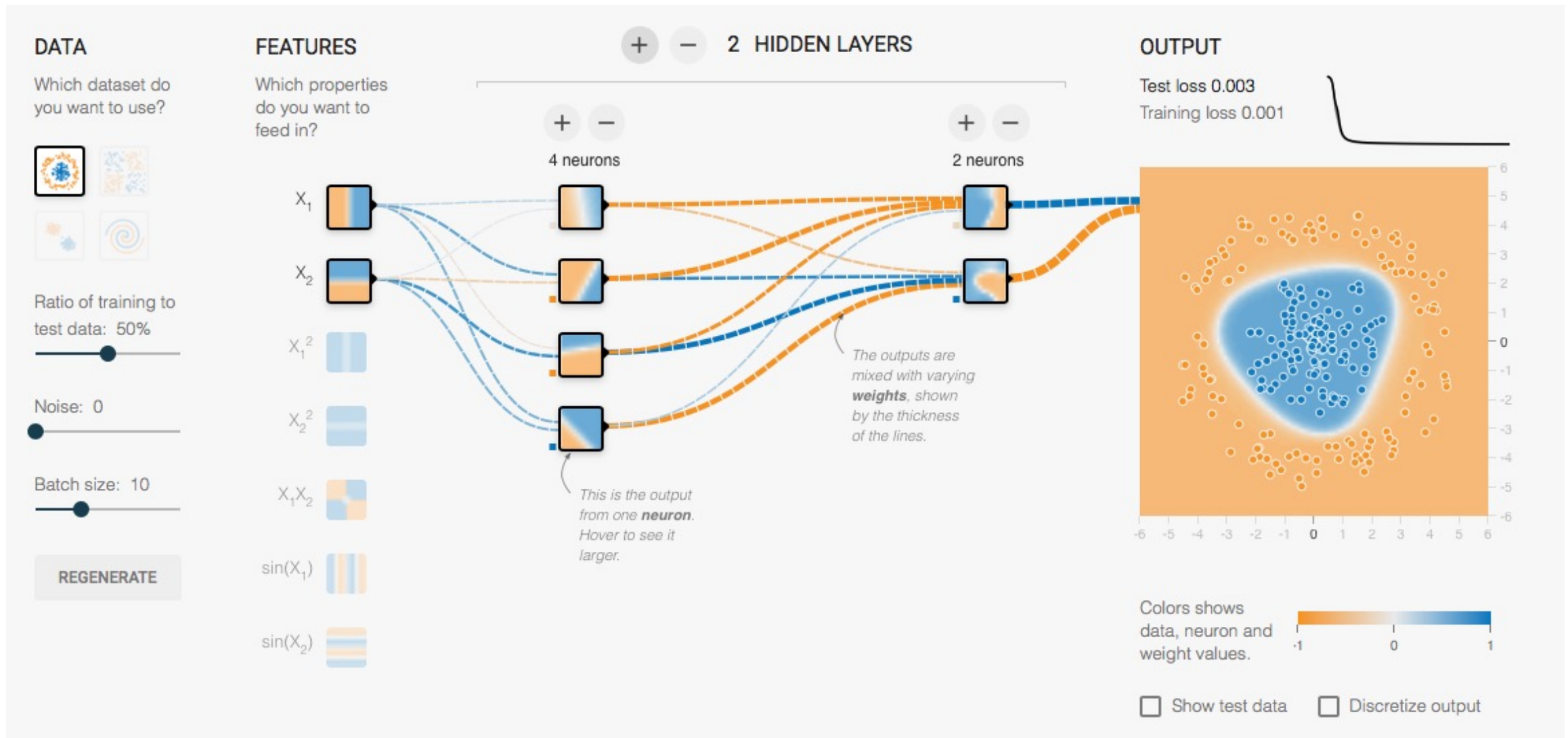


Lecture 2: Machine Learning Overview

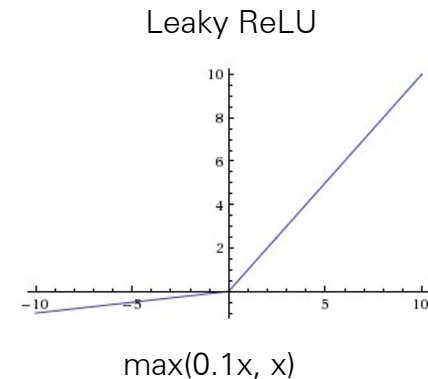
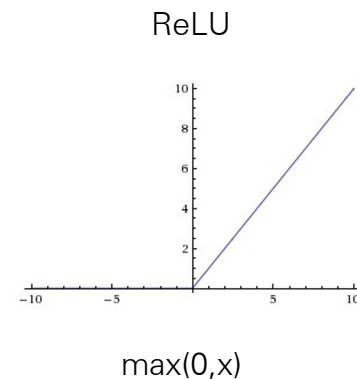
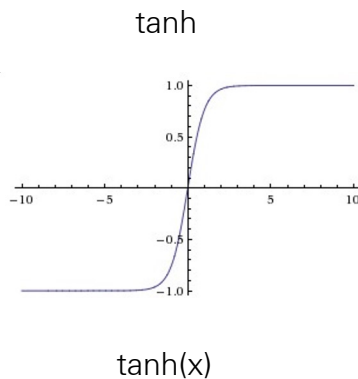
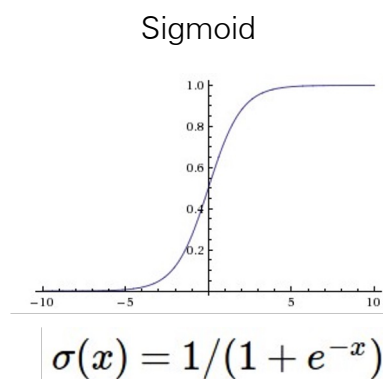
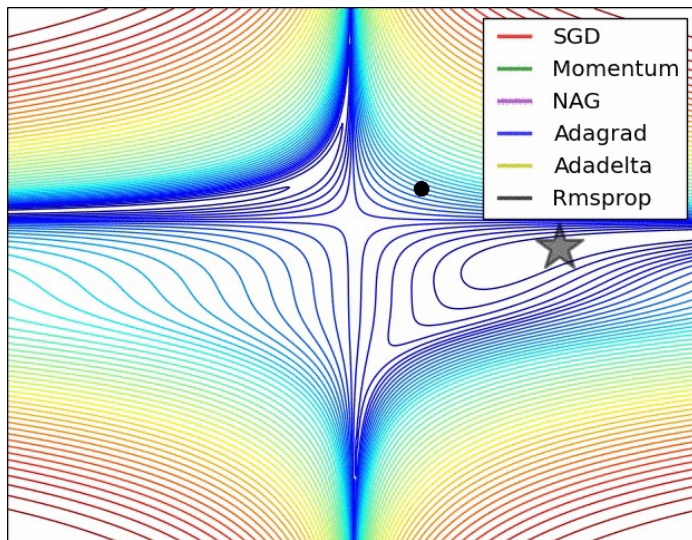


8	9	0	1	2	3	4	7	8	9	0	1	2	3	4	5	6	7	8	6
4	2	6	4	7	5	5	4	7	8	9	2	9	3	9	3	8	2	0	5
0	7	0	4	2	6	5	3	5	3	8	0	0	3	4	1	5	3	0	8
3	0	6	2	7	1	1	8	1	7	1	3	8	9	7	6	7	4	1	6
7	5	1	7	1	9	8	0	6	9	4	9	9	3	7	1	9	2	2	5
3	7	8	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	0
1	2	3	4	5	6	7	8	9	8	1	0	5	5	1	9	0	4	1	9
3	8	4	7	7	8	5	0	6	5	5	3	3	3	9	8	1	4	0	6
1	0	0	6	2	1	1	3	2	8	8	7	8	4	6	0	2	0	3	6
8	7	1	5	9	9	3	2	4	9	4	6	5	3	2	8	5	9	4	1
6	5	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7
8	9	0	1	2	3	4	5	6	7	8	9	6	4	2	6	4	7	5	5
4	7	8	9	2	9	3	9	3	8	2	0	9	8	0	5	6	0	1	0
4	2	6	5	5	5	4	3	4	1	5	3	0	8	3	0	6	2	7	1
1	8	1	7	1	3	8	5	4	2	0	9	7	6	7	4	1	6	8	4
7	5	1	2	6	7	1	9	8	0	6	9	4	9	9	6	2	3	7	1
9	2	2	5	3	7	8	0	1	2	3	4	5	6	7	8	0	1	2	3
4	5	6	7	8	0	1	2	3	4	5	6	7	8	9	2	1	2	1	3
9	9	8	5	3	7	0	7	7	5	7	9	9	4	7	0	3	4	1	4
4	7	5	8	1	4	8	4	1	8	6	6	4	6	3	5	7	2	5	9

Lecture 3: Multi-Layer Perceptrons

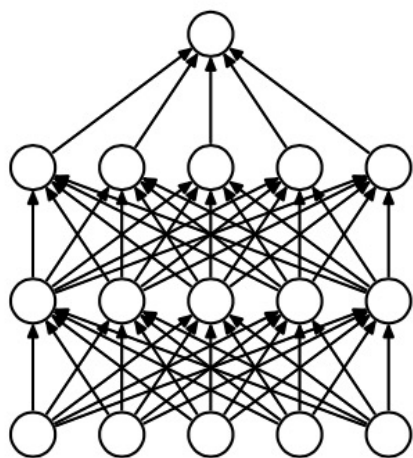


Lecture 4: Training Deep Neural Networks

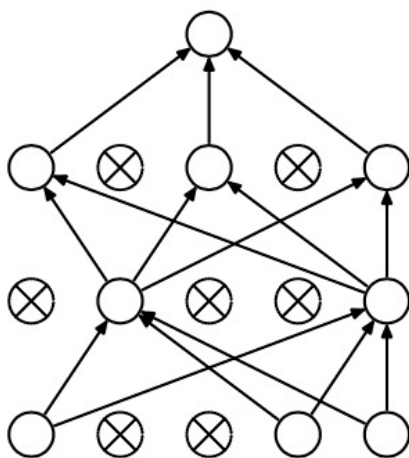


Activation Functions

Optimizers



(a) Standard Neural Net



(b) After applying dropout.

Dropout

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_1 \dots x_m\}$;

Parameters to be learned: γ, β

Output: $\{y_i = \text{BN}_{\gamma, \beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m} \sum_{i=1}^m x_i \quad // \text{ mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{\mathcal{B}})^2 \quad // \text{ mini-batch variance}$$

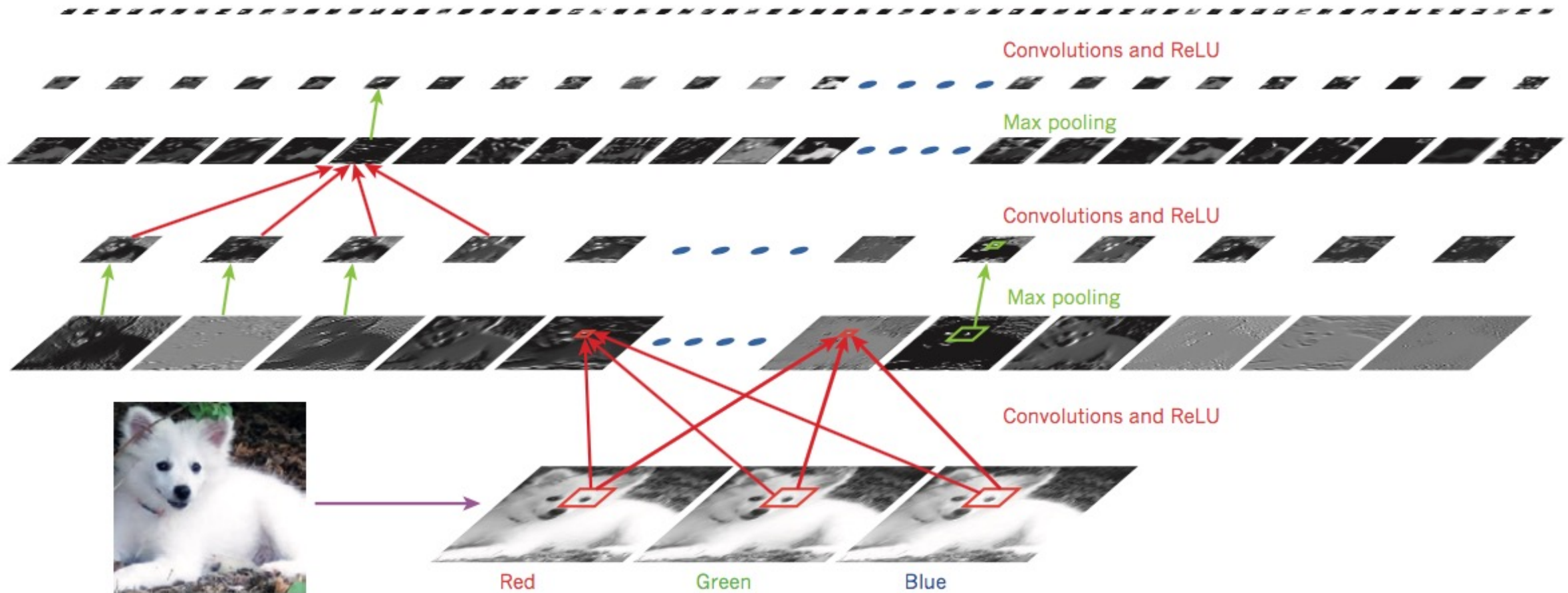
$$\hat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \quad // \text{ normalize}$$

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma, \beta}(x_i) \quad // \text{ scale and shift}$$

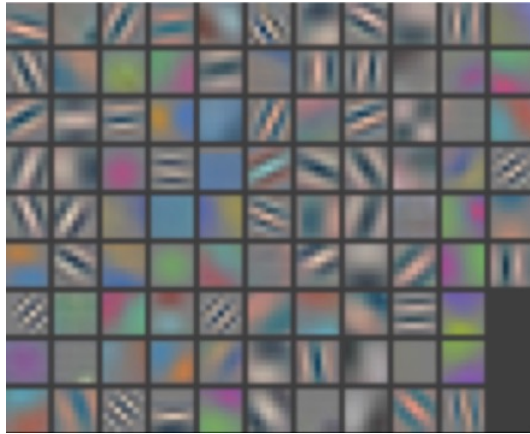
Batch Normalization

Lecture 5: Convolutional Neural Networks

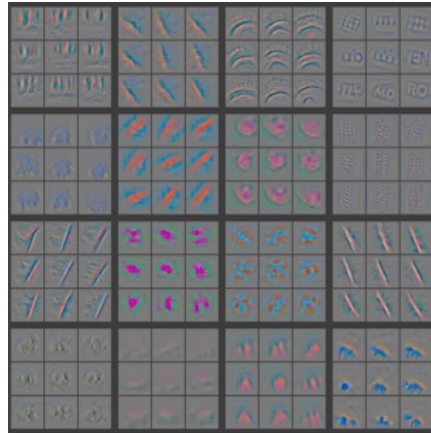
Samoyed (16); Papillon (5.7); Pomeranian (2.7); Arctic fox (1.0); Eskimo dog (0.6); white wolf (0.4); Siberian husky (0.4)



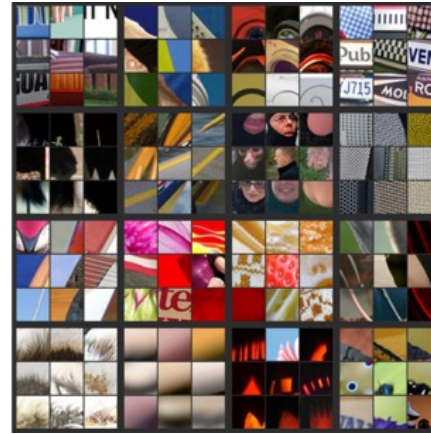
Lecture 6: Understanding and Visualizing CNNs



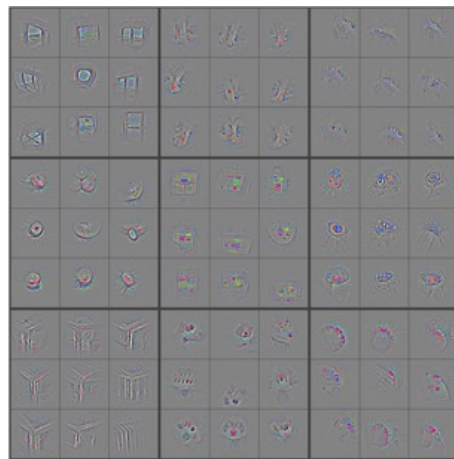
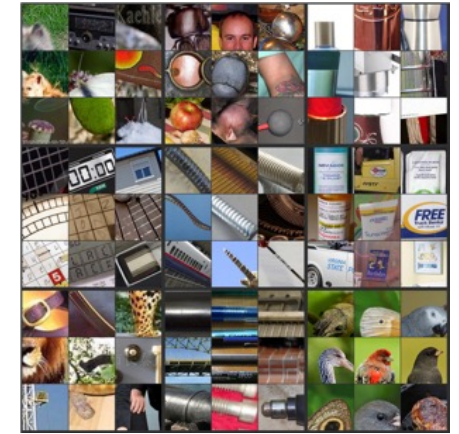
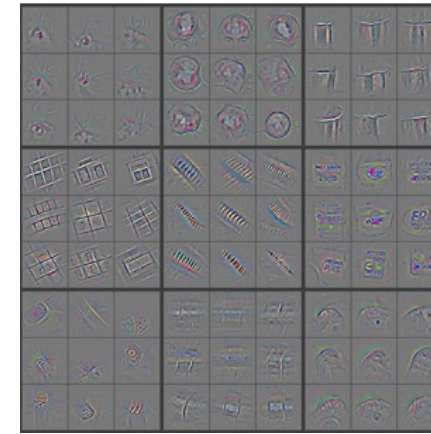
Layer 1



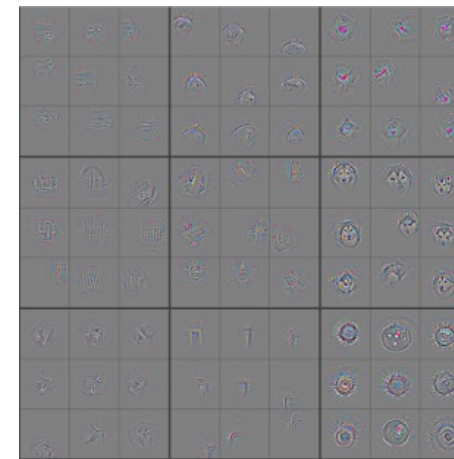
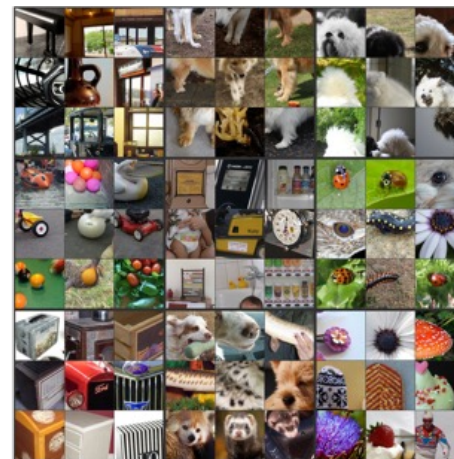
Layer 2



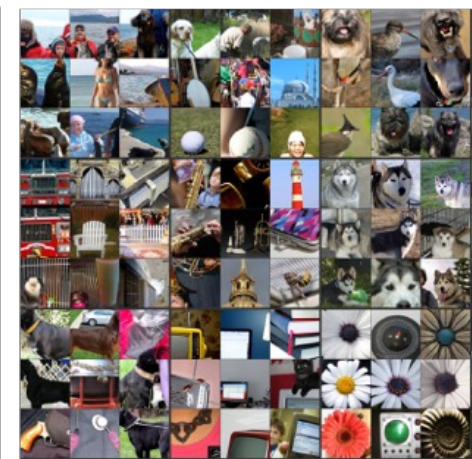
Layer 3



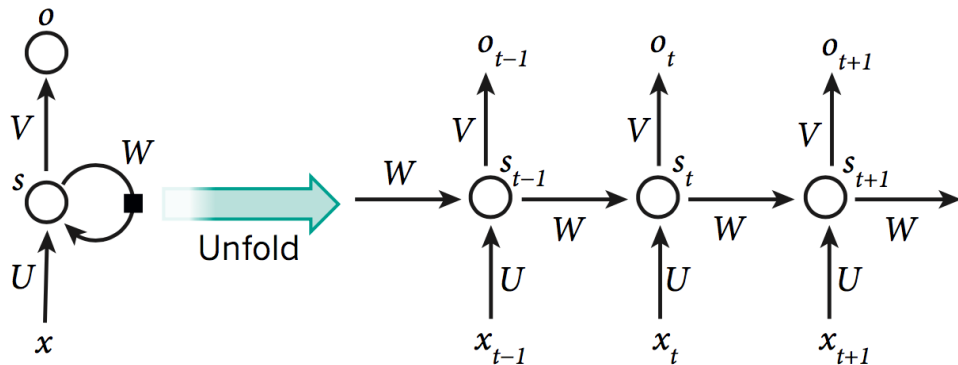
Layer 4



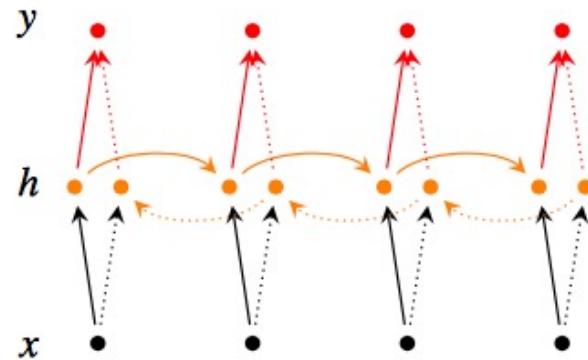
Layer 5



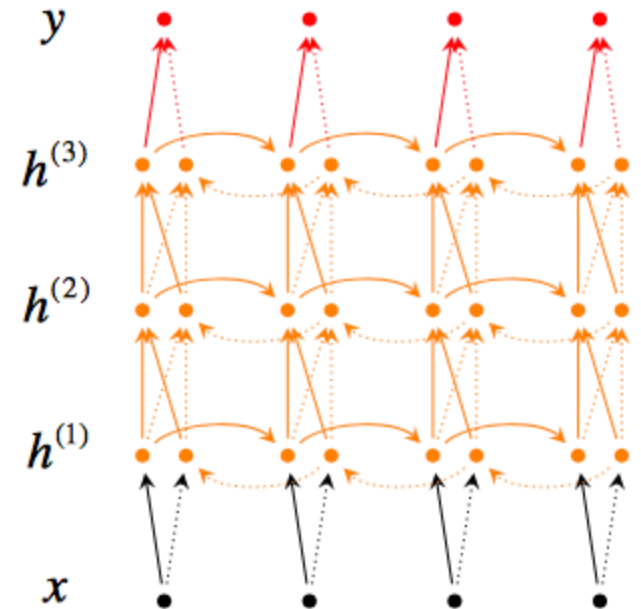
Lecture 7: Recurrent Neural Networks



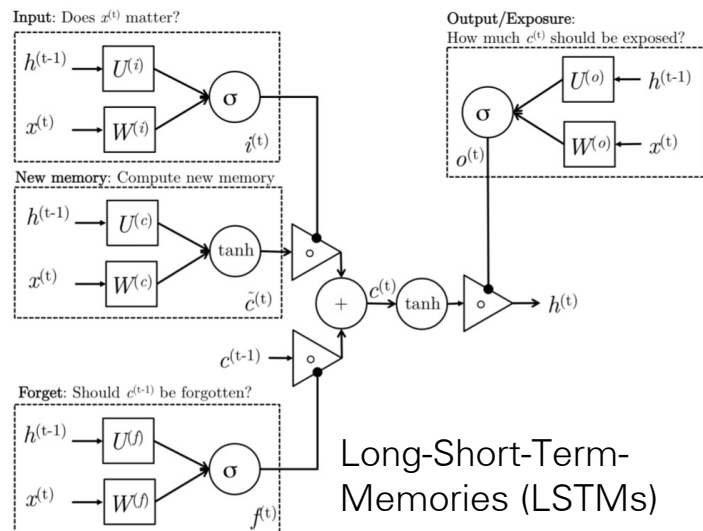
A Recurrent Neural Network (RNN)
(unfolded across time-steps)



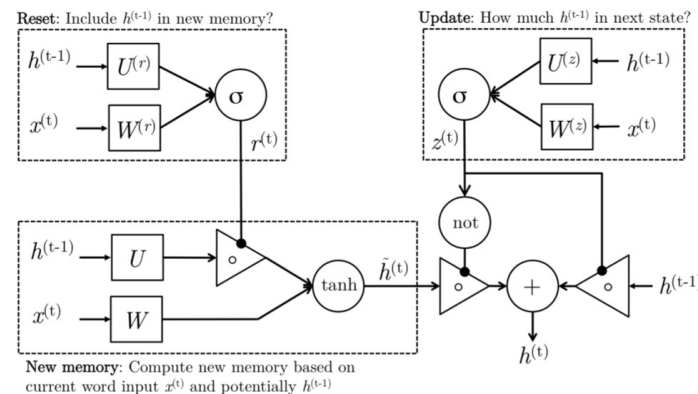
A bi-directional RNN



A deep bi-directional RNN



Long-Short-Term-Memories (LSTMs)



Gated Recurrent Units (GRUs)

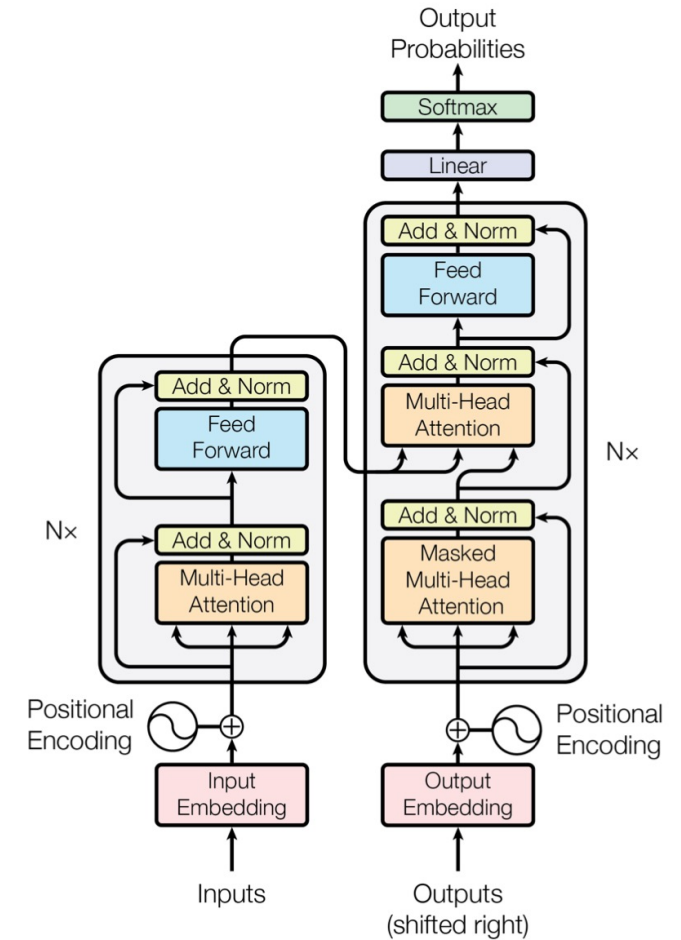
Lecture 8: Attention and Transformers



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.



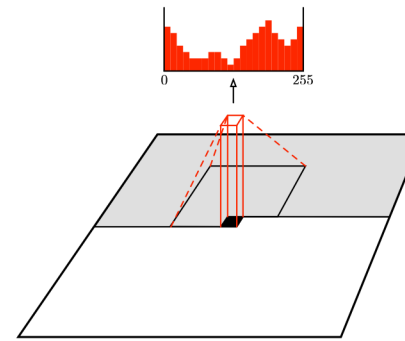
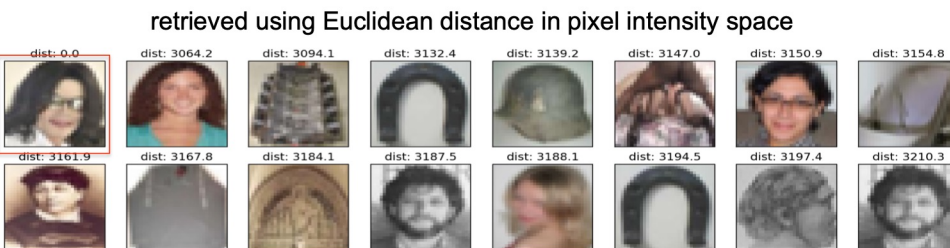
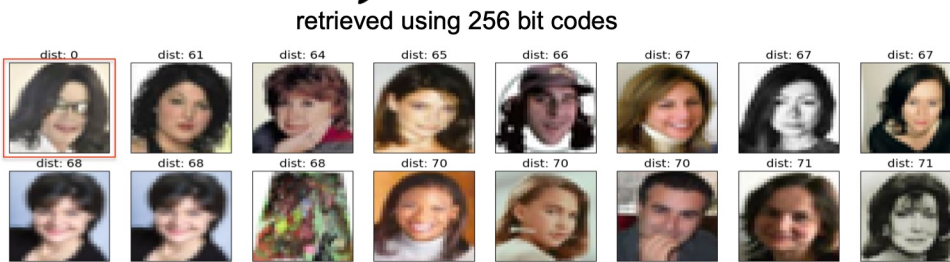
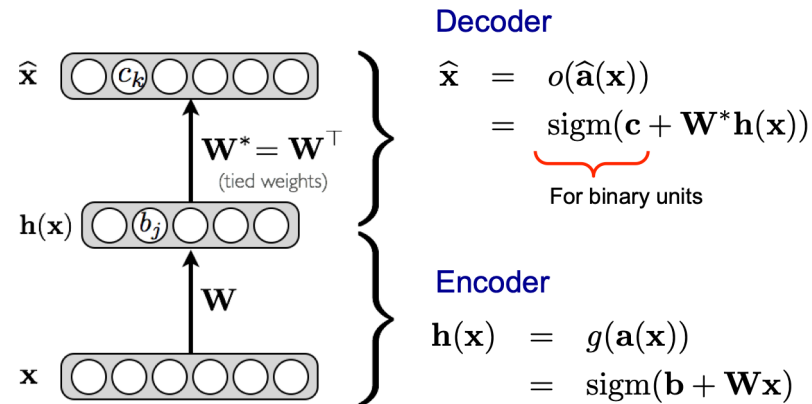
Transformer Architecture

K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015

C. Olah and S. Carter, "Attention and Augmented Recurrent Neural Networks", Distill, 2016

A. Vaswani et al. "Attention is All You Need", NeurIPS 2017.

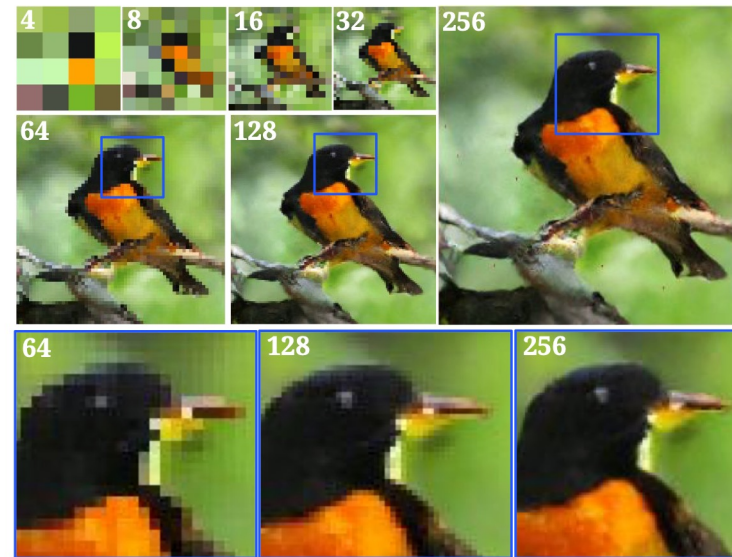
Lecture 9: Autoencoders and Deep Generative Models



PixelCNN



Class conditioned samples generated by PixelCNN

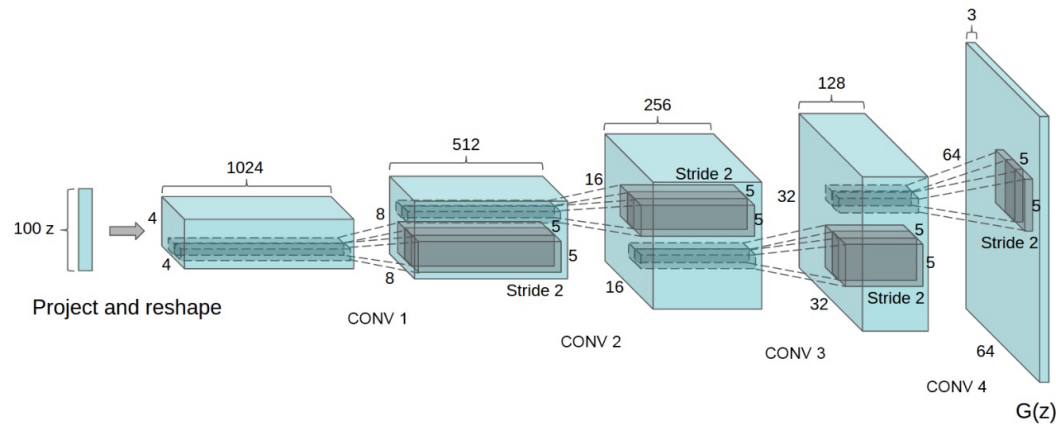


Text-to-image synthesis with Parallel Multiscale PixelCNNs

"A yellow bird with a black head, orange eyes and an orange bill."

