HELMHOLTZAI | ARTIFICIAL INTELLIGENCE COOPERATION UNIT

Image data and DL Family

Gunjan Joshi, Ph.D

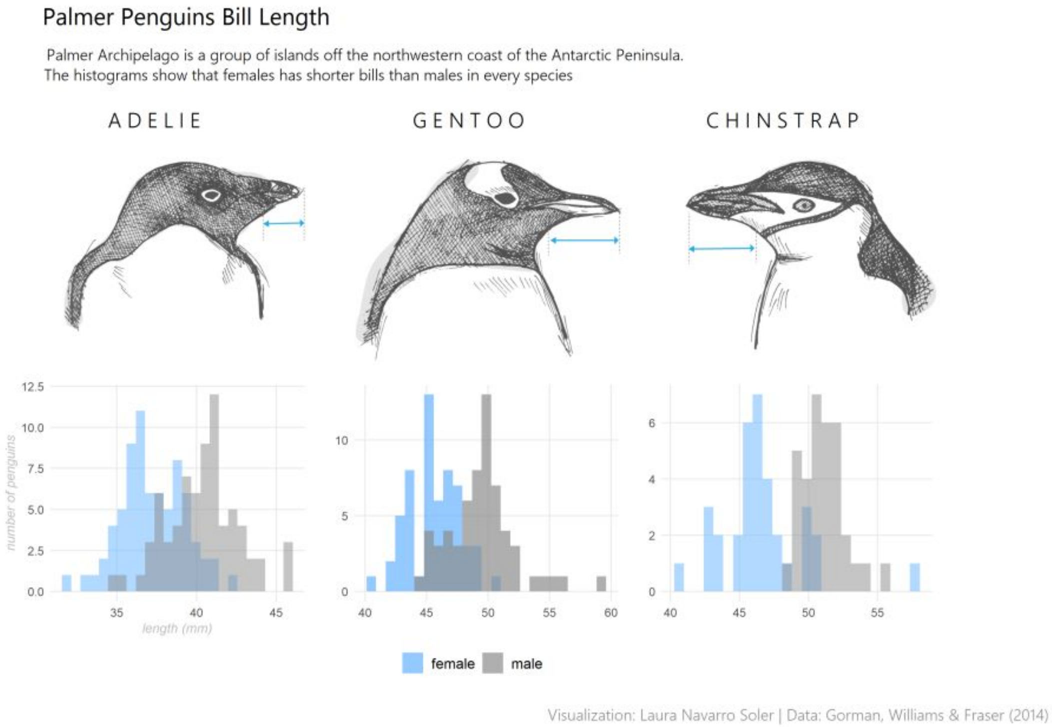
2025-11-04

HZDR HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF

Neural Networks = Data + Model

Data

Tabular data



https://lauranavarroviz.wordpress.com/2020/08/01/palmer-penguins/

Tabular data

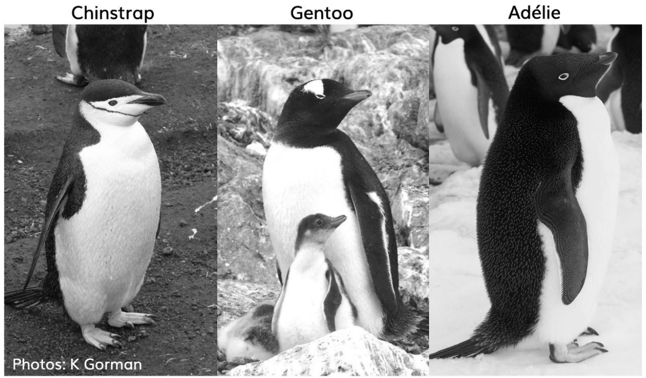
bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	species
39.1	18.7	181	3750	male	Adelie
39.5	17.4	186	3800	female	Adelie
40.3	18	195	3250	female	Adelie
36.7	19.3	193	3450	female	Adelie
39.3	20.6	190	3650	male	Adelie
55.9	17	228	5600	male	Gentoo
47.2	15.5	215	4975	female	Gentoo
49.1	15	228	5500	male	Gentoo
47.3	13.8	216	4725	male	Gentoo
46.8	16.1	215	5500	male	Gentoo
51.5	18.7	187	3250	male	Chinstrap
49.8	17.3	198	3675	female	Chinstrap
48.1	16.4	199	3325	female	Chinstrap
51.4	19	201	3950	male	Chinstrap
45.7	17.3	193	3600	female	Chinstrap
50.7	19.7	203	4050	male	?
35.9	19.2	189	3800	female	?
38.2	18.1	185	3950	male	?
38.8	17.2	180	3800	male	?
35.3	18.9	187	3800	female	?

