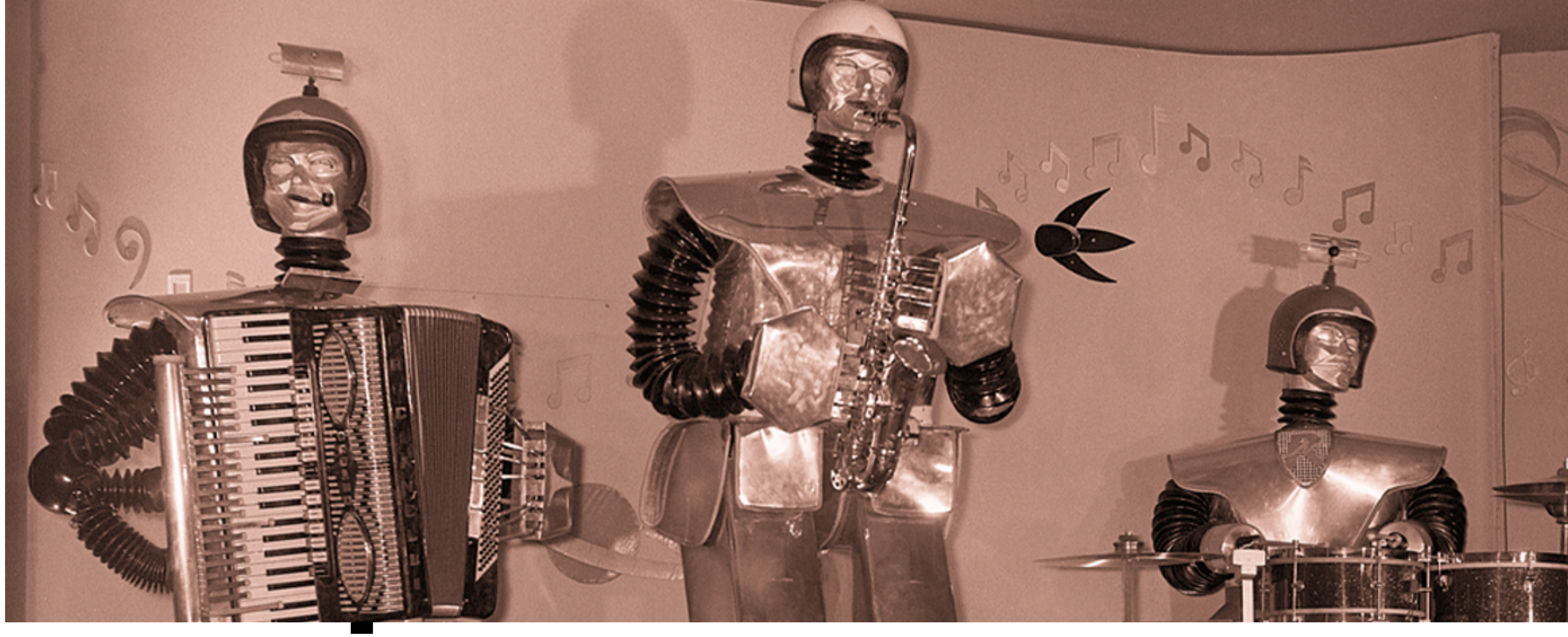


Artificial Neuroscience

metrology and engineering for Deep Learning using Linear Algebra

Mark Sandler 27 November 2024

Artificial what?

Aims of this talk

Artificial what?

Aims of this talk

- What’s the problem to solve here?

Artificial what?

Aims of this talk

- What's the problem to solve here?
  - The lack of rigour in much Applied AI research today!

Artificial what?

Aims of this talk

- What's the problem to solve here?
  - The lack of rigour in much Applied AI research today!
  - The lack of an engineering approach to building AI models

Artificial what?

Aims of this talk

- What’s the problem to solve here?
  - The lack of rigour in much Applied AI research today!
  - The lack of an engineering approach to building AI models
- “Artificial brains” are being created that

Artificial what?

