

# **TTIC 31230, Fundamentals of Deep Learning**

David McAllester, Autumn 2020

## **The History of Deep Learning**

**and Moore's Law of AI**

## Early History

**1943:** McCulloch and Pitts introduced the linear threshold “neuron”. Words in red are important — we will discuss them in detail in this class.

**1962:** Rosenblatt applies a “Hebbian” learning rule. Novikoff proved the perceptron convergence theorem.

**1969:** Minsky and Papert publish the book *Perceptrons*.

The Perceptrons book greatly discourages work in artificial neural networks. Symbolic methods dominate AI research through the 1970s.

## 80s Renaissance

**1980:** Fukushima introduces the neocognitron — a form of convolutional neural Network or CNN. CNNs created the deep revolution in 2012.

**1984:** Valiant defines PAC learnability and stimulates learning theory. Wins the Turing Award in 2010.

**1985:** Hinton and Sejnowski introduce the Boltzman machine

**1986:** Rummelhart, Hinton and Williams demonstrate empirical success with backpropagation (itself dating back to 1961).

## 90s and 00s: Research In the Shadows

**1997:** Schmidhuber et al. introduce LSTMs (a form of recurrent neural network or RNN).

**1998:** LeCunn draws attention to convolutional neural networks (CNNs) (LeNet).

**2003:** Bengio introduces neural language modeling.

## Current Era

**2012:** Alexnet dominates the Imagenet computer vision challenge.

Google speech recognition converts to deep learning.

Both developments come out of Hinton's group in Toronto.

**2013:** Refinement of AlexNet continues to dramatically improve computer vision.

## Current Era

**2014:** Neural machine translation appears (Seq2Seq models).

Variational auto-encoders (VAEs) appear.

Generative Adversarial Networks (GANs) appear.

Graph neural networks appear (GNNs) revolutionizing the prediction of molecular properties.

Dramatic improvement in computer vision and speech recognition continues.

## Current Era

**2015:** Google converts to neural machine translation leading to dramatic improvements.

Batch Normalization appears improving the performance of image classification.

Residual Connections appear. This makes yet another dramatic improvement in computer vision.

Diffusion Models are formulated which becomes important in 2021.

**2016:** Reinforcement Learning is used to develop Alphago which defeats Lee Sedol.

## Current Era

**2017:** AlphaZero learns both go and chess at super-human levels in a matter of hours entirely from self-play and advances computer go far beyond human abilities.

Unsupervised machine translation is demonstrated.

Progressive GANs demonstrate high resolution realistic face generation.

The **Transformer** appears greatly improving language modeling.

## Current Era

**2018:** Unsupervised pre-training significantly improves a broad range of NLP tasks including question answering (but dialogue remains unsolved).

Contrastive learning is formulated which ultimately becomes the foundation of various systems.

AlphaFold revolutionizes protein structure prediction.

**2019:** Vector quantized VAEs (VQ-VAE) demonstrate that VAEs can be competitive with GANs for high-resolution image generation.

Super-human performance is achieved on the GLUE natural language understanding benchmark.

## Current Era

**2020:** A neural language model (GPT-3) writes a fake blog post that landed in the No. 1 spot on Hacker News.

# Natural Language Understanding

Unsupervised pre-training leads to dramatic improvements on benchmarks for language understanding.

## GLUE: General Language Understanding Evaluation

ArXiv 1804.07461

Corpus	Train	Test	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	<b>1k</b>	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	<b>391k</b>	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	<b>20k</b>	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	<b>146</b>	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