- Rows and columns (structured data)
- → Easy to visualize and interpret (few features per sample)

Structured data

https://campus.datacamp.com/courses/understanding-artificial-intelligence/tasks-ai-can-solve?ex=8

Image data



- Images are grids of pixels (unstructured data)
- Each image has thousands of "features" (pixels × color channels)
- Spatial relationships between pixels carry meaning
- We use Convolutional Neural Networks (CNNs) to detect edges, shapes, textures
- Goal remains the same → predict a label

Unstructured data

https://education.rstudio.com/blog/2020/07/palmerpenguins-cran/
https://www.gabemednick.com/post/penguin/

Image data

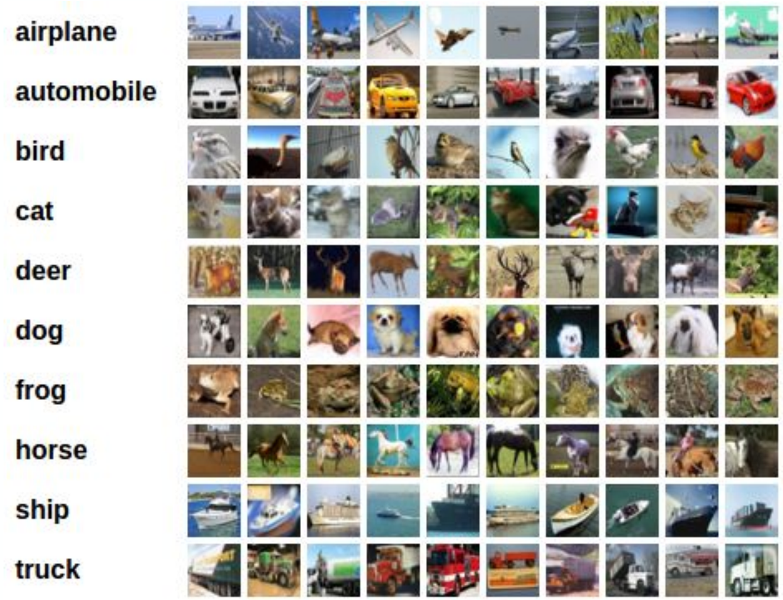
MNIST

https://datasets.activeloop.ai/docs/ml/datasets/mnist/

- What it is: Handwritten digits 0–9.
- Size: 70,000 grayscale images, each 28 × 28 pixels.
  60,000 for training, 10,000 for testing. Kaggle+1
- **Task:** “Which digit is this?”
- ***Why we like it:*** Super easy classification. You can get >99 percent accuracy with a tiny CNN.
- ***Limitation:*** It’s too clean and too easy, not very “real world.”

