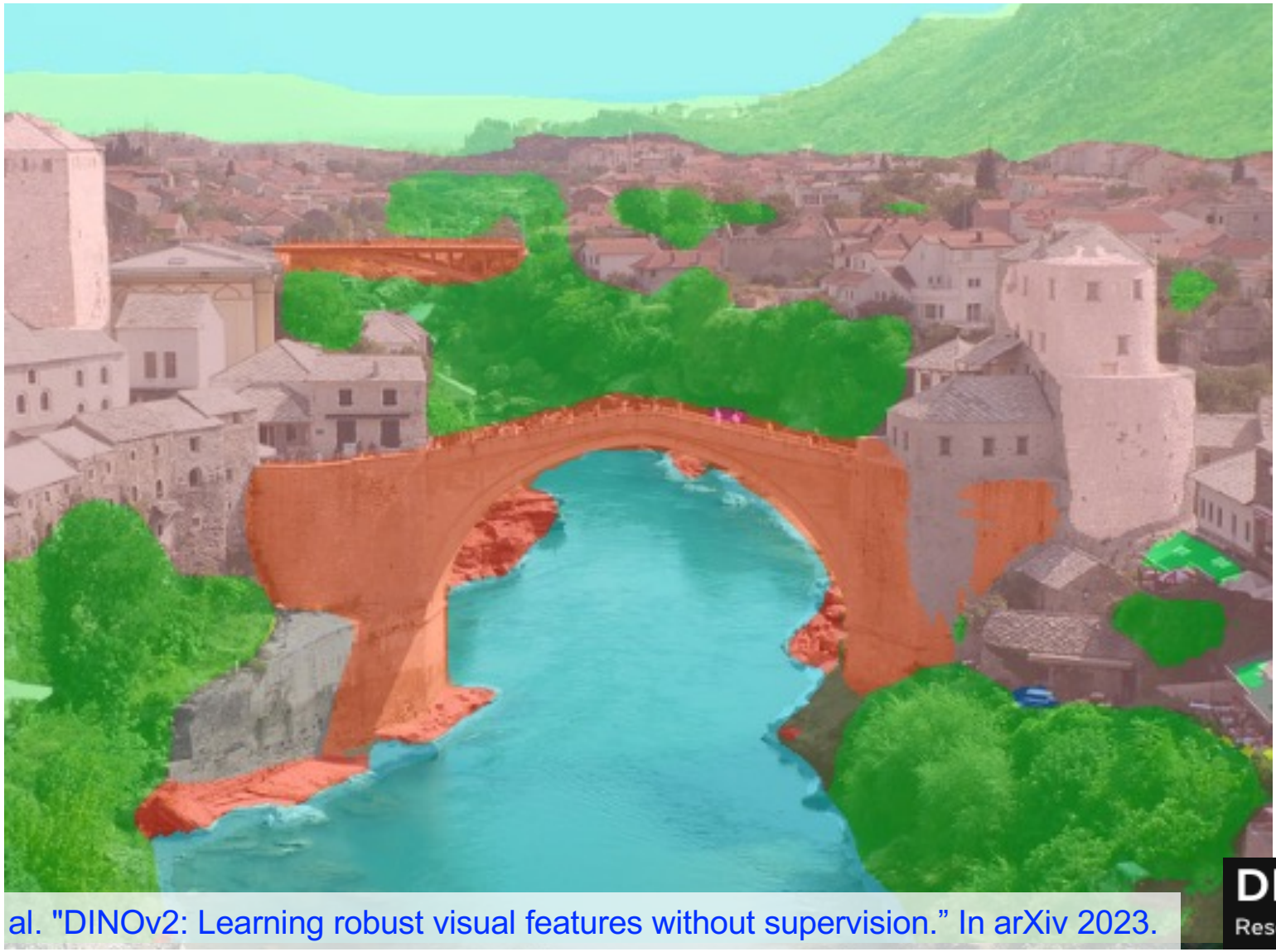
Why Foundational AI Matters?

Simon Lucey

THE UNIVERSITY of ADELAIDE

AUSTRALIAN INSTITUTE FOR MACHINE LEARNING



M. Oquab, et al. "DINOv2: Learning robust visual features without supervision." In arXiv 2023.

A. Kirillov, et al. "Segment Anything.” In CVPR 2023.

User: What is unusual about this image?

GPT4: 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.

TECH · ELON MUSK

Elon Musk's just fired up Colossus—the world's largest Nvidia GPU supercomputer built in just three months from start to finish

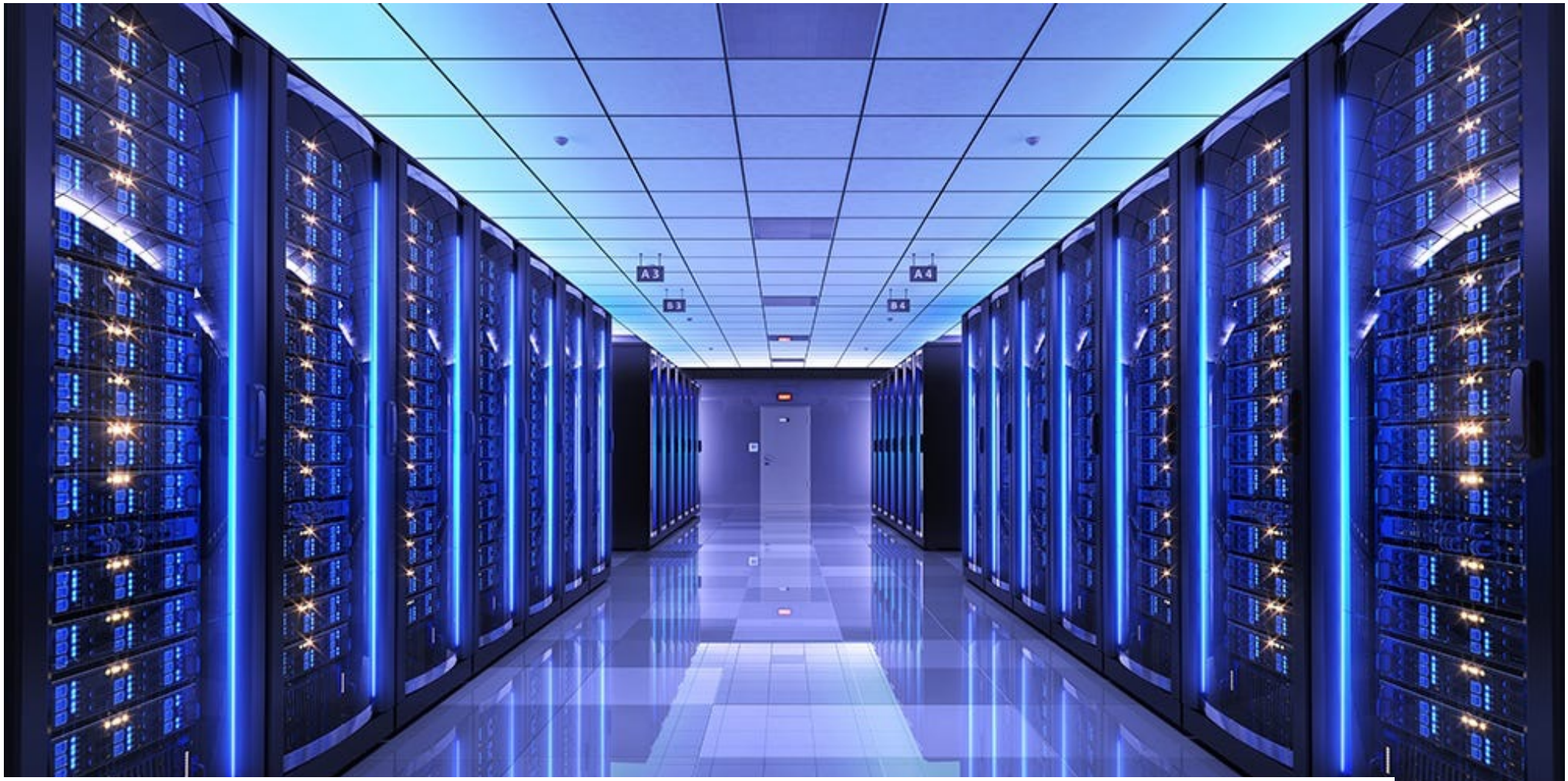
BY CHRISTIAAN HETZNER

September 3, 2024 at 11:48 PM GMT+9:30

xAI founder Elon Musk aims to double the capacity of his Memphis investors could end up benefiting as well thanks to Optimus.