# BERT and GLUE

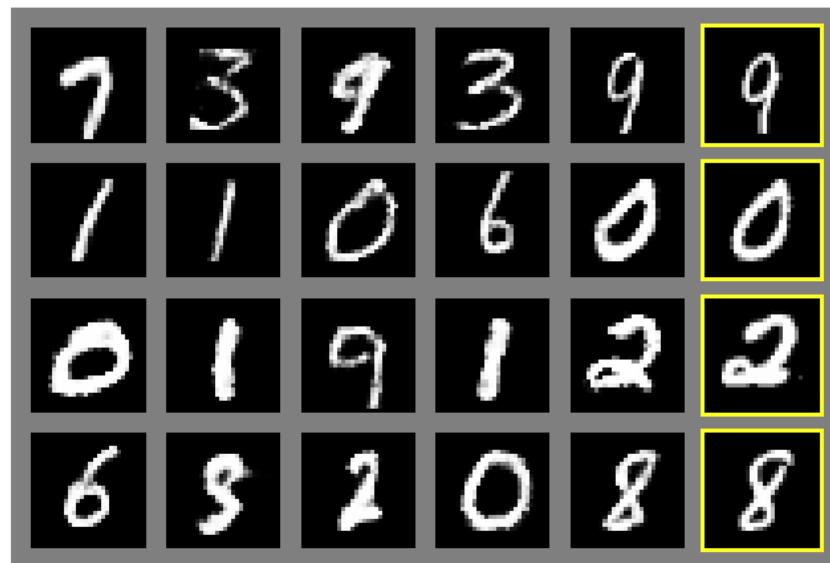
Rank	Name	Model	URL	Score
1	T5 Team - Google	T5		90.3
2	ERNIE Team - Baidu	ERNIE		90.1
3	Microsoft D365 AI & MSR AI & GATECH MT-DNN-SMART			89.9
4	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		89.7
5	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4
6	Junjie Yang	HIRE-RoBERTa		88.3
7	Facebook AI	RoBERTa		88.1