Image data

CIFAR-10

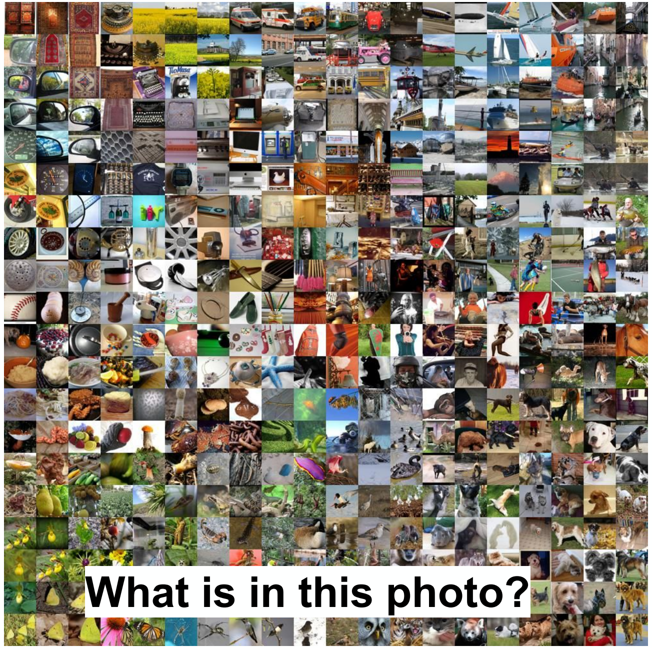


- **What it is:** Small color photos of everyday objects like airplane, cat, dog, ship, truck.
- **Size:** 60,000 images total at 32 × 32 RGB.
  50,000 training, 10,000 test.
  10 classes, 6,000 images per class.
  cs.toronto.edu+2Kaggle+2
- **Task:** “Which of these 10 classes is in the image?”
- **Why it matters**: It’s still tiny, but it's color and natural scenes, so it's more realistic than MNIST.
- **Limitation:** Very low resolution, so it’s not how real cameras see.

https://maucher.pages.mi.hdm-stuttgart.de/mlpytorch/03_cifar10_classification_mlp_cnn.html

Big benchmark datasets

ImageNet



What it is: Huge visual database of real-world objects.

Classic subset (ImageNet-1K / ILSVRC):

~1.2 million training images, 50,000 validation images, 100,000 test images.

1,000 object categories. Wikipedia+2Hugging Face+2

Full ImageNet (a larger version sometimes called ImageNet-21K):

Tens of thousands of categories and over 14 million images originally. Wikipedia

Task: Large-scale object classification (Which class is in this image?), plus localization (Where is it? bounding box).

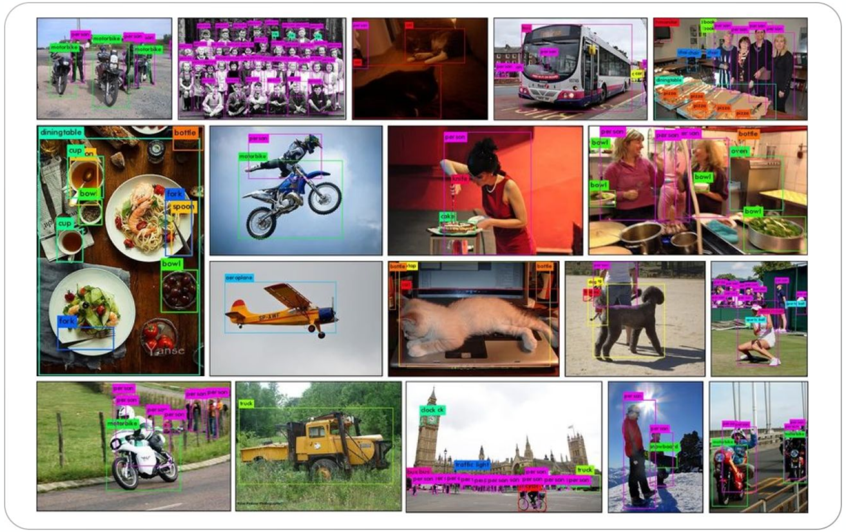
Why it matters:

- ImageNet is the reason deep learning blew up in vision in 2012: AlexNet (CNN) crushed this challenge and changed the field.
- Today, many pretrained CNNs (ResNet, EfficientNet, etc.) come "pretrained on ImageNet."

https://www.gopichandrakesan.com/day-27-imagenet-dataset-largest-visual-database/

Big benchmark datasets

COCO



Where are all the things, and what are they doing?