Aims of this talk

- What's the problem to solve here?
    - The lack of rigour in much Applied AI research today!
    - The lack of an engineering approach to building AI models
- "Artificial brains" are being created that
    - Aren't understood because their structures are so huge that many people just give up trying

Artificial what?

Aims of this talk

- What's the problem to solve here?
  - The lack of rigour in much Applied AI research today!
  - The lack of an engineering approach to building AI models
- "Artificial brains" are being created that
  - Aren't understood because their structures are so huge that many people just give up trying
  - Aren't evaluated properly, relying only on benchmarking, which is an engineering not a cognition procedure. What about ecological validity?

Artificial what?

Aims of this talk

- What’s the problem to solve here?
  - The lack of rigour in much Applied AI research today!
  - The lack of an engineering approach to building AI models
- “Artificial brains” are being created that
  - Aren’t understood because their structures are so huge that many people just give up trying
  - Aren’t evaluated properly, relying only on benchmarking, which is an engineering not a cognition procedure. What about ecological validity?
  - Are trained often with data that isn’t optimal for the task

Artificial what?

Aims of this talk

- What’s the problem to solve here?
  - The lack of rigour in much Applied AI research today!
  - The lack of an engineering approach to building AI models
- “Artificial brains” are being created that
  - Aren’t understood because their structures are so huge that many people just give up trying
  - Aren’t evaluated properly, relying only on benchmarking, which is an engineering not a cognition procedure. What about ecological validity?
  - Are trained often with data that isn’t optimal for the task
  - Interact with each other and with humans, proliferating without care and understanding

Artificial what?

Aims of this talk

- What's the problem to solve here?
  - The lack of rigour in much Applied AI research today!
  - The lack of an engineering approach to building AI models
- "Artificial brains" are being created that
  - Aren't understood because their structures are so huge that many people just give up trying
  - Aren't evaluated properly, relying only on benchmarking, which is an engineering not a cognition procedure. What about ecological validity?
  - Are trained often with data that isn't optimal for the task
  - Interact with each other and with humans, proliferating without care and understanding
  - Consume unsustainable amounts of energy

So, Artificial Neuroscience

What’s in a name?

So, Artificial Neuroscience

What's in a name?

- Neuroscience - encompassing various scientific disciplines dealing with the structure, development, function, chemistry, pharmacology, and pathology of the nervous system (the brain, spinal cord, and peripheral nervous system).

So, Artificial Neuroscience

What’s in a name?

- Neuroscience - encompassing various scientific disciplines dealing with the structure, development, function, chemistry, pharmacology, and pathology of the nervous system (the brain, spinal cord, and peripheral nervous system).
- It combines physiology, anatomy, molecular biology, developmental biology, cytology, psychology, physics, computer science, chemistry, medicine, statistics, and mathematical modeling to understand the fundamental and emergent properties of neurons, glia and neural circuits

So, Artificial Neuroscience

What's in a name?

- Neuroscience - encompassing various scientific disciplines dealing with the structure, development, function, chemistry, pharmacology, and pathology of the nervous system (the brain, spinal cord, and peripheral nervous system).
- It combines physiology, anatomy, molecular biology, developmental biology, cytology, psychology, physics, computer science, chemistry, medicine, statistics, and mathematical modeling to understand the fundamental and emergent properties of neurons, glia and neural circuits
- We need an equivalent definition and corresponding set of disciplines. This goes beyond Computer Science. Several branches of mathematics are vital. But also Engineering, Behavioural Sciences etc, and real application domain expertise

Holistic understanding …

- Artificial MRI
  - Linear Algebra and statistical mechanics for observing, measuring & understanding the learning and inference processes by observing and measuring
  - Mechanistic interpretability: exposing emergent structures and neural circuits
- Experimental Artificial Neuroscience
  - Beyond benchmarking: developing and testing behavioural hypotheses in ecologically valid experiments (incl. ablation and “surgery”)
  - Designing test data to fully probe behaviours
  - Exploring failure modes, not just accuracy
- Artificial Cognitive Development
  - Curriculum learning, transfer learning, domain adaptation, etc
- Machine Behavioural Science
  - Applying social sciences to collective behaviours of multiple AIs, AIs + humans

My journey..