8	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6
9	GLUE Human Baselines	GLUE Human Baselines		87.1

## BERT and SuperGLUE

Rank	Name	Model	URL	Score
1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8
2	T5 Team - Google	T5		89.3
3	Zhuiyi Technology	RoBERTa-mtl-adv		85.7
4	Facebook AI	RoBERTa		84.6
5	IBM Research AI	BERT-mtl		73.5

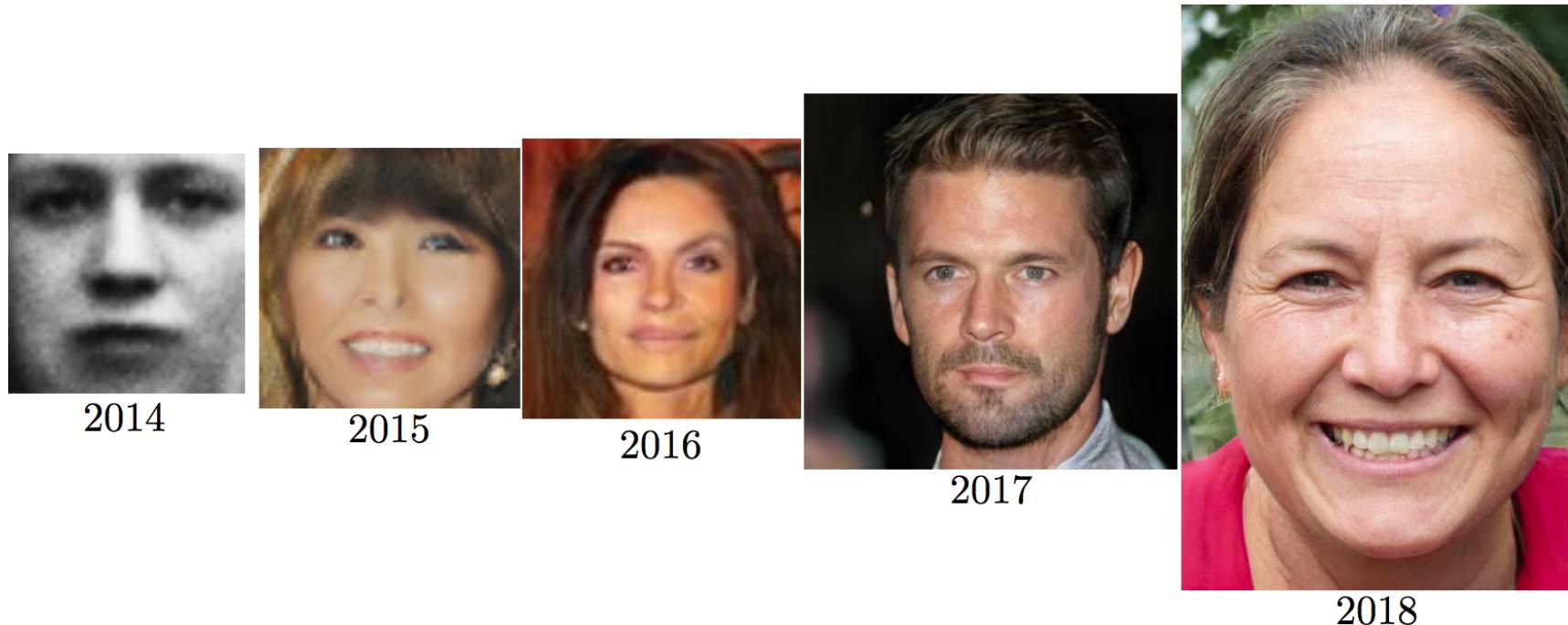
# Generative Adversarial Nets (GANs)