https://www.v7labs.com/blog/coco-dataset-guide

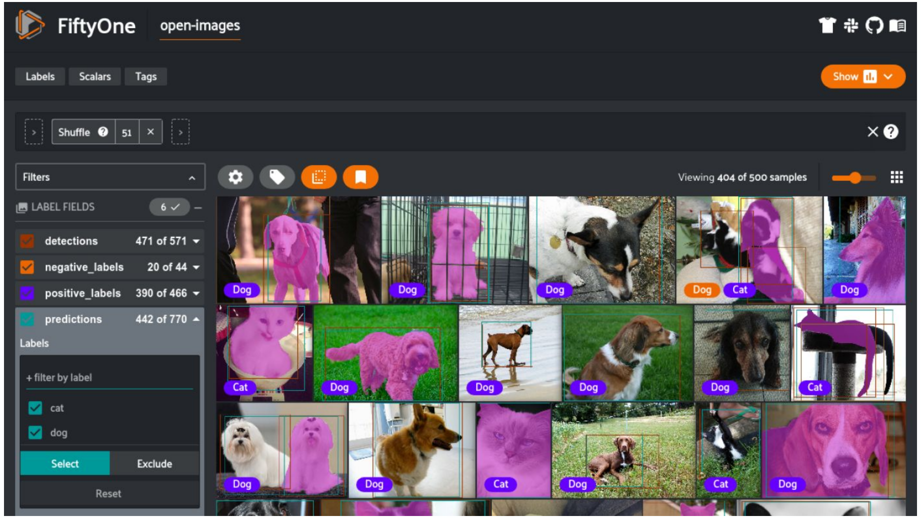
- **What it is:** Everyday scenes with multiple objects.
- **Size:** Roughly 330,000 images with dense annotations.
  80 object categories. cocodataset.org+2docs.ultralytics.com+2

- **Tasks:**
  a. Object detection: draw a box around each object.
  b. Instance segmentation: outline each object's exact shape.
  c. Image captioning: describe the scene in natural language.

- **Why it matters:**
  a. It's "objects in context," not just a single centered object. So it's much closer to the real world.
  b. Used to benchmark detection models like Faster R-CNN, YOLO, etc.

Very large / modern / multi-label datasets

Open Images (by Google)



Open Images is what you use when you want to train models that need to understand complicated real scenes at internet scale

Scale: Millions of real photos with tons of different labels.

<u>Annotations:</u>

- Image-level labels (what's present)
- Bounding boxes for objects
- Instance segmentation masks
- Relationships between objects ("man holding cup")
- Human-written localized narratives describing regions

How do I get these datasets ?

```
(train_images, train_labels), (test_images, test_labels) = keras.datasets.cifar10.load_data()
```

Custom image data (you collect and label it)

Workflow:

- Collect images yourself.
- Label them.
    - **For image classification**
        1. one label for the whole image
    - **For object detection**
        1. draw bounding boxes around each object
    - **For segmentation**
        1. label pixels
    - Tools for labeling include **VGG Image Annotator, ImageJ with plugins, COCO Annotator.**

Custom image data

Then you preprocess:

- **Resize** images to a consistent size so they all match (for example, make them all 32×32).
- **Augment** images (random flips, rotations, brightness changes) to make the model more robust.
- **Normalize** pixel values so they're on a stable numeric range (for example scale 0 to 255 down to 0 to 1).
- **Encode labels** as numbers the model can use.
- **Split** into train / validation / test sets.

Key idea: Images are just numbers. Each image is really an array (height × width × channels). Each pixel is one little square of color.

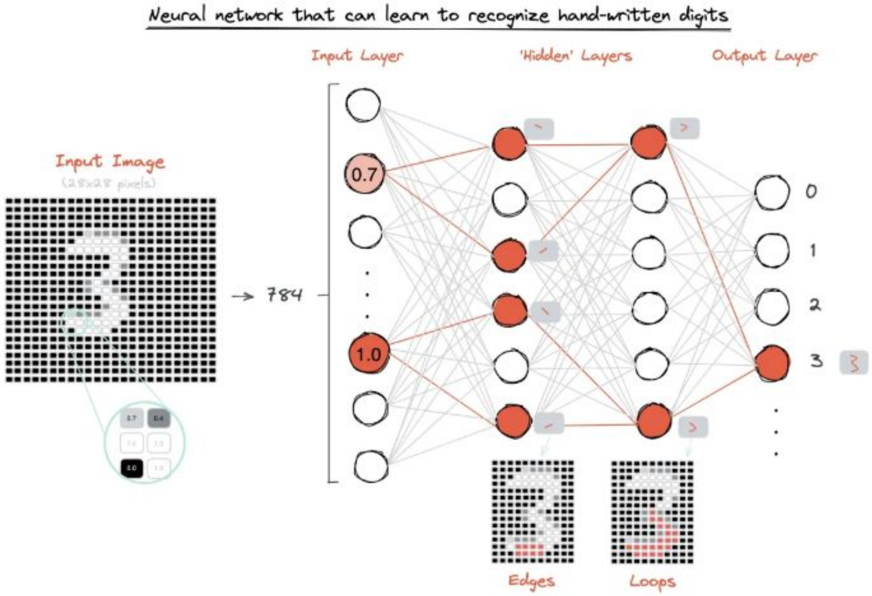
How much data do we need ?