- I recommended that my PhD student Rodrigo view DSP lectures from MIT (c.1978).
  - In return he recommended that I watch LA + DL lectures by Gilbert Strang. A revelation!
- From what I learned, I hypothesised about
  - Developing metrics for the internal state of a NN
  - Changing NN dynamics by tinkering with Singular Values
  - Initialising NNs using low rank layers
  - Speeding up training
  - Simplifying backprop computations
- I’ve been finding more and more evidence for all these things and recently started a UK-funded small project to build expertise

AI research harms the planet

- R. Couillet, D. Trystram and T. Ménissier, "The Submerged Part of the AI-Ceberg [Perspectives]," in IEEE Signal Processing Magazine, vol. 39, no. 5, pp. 10-17, Sept. 2022, doi: 10.1109/MSP.2022.3182938.
- Looking at energy consumption due to Deep Learning

Romain Couillet, Denis Trystram, and Thierry Ménissier

The Submerged Part of the AI-Ceberg

This article discusses the contradiction between the exploding energy demand of artificial intelligence (AI) and the information and communication (ICT) industry as a whole and the parallel strong request for energy sobriety imposed by the need to mitigate the impact of climate change and the anticipated collapse of civilization as we know it. Under the form of an open reflection on the goods and evils of AI, the article raises the suggestion of a drastic change in the AI paradigm, more in phase with the vital obligation to design a more resilient society.

Deep learning: The new Eldorado?

Over the past decade, the considerable growth of the digital world, propelled by AI, has had spectacular effects in a few scientific fields, such as computer vision and natural language processing, and given rise to many new technologies and consumer products. Today, this development even claims to revolutionize many other areas of our society. This revolution indeed concerns many aspects of our lives: we (and world with a few clicks, to name only a few [1], [2].

Deep neural network learning is at the forefront of this development and has spread rapidly, far beyond the confidential fields of its beginnings. In a matter of 10 years, this specific computer science tool—theorized as early as the 1980s [3]—has reached all levels of society: in companies, institutions, research laboratories, in virtually all engineering disciplines as well as life sciences. Easy to use as a black box thanks to an important software development effort—multiple "plug-and-play" solutions have been developed for engineers (and not only computer science experts), such as the popular TensorFlow library [4], [5]—deep learning has effectively replaced "conventional" tools (particularly in computer vision and natural language processing), imposing a form of radical monopoly on scientific domains. The radical monopoly of a tool is understood in the sense defined by Illich [34]: it alters the normative system of knowledge generation and sharing. Calls for projects, dedicated conferences, and job world really be on the way [6]? Of course, investing in deep learning and AI involves delegating to the machine the power of our decisions, which comes with many ethics and equity concerns [8]; as Stephen Hawking pessimistically stated in 2014, "The development of full artificial intelligence could spell the end of the human race.… It would take off on its own, and redesign itself at an ever-increasing rate. Humans, who are limited by slow biological evolution, couldn't compete, and would be superseded." [7] (As discussed next, this seemingly science-fictional statement is more profoundly explored by Illich [34] regarding the dangers of societal dependence on oil and machines, induced by an increasing loss of common knowledge and know-how that are moved from the population to computers and machines.) Yet, the many promises of AI clearly tip the scales toward increasingly more investment in the field [10]. Besides, researchers now deeply investigate the question of fairness in AI to smooth out these thorny angles [9].