RICHARD BORD—WIREIMAGE/GETTY IMAGES

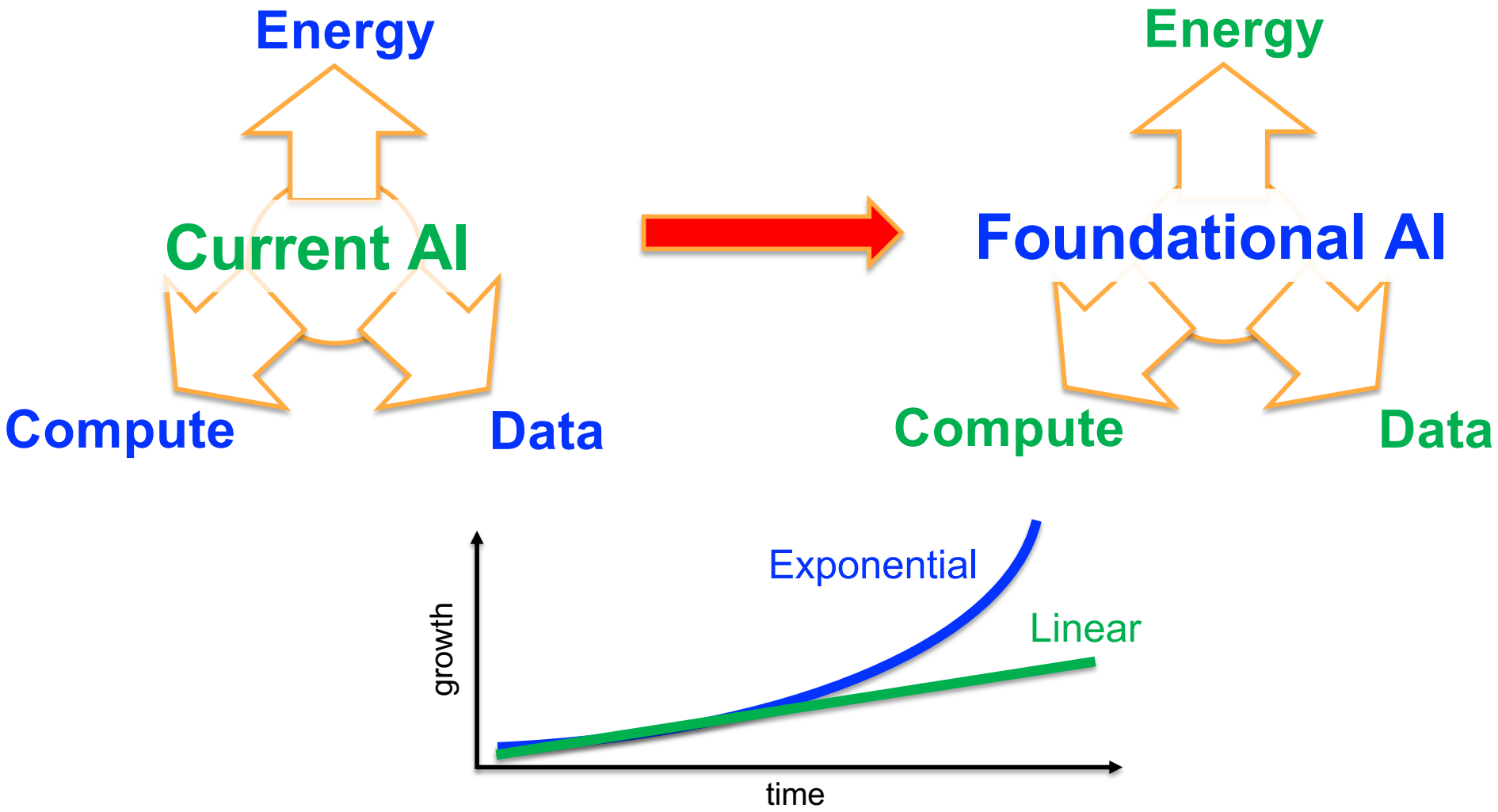
https://fortune.com/2024/09/03/elon-musk-xai-nvidia-colossus/



Contains 100,000 NVIDIA H100 processors, estimated cost $U3-4 billion.

Will use 1 million gallons of water per day to cool servers, and consume 150 megawatts of power.

https://www.reuters.com/markets/deals/constellation-inks-power-supply-deal-with-microsoft-2024-09-20/

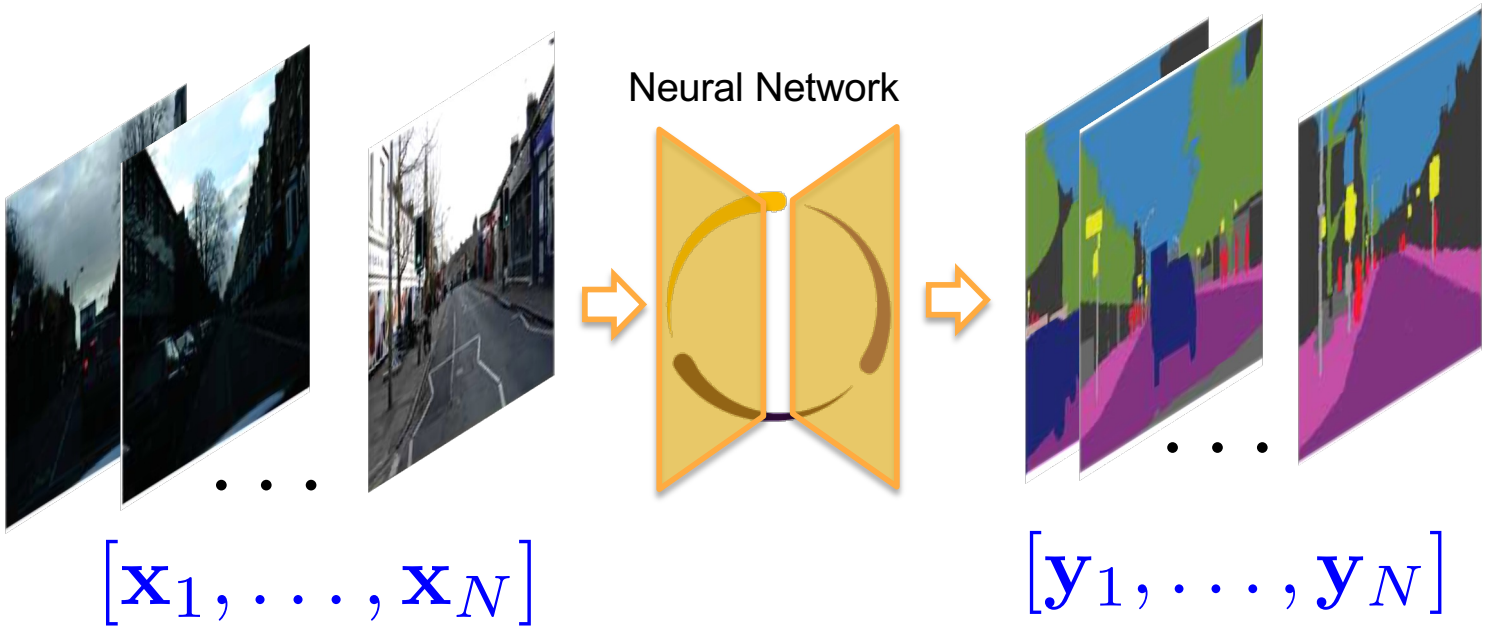

Energy
Current AI
Compute
Data
Energy
Foundational AI
Compute
Data
Exponential
Linear
growth
time

What is Foundational AI?


Pioneering the IntelligentEngine

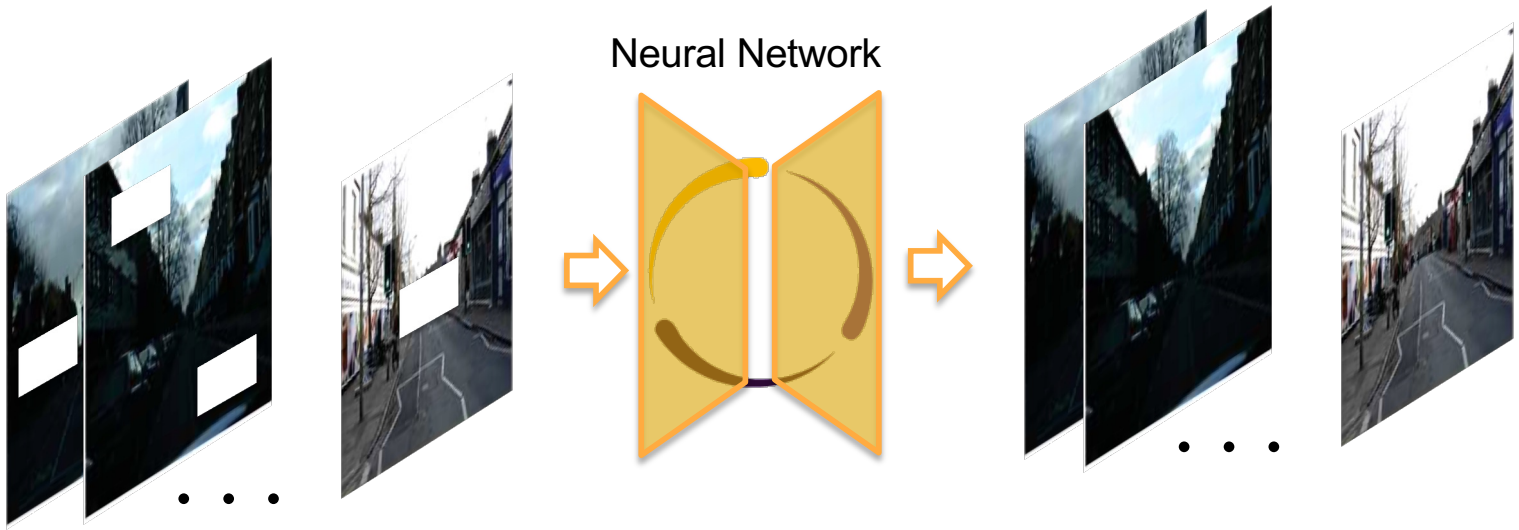
More DATA &
More Compute?