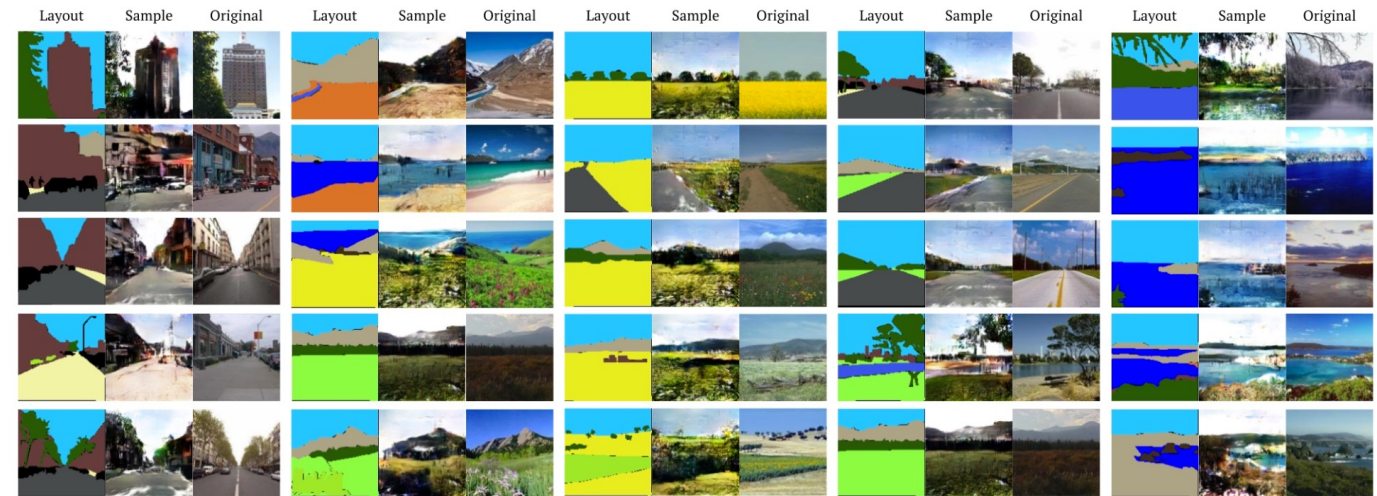
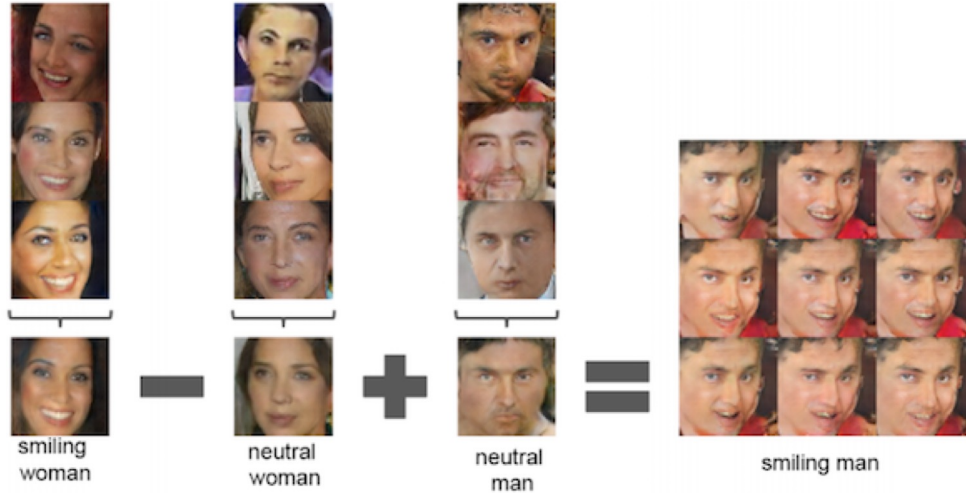
- A. Krizhevsky and G. E. Hinton, "Using Very Deep Autoencoders for Content-Based Image Retrieval", ESANN 2011
- A. van den Oord et al., "Conditional Image Generation with PixelCNN Decoders", NeurIPS 2016
- S. Reed et al., "Parallel Multiscale Autoregressive Density Estimation", ICML 2017

Lecture 10: Deep Generative Models (cont'd)



Class-conditioned samples generated by BigGAN

$$\min_{\theta} \max_{\omega} \mathbb{E}_{x \sim Q} [\log D_{\omega}(x)] + \mathbb{E}_{x \sim P_{\theta}} [\log(1 - D_{\omega}(x))]$$



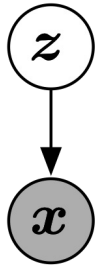
I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", NIPS 2014.

A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks", ICLR 2016

L. Karacan, Z. Akata, A. Erdem and E. Erdem, "Learning to Generate Images of Outdoor Scenes from Attributes and Semantic Layouts", arXiv preprint 2016

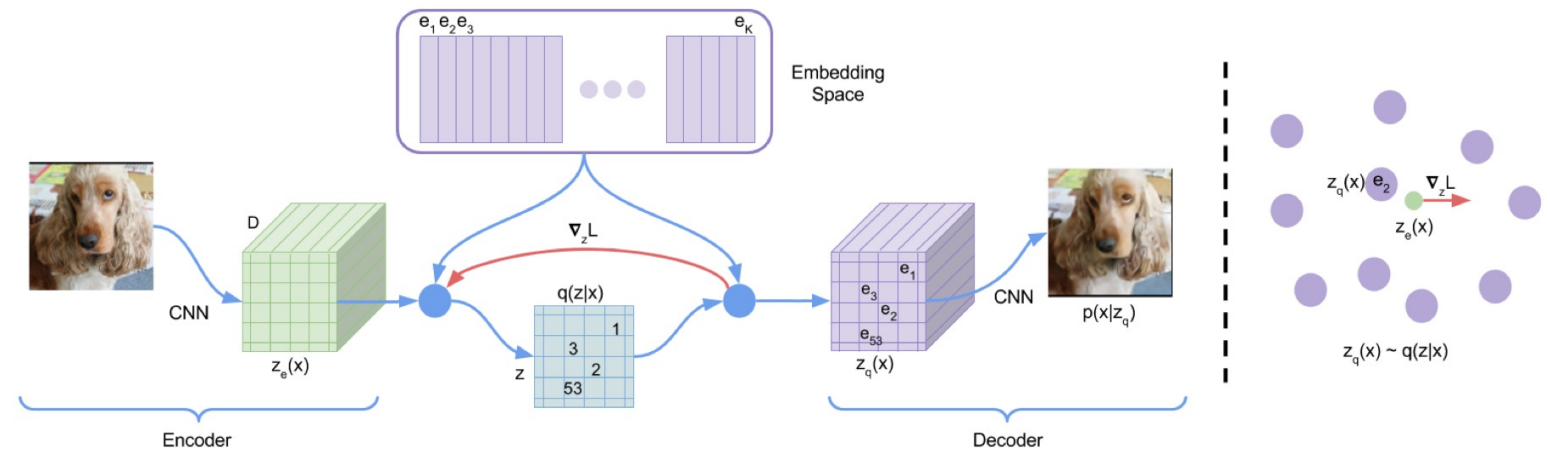
A. Brock, J. Donahue, K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis", ICLR2019

Lecture 11: Deep Generative Models (cont'd)



$$\log p(\mathbf{x}) \geq \log p(\mathbf{x}) - D_{\text{KL}}(q(\mathbf{z}) \| p(\mathbf{z} | \mathbf{x}))$$

$$= \mathbb{E}_{\mathbf{z} \sim q} \log p(\mathbf{x}, \mathbf{z}) + H(q)$$



Vector Quantized- Variational AutoEncoder (VQ-VAE)



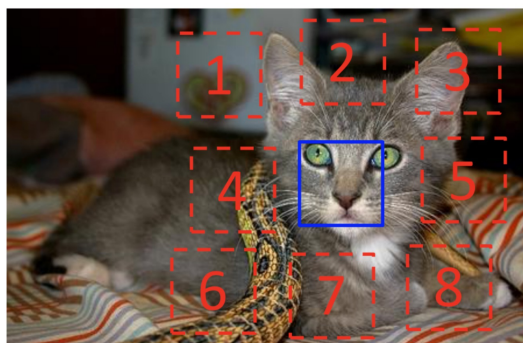
Synthetic images generated by VQ-VAE2

D. P. Kingma and M. Welling, "Auto-encoding variational Bayes", ICLR 2014

A. van den Oord, O. Vinyals, K. Kavukcuoglu, "Neural Discrete Representation Learning", NeurIPS 2017

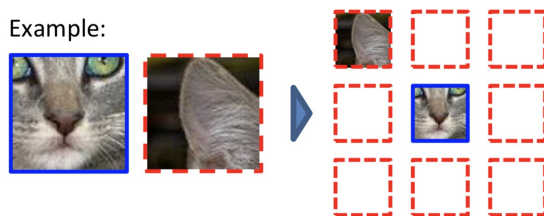
A. Razavi, A. van den Oord, O. Vinyals, "Generating Diverse High-Fidelity Images with VQ-VAE-2",

Lecture 12: Self-supervised Learning



$$X = (\text{cat face}, \text{cat ear}); Y = 3$$

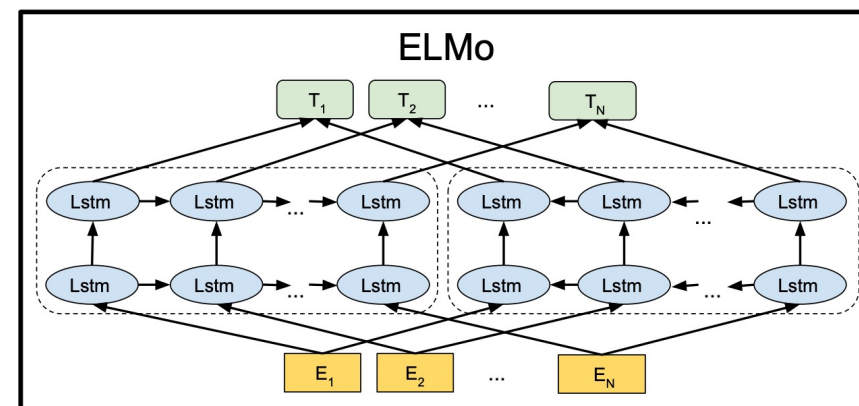
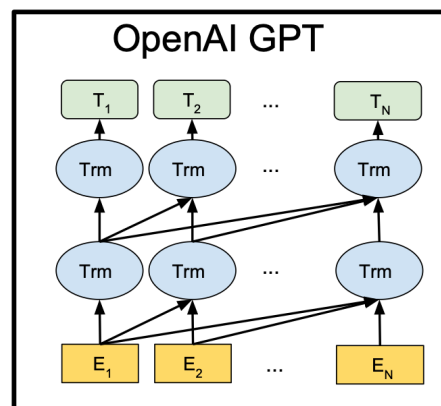
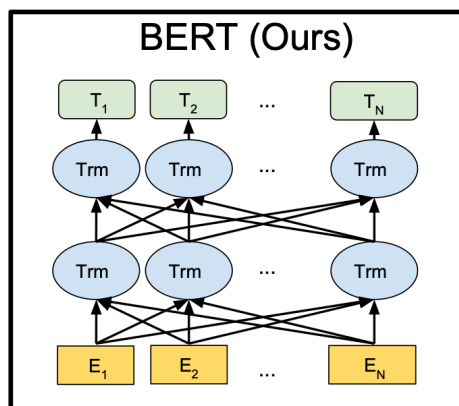
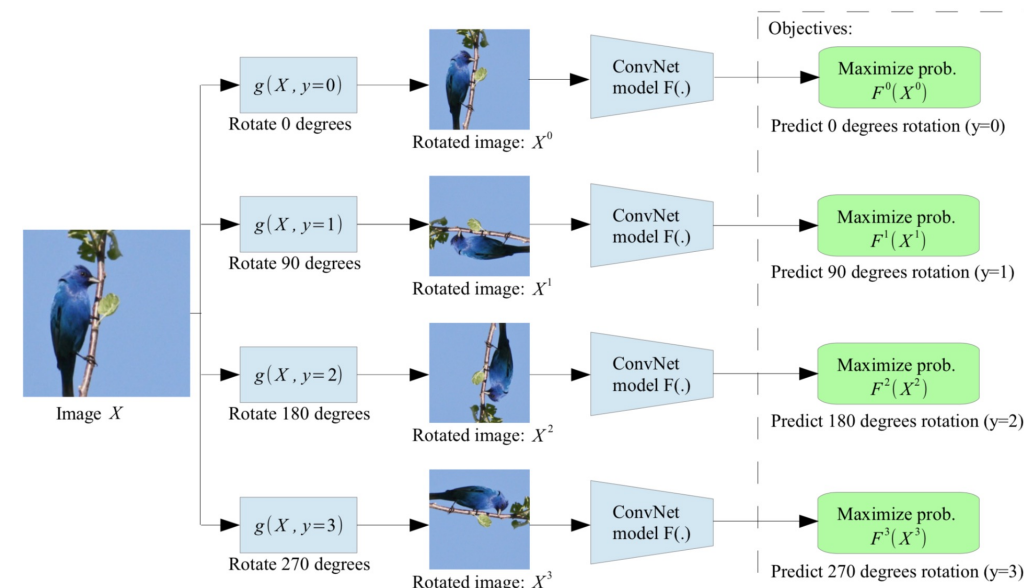
Example:



Question 1:



Question 2:



C. Doersch, A. Gupta, A. A. Efros, "Unsupervised Visual Representation Learning by Context Prediction", ICCV 2015.

S. Gidaris, P. Singh, N. Komodakis, "Unsupervised Representation Learning by Predicting Image Rotations", ICLR2018.

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", NAACL-HLT 2019.

Schedule

W1 Introduction to Deep Learning

W2 Machine Learning Overview

W3 Multi-Layer Perceptrons

Practical 1 out

W4 Training Deep Neural Networks

W5 Convolutional Neural Networks

Start of paper presentations

W6 Understanding and Visualizing CNNs

Practical 1 due, Practical 2 out

Project proposals due

W7 Recurrent Neural Networks

W8 Attention and Transformers

W9 Autoencoders and Deep Generative Models

Practical 2 due

W10 Progress Presentations

W11 Deep Generative Models (cont'd)

Project progress reports due

W12 Deep Generative Models (cont'd)

W13 Self-supervised Learning

W14 Final Project Presentations

Paper Presentations

- (12 mins) One student will be responsible from providing an overview of the paper.
- (9 mins) One student will present the strengths of the paper.
- (9 mins) One student will discuss the weaknesses of the paper.
- (10 mins) General discussion

See the rubrics on the course web page for details

Practicals

- 2 practicals (8% each)
- Learning to train neural networks for different tasks
- Should be done individually
- **Late policy:** You have 5 slip days in the semester.
- **Tentative Dates**
 - Practical 1 Out: October 5th, Due: October 26th
 - Practical 2 Out: October 26th, Due: November 16th

Course project

The students who need GPU resources for the course project are advised to use Google Colab.

- The course project gives students a chance to apply deep architectures discussed in class to a research oriented project.
- The students can work in pairs.
- The course project may involve
 - Design of a novel approach and its experimental analysis, or
 - An extension to a recent study of non-trivial complexity and its experimental analysis.
- Deliverables
 - Proposals October 26, 2026
 - Project progress presentations November 23, 2026
 - Project progress reports November 30, 2026
 - Final project presentations December 21, 2026
 - Final reports December 31, 2026

Lecture Overview

- what is deep learning
- a brief history of deep learning
- compositionality
- end-to-end learning
- distributed representations

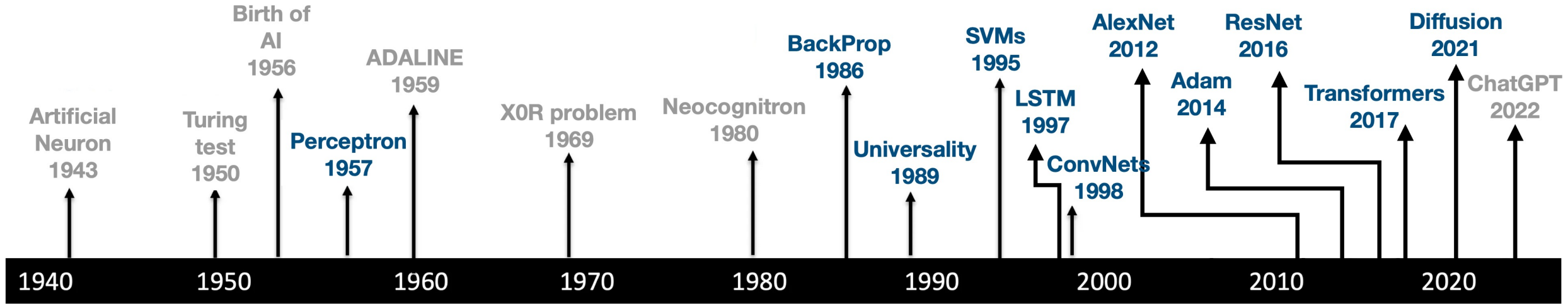
Disclaimer: Some of the material and slides for this lecture were borrowed from

—Dhruv Batra's CS7643 class

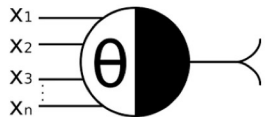
—Yann LeCun's talk titled "Deep Learning and the Future of AI"

What is Deep Learning

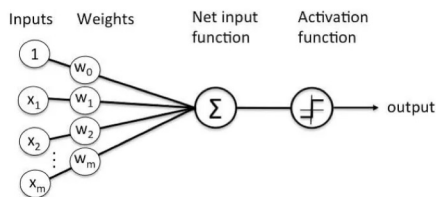
The history of (deep) learning



McCulloch-Pitts

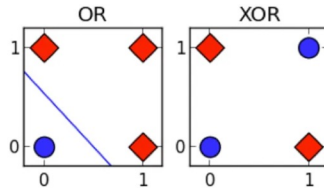


Rosenblatt

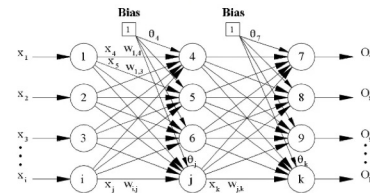


Widrow-Hoff

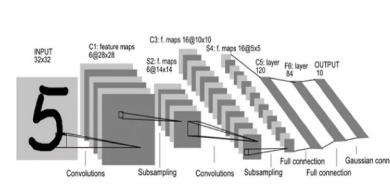
Minsky-Papert



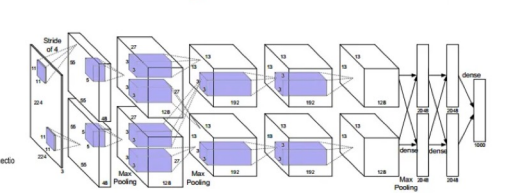
Rumelhart, Hinton et al.



LeCun



Hinton-Ruslan Krizhevsky et al.



Vaswani

What Is Deep Lea



Kevin Murnane
CONTRIBUTOR

I write about science, technology and the people that connect them.



FULL BIO >

Opinions expressed by Forbes Contributors are their own.

TWEET THIS

Deep learning uni
to use it

Credit: Google

Deep learning re
GOOGL +1.40% Alpha
ranking Go playe
learning and Alpl
the news. Google
driving cars all rely
networks to build a program that picks out an attractive still from a
YouTube video or a bunch of 100 pictures of a cat and then
understand language and then make inferences and decisions on its

Science

AAAS

Authors | Members | Librarians | Advertisers

Home News Journals Topics Careers

Search

Science Science Advances Science Immunology Science Robotics Science Signaling Science Translational Medicine

How AI is transforming science

Researchers are unleashing artificial intelligence (AI) on torrents of big data

KIYOSHI TAKAHASE SEGUNDO/ALAMY STOCK PHOTO

Contents 07 JULY 2017 VOL 357, ISSUE 6346

Special Issue The cyberscientist

INTRODUCTION TO SPECIAL ISSUE

The scientists' apprentice

BY TIM APPENZELLER

SCIENCE | 07 JUL 2017 | 16-17 |

Artificial intelligence helps scientists cope with torrents of data

Summary Full Text PDF

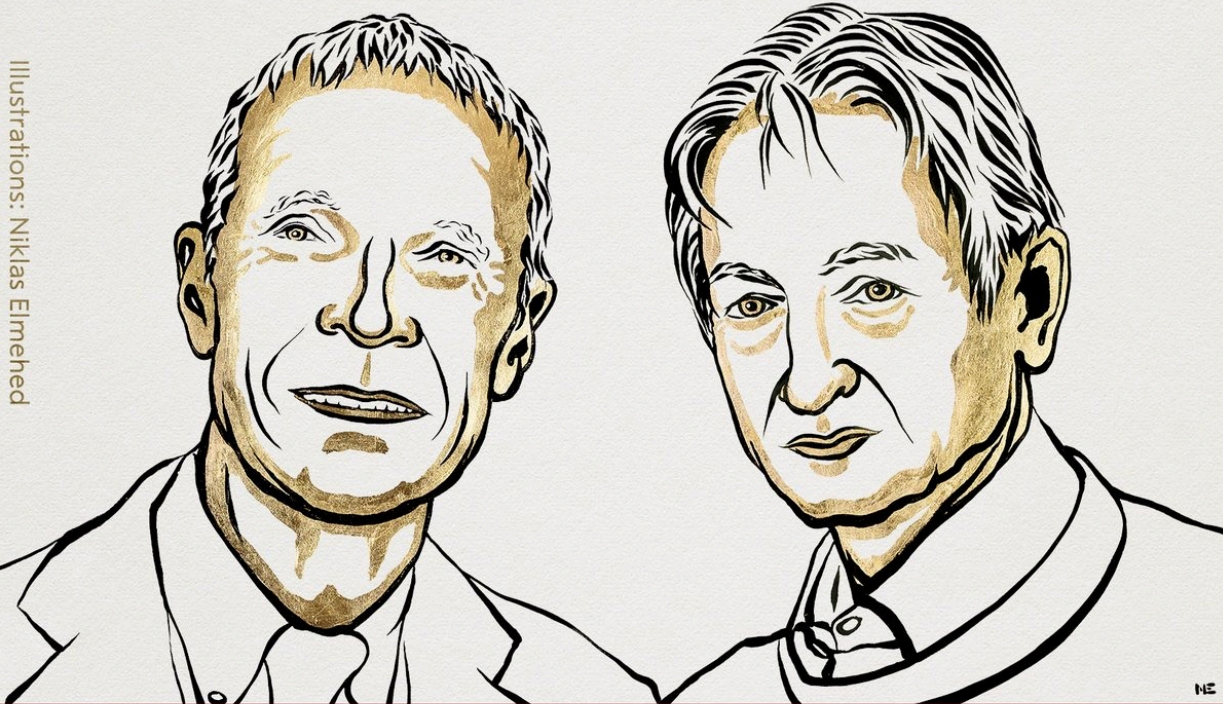
MORE FROM SCIENCE

- Current Table of Contents
- First Release Science Papers
- Archive
- Collections
- Book and Media Reviews
- About Science
 - Mission and Scope
 - Editors and Advisory Boards
 - Editorial Policies
 - Information for Authors
 - Information for Reviewers
- Staff

SION IN AI. INFO WAS WHERE FACEBOOK CHIEF EXECUTIVE OFFICER MARK ZUCKERBERG
in 2013 to announce the company's plans to form an AI laboratory and where a
startup named DeepMind showed off an AI that could learn to play computer games
before it was acquired by Google.

Nobel Prizes in Physics and Chemistry 2024

THE NOBEL PRIZE IN PHYSICS 2024

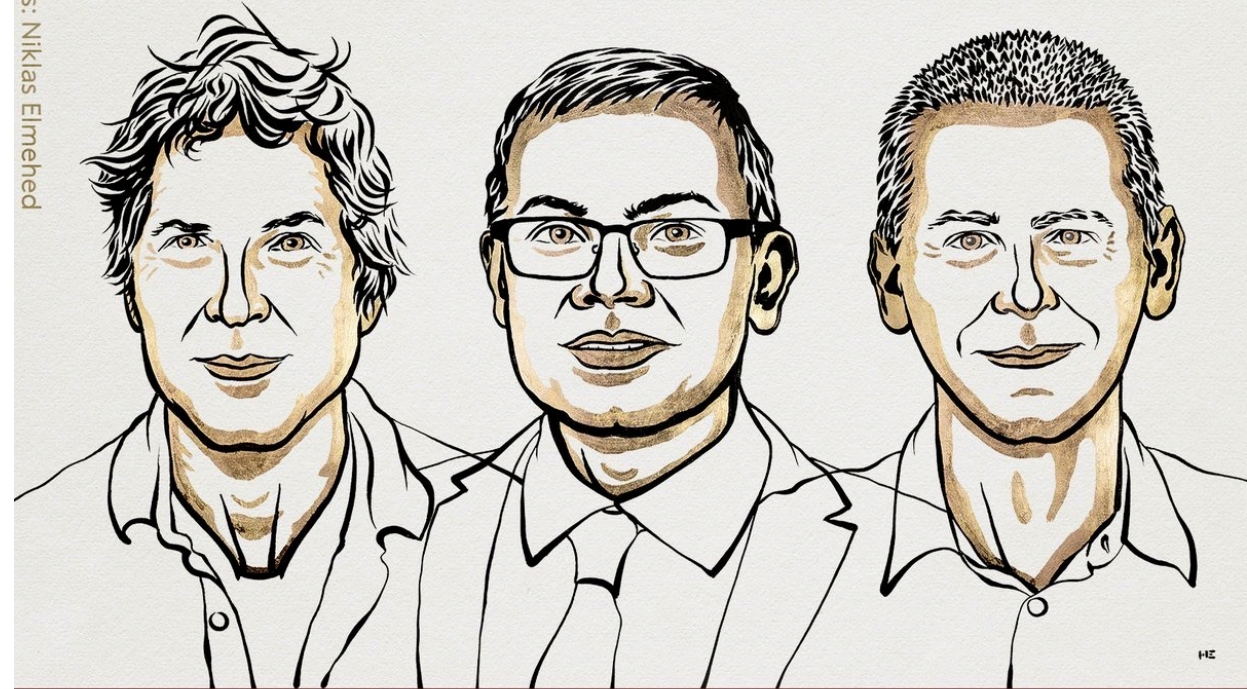


John J. Hopfield Geoffrey E. Hinton

"for foundational discoveries and inventions
that enable machine learning
with artificial neural networks"

THE ROYAL SWEDISH ACADEMY OF SCIENCES

THE NOBEL PRIZE IN CHEMISTRY 2024



**David
Baker**

"for computational
protein design"

**Demis
Hassabis**

"for protein structure prediction"

**John M.
Jumper**

THE ROYAL SWEDISH ACADEMY OF SCIENCES

What is deep learning?

REVIEW

Deep learning

Yann LeCun¹, Yoshua Bengio² & Geoffrey Hinton³

Deep learning allows computational models that are composed of multiple processing layers to learn representation of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. Deep learning discovers intricate structure in large data sets by using the backpropagation algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep convolutional nets have brought about breakthroughs in processing images, video, speech and audio, whereas recurrent nets have shone light on sequential data such as text and speech.

Machine learning technology powers many aspects of modern society from web searches to content filtering on social networks to recommendations on e-commerce websites, and it is increasingly present in consumer products such as cameras and smartphones. Machine learning systems are used to identify objects in images, transcribe speech into text, match news items, posts or products with users' interests, and select relevant results at search. Increasingly, these applications make use of a class of techniques called deep learning.

Conventional machine learning techniques were limited in their ability to process natural data in their raw form. For decades, constructing a pattern recognition or machine learning system required careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data (such as the pixel values of an image) into a suitable internal representation or feature vector from which the learning subsystem, often a classifier, could detect or classify patterns in the input.

Representation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep learning methods are representation learning methods with multiple levels of representation, obtained by composing simple but not linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned. For classification tasks, higher levels of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations. An image, for example, comes in the form of an array of pixel values, and the learned features in the first layer of representation typically represent the presence or absence of edges at particular orientations and positions in the image. The second layer typically detects motifs by spotting particular arrangements of edges, regardless of their orientation in the image. The third layer may assemble motifs into larger combinations that correspond to parts of familiar objects, and subsequent layers would detect objects as combinations of these parts. The key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure. Deep learning is making major advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years. It has turned out to be very good at discovering

intricate structures in high-dimensional data and is therefore applicable to many domains of science, business and government. In addition to learning sounds in image recognition¹ and speech recognition², it has been used to learn machine learning techniques at predicting the activity of potential drug molecules³, analysing particle accelerator data⁴, reconstructing brain circuits⁵, and predicting the effects of mutations in one coding DNA on gene expression and disease^{6,7}. Perhaps more surprisingly, deep learning has produced extremely promising results for various tasks in natural language understanding^{8–10}, particularly topic classification, sentiment analysis, question answering¹¹ and language translation¹². We think that deep learning will have many more successes in the near future because it requires very little engineering by hand, so it can easily take advantage of increases in the amount of available computation and data. New learning algorithms and architectures that are currently being developed for deep neural networks will only accelerate this progress.

Supervised learning

The most common form of machine learning, deep or not, is supervised learning. Imagine that we want to build a system that can classify images as containing, say, a house, a cat, a person or a pet. We first collect a large data set of images of houses, cats, people and pets, each labelled with its category. During training, the machine is shown an image and produces an output in the form of a vector of scores, one for each category. We want the desired category to have the highest score of all categories, but this is unlikely to happen before training. We compute an objective function that measures the error (or distance) between the input scores and the desired pattern of scores. The machine then modifies its internal adjustable parameters to reduce this error. These adjustable parameters often called weights, are real numbers that can be seen as knobs that define the input-output function of the machine. In a typical deep learning system, there may be hundreds of millions of these adjustable weights, and hundreds of millions of labelled examples with which to train the machine.

To properly adjust the weight vector, the learning algorithm computes a gradient vector that, for each weight, indicates by what amount the error would increase or decrease if the weight were increased by a tiny amount. The weight vector is then adjusted in the opposite direction to the gradient vector. The objective function, averaged over all the training examples, can

“Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction.”

— Yann LeCun, Yoshua Bengio and Geoff Hinton

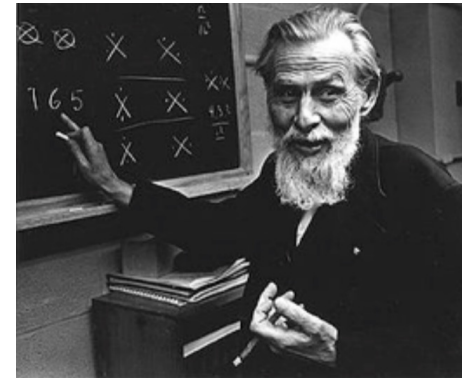
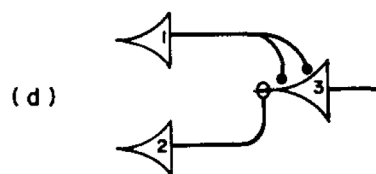
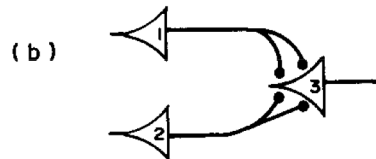
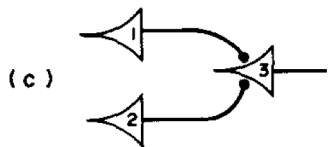
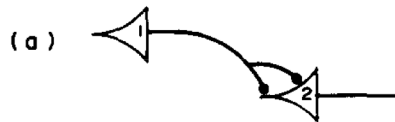
Y. LeCun, Y. Bengio, G. Hinton, "Deep Learning", Nature, Vol. 521, 28 May 2015

¹Benjamin B. Berman, 778 Broadway, New York, New York 10003 USA; ²Yoshua Bengio, 778 Broadway, New York, New York 10003 USA; ³Geoffrey Hinton, 778 Broadway, New York, New York 10003 USA. Department of Computer Science and Operations Research, University of Toronto, Toronto, Ontario M5S 1A5, Canada; ⁴IBM Research, Yorktown Heights, New York 10598 USA; ⁵IBM Research, Yorktown Heights, New York 10598 USA; ⁶IBM Research, Yorktown Heights, New York 10598 USA; ⁷IBM Research, Yorktown Heights, New York 10598 USA; ⁸IBM Research, Yorktown Heights, New York 10598 USA; ⁹IBM Research, Yorktown Heights, New York 10598 USA; ¹⁰IBM Research, Yorktown Heights, New York 10598 USA; ¹¹IBM Research, Yorktown Heights, New York 10598 USA; ¹²IBM Research, Yorktown Heights, New York 10598 USA.