- Deep learning models are data-hungry.
  - From-scratch training: thousands of labeled images per class
  - With transfer learning: much less data required
- Data quality and diversity are crucial.

Model

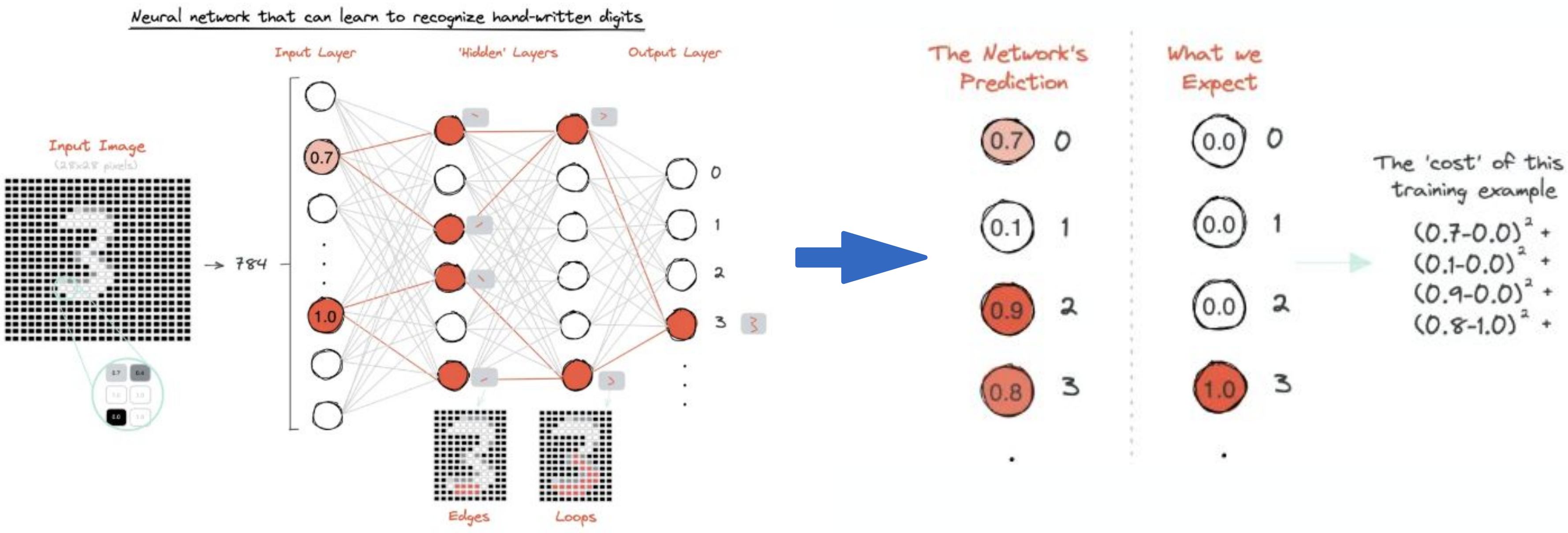
Members of the Deep Learning Family

"Plain Vanilla" Feed forward Neural Network

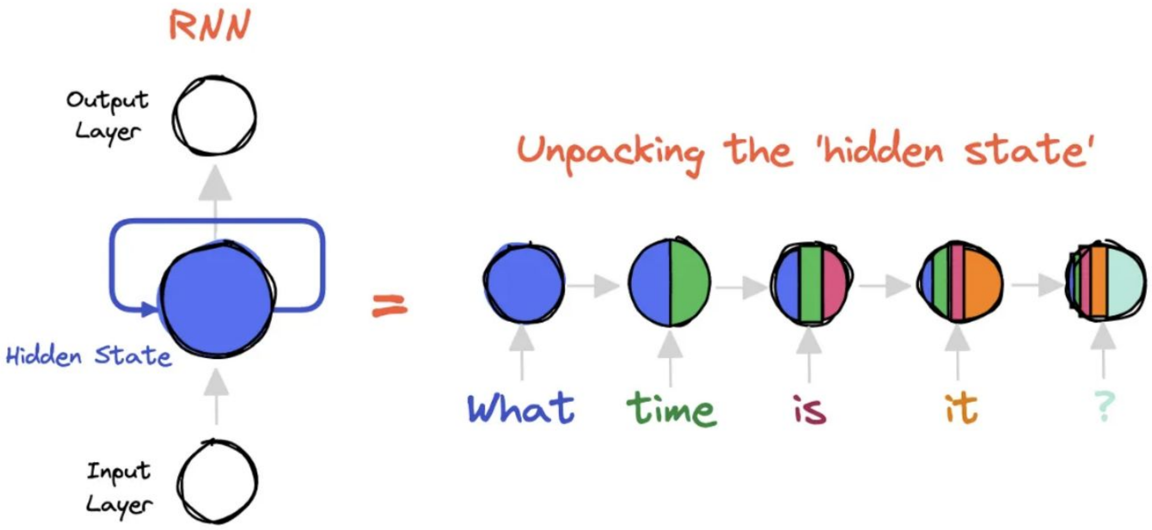


https://arcxhived.substack.com/p/foundation-models

“Plain Vanilla” Feed forward Neural Network