Where do I get the labels?



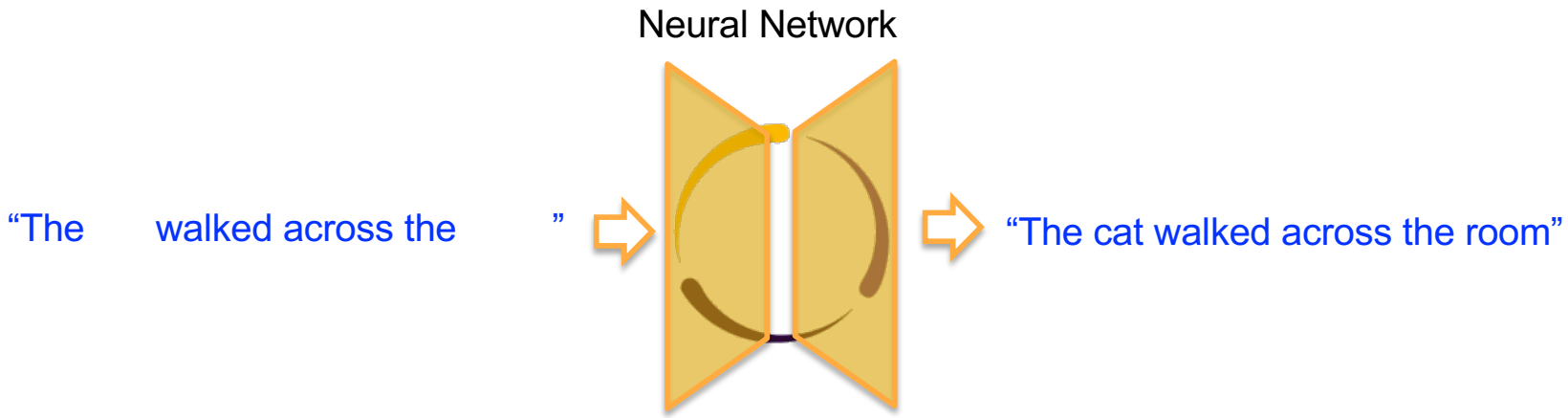
$[\mathbf{x}_1, \ldots, \mathbf{x}_N]$ $[\mathbf{y}_1, \ldots, \mathbf{y}_N]$

$$N \propto 10^6 - 10^9$$



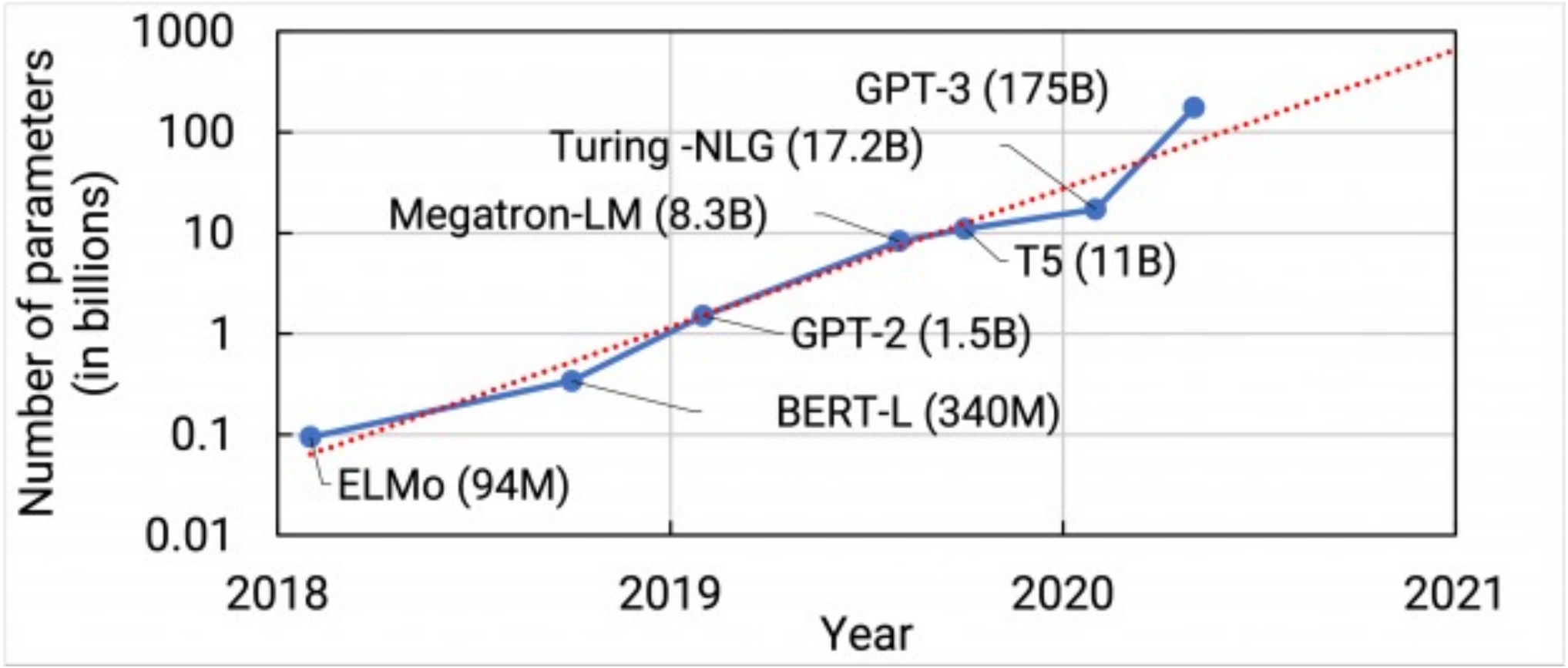
Self Supervision

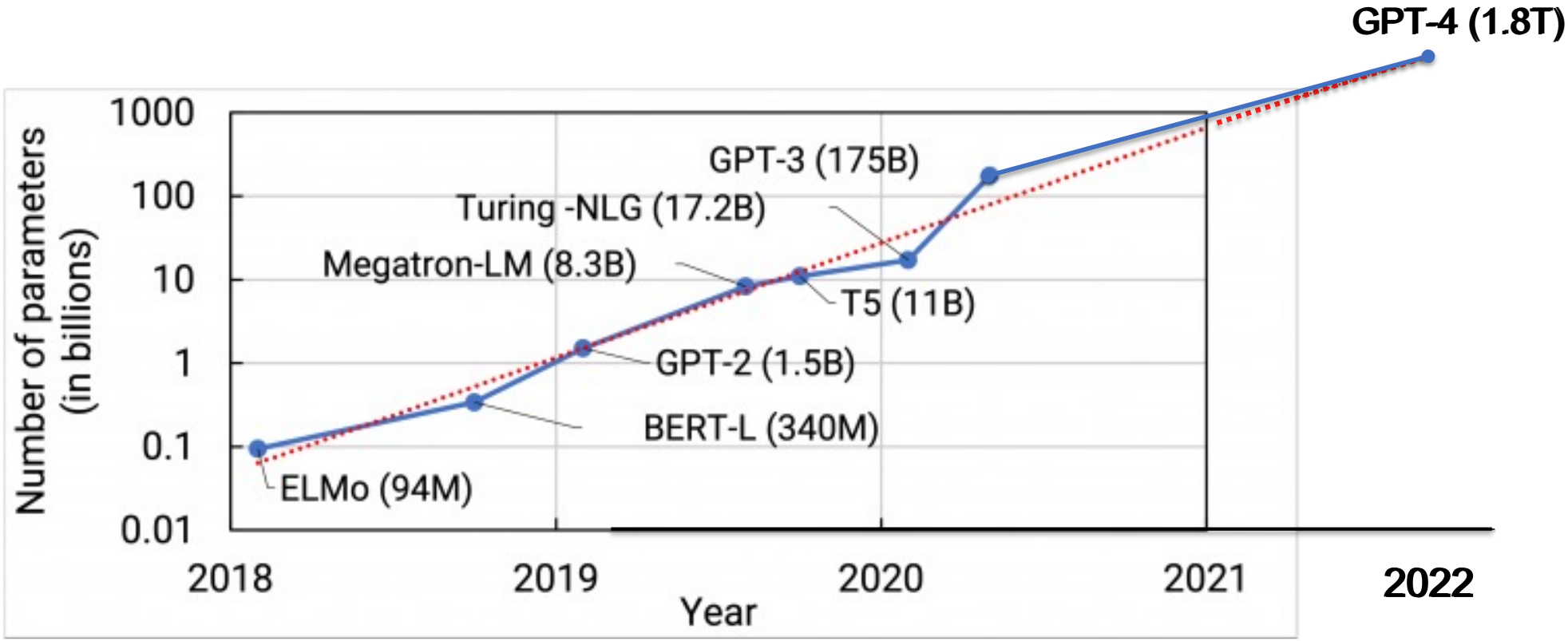
K. He, et al. "Masked Autoencoders Are Scalable Vision Learners." in CVPR 2022.