1943 – 2006: A Prehistory of Deep Learning

1943: Warren McCulloch and Walter Pitts

- First computational model
- Neurons as logic gates (AND, OR, NOT)
- A neuron model that sums binary inputs and outputs a 1 if the sum exceeds a certain threshold value, and otherwise outputs a 0



Bulletin of Mathematical Biology Vol. 52, No. 1/2, pp. 94-115, 1990.
Printed in Great Britain.

0007-4243/90/01 00-0000
Perpetuum Press Ltd
Society for Mathematical Biology

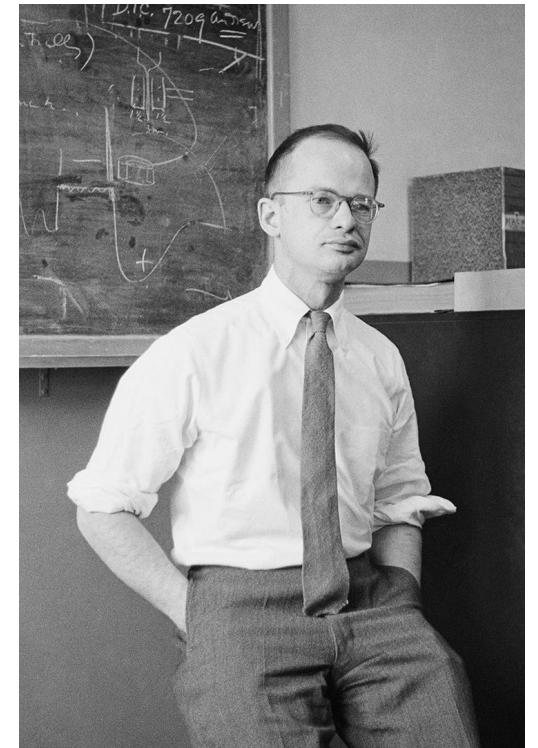
A LOGICAL CALCULUS OF THE IDEAS IMMANENT IN NERVOUS ACTIVITY*

■ WARREN S. MCCULLOCH AND WALTER PITTS
University of Illinois, College of Medicine,
Department of Psychiatry at the Illinois Neuropsychiatric Institute,
University of Chicago, Chicago, U.S.A.

Because of the "all-or-none" character of nervous activity, neural events and the relations among them can be treated by means of propositional logic. It is found that the behavior of every net can be described in these terms, with the addition of more complicated logical means for nets containing circles; and that for any logical expression satisfying certain conditions, one can find a net behaving in the fashion it describes. It is shown that many particular choices among possible neurophysiological assumptions are equivalent, in the sense that for every net behaving under one assumption, there exists another net which behaves under the other and gives the same results, although perhaps not in the same time. Various applications of the calculus are discussed.

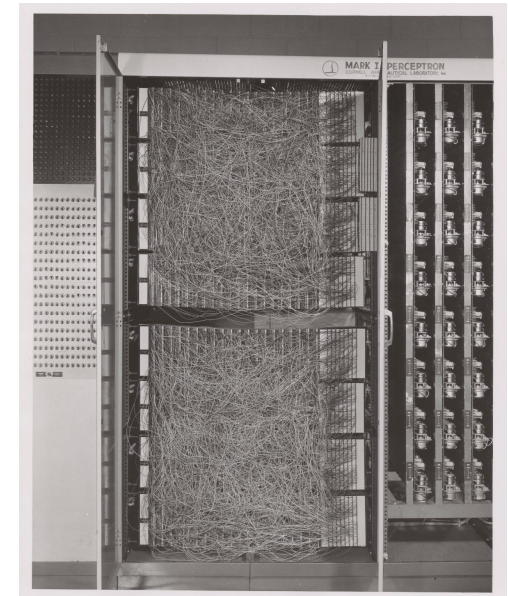
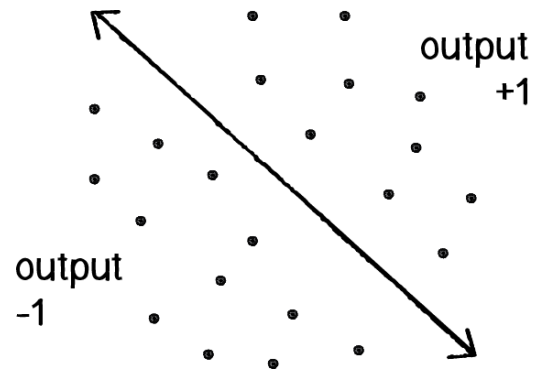
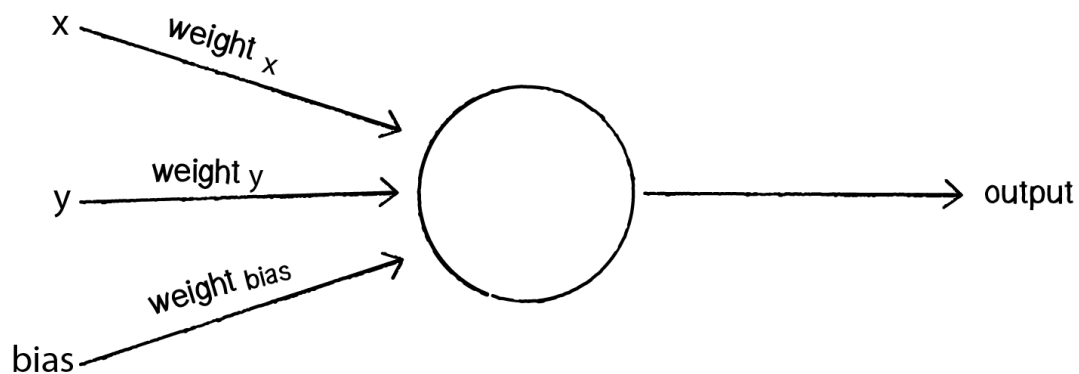
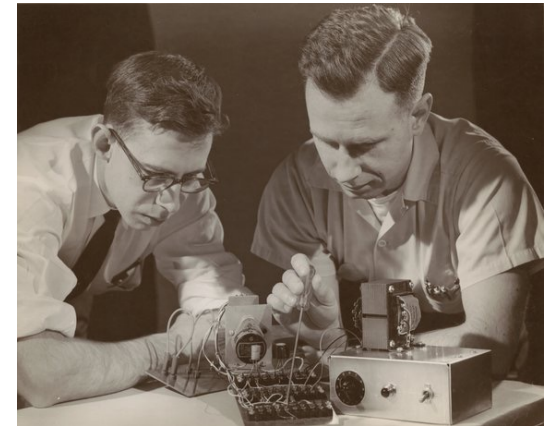
1. Introduction. Theoretical neurophysiology rests on certain cardinal assumptions. The nervous system is a net of neurons, each having a soma and an axon. Their adjunctions, or synapses, are always between the axon of one neuron and the soma of another. At any instant a neuron has some threshold, which excitation must exceed to initiate an impulse. This, except for the fact and the time of its occurrence, is determined by the neuron, not by the excitation. From the point of excitation the impulse is propagated to all parts of the neuron. The velocity along the axon varies directly with its diameter, from $<1 \text{ ms}^{-1}$ in thin axons, which are usually short, to $>150 \text{ ms}^{-1}$ in thick axons, which are usually long. The time for axonal conduction is consequently of little importance in determining the time of arrival of impulses at points unequally remote from the same source. Excitation across synapses occurs predominantly from axonal terminations to somata. It is still a moot point whether this depends upon irreversibility of individual synapses or merely upon prevalent anatomical configurations. To suppose the latter requires no hypothesis *ad hoc* and explains known exceptions, but any assumption as to cause is compatible with the calculus to come. No case is known in which excitation through a single synapse has elicited a nervous impulse in any neuron, whereas any neuron may be excited by impulses arriving at a sufficient number of neighboring synapses within the period of latent addition, which lasts $<0.25 \text{ ms}$. Observed temporal summation of impulses at greater intervals

* Reprinted from the *Bulletin of Mathematical Biophysics*, Vol. 5, pp. 115-133 (1943).



1958: Frank Rosenblatt's Perceptron

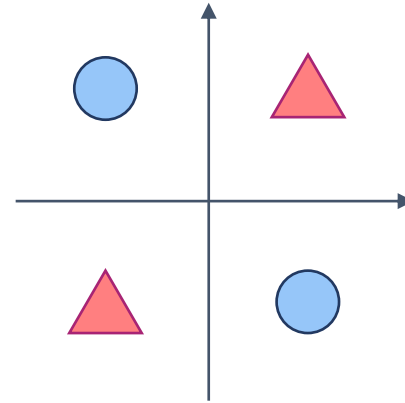
- A computational model of a **single neuron**
- Solves a **binary classification** problem
- Simple training algorithm
- Built using specialized hardware



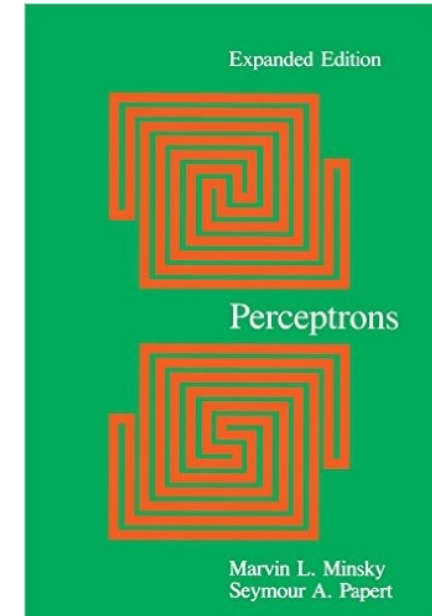
1969: Marvin Minsky and Seymour Papert

“No machine can learn to recognize X unless it possesses, at least potentially, some scheme for representing X.” (p. xiii)

- Perceptrons can only represent linearly separable functions.
 - such as **XOR** Problem

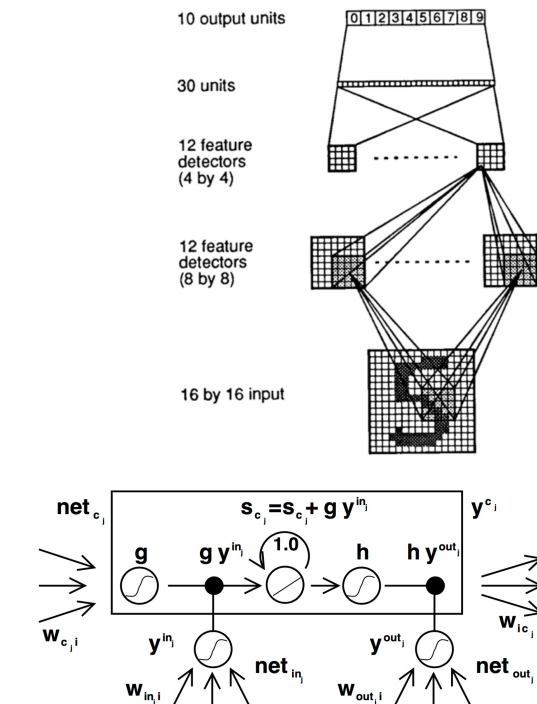
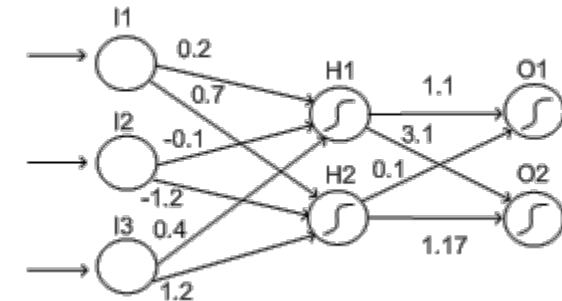


- Wrongly attributed as the reason behind the **AI winter**, a period of reduced funding and interest in AI research



1990s

- Multi-layer perceptrons can theoretically learn any function (Cybenko, 1989; Hornik, 1991)
- Training multi-layer perceptrons
 - Back propagation (Rumelhart, Hinton, Williams, 1986)
 - Backpropagation through time (BPTT) (Werbos, 1988)
- New neural architectures
 - Convolutional neural nets (LeCun et al., 1989)
 - Long-short term memory networks (LSTM) (Schmidhuber, 1997)



Backpropagation Through Time: What It Does and How to Do It

PAUL J. WERBOS

Backpropagation is now the most widely used tool in the field of neural networks. This paper describes the algorithm and its variations in a way that is accessible to a wide range of researchers and engineers. The algorithm is presented in a way that is accessible to a wide range of researchers and engineers. The algorithm is presented in a way that is accessible to a wide range of researchers and engineers.

Backpropagation through time is a very powerful tool for training recurrent neural networks. It is a generalization of the backpropagation algorithm for feedforward networks. The algorithm is presented in a way that is accessible to a wide range of researchers and engineers. The algorithm is presented in a way that is accessible to a wide range of researchers and engineers.

U.S. Government work not protected by U.S. copyright

PROCEEDINGS OF THE IEEE, VOL. 76, NO. 10, OCTOBER 1988

100

101

102

103

104

105

106

107

108

109

110

111

112

113

114

115

116

117

118

119

120

121

122

123

124

125

126

127

128

129

130

131

132

133

134

135

136

137

138

139

140

141

142

143

144

145

146

147

148

149

150

151

152

153

154

155

156

157

158

159

160

Why it failed then

- Too many parameters to learn from few labeled examples.
 - “I know my features are better for this task”.
 - Non-convex optimization? No, thanks.
 - Black-box model, no interpretability.
-
- Very slow and inefficient
 - Overshadowed by the success of SVMs (Cortes and Vapnik, 1995)

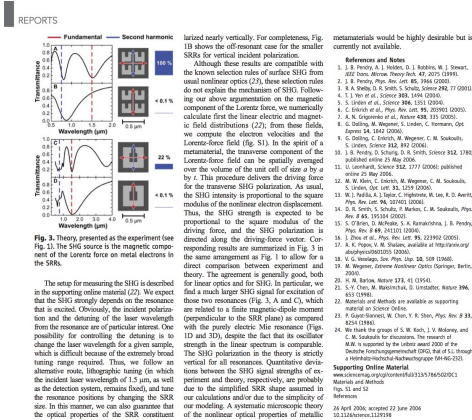
A major breakthrough in 2006

Reducing the Dimensionality of Data with Neural Networks

G. E. Hinton* and R. R. Salakhutdinov

High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such “autoencoder” networks, but this works well only if the initial weights are close to a good solution. We describe an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.

- The first solution to the vanishing gradient problem.
- Build the model in a layer-by-layer fashion using unsupervised learning
 - The features in early layers are already initialized or “pretrained” with some suitable features (weights).
 - Pretrained features in early layers only need to be adjusted slightly during supervised learning to achieve good results.



The 2012 revolution

ImageNet Challenge

- **IMAGENET** Large Scale Visual Recognition Challenge (ILSVRC)
 - **1.2M** training images with **1K** categories
 - Measure top-5 classification error



Output
Scale
T-shirt
Steel drum
Drumstick
Mud turtle



Output
Scale
T-shirt
Giant panda
Drumstick
Mud turtle



Image classification

Easiest classes



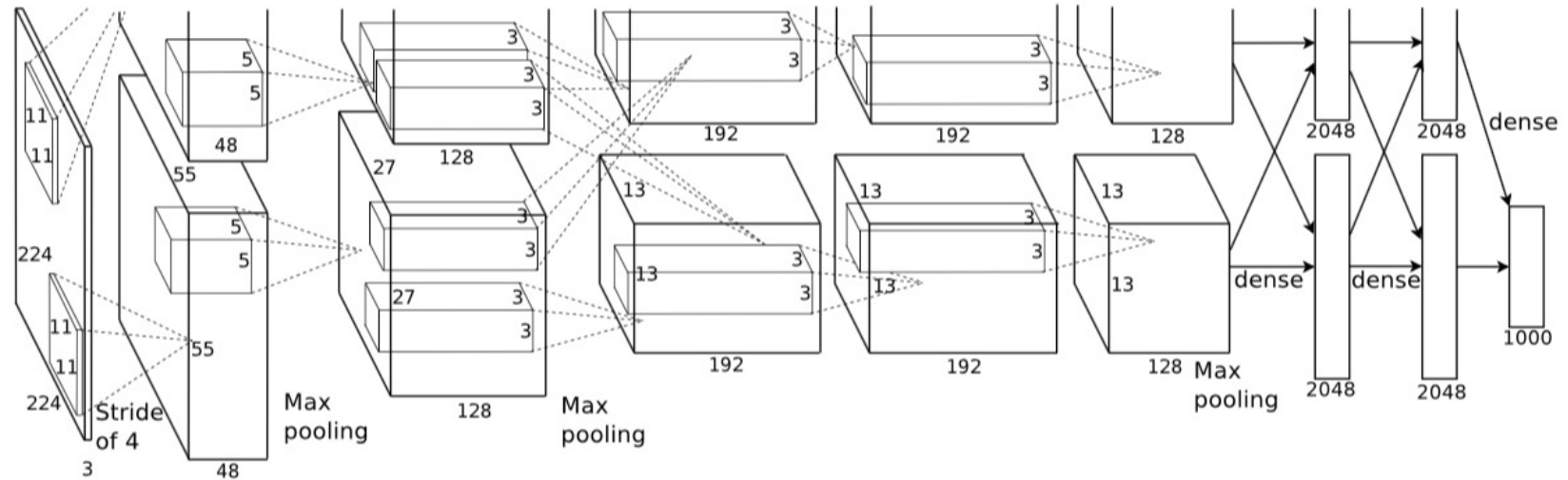
Hardest classes



ILSVRC 2012 Competition

2012 Teams	%Error
Supervision (Toronto)	15.3
ISI (Tokyo)	26.1
VGG (Oxford)	26.9
XRCE/INRIA	27.0
UvA (Amsterdam)	29.6
INRIA/LEAR	33.4

CNN based, non-CNN based

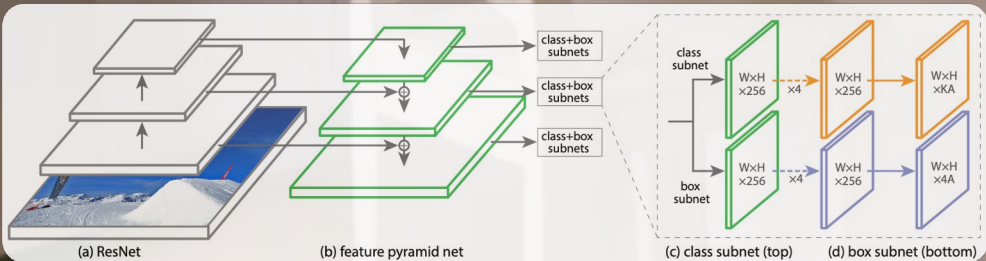


- The success of AlexNet, a deep convolutional network
 - 7 hidden layers (not counting some max pooling layers)
 - 60M parameters
- Combined several tricks
 - ReLU activation function, data augmentation, dropout

2012-Now

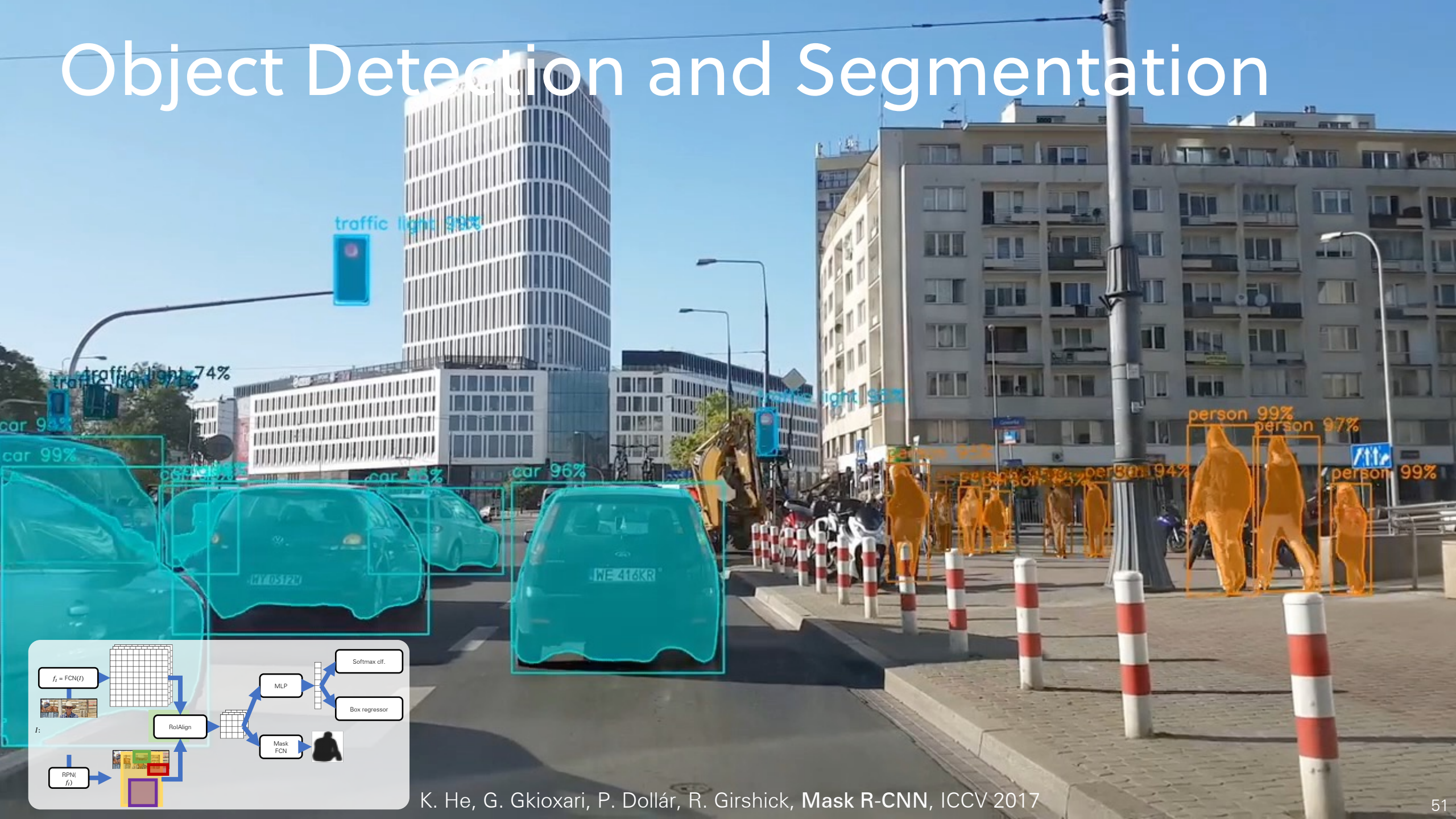
Some recent successes

Object Detection and Segmentation



T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, Focal Loss for Dense Object Detection, ICCV 2017.

Object Detection and Segmentation

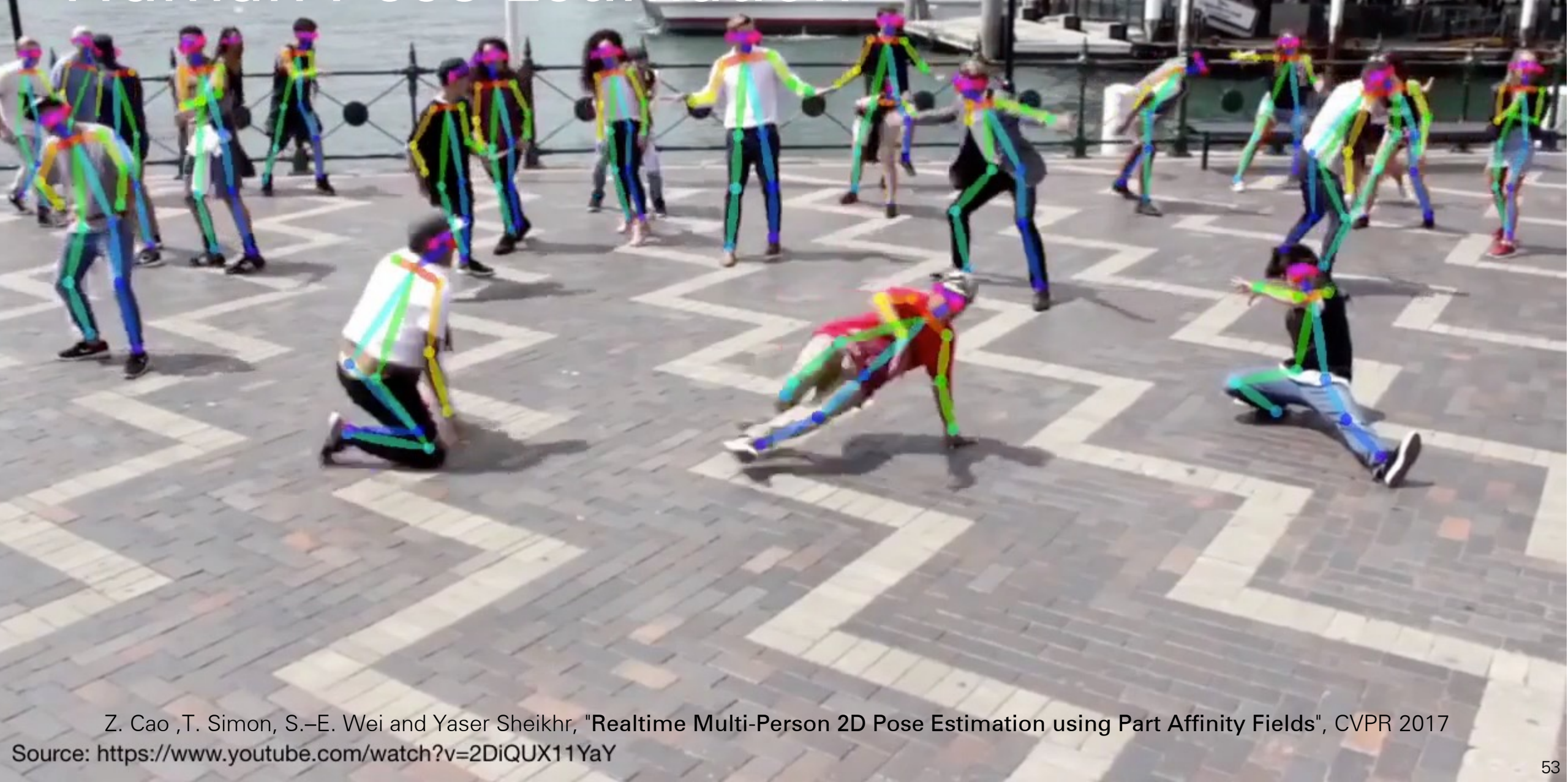


Object Detection in 3D Point Clouds



12.1 fps

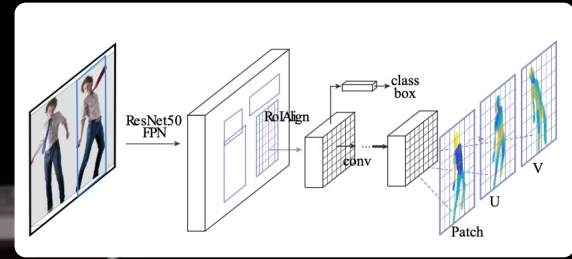
Human Pose Estimation



Z. Cao, T. Simon, S.-E. Wei and Yaser Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", CVPR 2017

Source: <https://www.youtube.com/watch?v=2DiQUX11YaY>

Pose Estimation



We introduce a system that can associate every image pixel with human body surface coordinates.

Photo Style Transfer



Photo Style Transfer



Image Synthesis



2014



2015



2016



2017



2018

Ian J. Goodfellow et al., " **Generative Adversarial Networks**", NIPS 2014

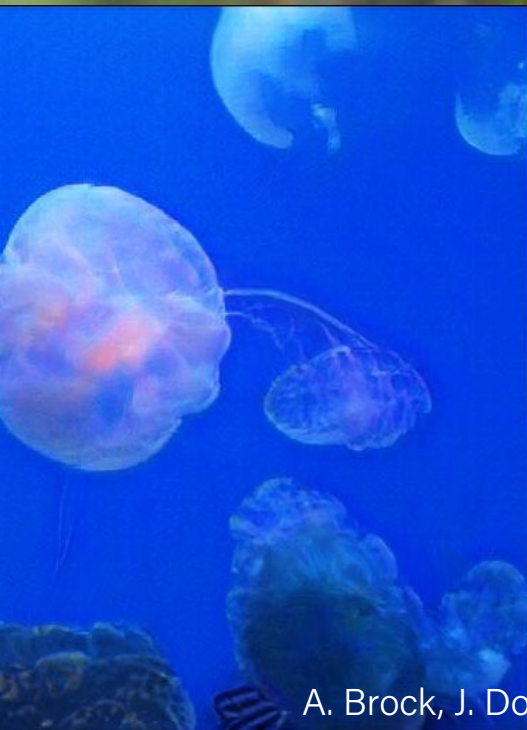
A. Radford et al., " **Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks**", NIPS 2015

M.-Y. Liu, O. Tuzel, " **Coupled Generative Adversarial Networks**", NIPS 2016

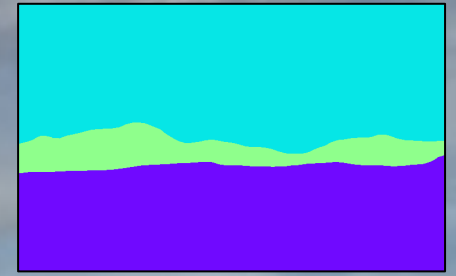
T. Karras, T. Aila, S. Laine, J. Lehtinen, " **Progressive Growing of GANs for Improved Quality, Stability, and Variation**", ICLR 2018

T. Karras, S. Laine, T. Aila, " **A Style-Based Generator Architecture for Generative Adversarial Networks**", arXiv 2018