Recurrent Neural Networks



The A.I Hacker - Illustrated Guide to Recurrent Neural Networks

Transformers

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly

Transformers

arXiv:1706.03762v7 [cs.CL] 2 Aug 2023

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

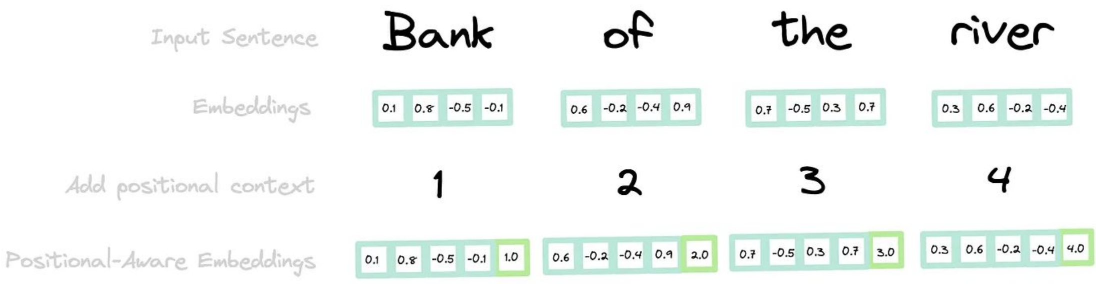
Abstract

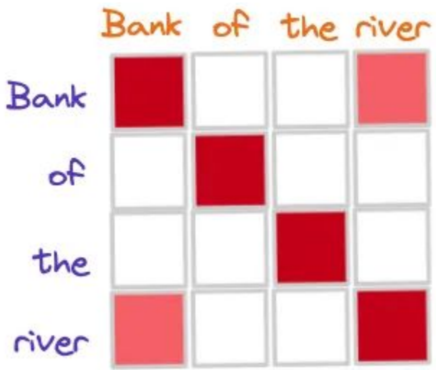
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly

Transformers brought two key innovations from its predecessor (RNNs)