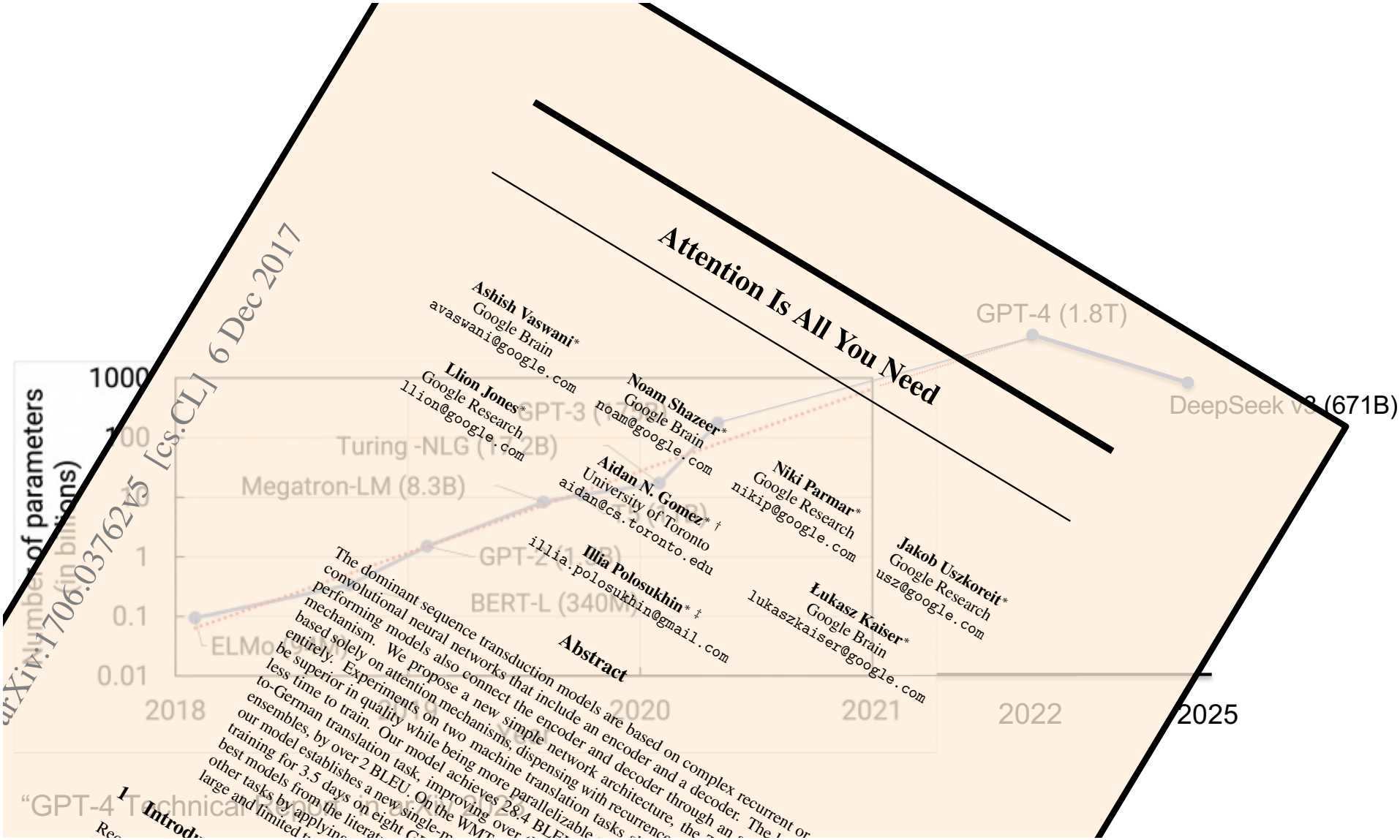
Self Supervision

J. Devlin, et al. “BERT: Pre-training of deep bidirectional transformers for language understanding” in arXiv 2018.


Number of parameters (in billions)
1000
100
10
1
0.1
0.01
2018
2019
2020
2021
Year
GPT-3 (175B)
Turing -NLG (17.2B)
Megatron-LM (8.3B)
T5 (11B)
GPT-2 (1.5B)
BERT-L (340M)
ELMo (94M)



"GPT-4 Technical Report" in arXiv 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 less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.
1 Introduction
[cs.CL] 6 Dec 2017
GPT-4 (1.8T)
DeepSeek v3 (671B)
Megatron-LM (8.3B)
BERT-L (340M)
2018
2020
2021
2022
2025
1000
100
1
0.1
0.01
of parameters
“GPT-4 Technical Report”

Is that IT?

T. Brookes, W. Peebles, et al. "Video generation models as world simulators." *in OpenAI Tech Report 2024.*

T. Brookes, W. Peebles, et al. "Video generation models as world simulators." *in OpenAI Tech Report 2024.*

Multiplying with GPT-4

Multiply 999 by 867

GPT-4

The product of multiplying 999 by 867 is 824,133

The correct answer is 866,133

GPT-4 achieves only 0% accuracy on 5 digit by 5 digit multiplication.

Faith and Fate: Limits of Transformers on Compositionality

Nouha Dziri[1*], Ximing Lu[1,2*], Melanie Sclar[2*], Xiang Lorraine Li[1†], Liwei Jiang[1,2†], Bill Yuchen Lin[3], Peter West[1,2], Chandra Bhagavatula[1], Ronan Le Bras[1], Jena D. Hwang[1], Soumya Sanyal[3], Sean Welleck[1,2], Xiang Ren[1,3], Allyson Ettinger[1,4], Zaid Harchaoui[1,2], Yejin Choi[1,2]

[1]Allen Institute for Artificial Intelligence [2]University of Washington
[3]University of Southern California [4]University of Chicago

nouhad@allenai.org, ximinglu@allenai.org, msclar@cs.washington.edu

Abstract

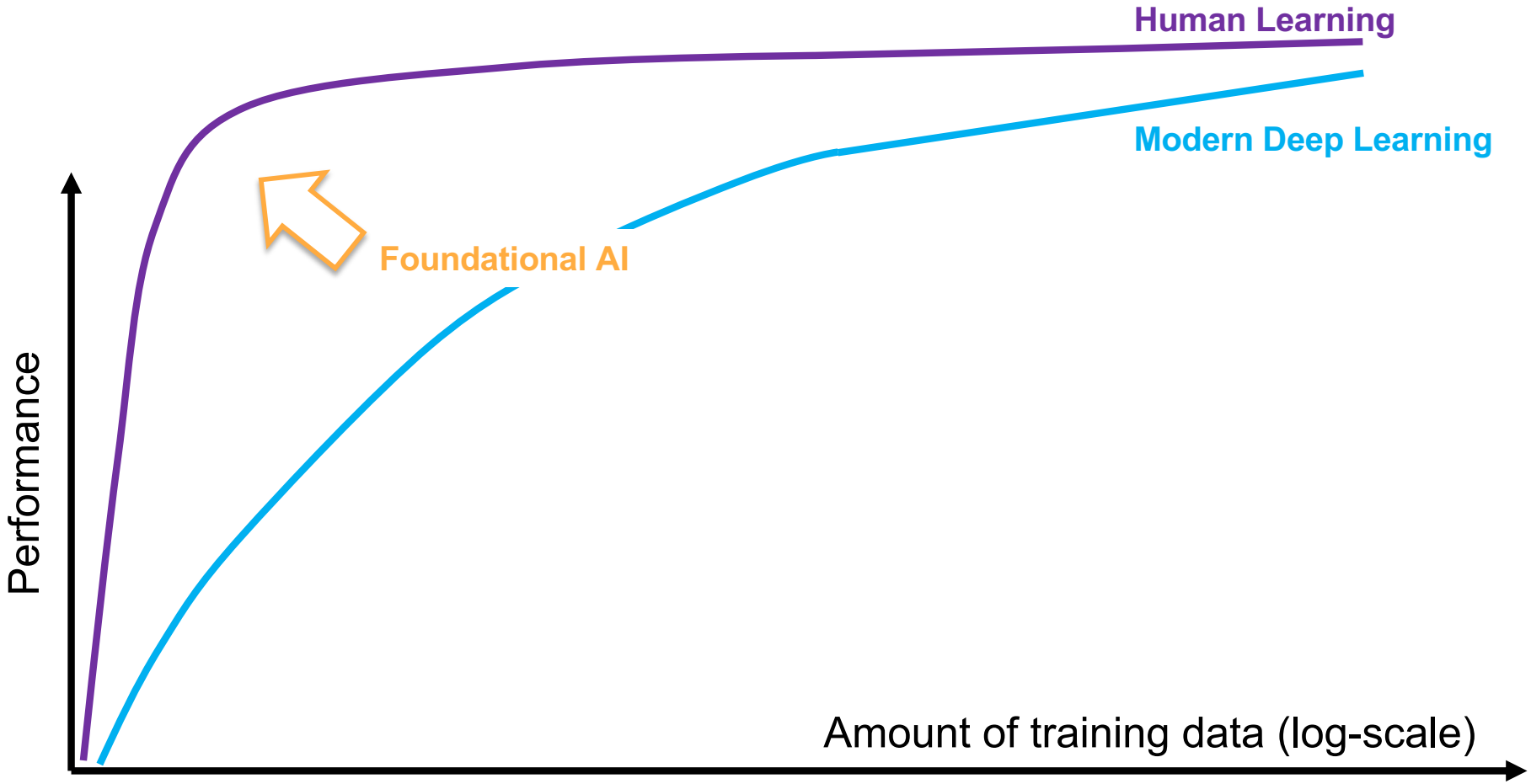
Transformer large language models (LLMs) have sparked admiration for their exceptional performance on tasks that demand intricate multi-step reasoning. Yet, these models simultaneously show failures on surprisingly trivial problems. This begs the question: Are these errors incidental, or do they signal more substantial limitations? In an attempt to demystify Transformers, we investigate the limits of these models across three representative *compositional* tasks—multi-digit multiplication, logic grid puzzles, and a classic dynamic programming problem. These tasks require breaking problems down into sub-steps and synthesizing these into a precise answer. We formulate compositional tasks as computation graphs to systematically quantify the level of complexity, and break down reasoning steps into intermediate sub-procedures. Our empirical findings suggest that Transformers solve compositional tasks by reducing multi-step compositional problems into linearized subgraph matching, without necessarily developing systematic problem solving skills. To round off our empirical study, we provide theoretical arguments on abstract multi-step reasoning problems that highlight how Transformers' performance will rapidly decay with increasing task complexity.