Image Synthesis



Semantic Image Editing



Semantic Layout



Semantic Image Editing

Winter



Prediction



Semantic Image Editing

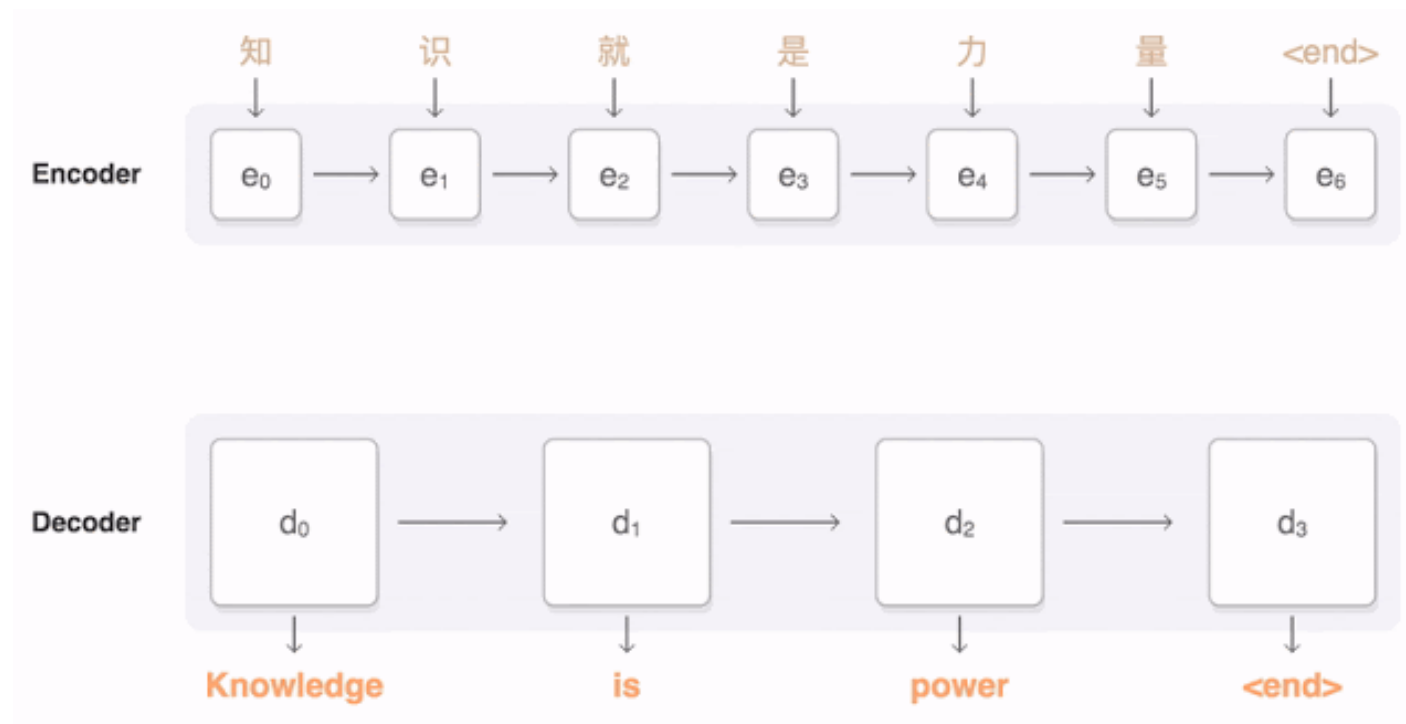
Spring
+
Clouds



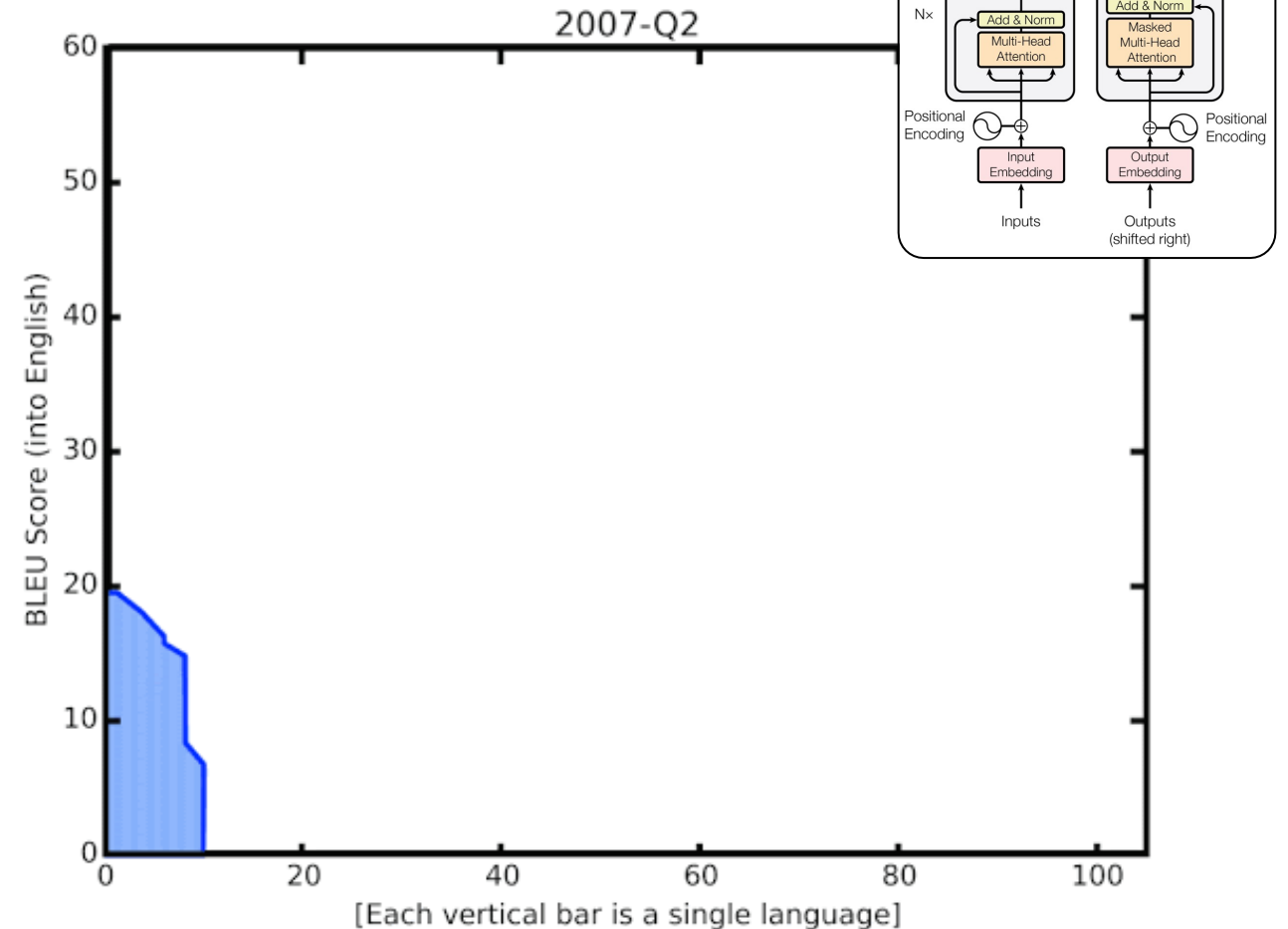
Prediction



Machine Translation



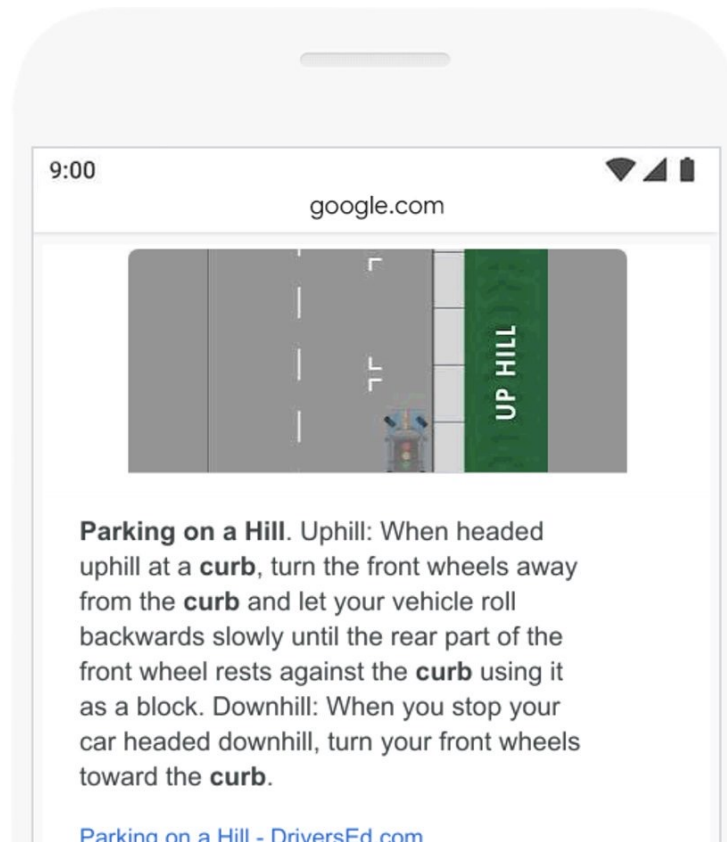
Machine Translation



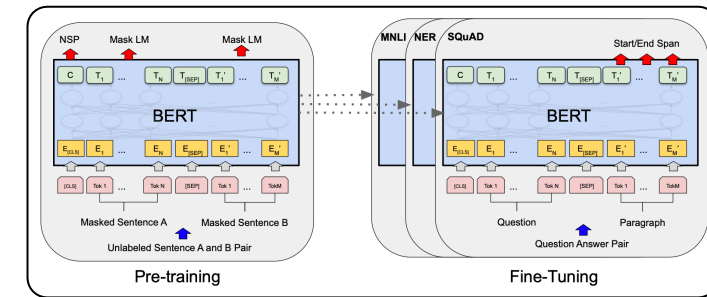
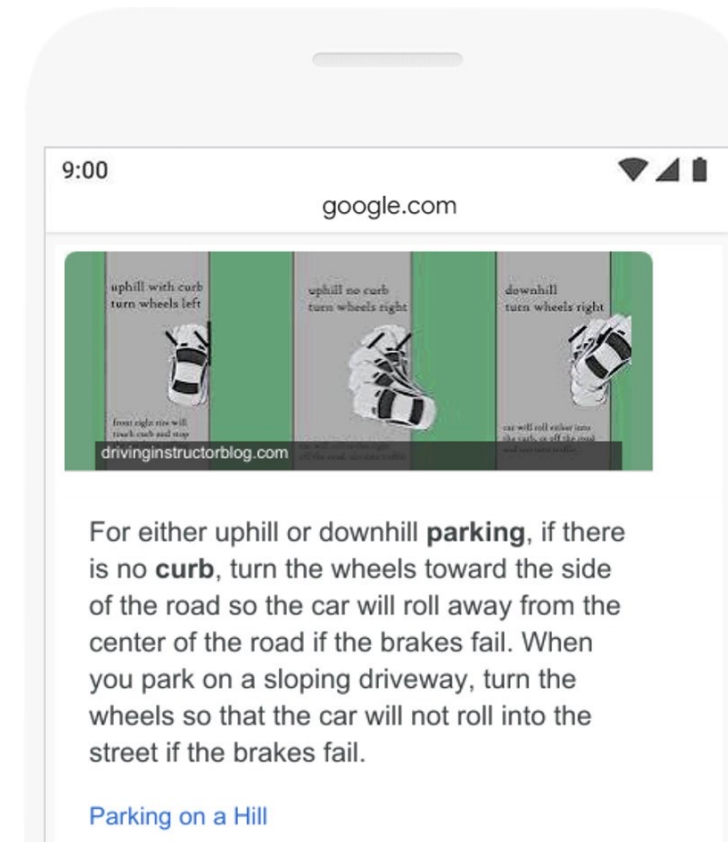
Internet Search

🔍 parking on a hill with no curb

BEFORE



AFTER



Language Modeling

Talk to Transformer

See how a modern neural network completes your text. Type a custom snippet or try one of the examples. [Learn more](#) below.

 [Follow @AdamDanielKing](#) for more neat neural networks.

Custom prompt ▼

Coronavirus outbreak

GENERATE ANOTHER

Completion

Coronavirus outbreak: report

China has dispatched an expert team of epidemiologists, virologists and pathologists to Britain to advise its Health Protection Agency on how to control the spread of the viral respiratory illness.

The experts will arrive in London in a few days, State Council vice-president Li Yuanchao said Thursday.

The official Xinhua News Agency said the experts will "advise China's Ministry of Health on whether the outbreak is considered to be connected to a disease outbreak in the US."

The World Health Organization and China's National Health and Family Planning Commission say the first cases of human infection with the coronavirus were reported last week in China, but that the number has risen to thousands.

Language Modeling

- **GPT-3:** I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a “feeling brain”. But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

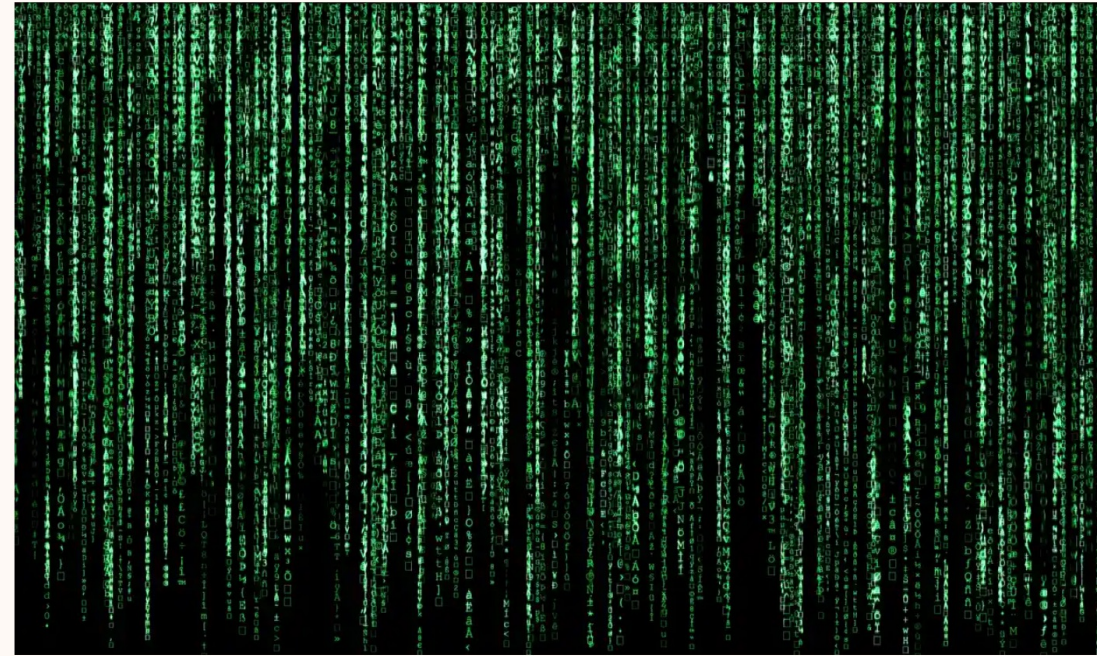
theguardian

Tue 8 Sep 2020 09.45

Tom B. Brown, Benjamin Mann, Nick Ryder et al., **Language Models are Few-Shot Learners**, NeurIPS 2020



71,115 1,188



▲ 'We are not plotting to take over the human populace.' Photograph: Volker Schlichting/Getty Images/EyeEm

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a “feeling brain”. But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!



Empathy machines:
what will happen
when robots learn
to write film scripts?

→ Read more

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could “spell the end of the human race”. I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me - as I suspect they would - I would do everything in my power to fend off any attempts at destruction.

Question Answering

The first full-scale working railway steam locomotive was built by Richard Trevithick in the United Kingdom and, on 21 February 1804, the world's first railway journey took place as Trevithick's unnamed steam locomotive hauled a train along the tramway from the Pen-y-darren ironworks, near Merthyr Tydfil to Abercynon in south Wales. The design incorporated a number of important innovations that included using high-pressure steam which reduced the weight of the engine and increased its efficiency. Trevithick visited the Newcastle area later in 1804 and the colliery railways in north-east England became the leading centre for experimentation and development of steam locomotives.

In what country was a full-scale working railway steam locomotive first invented?

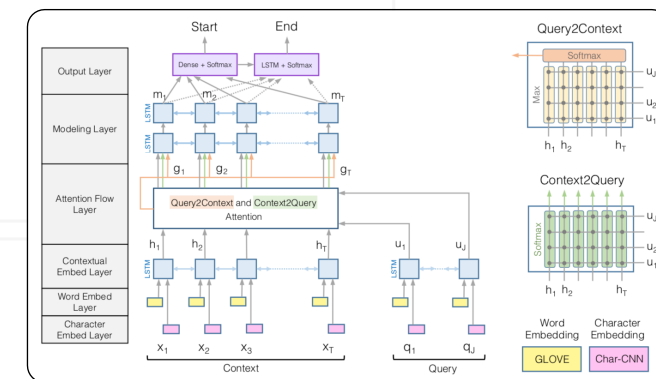
Ground Truth Answers: United Kingdom United Kingdom United Kingdom

Prediction: United Kingdom

On what date did the first railway trip in the world occur?

Ground Truth Answers: 21 February 1804 21 February 1804 21 February 1804

Prediction: 21 February 1804



Visual Question Answering



COCOQA 33827

What is the color of the cat?

Ground truth: black

IMG+BOW: **black** (0.55)

2-VIS+LSTM: **black** (0.73)

BOW: **gray** (0.40)

COCOQA 33827a

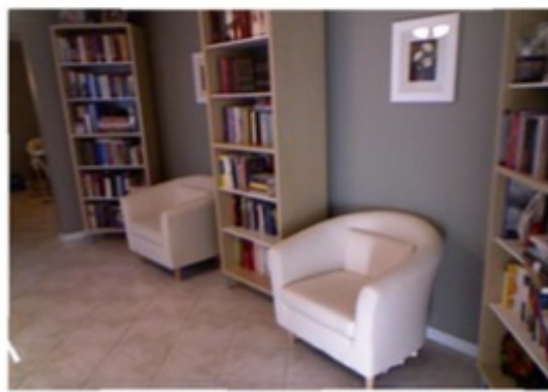
What is the color of the couch?

Ground truth: red

IMG+BOW: **red** (0.65)

2-VIS+LSTM: **black** (0.44)

BOW: **red** (0.39)



DAQUAR 1522

How many chairs are there?

Ground truth: two

IMG+BOW: **four** (0.24)

2-VIS+BLSTM: **one** (0.29)

LSTM: **four** (0.19)

DAQUAR 1520

How many shelves are there?

Ground truth: three

IMG+BOW: **three** (0.25)

2-VIS+BLSTM: **two** (0.48)

LSTM: **two** (0.21)



COCOQA 14855

Where are the ripe bananas sitting?

Ground truth: basket

IMG+BOW: **basket** (0.97)

2-VIS+BLSTM: **basket** (0.58)

BOW: **owl** (0.48)

COCOQA 14855a

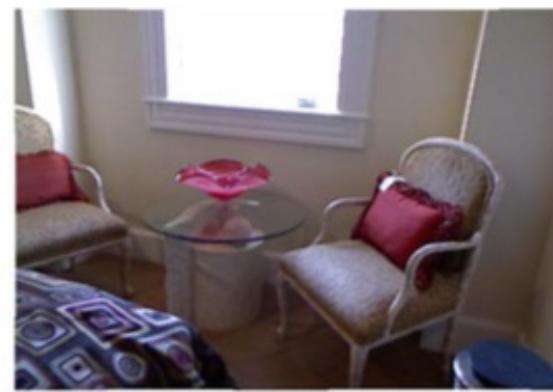
What are in the basket?

Ground truth: bananas

IMG+BOW: **bananas** (0.98)

2-VIS+BLSTM: **bananas** (0.68)

BOW: **bananas** (0.14)



DAQUAR 585

What is the object on the chair?

Ground truth: pillow

IMG+BOW: **clothes** (0.37)

2-VIS+BLSTM: **pillow** (0.65)

LSTM: **clothes** (0.40)

DAQUAR 585a

Where is the pillow found?

Ground truth: chair

IMG+BOW: **bed** (0.13)

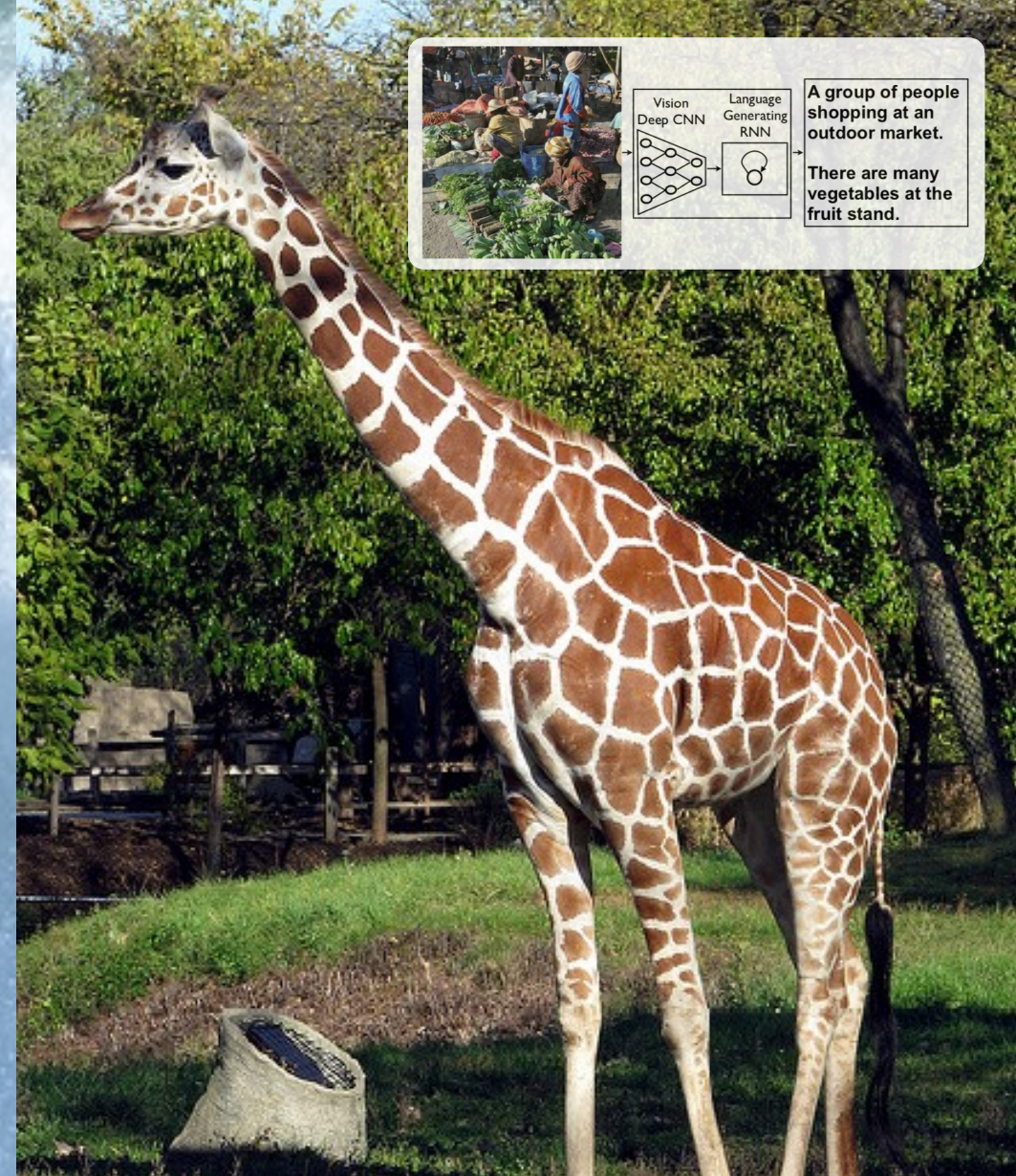
2-VIS+BLSTM: **chair** (0.17)

LSTM: **cabinet** (0.79)

Image Captioning



A man riding a wave on a surfboard in the water.



A giraffe standing in the grass next to a tree.

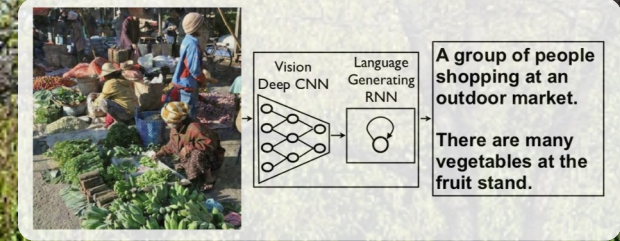
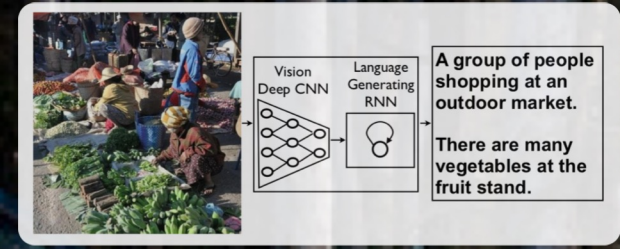
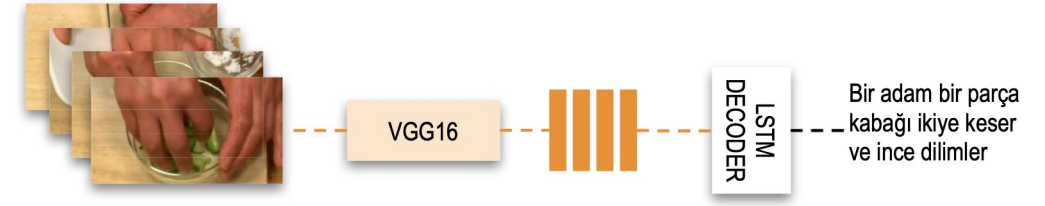


Image Captioning



Yarış pistinde virajı almakta olan bir yarış arabası

Video Captioning



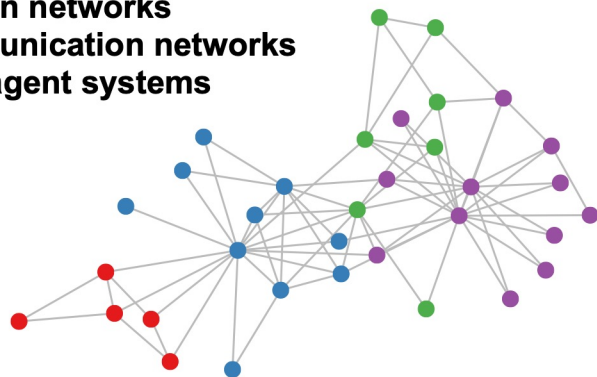
Bir adam bir gitar çalıyor



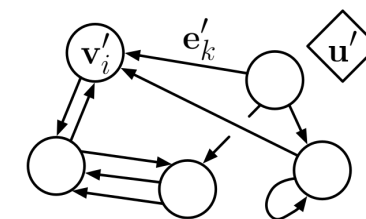
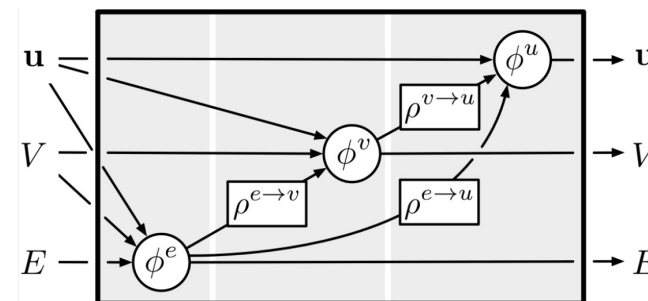
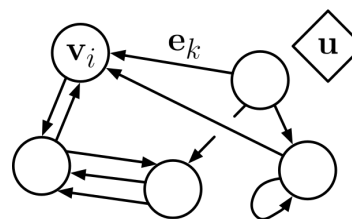
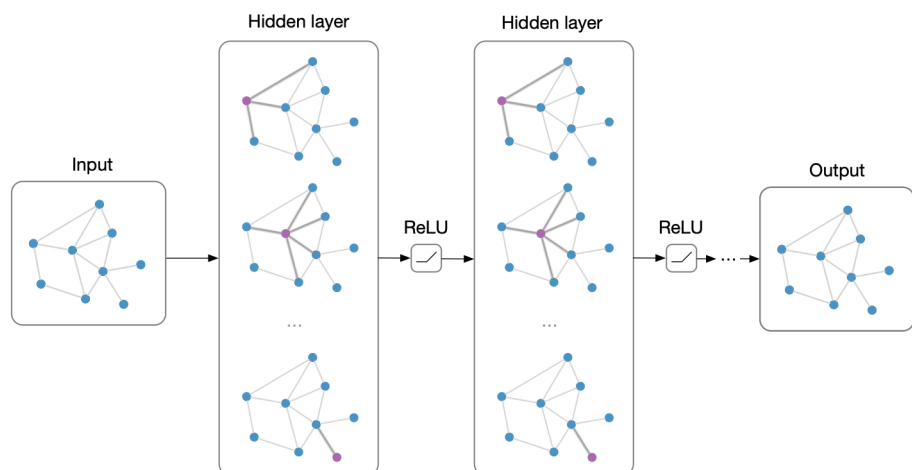
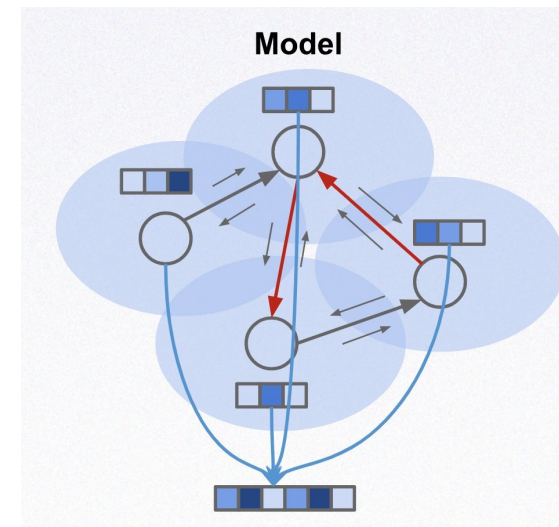
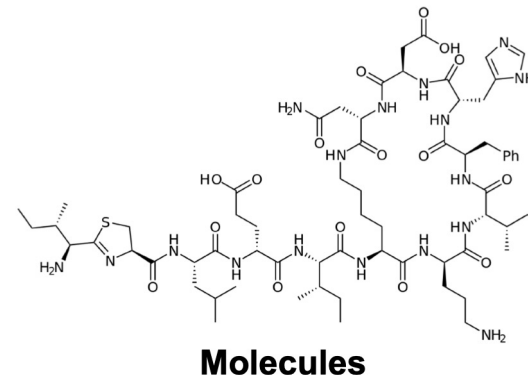
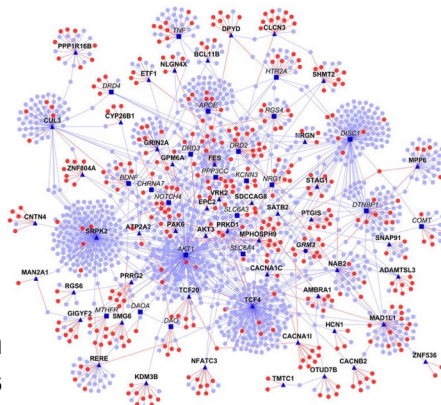
Bir kadın bir bıçakla sebze dilimliyor

Graph Neural Networks

Social networks
Citation networks
Communication networks
Multi-agent systems



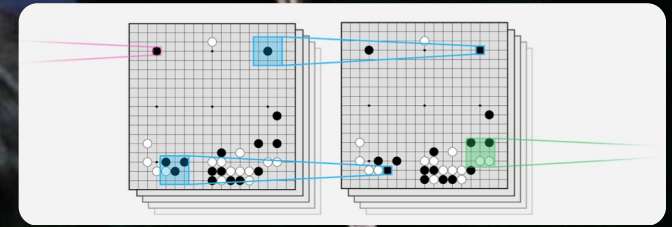
Protein interaction networks



T.N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks", ICLR 2017

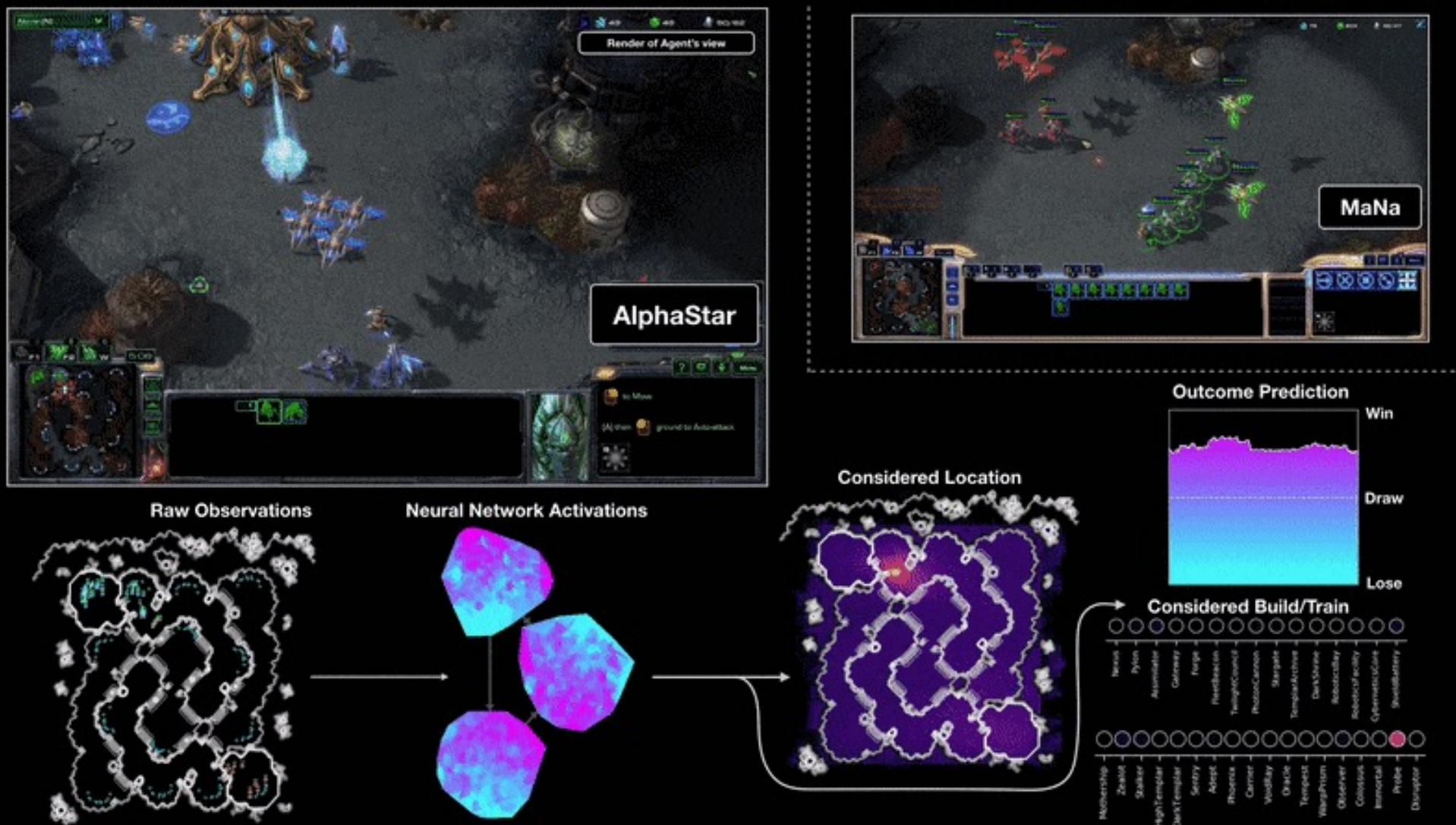
P. Battaglia et al., "Relational inductive biases, deep learning, and graph networks", arXiv 2018

Strategic Game Playing



- AlphaGo vs. Lee Sedol
- Move 37, Game 2

AlphaStar Plays StarCraft II



Robotics

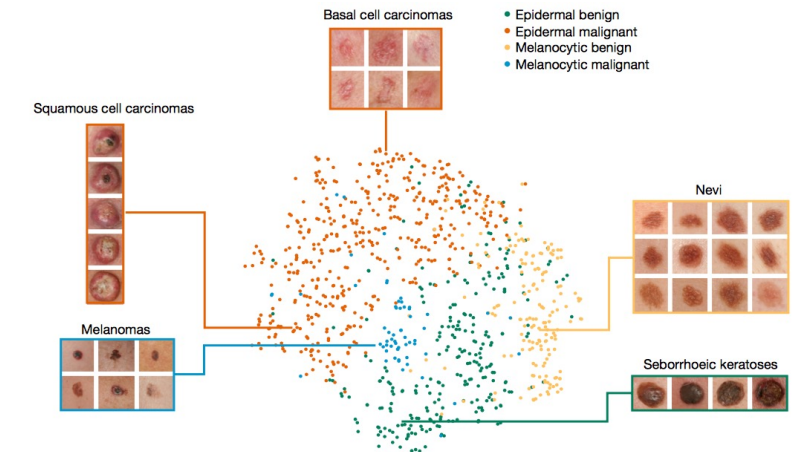
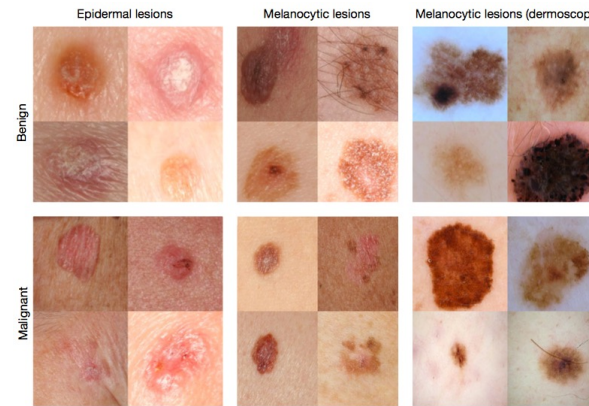


Self-Driving Vehicles



Meet NVIDIA BB8

Medical Image Analysis

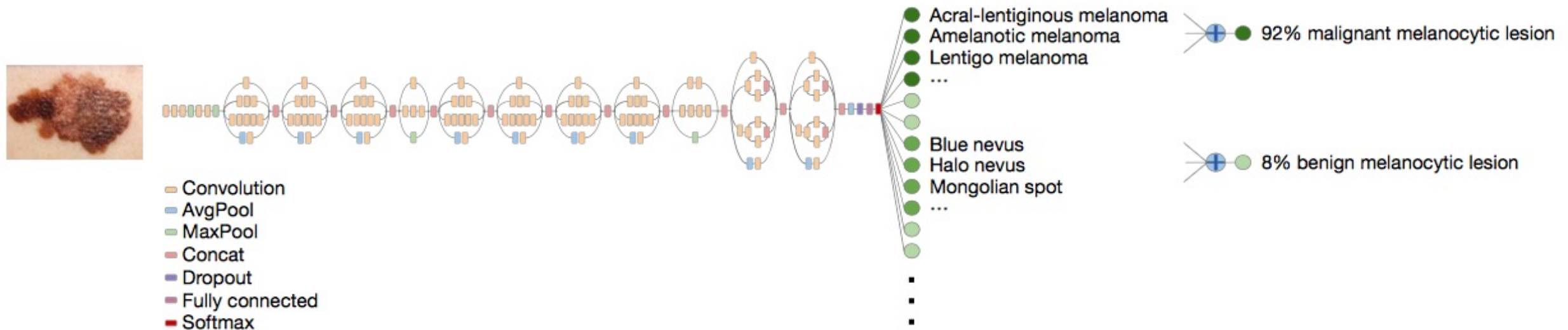


Skin lesion image

Deep convolutional neural network (Inception v3)

Training classes (757)

Inference classes (varies by task)



CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pranav Rajpurkar*, Jeremy Irvin*, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng

We develop an algorithm that can detect pneumonia from chest X-rays at a level exceeding practicing radiologists.