Goodfellow et al., 2014



# Moore's Law of AI

## 4.5 years of progress on faces



(Goodfellow 2019)

ArXiv 1406.2661, 1511.06434, 1607.07536, 1710.10196, 1812.04948  
Goodfellow, ICLR 2019 Invited Talk

# GANs for Imagenet



Odena et al  
2016



Miyato et al  
2017



Zhang et al  
2018



Brock et al  
2018

(Odena 2018)

# BigGANs, Brock et al., 2018



Figure 1: Class-conditional samples generated by our model.

## Variational Auto Encoders (VAEs, 2015)



[Alec Radford, 2015]

## VAEs in 2019



VQ-VAE-2, Razavi et al. June, 2019

## VAEs in 2019

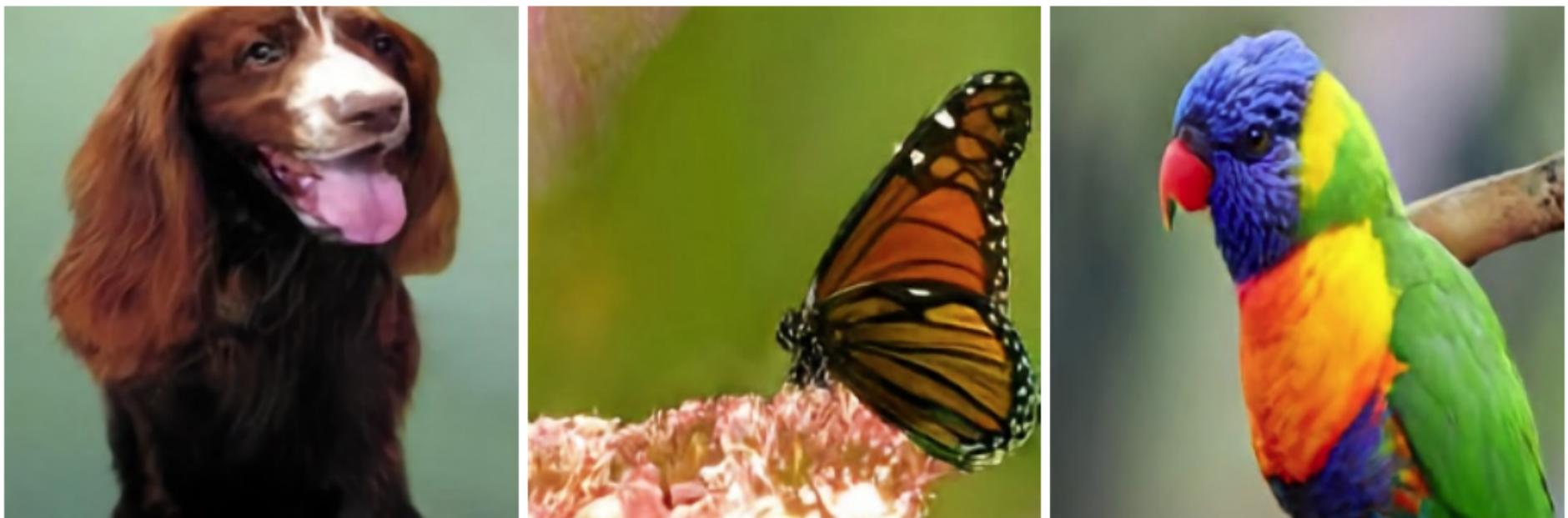


Figure 1: Class-conditional 256x256 image samples from a two-level model trained on ImageNet.

VQ-VAE-2, Razavi et al. June, 2019

## Wav2vec 2.0, June 2020, Facebook

Trained on 53k hours of unlabeled audio (no text) they use **contrastive learning** to convert speech to a sequence of discrete **quantized vectors** they call “pseudo-text units”.

By training on only one hour of human-transcribed audio, and using the Wav2vec transcription into pseudo-text, the outperform the previous state of the art in word error rate for 100 hours of human-transcribed text.

## GLSM, February 2021, Facebook

Generative Spoken Language Model (GSLM)

Using a form of **VQ-VAE** They then train a generative model of the sequences of pseudo-text units learned from unlabeled audio.

This model can continue speech from a speech prompt in much the same way that GPT-3 continues text from a text prompt.

Semantic and grammatical structure in a “unit language model” is recovered from speech alone.

## Codex, July 2021, OpenAI

Using an **unsupervised pretrained language model** they fine-tune on code, including comments, from public repositories.

Starting from an English prompt Codex continues with code — a form of automatic programming.

There is a published version (58 authors) and a production version that powers **GitHub Copilot**.

Copilot may supplant Stack Overflow for finding out how to do x in language y.

## CLIP, January 2021, OpenAI

CLIP: Contrastive Language-Image Pre-training.

Trained on images and associated text (such as image captions or hypertext links to images) CLIP computes embeddings of text and embeddings of images (“co-embeddings”) trained to capture the mutual information between the two.

This is done with contrastive learning.

## CLIP, January 2021, OpenAI

The model computes a probability of text given the co-embedding of the image.

It is then used for zero-shot image classification on various datasets.

One can classify an image by comparing the probabilities that the model assigns to “prompts”. There is a prompt for each class.