1 Introduction

"It was the epoch of b

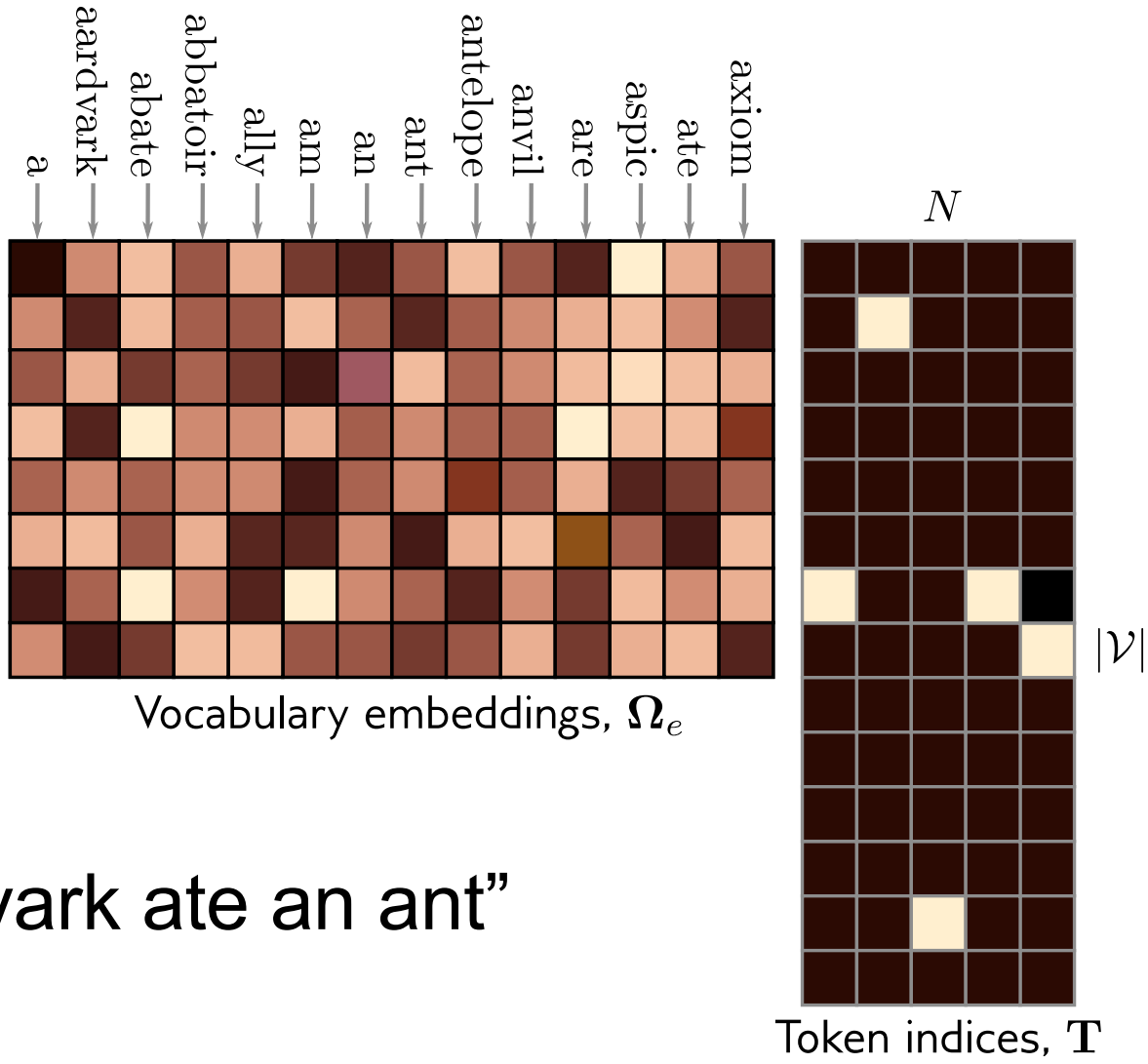
Large-scale T
bilities [4
some

Humans are data efficient learners……

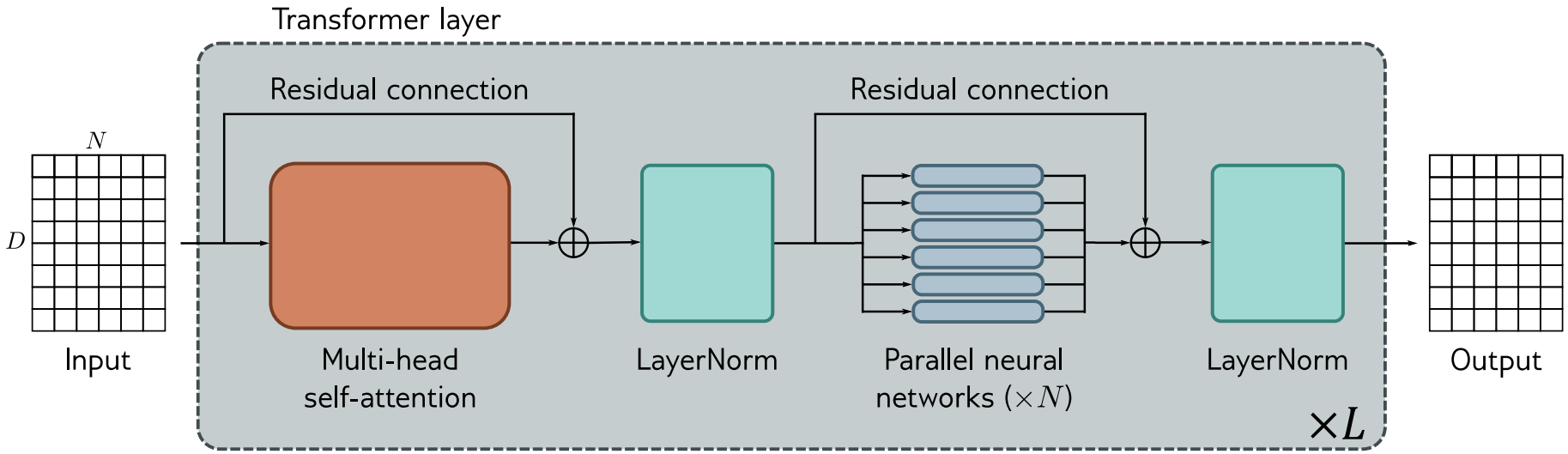


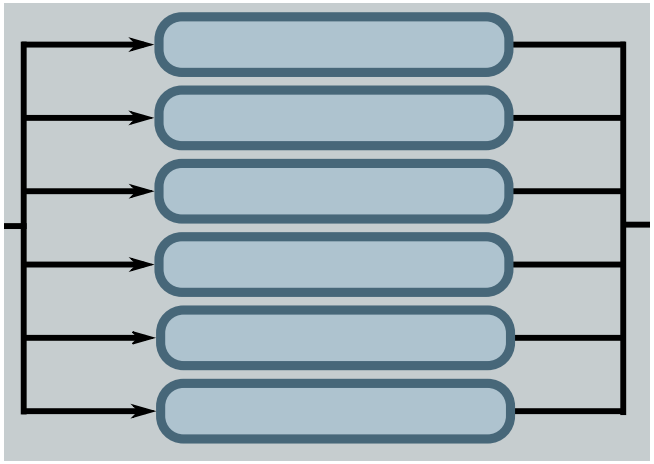
Data
CommBank AI research investment 'paid for itself' in three weeks
Justin Hendry
Editor
21 November 2024
Share
A $6 million investment in Australia's largest university-based machine learning research group by the Commonwealth Bank earlier this year paid for itself just three weeks into the five-year partnership.
That's according to the bank's chief information officer for technology Brendan Hopper, who has explained the thinking behind the recent investment in the Australian Institute for Machine Learning (AIML).
It represents the biggest ever investment from a private sector company in Australia into foundation AI, and follows several other investments by AIML's university and state government backers in recent months.
arXiv:2403.19243v1 [cs.LG] 28 Mar 2024
Sine Activated Low-Rank Matrices for Parameter Efficient Learning
Yiping Ji1*, Hemanth Saratchandran1*, Cameron Gordon1, Zeyu Zhang2, Simon Lucey1
1 Australian Institute for Machine Learning, The University of Adelaide
2 The Australian National University
yiping.ji@adelaide.edu.au
Abstract. Low-rank decomposition has emerged as a vital tool for enhancing parameter efficiency in neural network architectures, gaining traction across diverse applications in machine learning. These techniques significantly lower the number of parameters, striking a balance between compactness and performance. However, a common challenge has been the compromise between parameter efficiency and the accuracy of the model, where reduced parameters often lead to diminished accuracy compared to their full-rank counterparts. In this work, we propose a novel theoretical framework that integrates a sinusoidal function within the low-rank decomposition process. This approach not only preserves the benefits of the parameter efficiency characteristic of low-rank methods but also increases the decomposition's rank, thereby enhancing model accuracy. Our method proves to be an adaptable enhancement for existing low-rank models, as evidenced by its successful application in Vision Transformers (ViT), Large Language Models (LLMs), Neural Radiance Fields (NeRF), and 3D shape modeling. This demonstrates the wide-ranging potential and efficiency of our proposed
Keywords: Parameter Efficient · Low-rank
1 Introduction