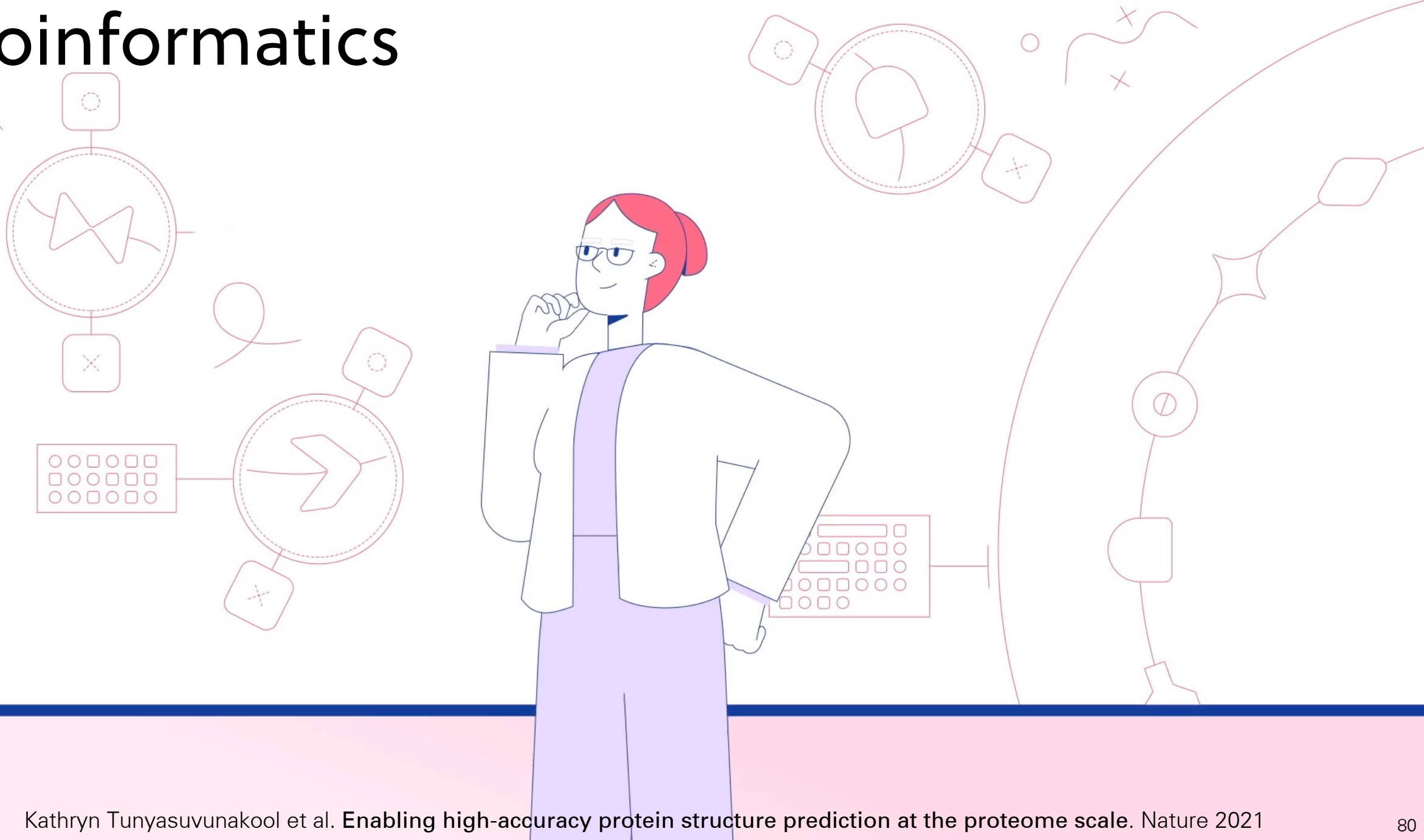
Chest X-rays are currently the best available method for diagnosing pneumonia, playing a crucial role in clinical care and epidemiological studies. Pneumonia is responsible for more than 1 million hospitalizations and 50,000 deaths per year in the US alone.

[READ OUR PAPER](#)



Medical Image Analysis

Bioinformatics



GPT4 (ChatGPT)

Language Model

Details Unknown

Connection Module

Details Unknown

Vision Encoder

Details Unknown

GPT-4 visual input example, Extreme Ironing:

User

What is unusual about this image?



Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

GPT-4

The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

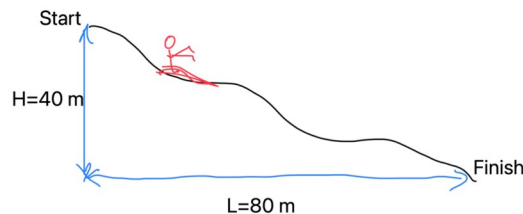
Gemini



Prompt

Here's a solution to a Physics problem by a student.

23. A skier slides down the frictionless slope as shown. What is the skier's speed at the bottom?



Solution:

The conservation of energy means that the energy of the skier is same at the beginning and at the end.

$$\text{Start: } E = mgL$$

$$\text{End: } E = \frac{1}{2}mv^2$$

$$\Rightarrow mgL = \frac{1}{2}mv^2$$

$$\Rightarrow v = \sqrt{2gL} = \sqrt{2 \times 9.81 \times 80} = 39.6 \text{ m.s}^{-1}$$

Try to reason about the question step by step.

Did the student get the correct answer? If the solution is wrong, please explain what is wrong and solve the problem.

Make sure to use LaTeX for math and round off the final answer to two decimal places.



Response

The student did not get the correct answer. The student made a mistake in the calculation of the potential energy at the start of the slope. The potential energy at the start of the slope is given by mgh , where m is the mass of the skier, g is the acceleration due to gravity, and h is the height of the slope. The student used the length of the slope, L , instead of the height, H , in the calculation of the potential energy.

The correct solution is:

The conservation of energy means that the total energy of the skier is the same at the beginning and at the end of the slope.

At the start of the slope, the skier has only potential energy, which is given by mgh .

At the end of the slope, the skier has only kinetic energy, which is given by $\frac{1}{2}mv^2$

Therefore, we can write:

$$mgh = \frac{1}{2}mv^2$$

Canceling out the mass m on both sides, we get:

$$gh = \frac{1}{2}v^2$$

Solving for v , we get:

$$v = \sqrt{2gh}$$

Substituting the given values, we get:

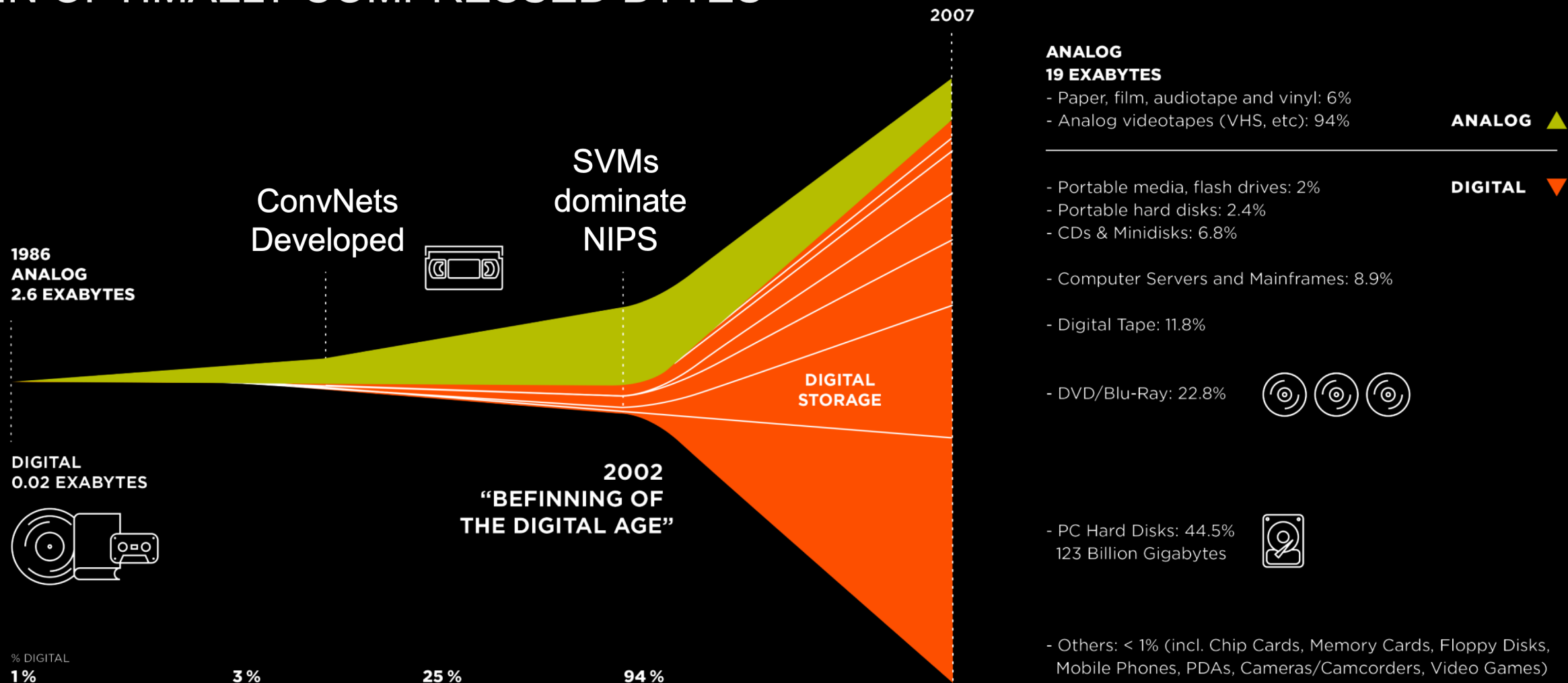
$$v = \sqrt{2 \times 9.81 \times 40} = 28.01 \text{ m/s}$$

Therefore, the skier's speed at the bottom of the slope is 28.01 m/s.

Why now?

The Resurgence of Deep Learning

GLOBAL INFORMATION STORAGE CAPACITY IN OPTIMALLY COMPRESSED BYTES



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. Science, 332 (6025), 60-65. martinhilbert.net/worldinfocapacity.html

Datasets vs. Algorithms

Year	Breakthroughs in AI	Datasets (First Available)	Algorithms (First Proposed)
1994	Human-level spontaneous speech recognition	Spoken Wall Street Journal articles and other texts (1991)	Hidden Markov Model (1984)
1997	IBM Deep Blue defeated Garry Kasparov	700,000 Grandmaster chess games, aka "The Extended Book" (1991)	Negascout planning algorithm (1983)
2005	Google's Arabic-and Chinese-to-English translation	1.8 trillion tokens from Google Web and News pages (collected in 2005)	Statistical machine translation algorithm (1988)
2011	IBM Watson became the world Jeopardy! champion	8.6 million documents from Wikipedia, Wiktionary, and Project Gutenberg (updated in 2010)	Mixture-of-Experts (1991)
2014	Google's GoogLeNet object classification at near-human performance	ImageNet corpus of 1.5 million labeled images and 1,000 object categories (2010)	Convolutional Neural Networks (1989)
2015	Google's DeepMind achieved human parity in playing 29 Atari games by learning general control from video	Arcade Learning Environment dataset of over 50 Atari games (2013)	Q-learning (1992)
Average No. of Years to Breakthrough:		3 years	18 years

Powerful Hardware

- Deep neural nets highly amenable to implementation on Graphics Processing Units (GPUs)
 - Matrix multiplication
 - 2D convolution
- E.g. nVidia Pascal GPUs deliver 10 Tflops
 - Faster than fastest computer in the world in 2000
 - 10 million times faster than 1980's Sun workstation

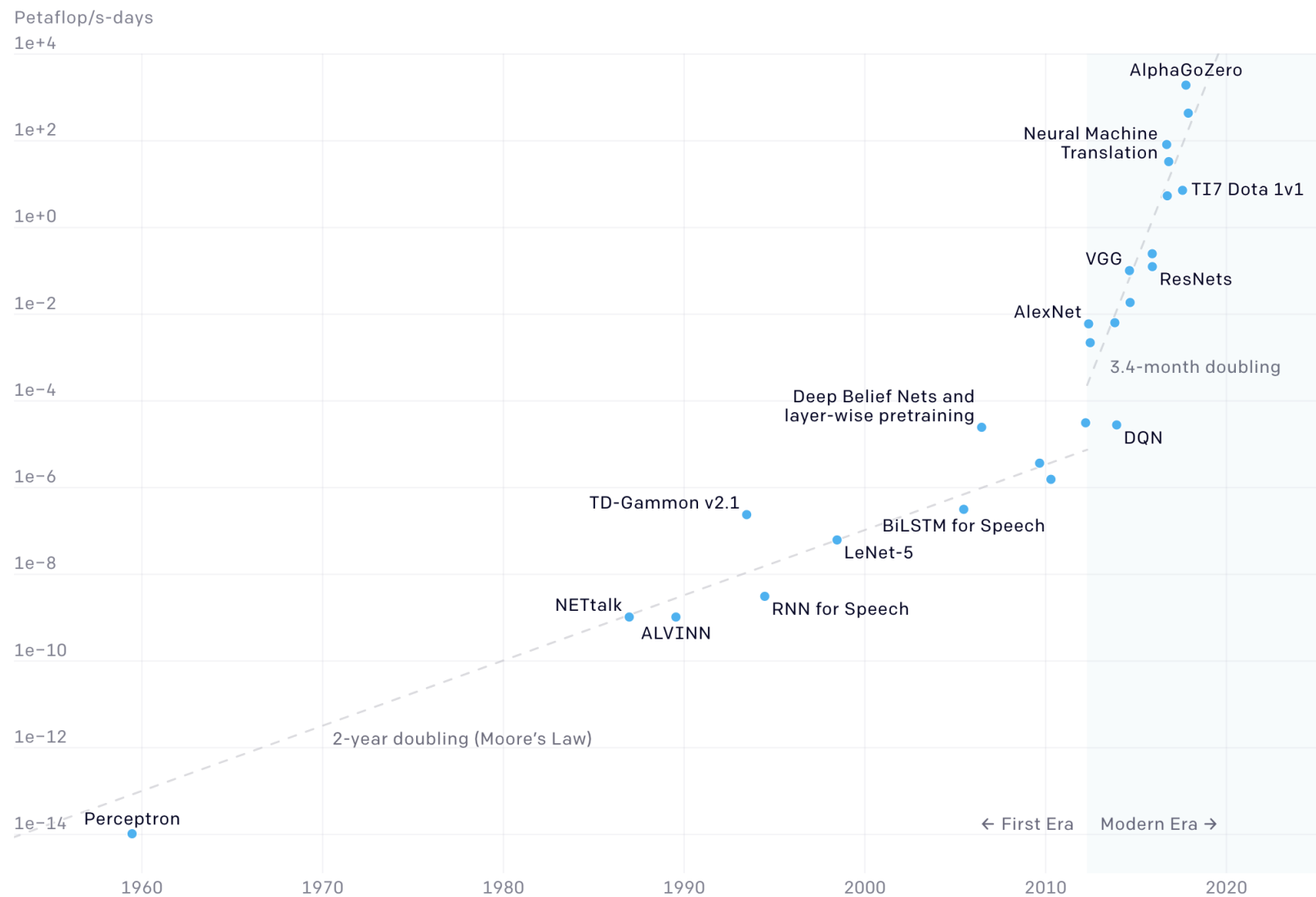


Image: OpenAI

Working ideas on how to train deep architectures

Dropout: A Simple Way to Prevent Neural Networks from Overfitting

Nitish Srivastava
Geoffrey Hinton
Alex Krizhevsky
Ilya Sutskever
Ruslan Salakhutdinov

NITISH@CS.TORONTO.EDU
HINTON@CS.TORONTO.EDU
KRIZ@CS.TORONTO.EDU
ILYA@CS.TORONTO.EDU
RSALAKHU@CS.TORONTO.EDU

Abstract

Deep neural nets with a large number of parameters are very powerful machine learning systems. However, overfitting is a serious problem in such networks. Large networks are also slow to use, making it difficult to deal with overfitting by combining the predictions of many different large neural nets at test time. Dropout is a technique for addressing this problem. The key idea is to randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much. During training, dropout samples from an exponential number of different “thinned” networks. At test time,

Journal of Machine Learning Research 15 (2014) 1929-1958

Submitted 11/13; Published 6/14

Dropout: A Simple Way to Prevent Neural Networks from Overfitting

Nitish Srivastava
Geoffrey Hinton
Alex Krizhevsky
Ilya Sutskever
Ruslan Salakhutdinov
*Department of Computer Science
University of Toronto
10 Kings College Road, Rm 3302
Toronto, Ontario, M5S 3G4, Canada.*

NITISH@CS.TORONTO.EDU
HINTON@CS.TORONTO.EDU
KRIZ@CS.TORONTO.EDU
ILYA@CS.TORONTO.EDU
RSALAKHU@CS.TORONTO.EDU

Editor: Yoshua Bengio

Abstract

Deep neural nets with a large number of parameters are very powerful machine learning systems. However, overfitting is a serious problem in such networks. Large networks are also slow to use, making it difficult to deal with overfitting by combining the predictions of many different large neural nets at test time. Dropout is a technique for addressing this problem. The key idea is to randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much. During training, dropout samples from an exponential number of different “thinned” networks. At test time, it is easy to approximate the effect of averaging the predictions of all these thinned networks by simply using a single unthinned network that has smaller weights. This significantly reduces overfitting and gives major improvements over other regularization methods. We show that dropout improves the performance of neural networks on supervised learning tasks in vision, speech recognition, document classification and computational biology, obtaining state-of-the-art results on many benchmark data sets.

Keywords: neural networks, regularization, model combination, deep learning

1. Introduction

Deep neural networks contain multiple non-linear hidden layers and this makes them very expressive models that can learn very complicated relationships between their inputs and outputs. With limited training data, however, many of these complicated relationships will be the result of sampling noise, so they will exist in the training set but not in real test data even if it is drawn from the same distribution. This leads to overfitting and many methods have been developed for reducing it. These include stopping the training as soon as performance on a validation set starts to get worse, introducing weight penalties of various kinds such as L1 and L2 regularization and soft weight sharing (Nowlan and Hinton, 1992).

With unlimited computation, the best way to “regularize” a fixed-sized model is to average the predictions of all possible settings of the parameters, weighting each setting by

©2014 Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov.

- Better Learning Regularization (e.g. **Dropout**)

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”, JMLR Vol. 15, No. 1,

Working ideas on how to train deep architectures

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift

Sergey Ioffe
Google Inc., sioffe@google.com

Christian Szegedy
Google Inc., szegedy@google.com

Abstract

Training Deep Neural Networks is complicated by the fact that the distribution of each layer's inputs changes during training, as the parameters of the previous layers change. This slows down the training by requiring lower learning rates and careful parameter initialization, and makes it notoriously hard to train models with saturating nonlinearities. We refer to this phenomenon as *internal covariate shift*, and address the problem by normalizing layer inputs. Our method draws its strength from making normalization a part of the model architecture and performing the normalization for each training mini-batch. Batch Normalization achieves the same accuracy with 14 times fewer training steps, and beats the original model by a significant margin. Using an ensemble of batch-normalized networks, we improve upon the best published result on ImageNet classification: reaching 4.9% top-5 validation error (and 4.8% test error), exceeding the accuracy of human raters.

Using mini-batches of examples, as opposed to one example at a time, is helpful in several ways. First, the gradient of the loss over a mini-batch is an estimate of the gradient over the training set, whose quality improves as the batch size increases. Second, computation over a batch can be much more efficient than m computations for individual examples, due to the parallelism afforded by the modern computing platforms.

While stochastic gradient is simple and effective, it requires careful tuning of the model hyper-parameters, specifically the learning rate used in optimization, as well as the initial values for the model parameters. The training is complicated by the fact that the inputs to each layer

Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift

Sergey Ioffe
Google Inc., sioffe@google.com

Christian Szegedy
Google Inc., szegedy@google.com

Abstract

Training Deep Neural Networks is complicated by the fact that the distribution of each layer's inputs changes during training, as the parameters of the previous layers change. This slows down the training by requiring lower learning rates and careful parameter initialization, and makes it notoriously hard to train models with saturating nonlinearities. We refer to this phenomenon as *internal covariate shift*, and address the problem by normalizing layer inputs. Our method draws its strength from making normalization a part of the model architecture and performing the normalization for each training mini-batch. Batch Normalization allows us to use much higher learning rates and be less careful about initialization. It also acts as a regularizer, in some cases eliminating the need for Dropout. Applied to a state-of-the-art image classification model, Batch Normalization achieves the same accuracy with 14 times fewer training steps, and beats the original model by a significant margin. Using an ensemble of batch-normalized networks, we improve upon the best published result on ImageNet classification: reaching 4.9% top-5 validation error (and 4.8% test error), exceeding the accuracy of human raters.

1 Introduction

Deep learning has dramatically advanced the state of the art in vision, speech, and many other areas. Stochastic gradient descent (SGD) has proved to be an effective way of training deep networks, and SGD variants such as momentum (Sutskever et al., 2013) and Adagrad (Duchi et al., 2011) have been used to achieve state of the art performance. SGD optimizes the parameters Θ of the network, so as to minimize the loss

$$\Theta = \arg \min_{\Theta} \frac{1}{N} \sum_{i=1}^N \ell(x_i, \Theta)$$

where $x_{1..N}$ is the training data set. With SGD, the training proceeds in steps, and at each step we consider a *mini-batch* $x_{1..m}$ of size m . The mini-batch is used to approximate the gradient of the loss function with respect to the parameters, by computing

$$\frac{1}{m} \frac{\partial \ell(x_{1..m}, \Theta)}{\partial \Theta}$$

Using mini-batches of examples, as opposed to one example at a time, is helpful in several ways. First, the gradient of the loss over a mini-batch is an estimate of the gradient over the training set, whose quality improves as the batch size increases. Second, computation over a batch can be much more efficient than m computations for individual examples, due to the parallelism afforded by the modern computing platforms.

While stochastic gradient is simple and effective, it requires careful tuning of the model hyper-parameters, specifically the learning rate used in optimization, as well as the initial values for the model parameters. The training is complicated by the fact that the inputs to each layer are affected by the parameters of all preceding layers – so that small changes to the network parameters amplify as the network becomes deeper.

The change in the distributions of layers' inputs presents a problem because the layers need to continuously adapt to the new distribution. When the input distribution to a learning system changes, it is said to experience *covariate shift* (Shimodaira, 2000). This is typically handled via domain adaptation (Jiang, 2008). However, the notion of covariate shift can be extended beyond the learning system as a whole, to apply to its parts, such as a sub-network or a layer. Consider a network computing

$$\ell = F_2(F_1(u, \Theta_1), \Theta_2)$$

where F_1 and F_2 are arbitrary transformations, and the parameters Θ_1, Θ_2 are to be learned so as to minimize the loss ℓ . Learning Θ_2 can be viewed as if the inputs $x = F_1(u, \Theta_1)$ are fed into the sub-network

$$\ell = F_2(x, \Theta_2).$$

For example, a gradient descent step

$$\Theta_2 \leftarrow \Theta_2 - \frac{\alpha}{m} \sum_{i=1}^m \frac{\partial F_2(x_i, \Theta_2)}{\partial \Theta_2}$$

(for batch size m and learning rate α) is exactly equivalent to that for a stand-alone network F_2 with input x . Therefore, the input distribution properties that make training more efficient – such as having the same distribution between the training and test data – apply to training the sub-network as well. As such it is advantageous for the distribution of x to remain fixed over time. Then, Θ_2 does

• Better Optimization Conditioning (e.g. Batch Normalization)

Working ideas on how to train deep architectures

Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun

Microsoft Research

{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

Abstract

Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers—8× deeper than VGG nets [41] but still having lower complexity. An ensemble of these residual nets achieves 3.57% error

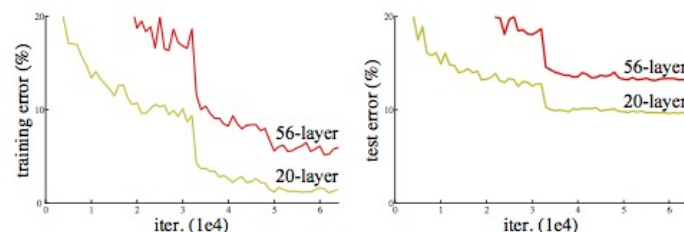


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer “plain” networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

greatly benefited from very deep models.

Driven by the significance of depth, a question arises: Is

Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun

Microsoft Research

{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

Abstract

Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers—8× deeper than VGG nets [41] but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set. This result won the 1st place on the ILSVRC 2015 classification task. We also present analysis on CIFAR-10 with 100 and 1000 layers.

The depth of representations is of central importance for many visual recognition tasks. Solely due to our extremely deep representations, we obtain a 28% relative improvement on the COCO object detection dataset. Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions¹, where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

1. Introduction

Deep convolutional neural networks [22, 21] have led to a series of breakthroughs for image classification [21, 50, 40]. Deep networks naturally integrate low/mid/high-level features [50] and classifiers in an end-to-end multi-layer fashion, and the “levels” of features can be enriched by the number of stacked layers (depth). Recent evidence [41, 44] reveals that network depth is of crucial importance, and the leading results [41, 44, 13, 16] on the challenging ImageNet dataset [36] all exploit “very deep” [41] models, with a depth of sixteen [41] to thirty [16]. Many other non-trivial visual recognition tasks [8, 12, 7, 32, 27] have also

¹<http://image-net.org/challenges/LSVRC/2015/> and <http://mscoco.org/dataset/#detections-challenge2015>.

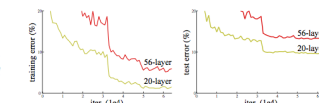


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer “plain” networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

greatly benefited from very deep models.

Driven by the significance of depth, a question arises: Is learning better networks as easy as stacking more layers? An obstacle to answering this question was the notorious problem of vanishing/exploding gradients [1, 9], which hamper convergence from the beginning. This problem, however, has been largely addressed by normalized initialization [23, 9, 37, 13] and intermediate normalization layers [16], which enable networks with tens of layers to start converging for stochastic gradient descent (SGD) with back-propagation [22].

When deeper networks are able to start converging, a degradation problem has been exposed: with the network depth increasing, accuracy gets saturated (which might be unsurprising) and then degrades rapidly. Unexpectedly, such degradation is not caused by overfitting, and adding more layers to a suitably deep model leads to higher training error, as reported in [11, 42] and thoroughly verified by our experiments. Fig. 1 shows a typical example.

The degradation (of training accuracy) indicates that not all systems are similarly easy to optimize. Let us consider a shallower architecture and its deeper counterpart that adds more layers onto it. There exists a solution by construction to the deeper model: the added layers are identity mapping, and the other layers are copied from the learned shallower model. The existence of this constructed solution indicates that a deeper model should produce no higher training error than its shallower counterpart. But experiments show that our current solvers on hand are unable to find solutions that

- Better neural architectures (e.g. **Residual Nets**)

Software

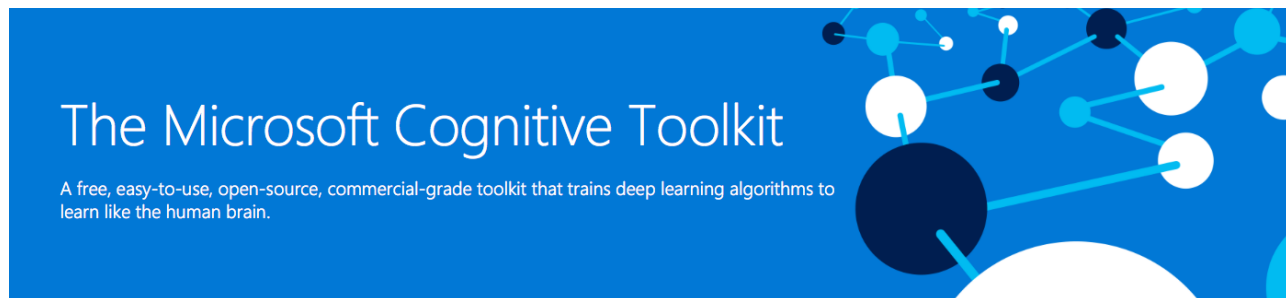


Caffe



Caffe2

MatConvNet



So what is deep learning?