# Zero-Shot Image Classification

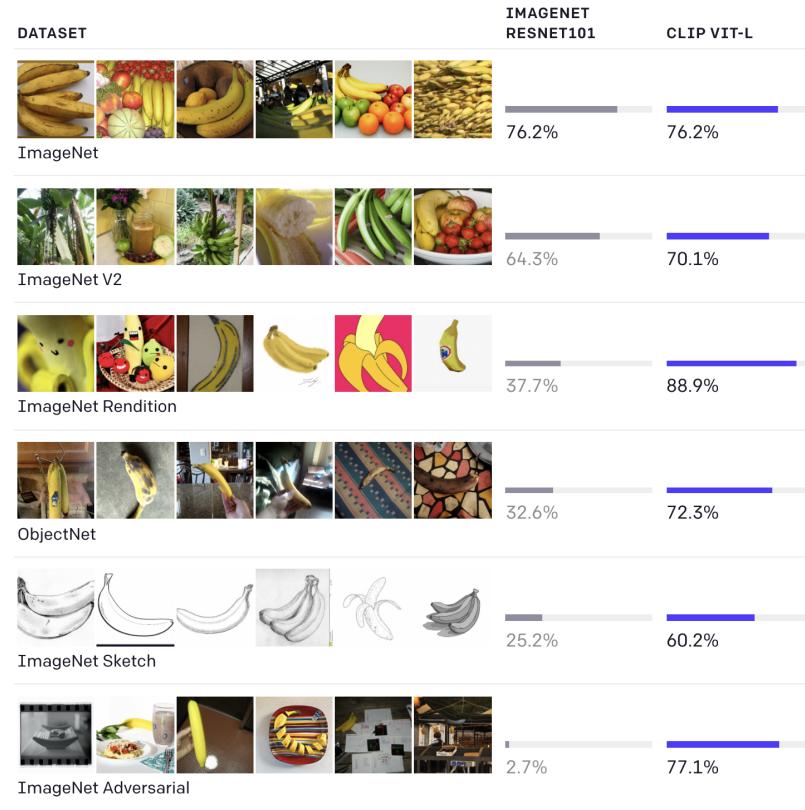
FOOD101

**guacamole** (90.1%) Ranked 1 out of 101 labels



- ✓ a photo of **guacamole**, a type of food.
- ✗ a photo of **ceviche**, a type of food.
- ✗ a photo of **edamame**, a type of food.
- ✗ a photo of **tuna tartare**, a type of food.
- ✗ a photo of **hummus**, a type of food.

# Zero-Shot Image Classification



Although both models have the same accuracy on the ImageNet test set, CLIP's performance is much more representative of how it will fare on datasets that measure accuracy in different, non-ImageNet settings. For instance, ObjectNet checks a model's ability to recognize objects in many different poses and with many different backgrounds inside homes while ImageNet Rendition and ImageNet Sketch check a model's ability to recognize more abstract depictions of objects.

**DALL·E, January 2021, OpenAI**

**DALL·E-2, April 2022**

The name DALL·E is simply some kind of homage to the painter Dali and the Disney character WALL·E.

Both versions of DALL·E uses CLIP's co-embeddings of images and text.

Given text, DALL·E generates an image using a **diffusion model**.

# DALL·E-1 Zero-Shot Image Rendering from Language

TEXT PROMPT    an illustration of a baby daikon radish in a tutu walking a dog

AI-GENERATED  
IMAGES



[Edit prompt or view more images↓](#)

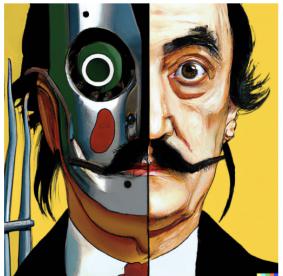
TEXT PROMPT    an armchair in the shape of an avocado....

AI-GENERATED  
IMAGES



[Edit prompt or view more images↓](#)

# DALL·E-2



vibrant portrait painting of Salvador Dalí 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



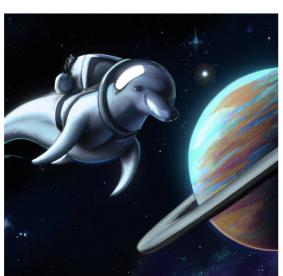
an espresso machine that makes coffee from human souls, artstation



panda mad scientist mixing sparkling chemicals, artstation



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



a dolphin in an astronaut suit on saturn, artstation



a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese



a teddy bear on a skateboard in times square

## Diffusion Models

DALL·E-2 uses diffusion models. Although originally defined in 2015, they have become very prominent in the last year.

# Chain of Thought Prompting, January 2022

Give examples of “chains of thought” for few shot learning of reasoning steps.