“an aardvark ate an ant”



"an aardvark ate an ant"


Transformer layer
Residual connection
Residual connection
N
D
Input
Multi-head
self-attention
LayerNorm
Parallel neural
networks (×N)
LayerNorm
Output
×L

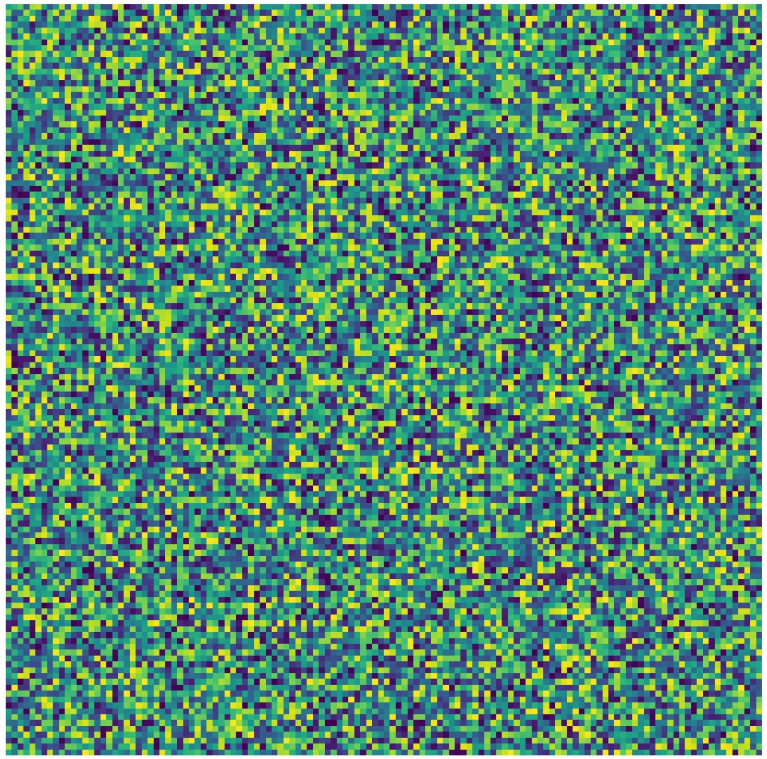


$$(W \not+ \Delta W)x$$

pre-trained weights

fine-tuned weights

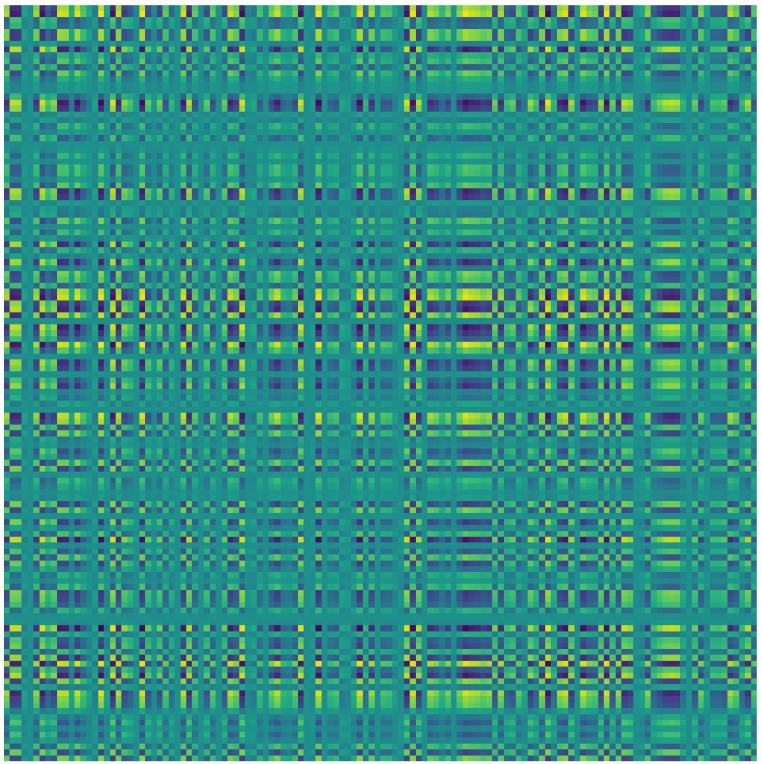
E. J. Hu, et al. “LoRA: Low-rank adaptation of large language models.” In ICLR 2022.



ΔW

- Ideally, full-rank.
- However, too many parameters.

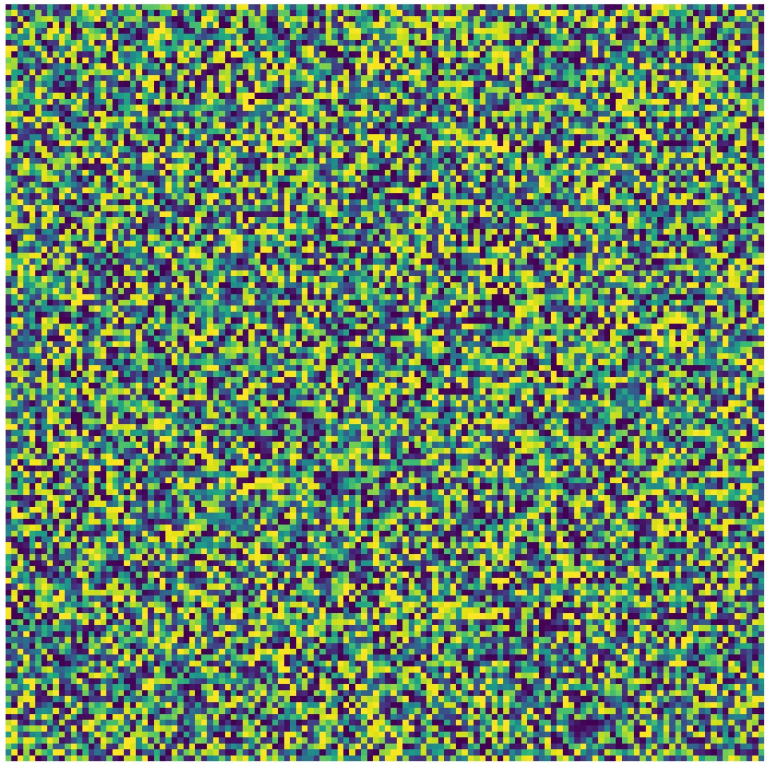
E. J. Hu, et al. “LoRA: Low-rank adaptation of large language models.” In ICLR 2022.



$\mathbf{uv}^T$

- Use low-rank approximation.
- However, lacks expressibility.

E. J. Hu, et al. “LoRA: Low-rank adaptation of large language models.” In ICLR 2022.



$$\sin(\omega \cdot \mathbf{u}\mathbf{v}^T)$$

- Foundational insight.
- Gives both parameter efficiency and expressability.

Y. Ji, S. Lucey, et al. “Sine Activated Low-Rank Matrices for Parameter Efficient Learning.” In ICLR 2025.

CBA's Adelaide Uni partnership spins out potentially game-changing AI capability

Patrick Buncsi 30 October 2024

Just two weeks after securing its partnership with Adelaide University's Australian Institute for Machine Learning (AIML), CommBank and the AIML built a potentially transformative "deep learning" capability expected to halve the speed for critical data processing functions within the bank.