Three key ideas

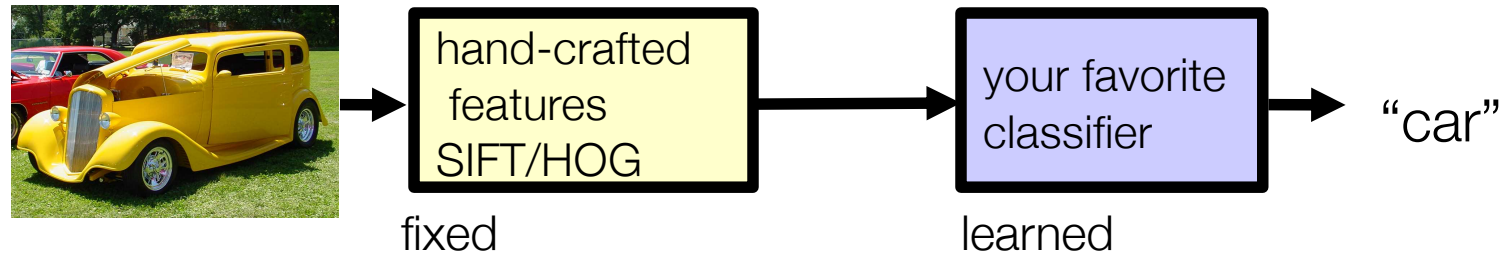
- (Hierarchical) Compositionality
- End-to-End Learning
- Distributed Representations

Three key ideas

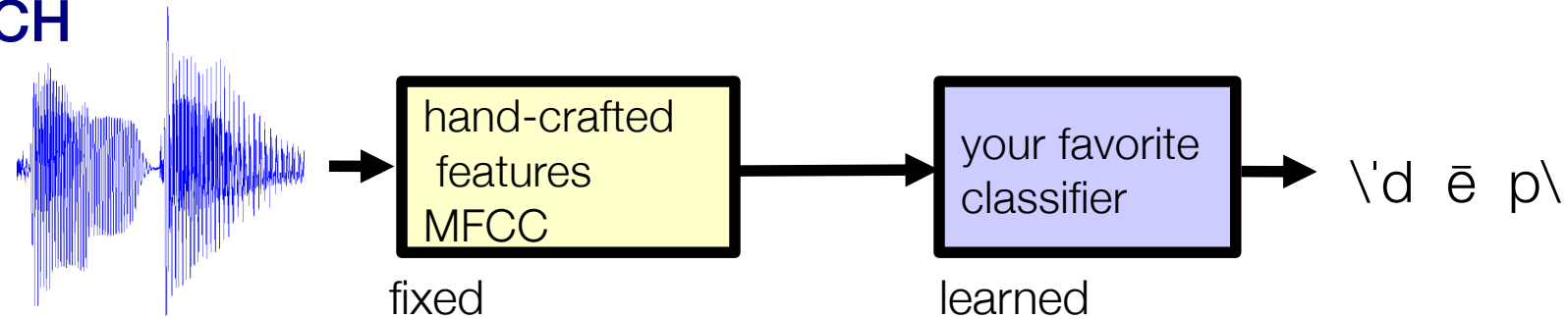
- **(Hierarchical) Compositionality**
 - Cascade of non-linear transformations
 - Multiple layers of representations
- End-to-End Learning
 - Learning (goal-driven) representations
 - Learning to feature extract
- Distributed Representations
 - No single neuron “encodes” everything
 - Groups of neurons work together

Traditional Machine Learning

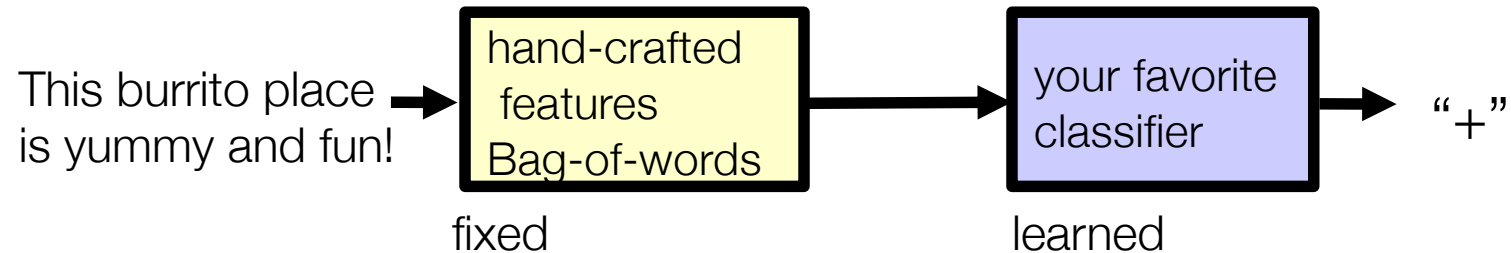
VISION



SPEECH

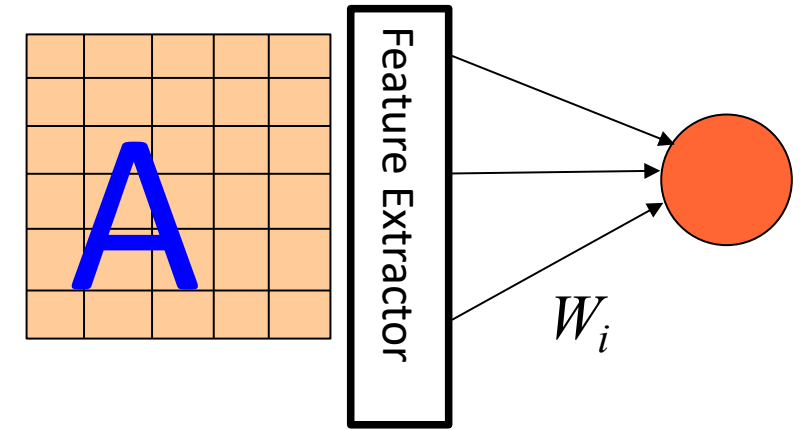


NLP

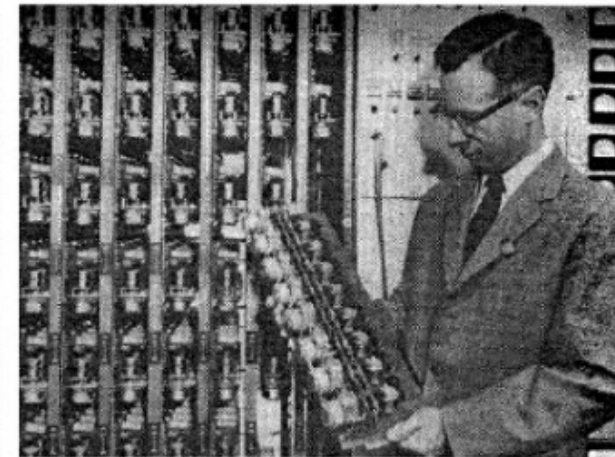
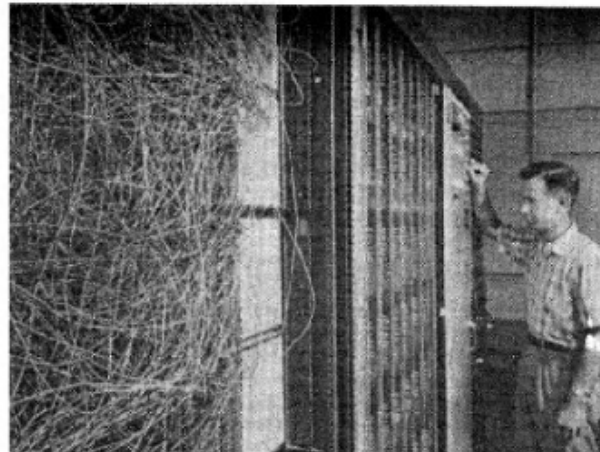
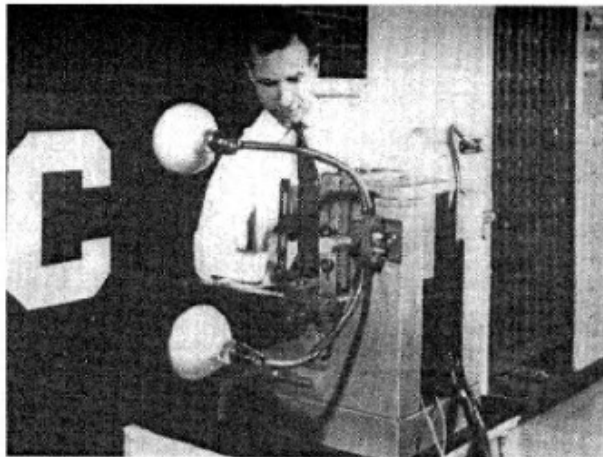


It's an old paradigm

- The first learning machine:
the **Perceptron**
 - Built at Cornell in 1960
- The Perceptron was a **linear classifier** on top of a simple **feature extractor**
- The vast majority of practical applications of ML today use glorified **linear classifiers** or glorified template matching.
- Designing a feature extractor requires considerable efforts by experts.



$$y = \text{sign} \left(\sum_i^N W_i F_i(X) + b \right)$$



Hierarchical Compositionality

VISION

pixels → edge → texon → motif → part → object

SPEECH

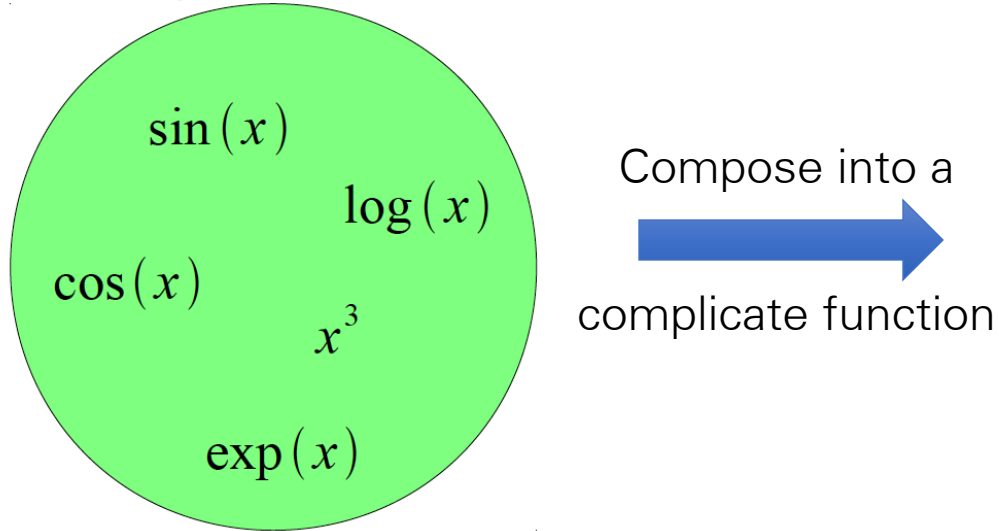
sample → spectral
band → formant → motif → phone → word

NLP

character → word → NP/VP/.. → clause → sentence → story

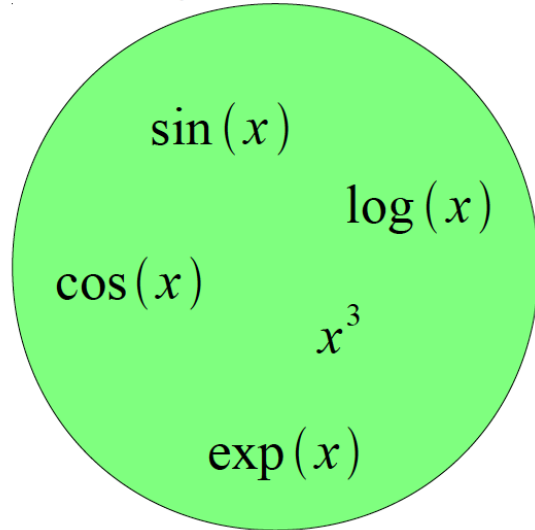
Building A Complicated Function

Given a library of simple functions



Building A Complicated Function

Given a library of simple functions

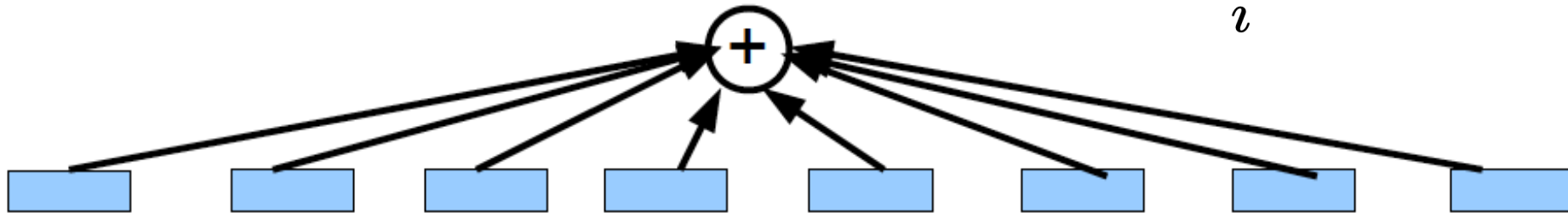


Compose into a
→
complicate function

Idea 1: Linear Combinations

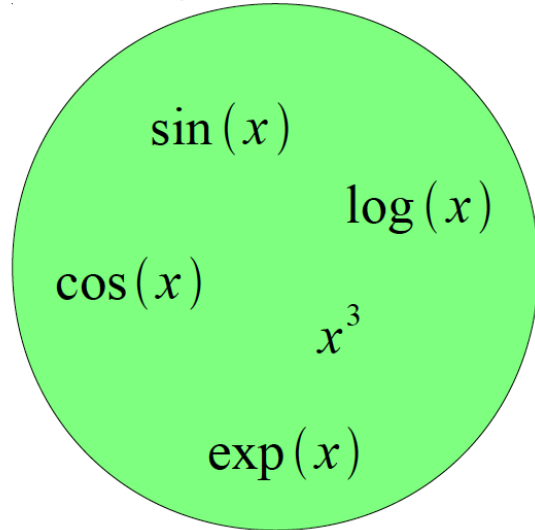
- Boosting
- Kernels
- ...

$$f(x) = \sum_i \alpha_i g_i(x)$$



Building A Complicated Function

Given a library of simple functions



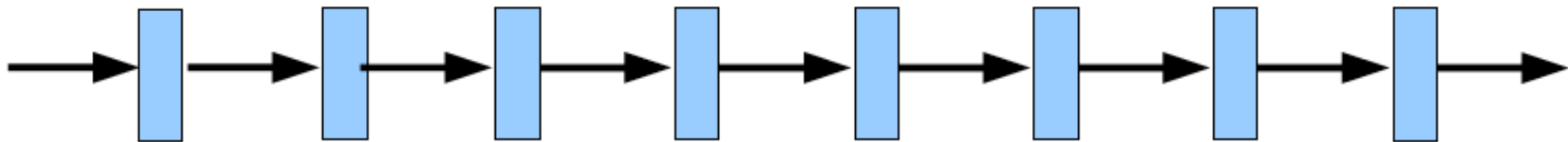
Compose into a

complicate function

Idea 2: Compositions

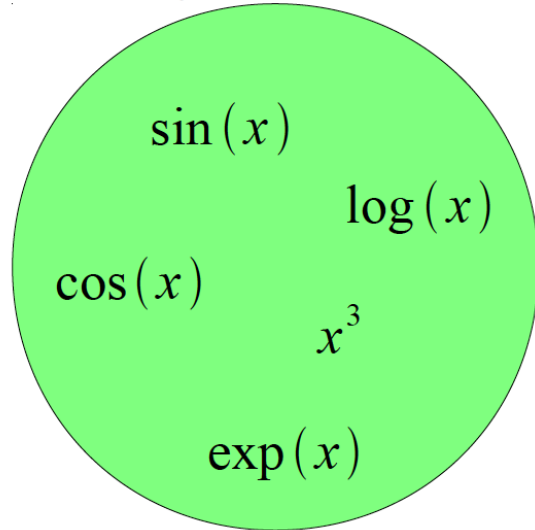
- Deep Learning
- Grammar models
- Scattering transforms...

$$f(x) = g_1(g_2(\dots(g_n(x)\dots)))$$



Building A Complicated Function

Given a library of simple functions

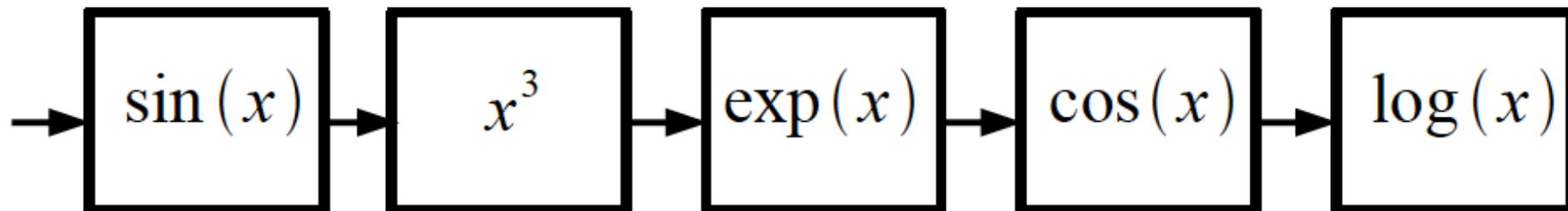


Compose into a
→
complicate function

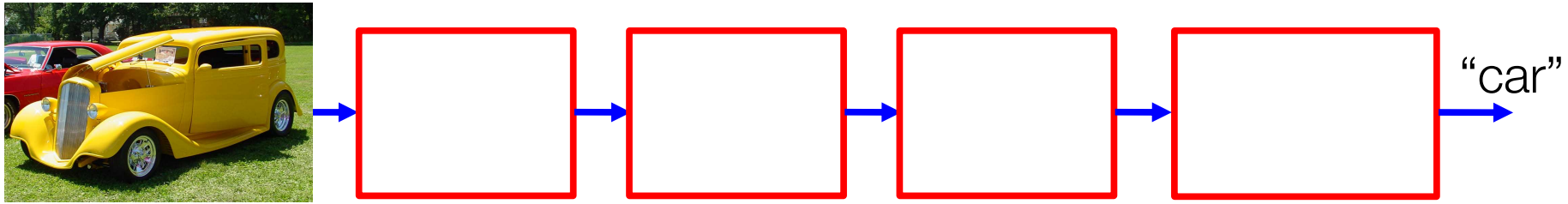
Idea 2: Compositions

- Deep Learning
- Grammar models
- Scattering transforms...

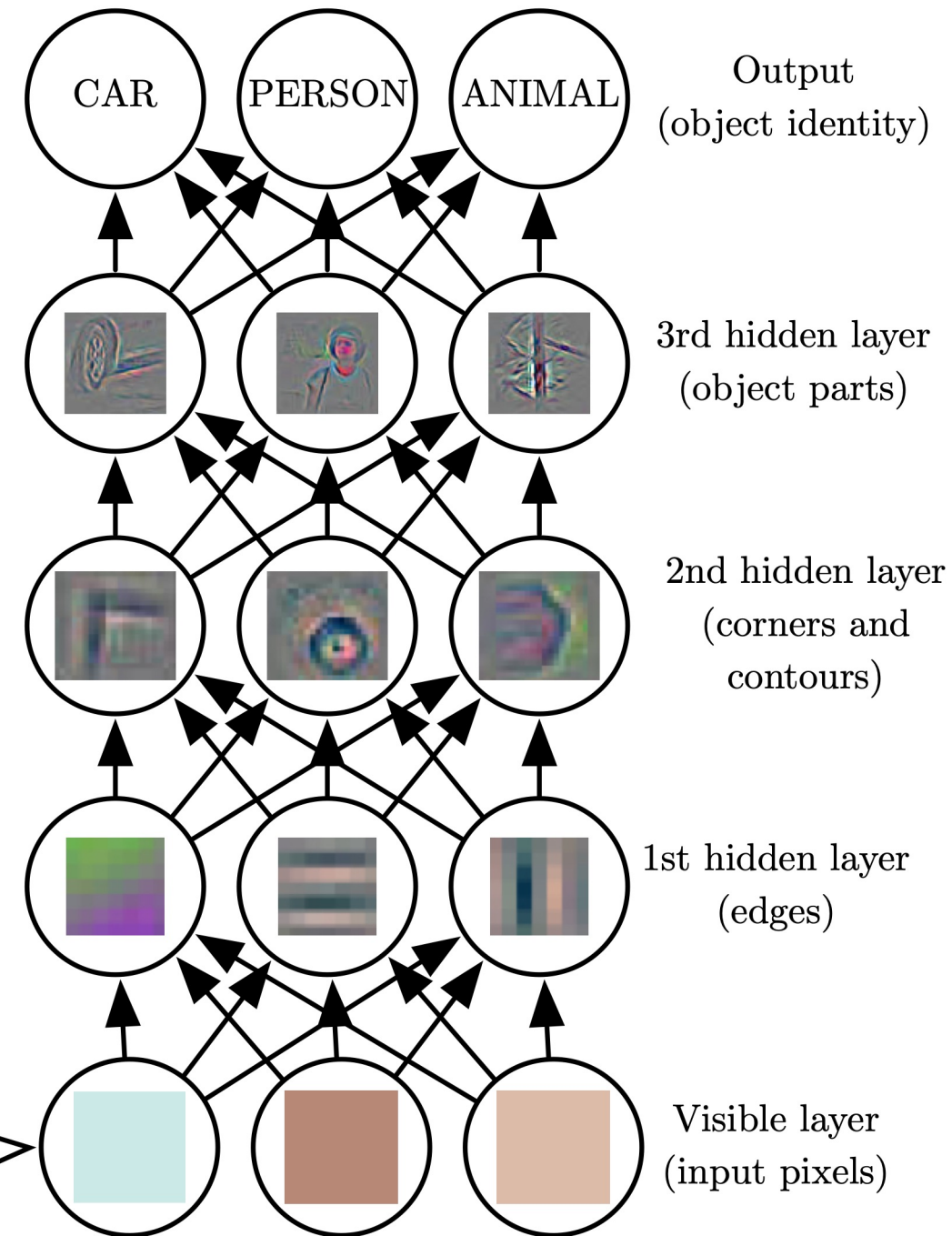
$$f(x) = \log(\cos(\exp(\sin^3(x))))$$



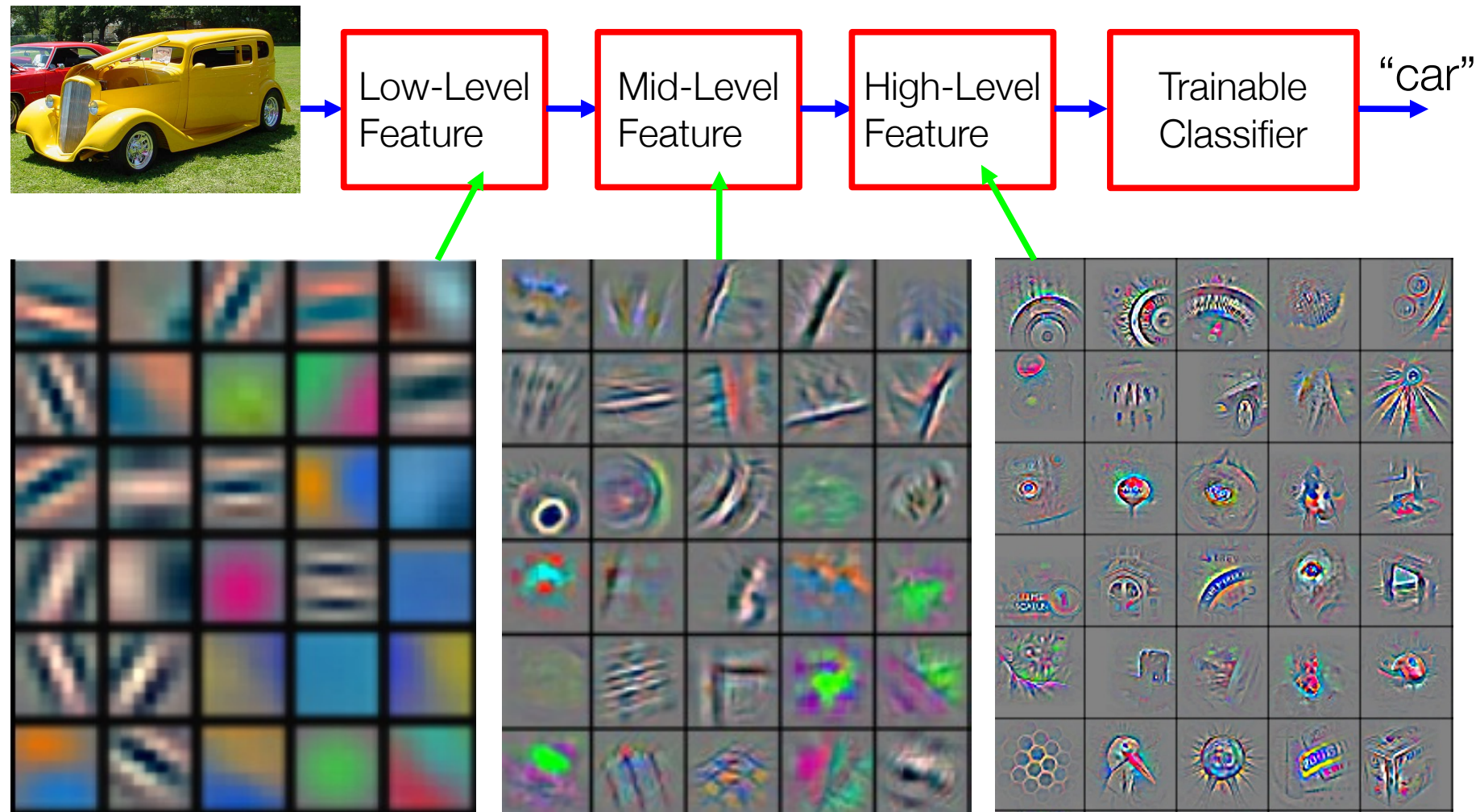
Deep Learning = Hierarchical Compositionality



Deep Learning = Hierarchical Compositionality



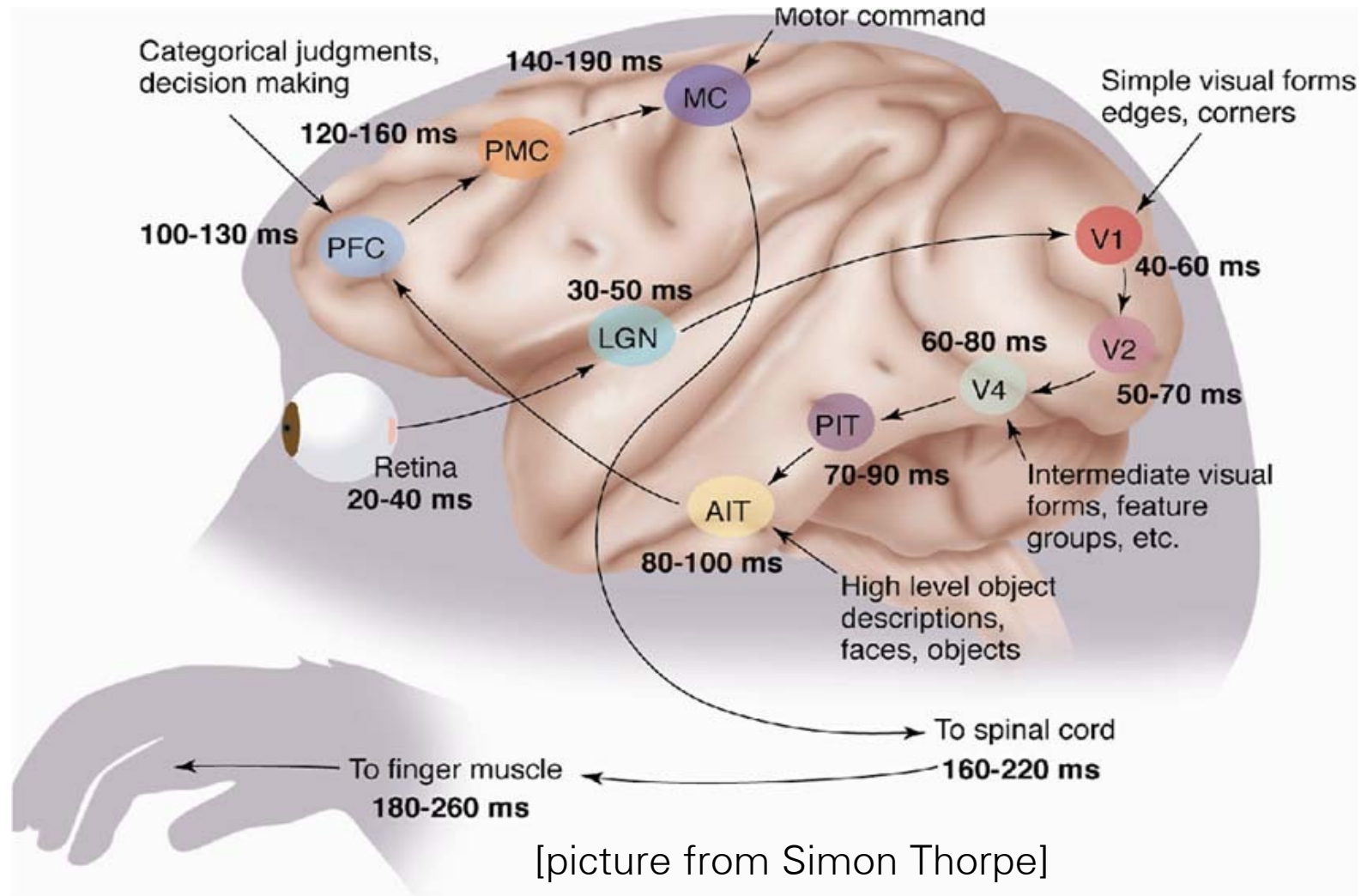
Deep Learning = Hierarchical Compositionality



M.D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks", In ECCV 2014

The Mammalian Visual Cortex is Hierarchical

- The ventral (recognition) pathway in the visual cortex

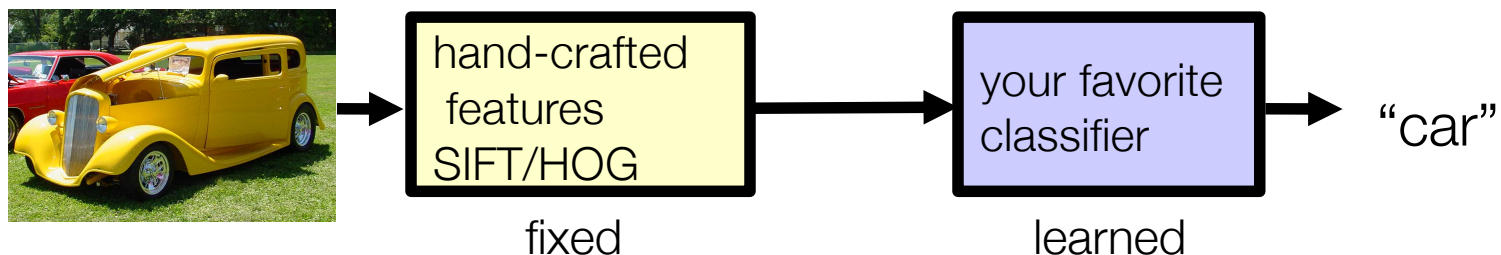


Three key ideas

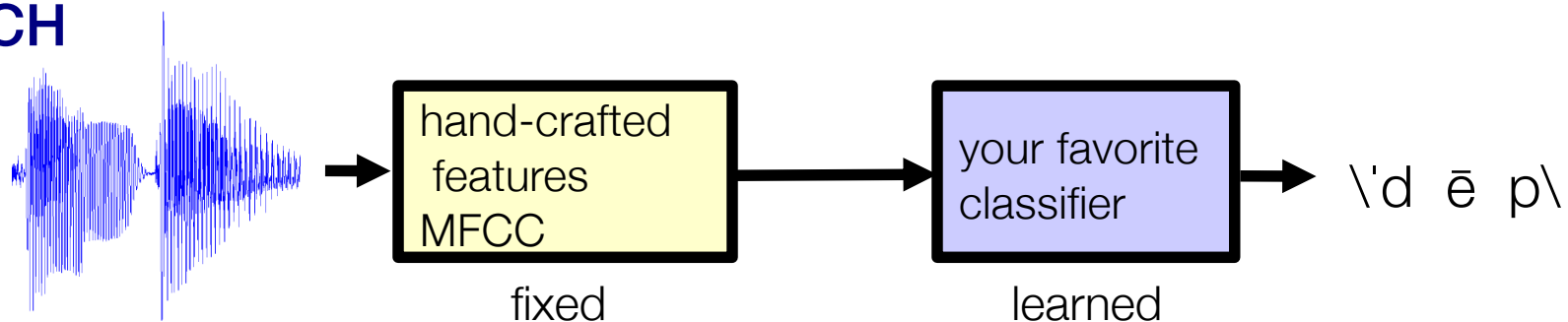
- (Hierarchical) Compositionality
 - Cascade of non-linear transformations
 - Multiple layers of representations
- **End-to-End Learning**
 - Learning (goal-driven) representations
 - Learning to feature extract
- Distributed Representations
 - No single neuron “encodes” everything
 - Groups of neurons work together