# Naive Prompting

## Standard Prompting

Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

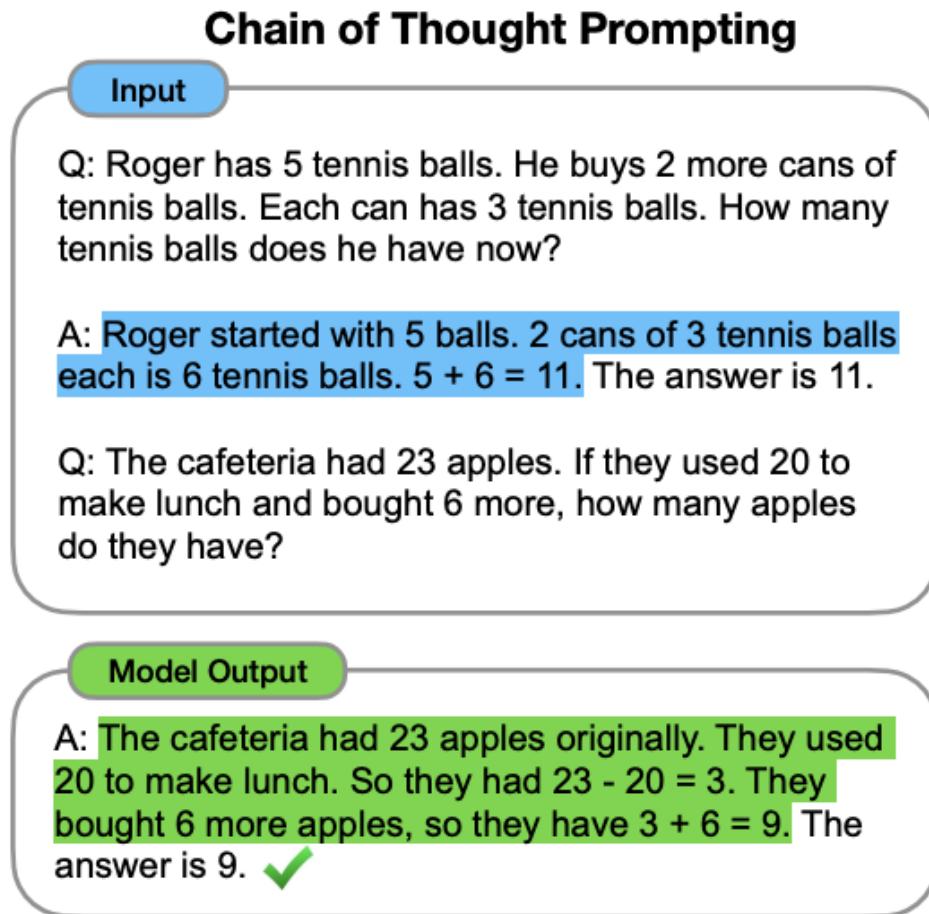
A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. 

# Chain of Thought Prompting, January 2022



## **Step by Step Prompting, June 2022**

It turns out that adding the simple instruction “take it step by step” elicits powerful chain of thought reasoning in GPT-3.

## **Humanoid Soccer, September 2022, Deep Mind**

Deep mind demonstrated two-on-two humanoid soccer in simulation (MuJoCo).

This a startling advance in the state of the art in humanoid humanoid control.

It also continues Deep Mind's effort in reinforcement learning.

## **Application Advancements vs. Architecture Advancements**

Advancements in the general principles of learning are having applications over very diverse applications.

When considering Moore's law of AI it seems worth distinguishing architectural advancements (new general learning methods) from new applications of established architectures.

This course will focus on general, architectural, ideas.

## Architectural Ideas

- Linear Threshold “neuron”
- Convolutional Neural Network or CNN
- Backpropagation
- Recurrent Neural Network or RNN
- Neural Language Modeling
- Variational Auto-Encoders (VAEs)
- Generative Adversarial Networks (GANs)
- Graph Neural Networks
- Batch Normalization
- Residual Connections
- Diffusion Models

- Reinforcement Learning
- The Transformer
- Unsupervised pre-training
- Contrastive Learning
- Prompting

**END**