- **Positional Encodings**
- **Self-Attention**

Transformers

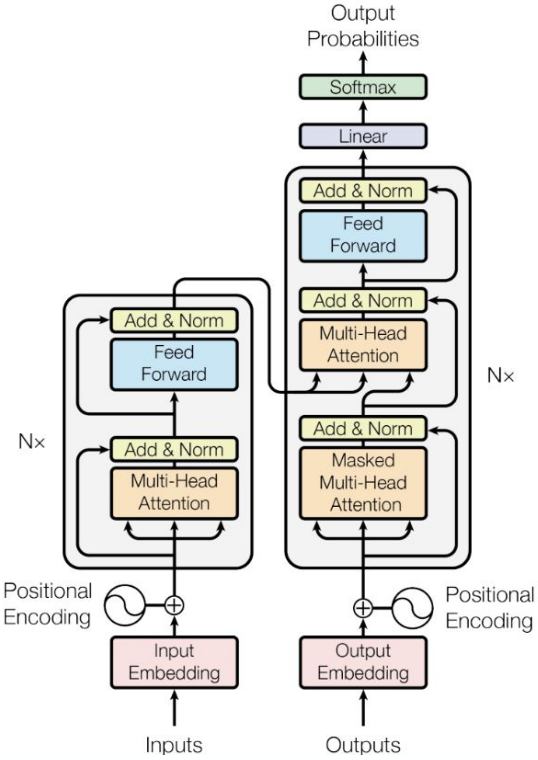




Transformers

BERT

Encoder



GPT

Decoder

Transformers

BERT

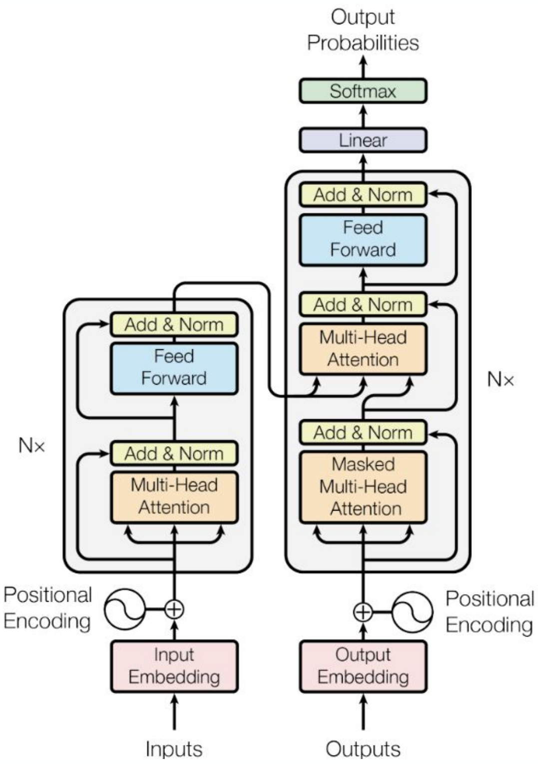
BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the



GPT

Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com
Karthik Narasimhan OpenAI karthikn@openai.com
Tim Salimans OpenAI tim@openai.com
Ilya Sutskever OpenAI ilyasu@openai.com

Abstract

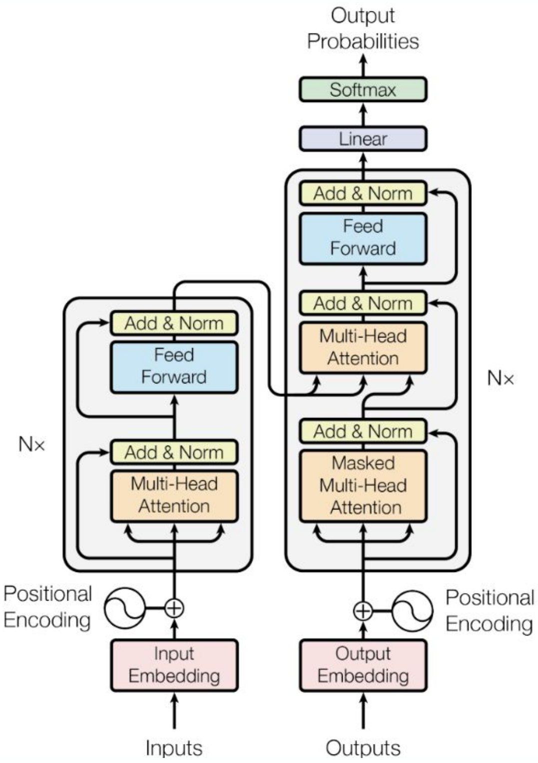
Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification. Although large unlabeled text corpora are abundant, labeled data for learning these specific tasks is scarce, making it challenging for discriminatively trained models to perform adequately. We demonstrate that large gains on these tasks can be realized by *generative pre-training* of a language model on a diverse corpus of unlabeled text, followed by *discriminative fine-tuning* on each specific task. In contrast to previous approaches, we make use of task-aware input transformations during fine-tuning to achieve effective transfer while requiring minimal changes to the model architecture. We demonstrate the effectiveness of

Transformers

BERT

Encoder

use transfer learning to continue learning from its existing data when adding user-specific tasks and layer



GPT

Decoder

decodes from its massive pre-learned embeddings to present output that matches user prompts