Traditional Machine Learning

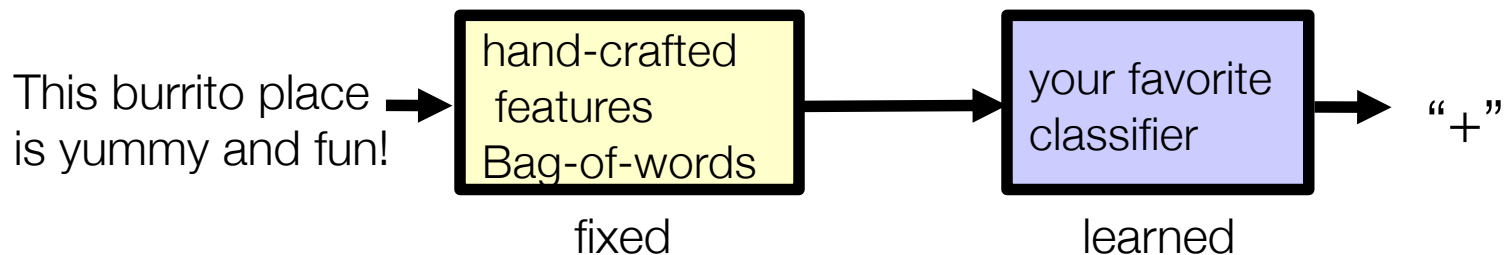
VISION



SPEECH

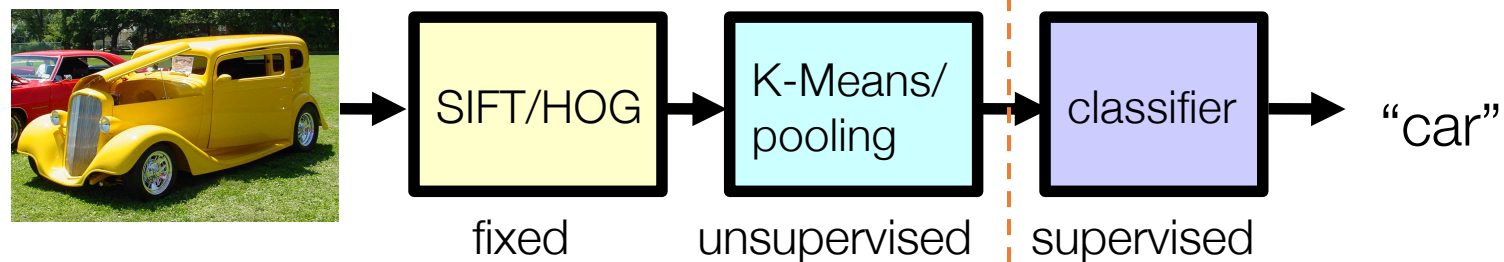


NLP

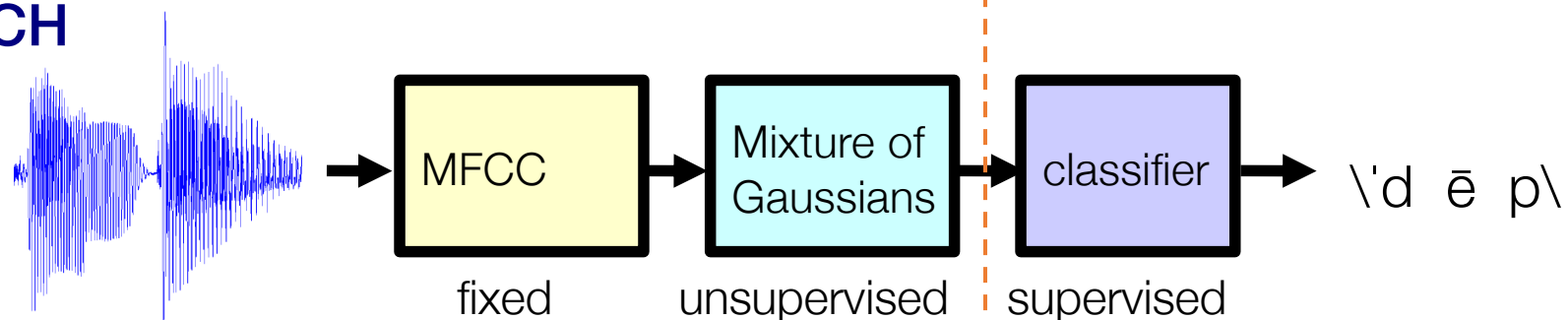


More accurate version

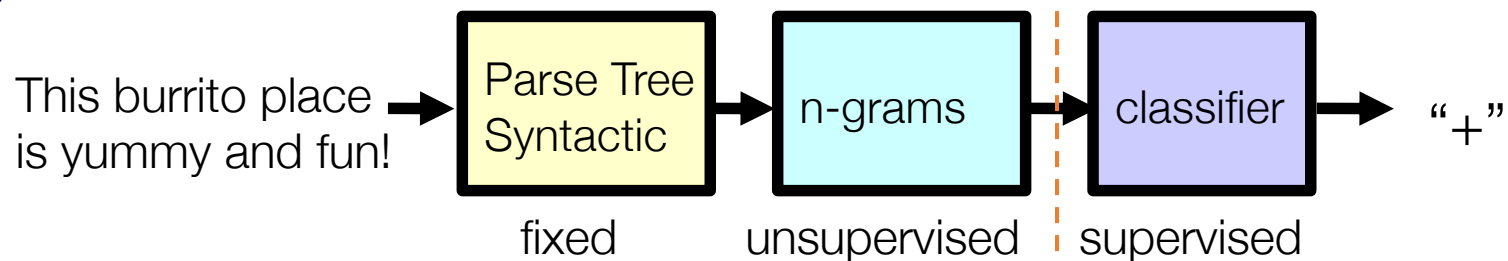
VISION



SPEECH

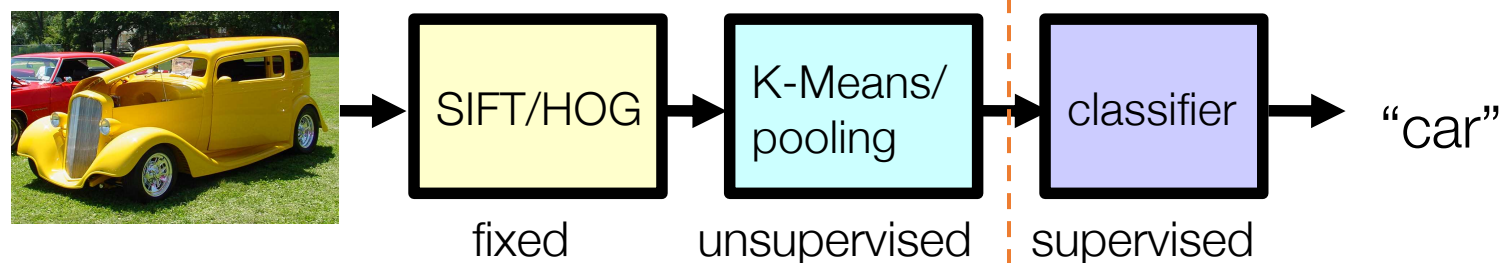


NLP

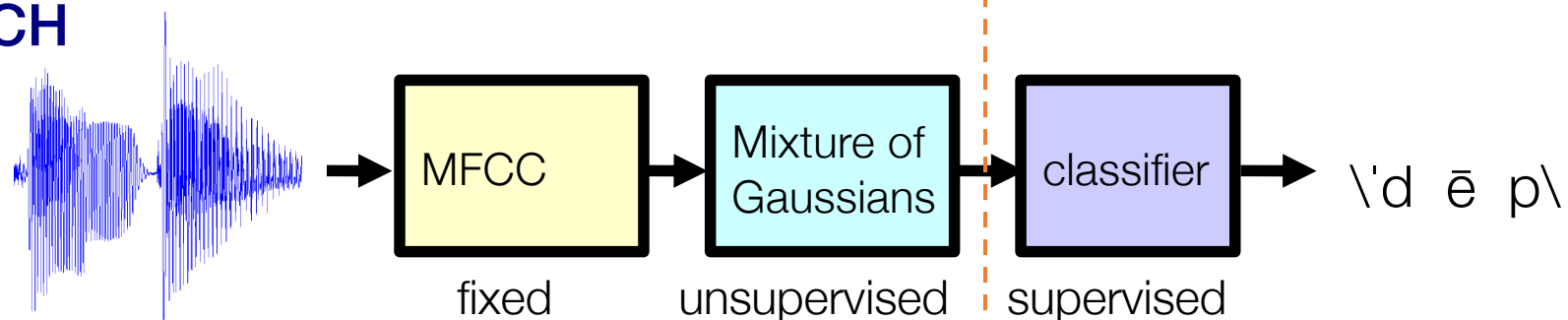


Deep Learning = End-to-End Learning

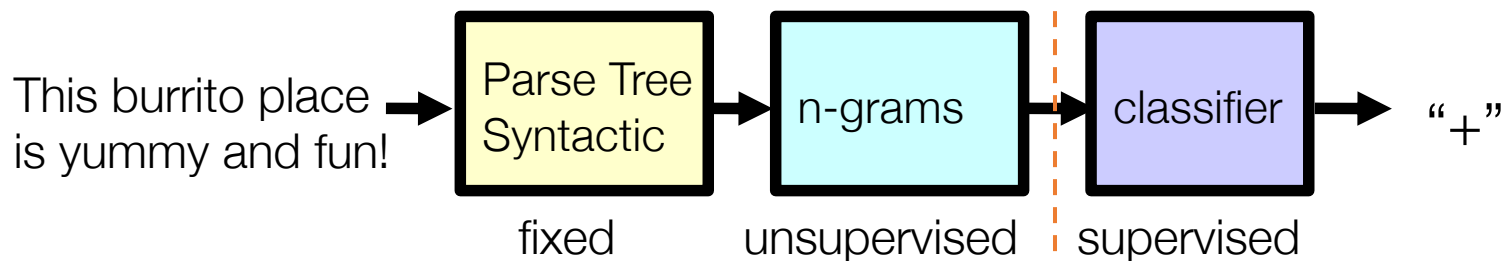
VISION



SPEECH

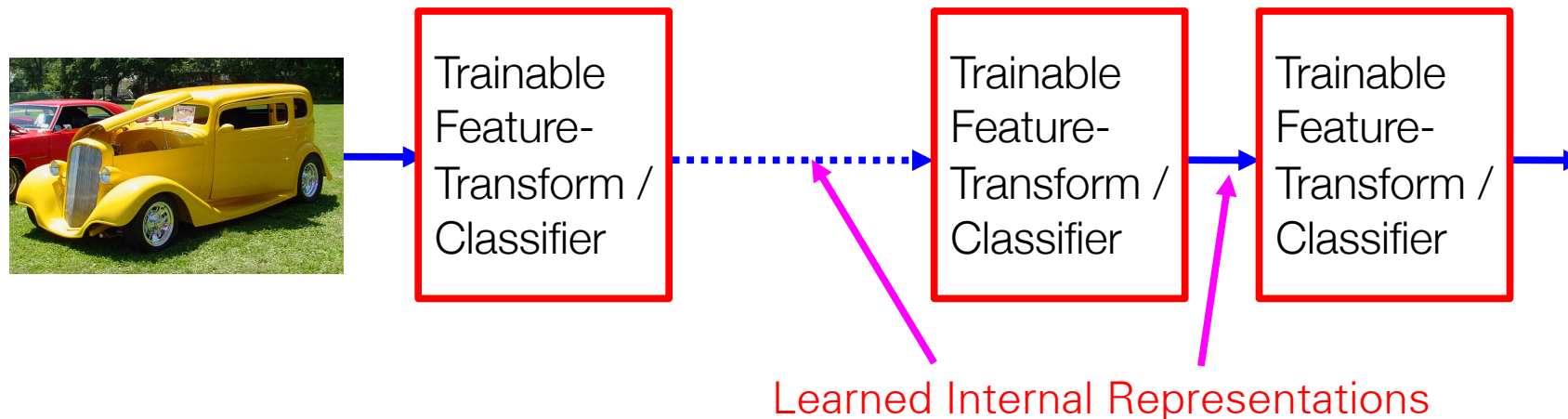


NLP



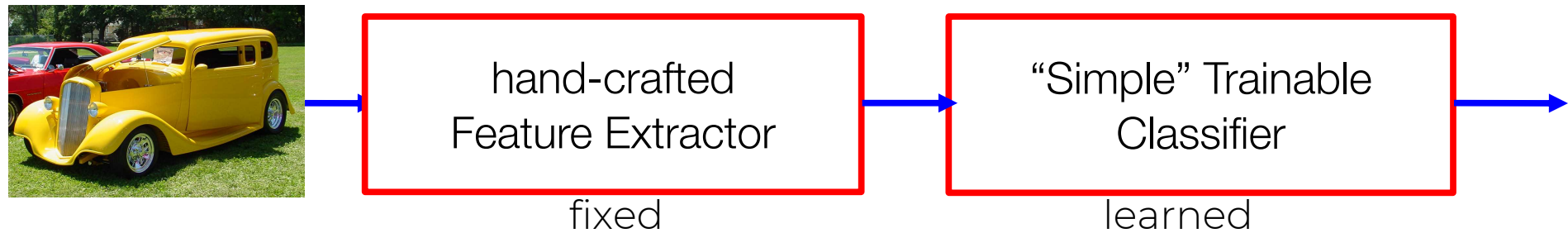
Deep Learning = End-to-End Learning

- A hierarchy of trainable feature transforms
 - Each module transforms its input representation into a higher-level one.
 - High-level features are more global and more invariant
 - Low-level features are shared among categories

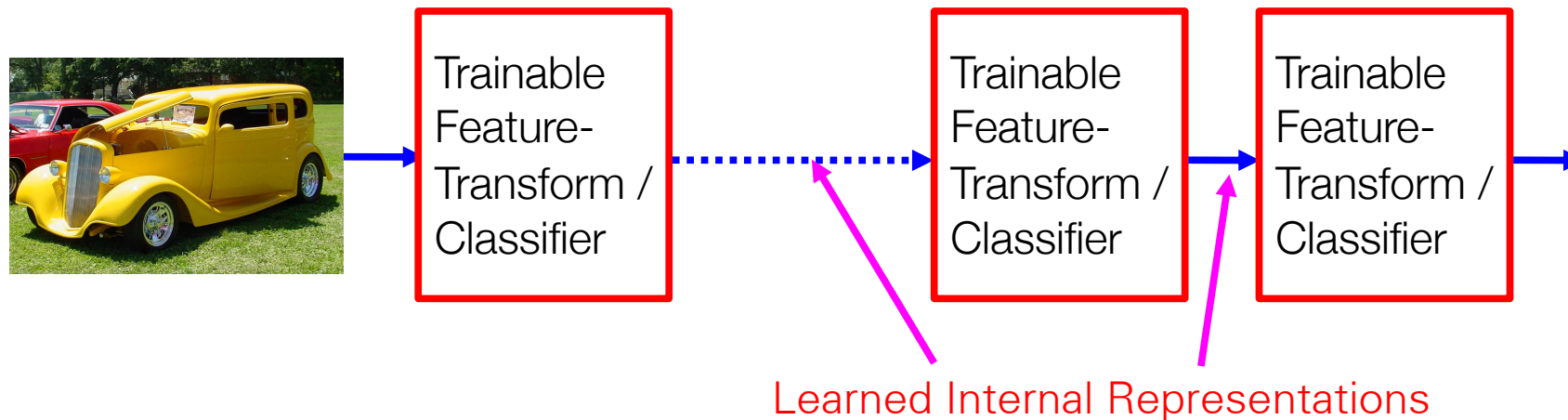


"Shallow" vs Deep Learning

- "Shallow" models



- Deep models



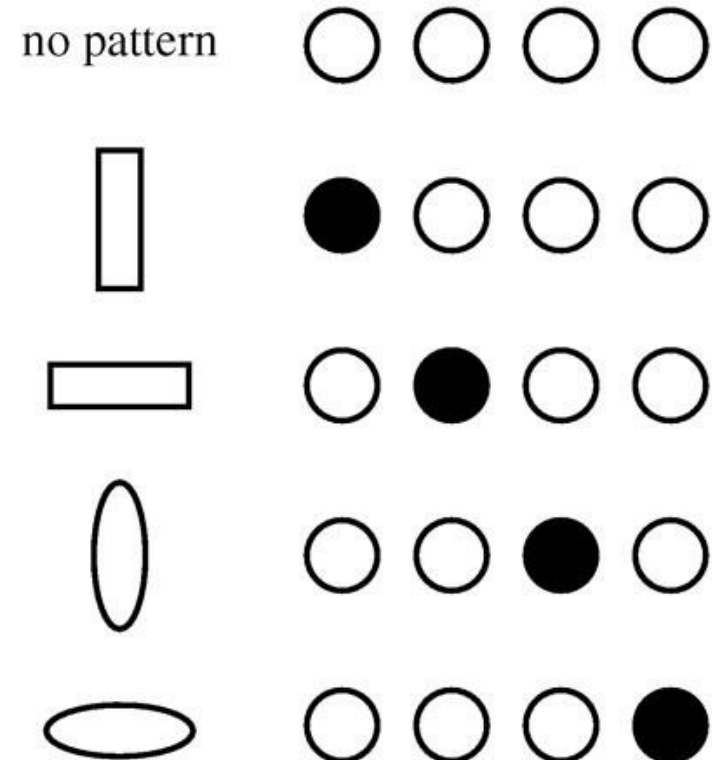
Three key ideas

- (Hierarchical) Compositionality
 - Cascade of non-linear transformations
 - Multiple layers of representations
- End-to-End Learning
 - Learning (goal-driven) representations
 - Learning to feature extract
- **Distributed Representations**
 - No single neuron “encodes” everything
 - Groups of neurons work together

Localist representations

- The simplest way to represent things with neural networks is to **dedicate one neuron to each thing**.
 - Easy to understand.
 - Easy to code by hand
 - Often used to represent inputs to a net
 - Easy to learn
 - This is what mixture models do.
 - Each cluster corresponds to one neuron
 - Easy to associate with other representations or responses.
- But localist models are very inefficient whenever the data has componential structure.

(a)

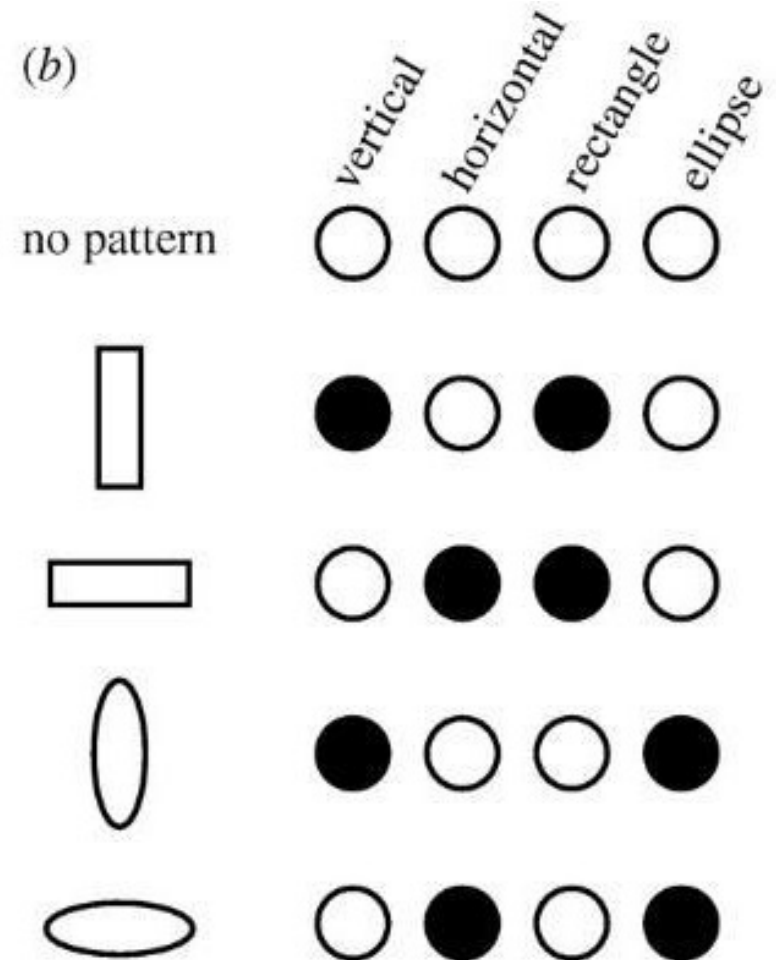


Distributed Representations

- Each neuron must represent something, so this must be a local representation.
- **Distributed representation** means a many-to-many relationship between two types of representation (such as concepts and neurons).
 - Each concept is represented by many neurons
 - Each neuron participates in the representation of many concepts

Local ● ● ○ ● = VR + HR + HE = ?

Distributed ● ● ○ ● = V + H + E ≈ ○



Power of distributed representations!

Scene Classification

bedroom

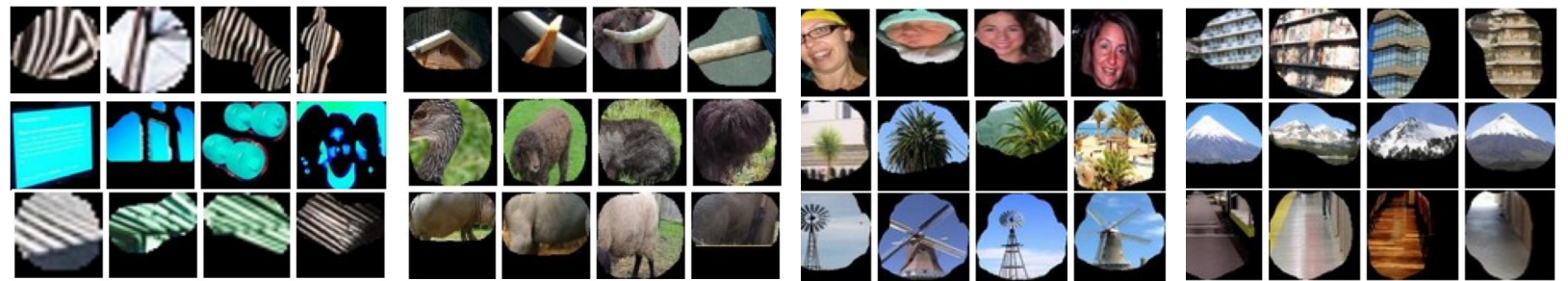


mountain



- Possible internal representations:

- Objects
- Scene attributes
- Object parts
- Textures



Simple elements & colors

Object part

Object

Scene

Three key ideas of deep learning

- **(Hierarchical) Compositionality**
 - Cascade of non-linear transformations
 - Multiple layers of representations
- **End-to-End Learning**
 - Learning (goal-driven) representations
 - Learning to feature extract
- **Distributed Representations**
 - No single neuron “encodes” everything
 - Groups of neurons work together

Three key ideas of deep learning

- **(Hierarchical) Compositionality**

- Cascade of non-linear transformations
- Multiple layers of representations

- ~~End-to-End Learning~~

- ~~Learning (goal-driven) - Not anymore~~
- ~~Learning to feature extract~~

- **Distributed Representations**

- No single neuron “encodes” everything
- Groups of neurons work together

Era of Foundation Models

- We are now in the era of foundation models.
- No more training from scratch!
- Foundation models per domain.
- Take a foundation model and tune it or use it for your own purposes.

Benefits of Deep/Representation Learning

- (Usually) Better Performance
 - “Because gradient descent is better than you”
Yann LeCun
- New domains without “experts”
 - RGBD
 - Multi-spectral data
 - Gene-expression data
 - Unclear how to hand-engineer

Problems with Deep Learning

- **Problem#1: Non-Convex! Non-Convex! Non-Convex!**

- Depth ≥ 3 : most losses non-convex in parameters
- Theoretically, all bets are off
- Leads to stochasticity
 - different initializations $\rightarrow$ different local minima

- Standard response #1

- “Yes, but all interesting learning problems are non-convex”
- For example, human learning
 - Order matters $\rightarrow$ wave hands $\rightarrow$ non-convexity

- Standard response #2

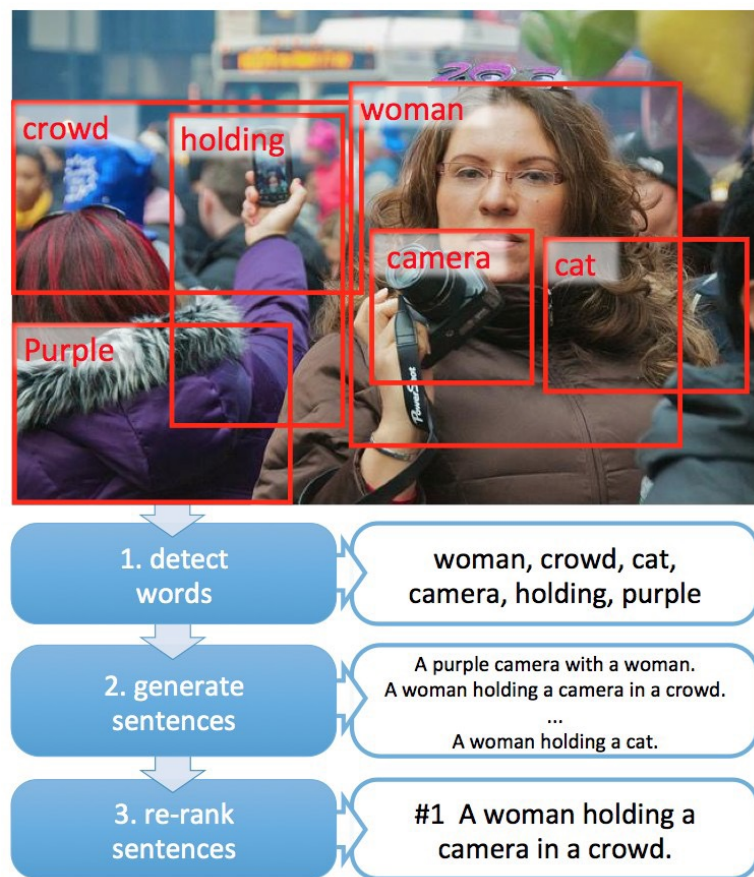
- “Yes, but it often works!”

Problems with Deep Learning

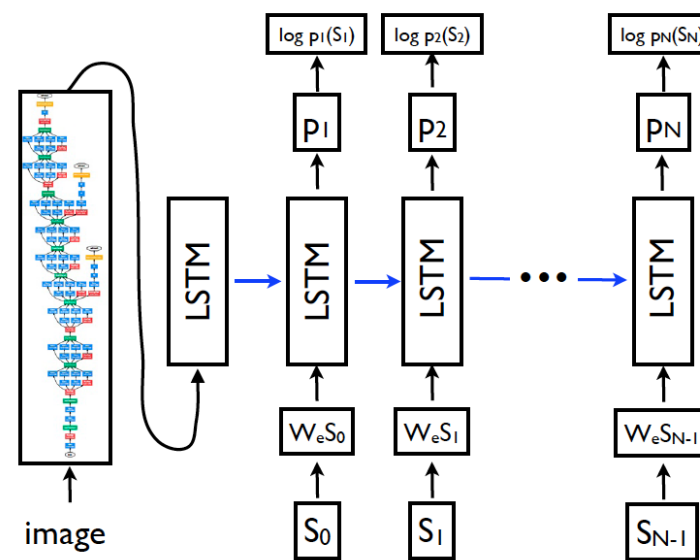
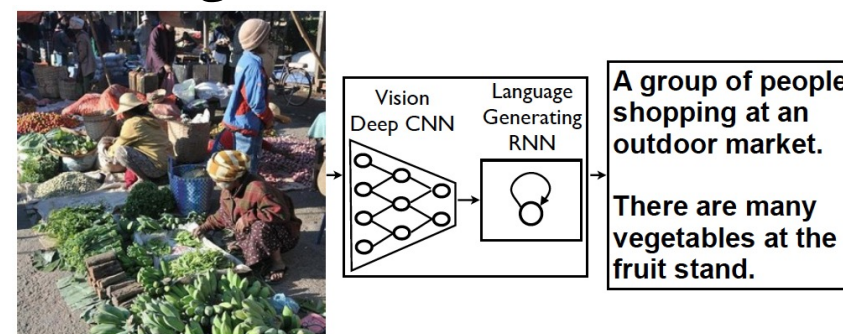
- **Problem#2: Hard to track down what's failing**
 - Pipeline systems have “oracle” performances at each step
 - In end-to-end systems, it's hard to know why things are not working

Problems with Deep Learning

- **Problem#2: Hard to track down what's failing**



[Fang et al. CVPR15]



[Vinyals et al. CVPR15]

Pipeline

End-to-End

Problems with Deep Learning

- **Problem#2: Hard to track down what's failing**

- Pipeline systems have “oracle” performances at each step
- In end-to-end systems, it's hard to know why things are not working

- Standard response #1

- Tricks of the trade: visualize features, add losses at different layers, pre-train to avoid degenerate initializations...
- “We're working on it”

- Standard response #2

- “Yes, but it often works!”

Problems with Deep Learning

- **Problem#3: Lack of easy reproducibility**
 - Direct consequence of stochasticity & non-convexity
- Standard response #1
 - It's getting much better
 - Standard toolkits/libraries/frameworks now available
- Standard response #2
 - "Yes, but it often works!"

NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo
of Computer Designed to
Read and Grow Wiser

WASHINGTON, July 7 (UPI)—The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's \$2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of \$100,000.

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human be-

ings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

Without Human Controls

The Navy said the perceptron would be the first non-living mechanism "capable of receiving, recognizing and identifying its surroundings without any human training or control."

The "brain" is designed to remember images and information it has perceived itself. Ordinary computers remember only what is fed into them on punch cards or magnetic tape.

Later Perceptrons will be able to recognize people and call out their names and instantly translate speech in one language to speech or writing in another language, it was predicted.

Mr. Rosenblatt said in principle it would be possible to build brains that could reproduce themselves on an assembly line and which would be conscious of their existence.

1958 New York Times...

In today's demonstration, the "704" was fed two cards, one with squares marked on the left side and the other with squares on the right side.

Learns by Doing

In the first fifty trials, the machine made no distinction between them. It then started registering a "Q" for the left squares and "O" for the right squares.

Dr. Rosenblatt said he could explain why the machine learned only in highly technical terms. But he said the computer had undergone a "self-induced change in the wiring diagram."

The first Perceptron will have about 1,000 electronic "association cells" receiving electrical impulses from an eye-like scanning device with 400 photo-cells. The human brain has 10,000,000,000 responsive cells, including 100,000,000 connections with the eyes.

The New York Times

Science

[WORLD](#)[U.S.](#)[N.Y. / REGION](#)[BUSINESS](#)[TECHNOLOGY](#)[SCIENCE](#)[HEALTH](#)[SPORTS](#)[OPINION](#)[ENVIRONMENT](#) [SPACE & COSMOS](#)

COMPUTER SCIENTISTS STYMIED IN THEIR QUEST TO MATCH HUMAN VISION

By WILLIAM J. BROAD

Published: September 25, 1984

EXPERTS pursuing one of man's most audacious dreams - to create machines that think - have stumbled while taking what seemed to be an elementary first step. They have failed to master vision.

After two decades of research, they have yet to teach machines the seemingly simple act of being able to recognize everyday objects and to distinguish one from another.

Instead, they have developed a profound new respect for the sophistication of human sight and have scoured such fields as mathematics, physics, biology and psychology for clues to help them achieve the goal of machine vision.

[FACEBOOK](#)[TWITTER](#)[GOOGLE+](#)[EMAIL](#)[SHARE](#)[PRINT](#)[REPRINTS](#)

SCIENCE

Researchers Announce Advance in Image-Recognition Software

By JOHN MARKOFF NOV. 17, 2014

 Email

 Share

 Tweet

 Save

 More

MOUNTAIN VIEW, Calif. — Two groups of scientists, working independently, have created artificial intelligence software capable of recognizing and describing the content of photographs and videos with far greater accuracy than ever before, sometimes even mimicking human levels of understanding.

Until now, so-called computer vision has largely been limited to recognizing individual objects. The new software, described on Monday by researchers at Google and at [Stanford University](#), teaches itself to identify entire scenes: a group of young men playing Frisbee, for example, or a herd of elephants marching on a grassy plain.

The software then writes a caption in English describing the picture. Compared with human observations, the researchers found, the computer-written descriptions are surprisingly accurate.

Captioned by Human and by Google's Experimental Program



Human: "A group of men playing Frisbee in the park."

Computer model: "A group of young people playing a game of Frisbee."

TWEETS
587

FOLLOWING
18

FOLLOWERS
746

FAVORITES
13



INTERESTING.JPG @INTERESTING_JPG · 10h

a man holding a mirror up to his face .



[View more photos and videos](#)

Results from @INTERESTING_JPG via <http://deeplearning.cs.toronto.edu/i2t>

TWEETS
587

FOLLOWING
18

FOLLOWERS
746

FAVORITES
13



INTERESTING.JPG @INTERESTING_JPG · 18h

a man carrying a bucket of his hands in a yard .



[View more photos and videos](#)

Results from @INTERESTING_JPG via <http://deeplearning.cs.toronto.edu/i2t>

TWEETS
587

FOLLOWING
18

FOLLOWERS
746

FAVORITES
13



INTERESTING.JPG @INTERESTING_JPG · Feb 20

a surfboard attached to the top of a car .



[View more photos and videos](#)

TWEETS
587

FOLLOWING
18

FOLLOWERS
746

FAVORITES
13



INTERESTING.JPG @INTERESTING_JPG · Feb 19

a man dressed in uniform is looking at his cell phone .



[View more photos and videos](#)

Results from @INTERESTING_JPG via <http://deeplearning.cs.toronto.edu/i2t>

TWEETS
587

FOLLOWING
18

FOLLOWERS
746

FAVORITES
13



INTERESTING.JPG @INTERESTING_JPG · 16h

this appears to be a small bedroom in the snow .