Related News


Delivering on a digital tomorrow – Sam Shaheen, Australia Post


ASIC warns licensees governance mus AI adoption


AI Accelerator expands Westpac's AI to

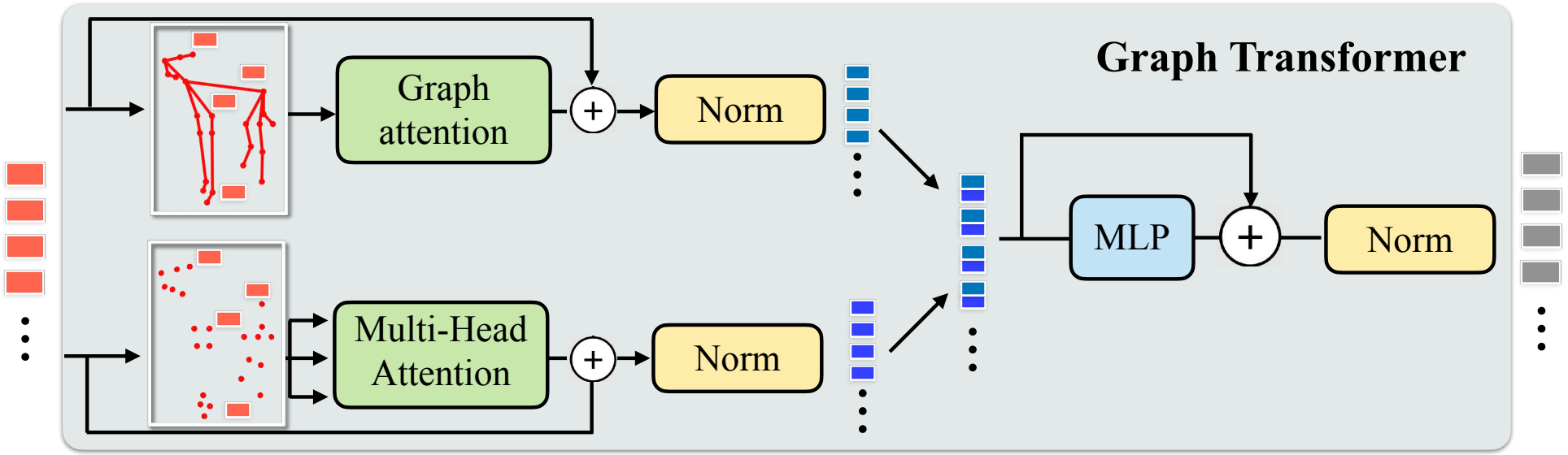

Unmasking the deepfake threat: Guida FSI execs, boards


Beyond Bank taps Boomi for loan syste revamp

https://fst.net.au/financial-services-news/cbas-adelaide-uni-partnership-spins-out-critical-ai-innovation/

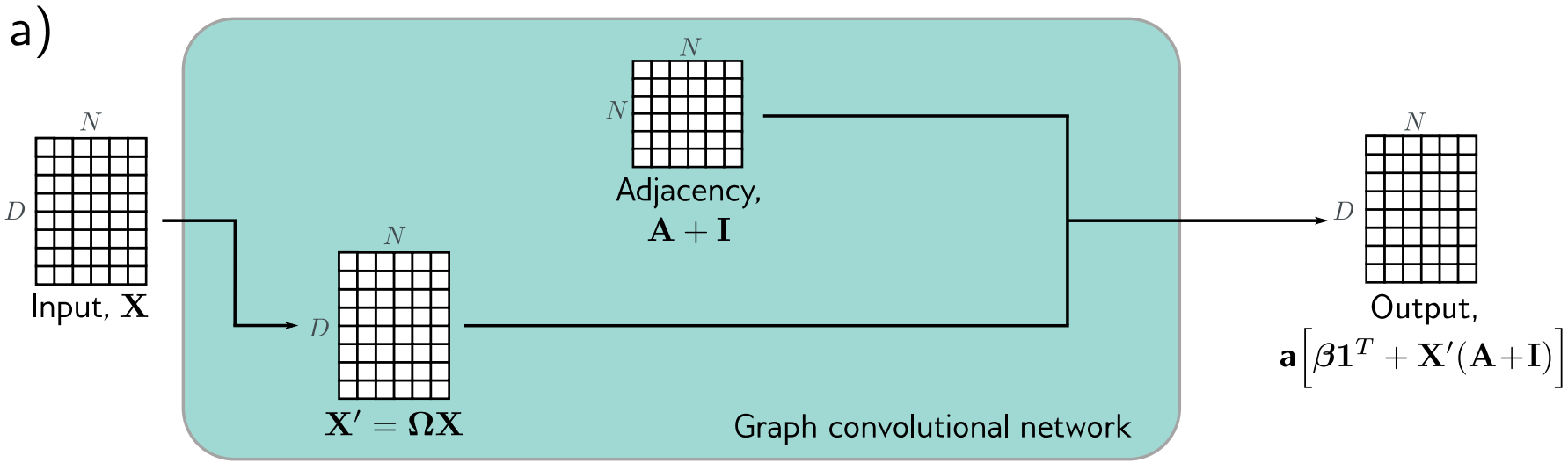
Beyond Vision and Language

3D Lifting Foundation Model (3D-LFM)



M. Dabhi, L. A. Jeni, and S. Lucey. "3D-LFM: Lifting Foundation Model." In CVPR 2024.

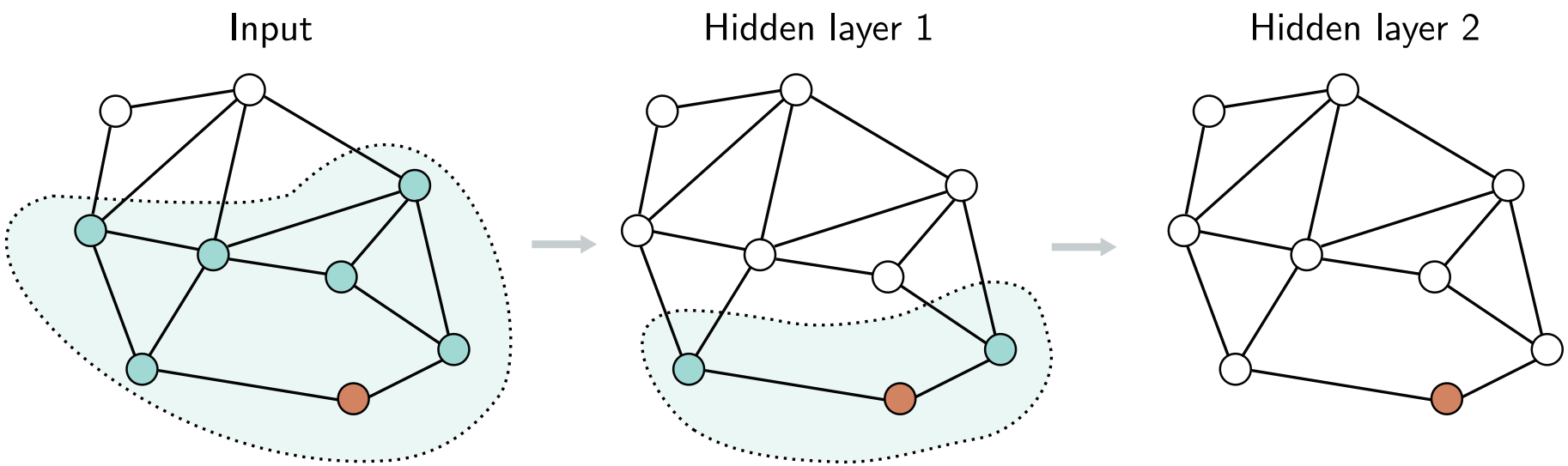
Graph Convolution Network



Only Ω and β are learned!!!

Prince “Understanding Deep Learning” 2023.

Graph Convolution Network



Prince “Understanding Deep Learning” 2023.

Graph Attention Networks

GRAPH ATTENTION NETWORKS

Petar Veličković*
Department of Computer Science and Technology
University of Cambridge
petar.velickovic@cst.cam.ac.uk

Guillem Cucurull*
Centre de Visió per Computador, UAB
gcucurull@gmail.com

Arantxa Casanova*
Centre de Visió per Computador, UAB
ar.casanova.8@gmail.com

Adriana Romero
Montréal Institute for Learning Algorithms
adriana.romero.soriano@umontreal.ca

Pietro Liò
Department of Computer Science and Technology
University of Cambridge
pietro.lio@cst.cam.ac.uk

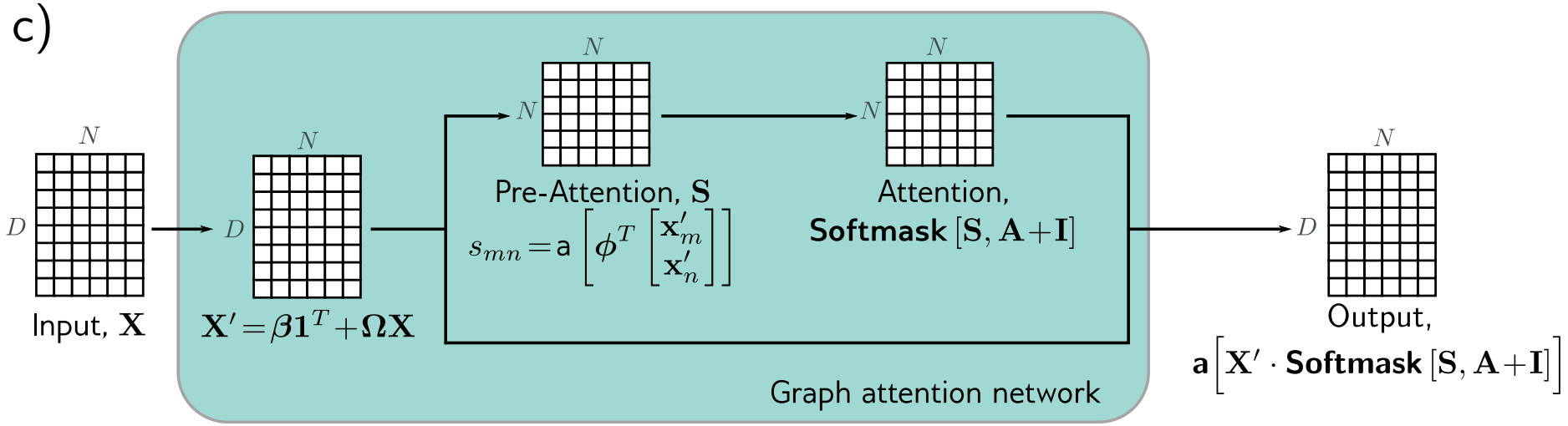
Yoshua Bengio
Montréal Institute for Learning Algorithms
yoshua.umontreal@gmail.com