Transformers

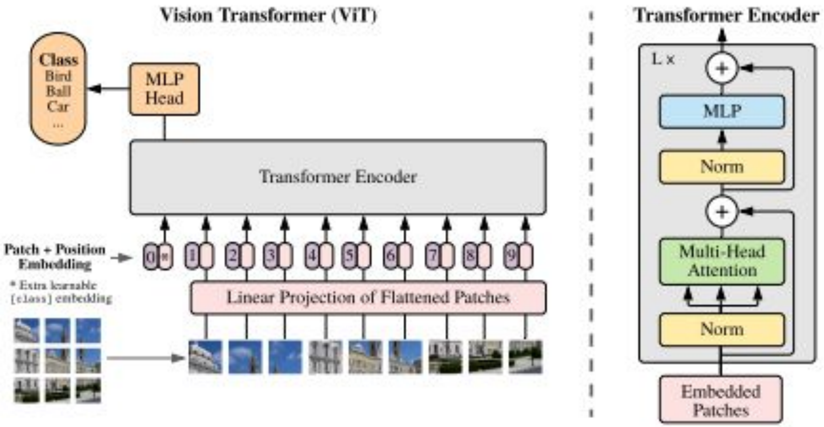
AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

Alexey Dosovitskiy[*,†], Lucas Beyer[*], Alexander Kolesnikov[*], Dirk Weissenborn[*], Xiaohua Zhai[*], Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby[*,†]
[*]equal technical contribution, [†]equal advising
Google Research, Brain Team
{adosovitskiy, neilhoulsby}@google.com

ABSTRACT

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

Treat an image like a **sequence of patch tokens**, just like words in a sentence and use the same transformer architecture from NLP.



Foundation Model

On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*[1]

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks.

1.1.1 Naming.

We introduce the term *foundation models* to fill a void in describing the paradigm shift we are witnessing; we briefly recount some of our reasoning for this decision. Existing terms (e.g., *pretrained model, self-supervised model*) partially capture the technical dimension of these models, but fail to capture the significance of the paradigm shift in an accessible manner for those beyond machine learning. In particular, foundation model designates a model class that are distinctive in their sociological impact and how they have conferred a broad shift in AI research and deployment. In contrast, forms of pretraining and self-supervision that technically foreshadowed foundation models fail to clarify the shift in practices we hope to highlight.

Foundation Model

On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*[1]

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks.

1.1.1 Naming.

We introduce the term *foundation models* to fill a void in describing the paradigm shift we are witnessing; we briefly recount some of our reasoning for this decision. Existing terms (e.g., *pretrained model, self-supervised model*) partially capture the technical dimension of these models, but fail to capture the significance of the paradigm shift in an accessible manner for those beyond machine learning. In particular, foundation model designates a model class that are distinctive in their sociological impact and how they have conferred a broad shift in AI research and deployment. In contrast, forms of pretraining and self-supervision that technically foreshadowed foundation models fail to clarify the shift in practices we hope to highlight.

Stanford defines foundation models as:

*"**Models** trained on **broad data** (generally using **self supervision** at scale) that can be **adapted** (fine-tuned) to a wide range of downstream tasks"*

HELMHOLTZ AI Artificial Intelligence Cooperation Unit

Foundation Model

On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi
Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
Percy Liang*[1]

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

Xiv:2108.07258v3 [cs.LG] 12 Jul 2022

AI is undergoing a paradigm shift with the rise of models (e.g., BERT, DALL-E, GPT-3) trained on broad data (generally using self-supervision at scale) that can be adapted to a wide range of downstream tasks.

1.1.1 Naming.

We introduce the term *foundation models* to fill a void in describing the paradigm shift we are witnessing; we briefly recount some of our reasoning for this decision. Existing terms (e.g., *pretrained model, self-supervised model*) partially capture the technical dimension of these models, but fail to capture the significance of the paradigm shift in an accessible manner for those beyond machine learning. In particular, foundation model designates a model class that are distinctive in their sociological impact and how they have conferred a broad shift in AI research and deployment. In contrast, forms of pretraining and self-supervision that technically foreshadowed foundation models fail to clarify the shift in practices we hope to highlight.

Stanford defines foundation models as:

*"**Models** trained on **broad data** (generally using **self supervision** at scale) that can be **adapted** (fine-tuned) to a wide range of downstream tasks"*

Foundation Model = (Large corpus of (unlabeled) data + Scale + SSL) → Transfer learning capability

Foundation Model