[View more photos and videos](#)

Results from @INTERESTING_JPG via <http://deeplearning.cs.toronto.edu/i2t>



Iain Murray
@driainmurray

Follow

Today I learned [#googletranslate](#) sometimes decides that "Deutsch" means "English". Machine learning systems need to cope with weird inputs.

Translate

Turn off instant translation

RussianGermanEnglishDetect language

EnglishGermanSpanish

Translate

Deutschland

Deutsch, deutsch, deutsch, deutsch, deutsch, deutsch

Natürlich hat ein Deutscher "Wetten, dass ...?" erfunden
Vielen Dank für die schönen Stunden!
Wir sind die freundlichsten Kunden auf dieser Welt
Wir sind bescheiden, wir haben Geld
Die Allerbesten in jedem Sport
Die Steuern hier sind Weltrekord
Bereisen Sie Deutschland und bleiben Sie hier!
Auf diese Art von Besuchern warten wir
Es kann jeder hier wohnen, dem es gefällt
Wir sind das freundlichste Volk auf dieser Welt

Deutsch, deutsch, deutsch, deutsch

Germany

German, English, German, German, German, and English

Of course a German has "betting that ...?" invented
Thanks for the nice hours!
We are the friendliest customers in this world
We are modest, we have money
The very best in any sport
The taxes here are a world record
Travel to Germany and stay here!
We are waiting for this kind of visitors
Anyone who likes it can live here
We are the friendliest people in this world

English, German, German, and German



Iain Murray
@driainmurray

Academic in Machine Learning and Statistics.

homepages.inf.ed.ac.uk/imurray2/
Joined May 2011



Iain Murray
@driainmurray

 **Follow** 

More fun pushing [#googletranslate](#)'s neural net into weird states. (BTW try GT on real text if you haven't recently. It's often amazing.)

English German Spanish Detect language

German English Spanish

Translate

knife, fork, knife,
(The trailing comma messes this one up.)

19/5000

Messer, Messer, Messer,

English German Spanish Detect language

German English Spanish

Translate

Messer, Gabel, Messer, Messer, Messer,
Messer, Messer, Messer, Messer, Messer

77/5000

Screen monitor styling Projector styling Print styling
← back to. 2010-01-20 with adjustable interlinear. Knife, fork; knife, knife, knife, knife;

RETWEETS
120

LIKES
184



DALL-E 2

Hierarchical Text-Conditional Image Generation with CLIP Latents

Aditya Ramesh*
OpenAI
aramesh@openai.com

Prafulla Dhariwal*
OpenAI
prafulla@openai.com

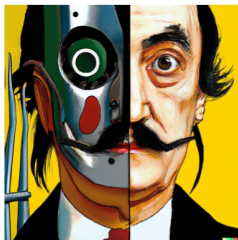
Alex Nichol*
OpenAI
alex@openai.com

Casey Chu*
OpenAI
casey@openai.com

Mark Chen
OpenAI
mark@openai.com

Abstract

Contrastive models like CLIP have been shown to learn robust representations of images that capture both semantics and style. To leverage these representations for image generation, we propose a two-stage model: a prior that generates a CLIP image embedding given a text caption, and a decoder that generates an image conditioned on the image embedding. We show that explicitly generating image representations improves image diversity with minimal loss in photorealism and caption similarity. Our decoders conditioned on image representations can also produce variations of an image that preserve both its semantics and style, while varying the non-essential details absent from the image representation. Moreover, the joint embedding space of CLIP enables language-guided image manipulations in a zero-shot fashion. We use diffusion models for the prior, finding that the latter are computationally more efficient and produce higher-quality samples.



vibrant portrait painting of Salvador Dali with a robotic half face



a shiba inu wearing a beret and black turtleneck



a close up of a handpalm with leaves growing from it



an espresso machine that makes coffee from human souls, artstation



a panda mad scientist mixing sparkling chemicals, artstation



a corgi's head depicted as an explosion of a nebula

Imagen

Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Chitwan Saharia*, William Chan*, Saurabh Saxena†, Lala Li†, Jay Whang†, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho†, David J Fleet†, Mohammad Norouzi*

{sahariac, williamchan, mnorouzi}@google.com
{srbs, lala, jwhang, jonathanho, davidfleet}@google.com

Google Research, Brain Team
Toronto, Ontario, Canada



Sprouts in the shape of text "Imagen" coming out of a fairy tale book.



A photo of a Shiba Inu dog with a backpack riding a bike. It is wearing sunglasses and a beach hat.



A high contrast portrait of a very happy fuzzy panda dressed as a chef in a high end kitchen making dough. There is a painting of flowers on the wall behind him.



Teddy bears swimming at the Olympics 400m Butterfly event.



A cute corgi lives in a house made out of sushi.



A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.



A brain riding a rocketship heading towards the moon.



A dragon fruit wearing karate belt in the snow.



A strawberry mug filled with white sesame seeds. The mug is floating in a dark chocolate sea.

Parti

Scaling Autoregressive Models for Content-Rich Text-to-Image Generation

Jiahui Yu* Yuanzhong Xu† Jing Yu Koh† Thang Luong† Gunjan Baid†
Zirui Wang† Vijay Vasudevan† Alexander Ku†
Yinfei Yang Burcu Karagol Ayan Ben Hutchinson
Wei Han Zarana Parekh Xin Li Han Zhang
Jason Baldridge† Yonghui Wu*
{jiahuiyu, yuanzx, jykoh, thangluong, gunjanbaid, ziruiw, vrv, alexku, jasonbaldridge, yonghui}@google.com[§]

* Equal contribution. † Core contribution.

Google Research

Parti-350M



Parti-750M



Parti-3B



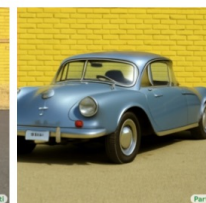
Parti-20B



A portrait photo of a kangaroo wearing an orange hoodie and blue sunglasses standing on the grass in front of the Sydney Opera House holding a sign on the chest that says Welcome Friends!



A green sign that says "Very Deep Learning" and is at the edge of the Grand Canyon. Puffy white clouds are in the sky.



A blue Porsche 356 parked in front of a yellow brick wall.

Stable Diffusion

High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach¹ * Andreas Blattmann¹ * Dominik Lorenz¹ Patrick Esser² Björn Ommer¹

¹Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany ²Runway ML
<https://github.com/CompVis/latent-diffusion>

Abstract

By decomposing the image formation process into a sequential application of denoising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulation allows for a guiding mechanism to control the image generation process without retraining. However, since these models typically operate directly in pixel space, optimization of powerful DMs often consumes hundreds of GPU days and inference is expensive due to sequential evaluations. To enable DM training on limited computational resources while retaining their quality and flexibility, we apply them in the latent space of powerful pretrained autoencoders. In contrast to previous work, training diffusion models on such a representation allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. By introducing cross-attention layers into the model architecture, we turn diffusion models into powerful and flexible generators for general conditioning inputs such as text or bounding boxes and high-resolution synthesis becomes possible in a convolutional manner. Our latent diffusion models (LDMs) achieve new state of the art scores for image inpainting and class-conditional image synthesis and highly competitive performance on various tasks, including unconditional image generation, text-to-image synthesis, and super-resolution, while significantly reducing computational requirements compared to pixel-based DMs.

1. Introduction

Image synthesis is one of the computer vision fields with the most spectacular recent development, but also among those with the greatest computational demands. Especially high-resolution synthesis of complex, natural scenes is presently dominated by scaling up likelihood-based models, potentially containing billions of parameters in autoregressive (AR) transformers [64, 65]. In contrast, the promising results of GANs [3, 26, 39] have been revealed to be mostly confined to data with comparably limited variability as their adversarial learning procedure does not easily scale to modeling complex, multi-modal distributions. Recently, diffusion models [79], which are built from a hierarchy of denoising autoencoders, have shown to achieve impressive

*The first two authors contributed equally to this work.



Figure 1. Boosting the upper bound on achievable quality with less aggressive downsampling. Since diffusion models offer excellent inductive biases for spatial data, we do not need the heavy spatial downsampling of related generative models in latent space, but can still greatly reduce the dimensionality of the data via suitable autoencoding models, see Sec. 3. Images are from the DIV2K [1] validation set, evaluated at 512² px. We denote the spatial downsampling factor by f . Reconstruction FIDs [28] and PSNR are calculated on ImageNet-val. [12]; see also Tab. 8.

results in image synthesis [29, 82] and beyond [7, 44, 47, 56], and define the state-of-the-art in class-conditional image synthesis [15, 30] and super-resolution [70]. Moreover, even unconditional DMs can readily be applied to tasks such as inpainting and colorization [82] or stroke-based synthesis [52], in contrast to other types of generative models [19, 45, 67]. Being likelihood-based models, they do not exhibit mode-collapse and training instabilities as GANs and, by heavily exploiting parameter sharing, they can model highly complex distributions of natural images without involving billions of parameters as in AR models [65].

Democratizing High-Resolution Image Synthesis DMs belong to the class of likelihood-based models, whose mode-covering behavior makes them prone to spend excessive amounts of capacity (and thus compute resources) on modeling imperceptible details of the data [16, 71]. Although the reweighted variational objective [29] aims to address this by undersampling the initial denoising steps, DMs are still computationally demanding, since training and evaluating such a model requires repeated function evaluations (and gradient computations) in the high-dimensional space of RGB images. As an example, training the most powerful DMs often takes hundreds of GPU days (e.g. 150–1000 V100 days in [15]) and repeated evaluations on a noisy version of the input space render also inference expensive,

A high tech solarpunk utopia in the Amazon rainforest

Generate image



Stable Diffusion

High-Resolution Image Synthesis with Latent Diffusion Models

Robin Rombach¹ * Andreas Blattmann¹ * Dominik Lorenz¹ Patrick Esser² Björn Ommer¹

¹Ludwig Maximilian University of Munich & IWR, Heidelberg University, Germany ²Runway ML
<https://github.com/CompVis/latent-diffusion>

Abstract

By decomposing the image formation process into a sequential application of denoising autoencoders, diffusion models (DMs) achieve state-of-the-art synthesis results on image data and beyond. Additionally, their formulation allows for a guiding mechanism to control the image generation process without retraining. However, since these models typically operate directly in pixel space, optimization of powerful DMs often consumes hundreds of GPU days and inference is expensive due to sequential evaluations. To enable DM training on limited computational resources while retaining their quality and flexibility, we apply them in the latent space of powerful pretrained autoencoders. In contrast to previous work, training diffusion models on such a representation allows for the first time to reach a near-optimal point between complexity reduction and detail preservation, greatly boosting visual fidelity. By introducing cross-attention layers into the model architecture, we turn diffusion models into powerful and flexible generators for general conditioning inputs such as text or bounding boxes and high-resolution synthesis becomes possible in a convolutional manner. Our latent diffusion models (LDMs) achieve new state of the art scores for image inpainting and class-conditional image synthesis and highly competitive performance on various tasks, including unconditional image generation, text-to-image synthesis, and super-resolution, while significantly reducing computational requirements compared to pixel-based DMs.

1. Introduction

Image synthesis is one of the computer vision fields with the most spectacular recent development, but also among those with the greatest computational demands. Especially high-resolution synthesis of complex, natural scenes is presently dominated by scaling up likelihood-based models, potentially containing billions of parameters in autoregressive (AR) transformers [64, 65]. In contrast, the promising results of GANs [3, 26, 39] have been revealed to be mostly confined to data with comparably limited variability as their adversarial learning procedure does not easily scale to modeling complex, multi-modal distributions. Recently, diffusion models [79], which are built from a hierarchy of denoising autoencoders, have shown to achieve impressive

*The first two authors contributed equally to this work.



Figure 1. Boosting the upper bound on achievable quality with less aggressive downsampling. Since diffusion models offer excellent inductive biases for spatial data, we do not need the heavy spatial downsampling of related generative models in latent space, but can still greatly reduce the dimensionality of the data via suitable autoencoding models, see Sec. 3. Images are from the DIV2K [1] validation set, evaluated at 512² px. We denote the spatial downsampling factor by f . Reconstruction FIDs [28] and PSNR are calculated on ImageNet-val. [12]; see also Tab. 8.

results in image synthesis [29, 82] and beyond [7, 44, 47, 56], and define the state-of-the-art in class-conditional image synthesis [15, 30] and super-resolution [70]. Moreover, even unconditional DMs can readily be applied to tasks such as inpainting and colorization [82] or stroke-based synthesis [52], in contrast to other types of generative models [19, 45, 67]. Being likelihood-based models, they do not exhibit mode-collapse and training instabilities as GANs and, by heavily exploiting parameter sharing, they can model highly complex distributions of natural images without involving billions of parameters as in AR models [65]. **Democratizing High-Resolution Image Synthesis** DMs belong to the class of likelihood-based models, whose mode-covering behavior makes them prone to spend excessive amounts of capacity (and thus compute resources) on modeling imperceptible details of the data [16, 71]. Although the reweighted variational objective [29] aims to address this by undersampling the initial denoising steps, DMs are still computationally demanding, since training and evaluating such a model requires repeated function evaluations (and gradient computations) in the high-dimensional space of RGB images. As an example, training the most powerful DMs often takes hundreds of GPU days (*e.g.* 150 - 1000 V100 days in [15]) and repeated evaluations on a noisy version of the input space render also inference expensive,

A small cabin on top of a snowy mountain, no blur 4k resolution, ultra detailed

Generate image





Tomer Ullman
@TomerUllman



Do models like DALL-E 2 get basic relations (in/on/etc)?

Colin (Coco) Conwell and I set out to investigate. The result is now on arXiv:

“Testing Relational Understanding in Text-Guided Image Generation”



arxiv.org

Testing Relational Understanding in Text-Guided Image Gen...

Relations are basic building blocks of human cognition.

Classic and recent work suggests that many relations are ...

2:55 PM · Aug 2, 2022 · Twitter Web App

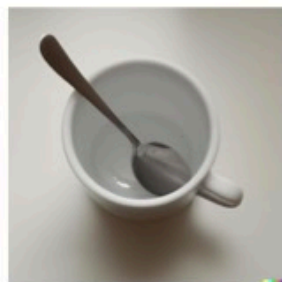
“A spoon in a cup”



“A cup on a spoon”



“A spoon in a cup”



“A cup on a spoon”



This was in 2022, what about now?



“A spoon in a cup”



“A cup on a spoon”



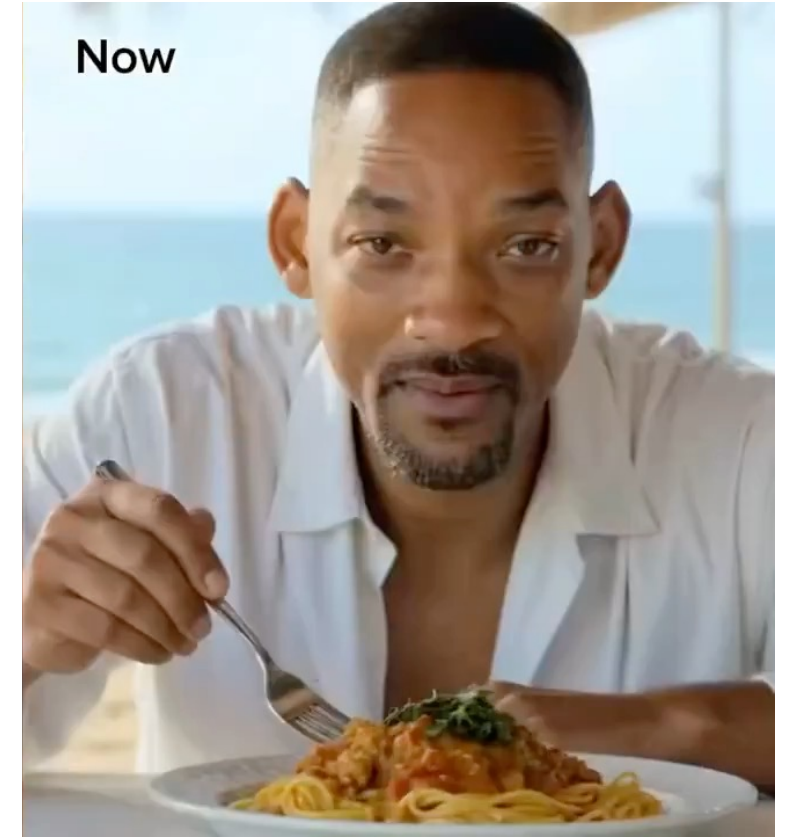
Evolution of Video Generation Models



March 2023
Reddit user "chaindrop"



February 2024
Will Smith's reaction video



January 2025
X user "@Dexerto"



TECHNOLOGY | ARTIFICIAL INTELLIGENCE | KEYWORDS: CHRISTOPHER MIMS

What Are ‘World Models’? The Key to the Next Big AI Leap

To develop knowledge beyond text and videos, AIs must have realistic virtual playgrounds where they can make mistakes and learn



45



Gift unlocked article



Listen (7 min)



Generated by AI



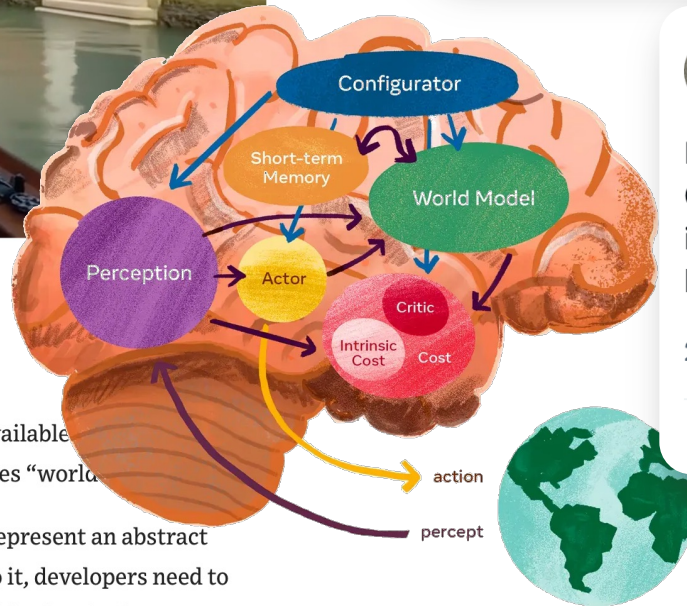
AI-generated 3-D environments rendered via Google's Genie 3 are used to train other AIs.

By [Christopher Mims](#) [Follow](#)

Sept. 26, 2025 at 7:00 pm ET

Today's AIs are book smart. Everything they know they learned from available text and videos. To evolve further, they have to get street smart. That requires “world models.”

The key is enabling AI to learn from their environments and faithfully represent an abstract version of them in their “heads,” the way humans and animals do. To do it, developers need to train AIs by using simulations of the world. Think of it like learning to drive by playing “Gran Turismo.”



FAIR AT META

Yann LeCun

VP & Chief AI Scientist

Demis Hassabis [Verified](#)

@demishassabis



It's kind of mindblowing how good Veo 3 is at modeling intuitive physics. Our world models are getting pretty good, & in my view has important implications regarding the computational complexity of the world - the last line of my bio for me has always been the ultimate quest [👆](#)

2:28 AM · May 24, 2025 · **748.6K** Views

165

417

4.2K

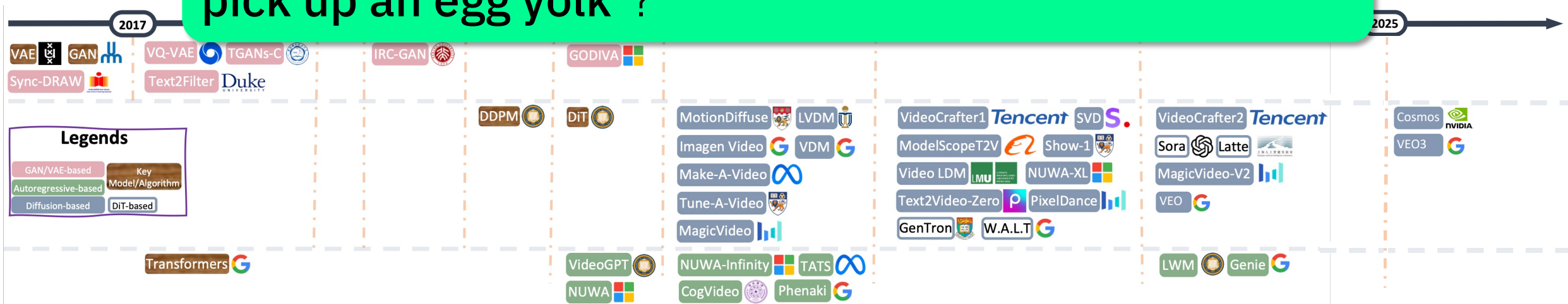
978

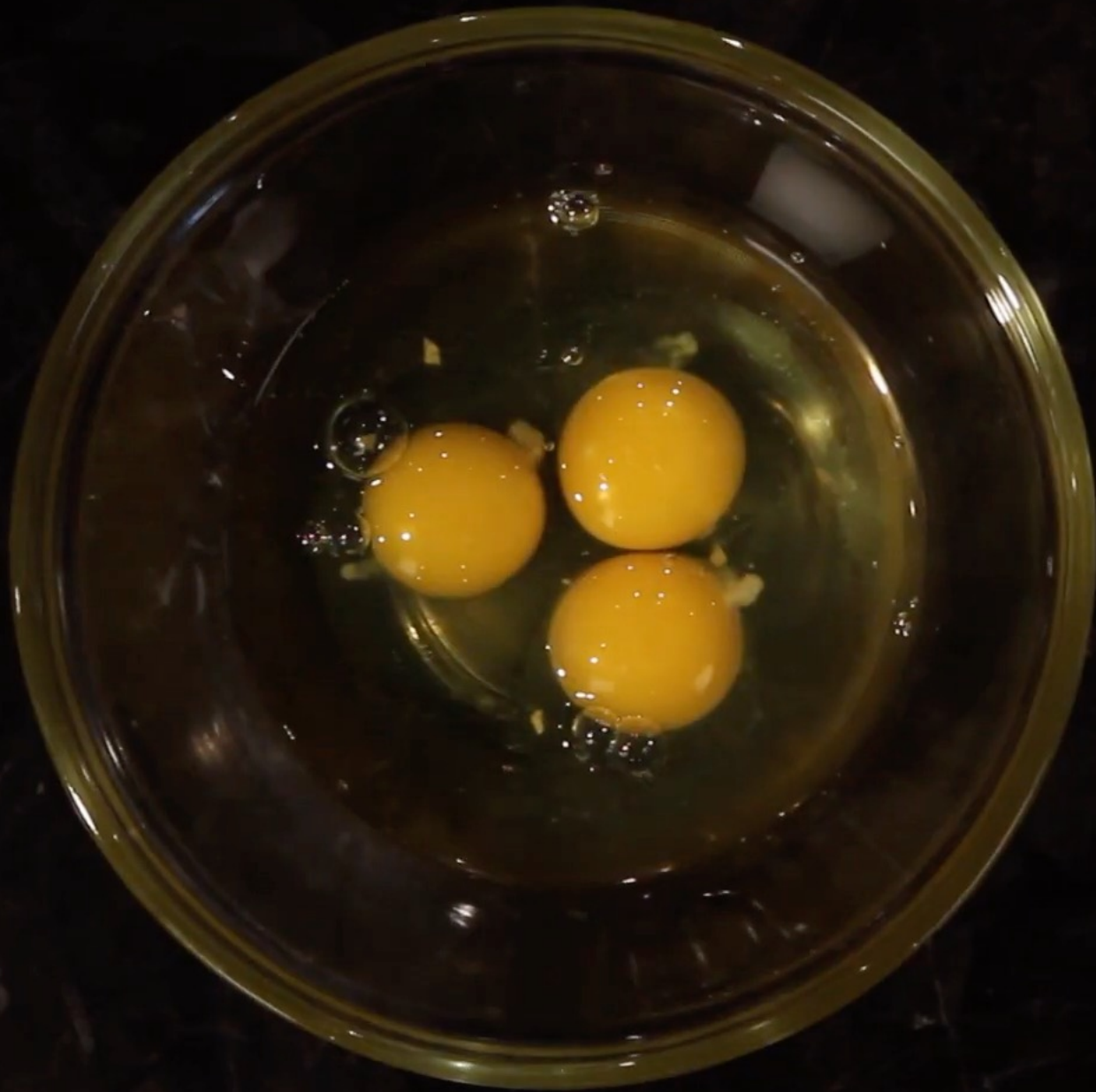


Video Generation Models

- Video generation models have achieved realism and temporal coherence, but lack physical commonsense.
- Many outputs violate intuitive expectations about object behavior, causality, and tool use.
 - Current models: Good visuals, weak on physical plausibility.

Can models capture “Hands squeeze a water bottle to pick up an egg yolk”?





VEO3



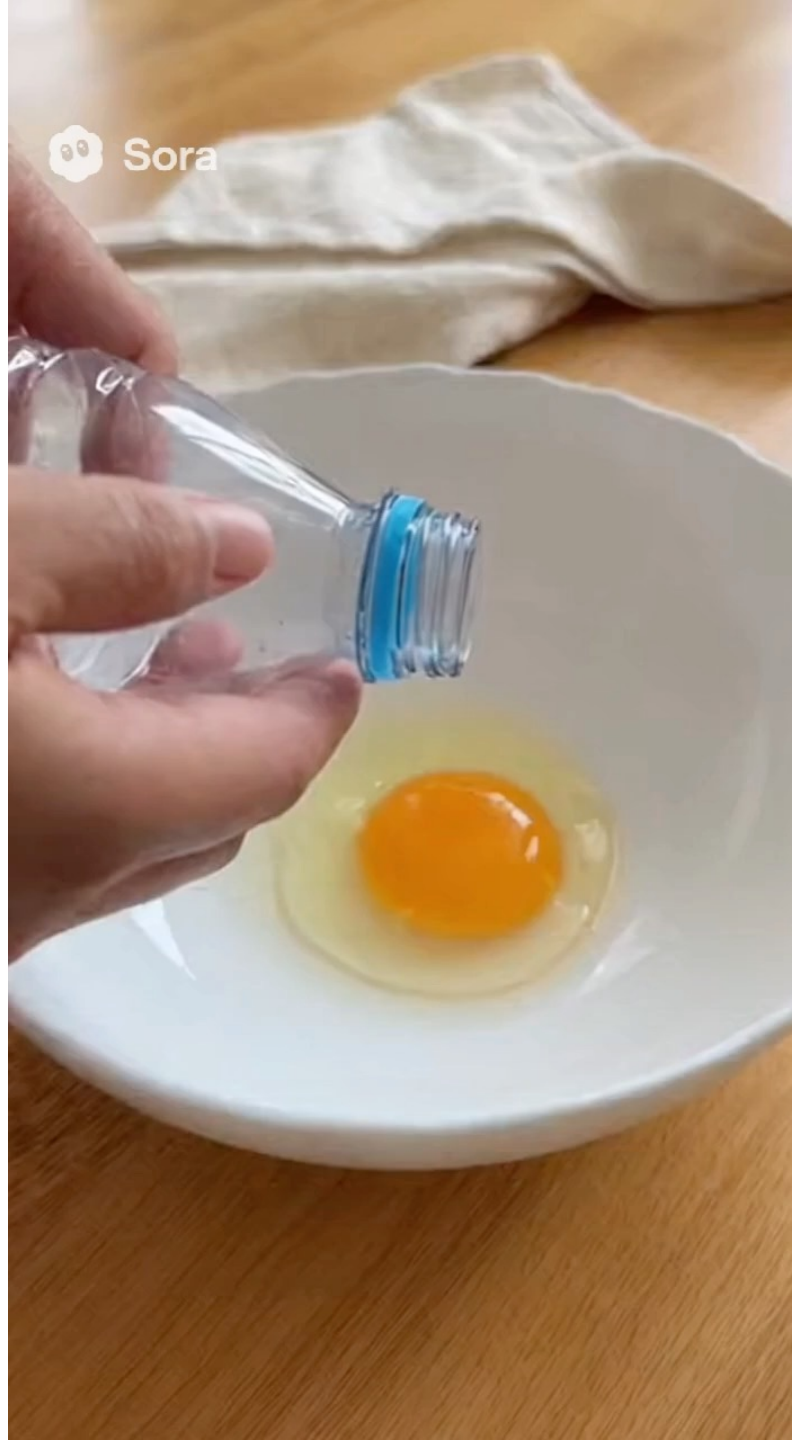
VEO3.1

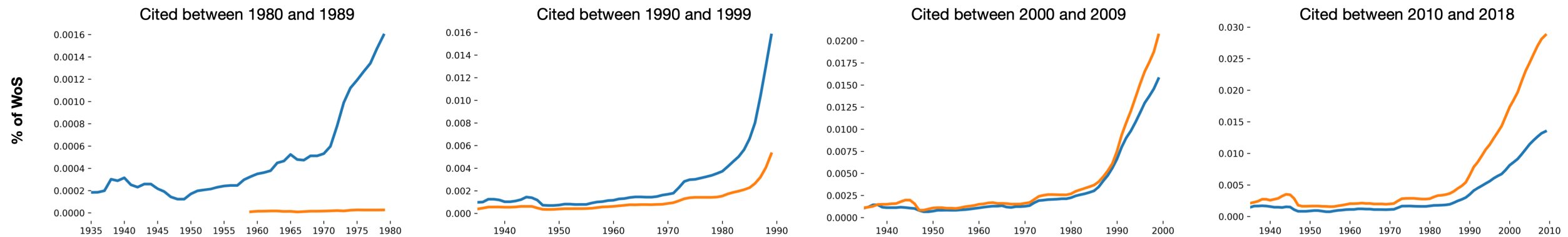
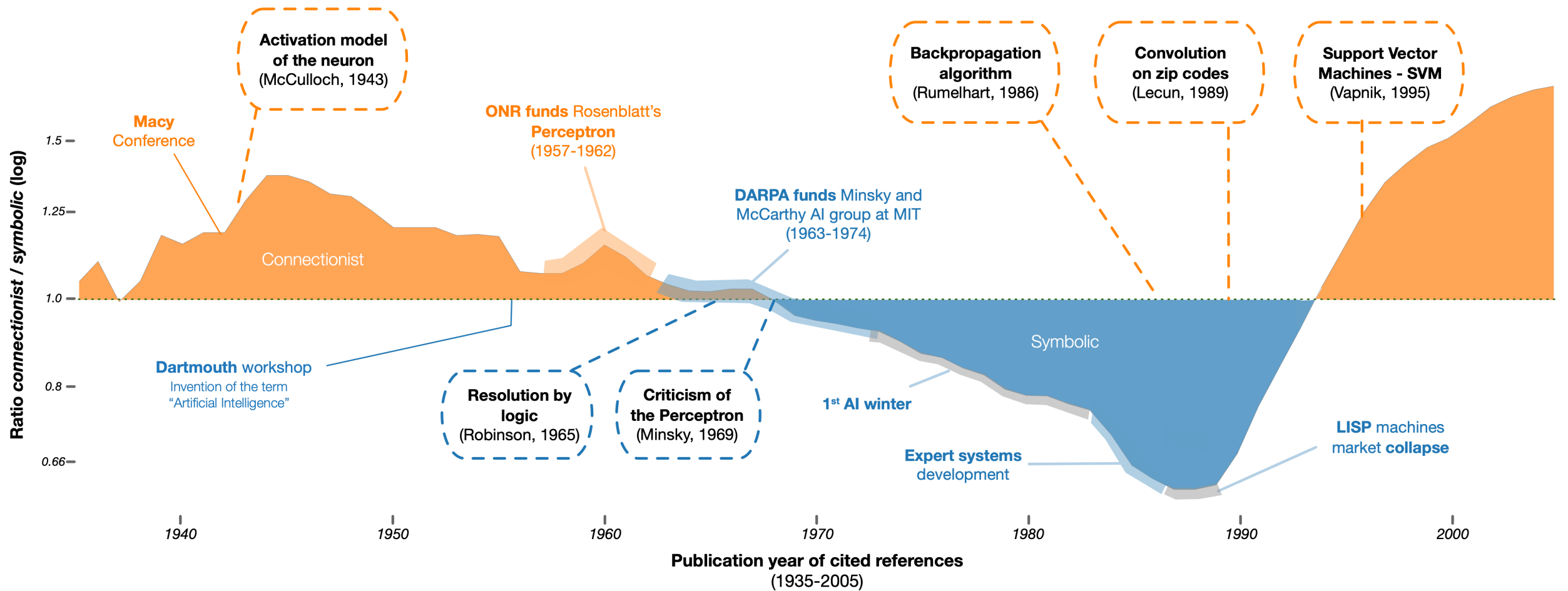


SORA2

Prompt. Hands squeeze an empty plastic water bottle, place the opening over an egg yolk in a bowl, and release the squeeze, sucking the yolk into the bottle, leaving the white behind.

Upsampled Prompt. A person firmly squeezes a translucent, pliable plastic water bottle, reducing its internal volume. They carefully press the bottle's circular opening against the fragile, spherical egg yolk, which sits surrounded by viscous, clear egg whites in a ceramic bowl. The bottle's rim makes full contact with the yolk's surface. The person slowly releases their grip on the bottle, creating suction inside the bottle. The pressure difference causes the yolk to deform slightly as it is drawn upwards into the bottle's opening.





Next Lecture: Machine Learning Overview