ABSTRACT

We present graph attention networks (GATs), novel neural network architectures that operate on graph-structured data, leveraging masked self-attentional layers to address the shortcomings of prior methods based on graph convolutions or their approximations. By stacking layers in which nodes are able to attend over their neighborhoods' features, we enable (implicitly) specifying different weights to different nodes in a neighborhood, without requiring any kind of costly matrix operation (such as inversion) or depending on knowing the graph structure upfront. In this way, we address several key challenges of spectral-based graph neural networks simultaneously, and make our model readily applicable to inductive as well as transductive problems. Our GAT models have achieved or matched state-of-the-art results across four established transductive and inductive graph benchmarks: the *Cora*, *Citeseer* and *Pubmed* citation network datasets, as well as a *protein-protein interaction* dataset (wherein test graphs remain unseen during training).

A. Velickovic et al. “Graph Attention Networks”, ICLR 2018.

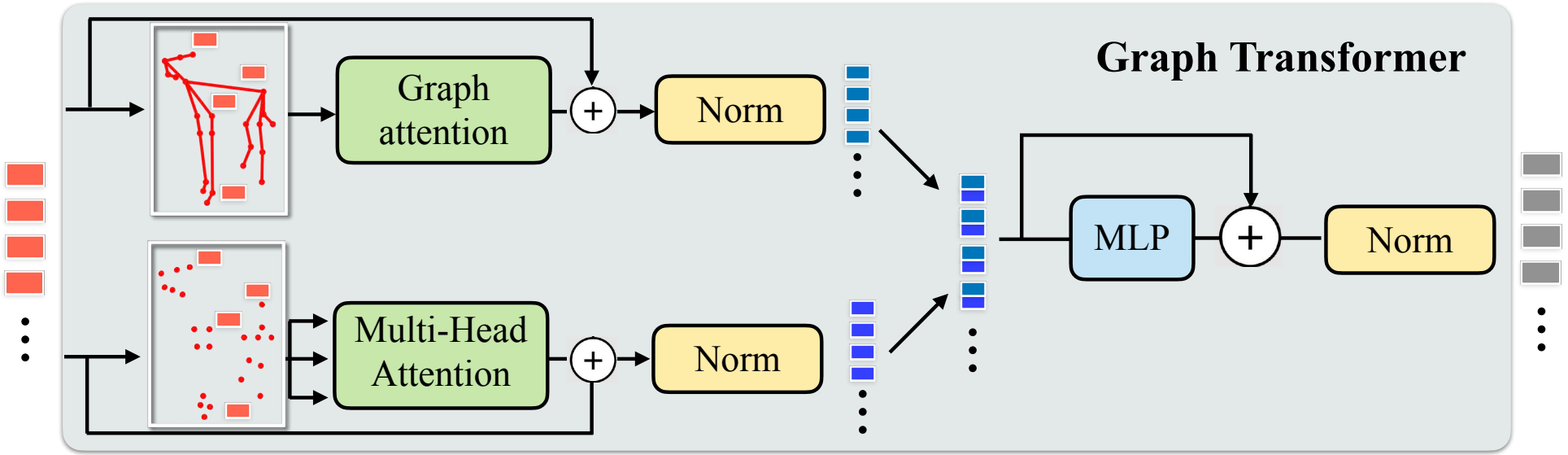
Graph Attention Networks



Ω, ϕ and β are learned!!!

Prince “Understanding Deep Learning” 2023.

Reminder: 3D Lifting Foundation Model (3D-LFM)



M. Dabhi, L. A. Jeni, and S. Lucey. "3D-LFM: Lifting Foundation Model." In CVPR 2024.

Flawed Original video

Corrected Rendered video

Error! Observe the right leg flip!

Our corrected re-render 🙏 to…

$20m AI research centre set up in South Australia

Justin Hendry

Editor

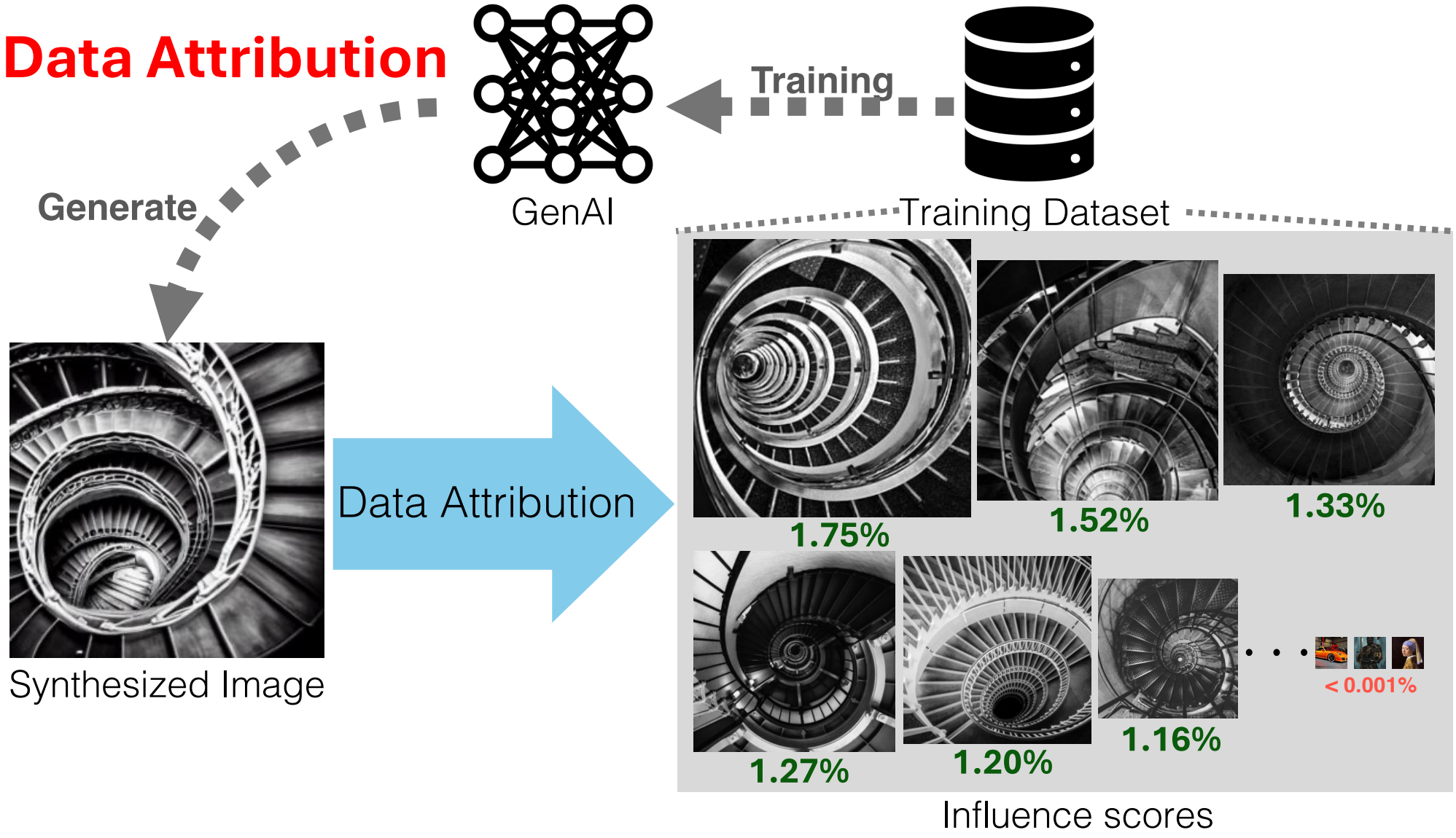
9 December 2024

Share

A new AI research centre has been set up in South Australia, bringing together experts from the CSIRO's digital arm Data61 and the Australian Institute for Machine Learning (AIML) to work on adoption challenges.

Jointly funded by the University of Adelaide, CSIRO and the South Australian government, the Responsible AI Research (RAIR) Centre aims to tackle issues of AI explainability, hallucinations and misinformation.

The $20 million centre will focus on building foundational models and better understanding the inner workings of AI systems when it becomes fully operational at Adelaide's Lot Fourteen early next year.


Data Attribution
Training
Generate
GenAI
Training Dataset
Synthesized Image
Data Attribution
1.75%
1.52%
1.33%
1.27%
1.20%
1.16%
< 0.001%
Influence scores

Summary

- AI growth is dependent upon exponential energy, data and compute.
- Understanding the basic mechanics of deep learning has the potential to unlock billions of dollars in efficiency and sovereign capability.
- Australia is already building foundational models from scratch!!