Artificial Psychology

- Critiques familiar practice in DL research
- Proposes methodologies and roles for psychologists
- Appropriate experimentation delivers insights into black-box systems -> XAI

Psychonomic Bulletin and Review (2021) 28:454–475
https://doi.org/10.3758/s13423-020-01825-5

THEORETICAL REVIEW

Artificial cognition: How experimental psychology can help generate explainable artificial intelligence

J. Eric T. Taylor[1,2] · Graham W. Taylor[1,2]

Accepted: 2 October 2020 / Published online: 6 November 2020

Abstract
Artificial intelligence powered by deep neural networks has reached a level of complexity where it can be difficult or impossible to express how a model makes its decisions. This black-box problem is especially concerning when the model makes decisions with consequences for human well-being. In response, an emerging field called explainable artificial intelligence (XAI) aims to increase the interpretability, fairness, and transparency of machine learning. In this paper, we describe how cognitive psychologists can make contributions to XAI. The human mind is also a black box, and cognitive psychologists have over 150 years of experience modeling it through experimentation. We ought to translate the methods and rigor of cognitive psychology to the study of artificial black boxes in the service of explainability. We provide a review of XAI for psychologists, arguing that current methods possess a blind spot that can be complemented by the experimental cognitive tradition. We also provide a framework for research in XAI, highlight exemplary cases of experimentation within XAI inspired by psychological science, and provide a tutorial on experimenting with machines. We end by noting the advantages of an experimental approach and invite other psychologists to conduct research in this exciting new field.

Artificial Psychology #2

- discover shape bias in a Comp Vis system by applying Cog Psych to a DNN.
- hence possibilities of ‘exposing hidden computational properties of DNN’
- Proceedings of the 34 th International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017

Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study

Samuel Ritter [*1] David G.T. Barrett [*1] Adam Santoro [1] Matt M. Botvinick [1]

Abstract

Deep neural networks (DNNs) have advanced performance on a wide range of complex tasks, rapidly outpacing our understanding of the nature of their solutions. While past work sought to advance our understanding of these models, none has made use of the rich history of problem descriptions, theories, and experimental methods developed by cognitive psychologists to study the human mind. To explore the potential value of these tools, we chose a well-established analysis from developmental psychology that explains how children learn word labels for objects, and applied that analysis to DNNs. Using datasets of stimuli inspired by the original cognitive psychology experiments, we find that state-of-the-art one shot learning models trained on ImageNet exhibit a similar bias to that observed in humans: they prefer to categorize objects according to shape rather than color. The magnitude of this shape bias varies greatly among architecturally identical, but differently seeded models, and even fluctuates within seeds throughout training, despite nearly equivalent classification performance. These results demonstrate the capability of tools from cognitive psychology for exposing hidden computational properties of DNNs, while concurrently providing us with a computational model for human word learning.

Machine Behaviour

- Paper has many citations
- Argues for social science techniques to be applied to machine intelligence
- Out of MIT.
  - Lovely web site, though no changes since 2019.

REVIEW

https://doi.org/10.1038/s41586-019-1138-y

Machine behaviour

Iyad Rahwan[1,2,3,34]*, Manuel Cebrian[1,34], Nick Obradovich[1,34], Josh Bongard[4], Jean-François Bonnefon[5], Cynthia Breazeal[1], Jacob W. Crandall[6], Nicholas A. Christakis[7,8,9,10], Iain D. Couzin[11,12,13], Matthew O. Jackson[14,15,16], Nicholas R. Jennings[17,18], Ece Kamar[19], Isabel M. Kloumann[20], Hugo Larochelle[21], David Lazer[22,23,24], Richard McElreath[25,26], Alan Mislove[27], David C. Parkes[28,29], Alex 'Sandy' Pentland[1], Margaret E. Roberts[30], Azim Shariff[31], Joshua B. Tenenbaum[32] & Michael Wellman[33]

Machines powered by artificial intelligence increasingly mediate our social, cultural, economic and political interactions. Understanding the behaviour of artificial intelligence systems is essential to our ability to control their actions, reap their benefits and minimize their harms. Here we argue that this necessitates a broad scientific research agenda to study machine behaviour that incorporates and expands upon the discipline of computer science and includes insights from across the sciences. We first outline a set of questions that are fundamental to this emerging field and then explore the technical, legal and institutional constraints on the study of machine behaviour.

Interpretability

http://arxiv.org/abs/2208.06894

- Aware of visualisation and auralisation of layers and weights
- Improves on this using formal methods from Linear Algebra
- Links to interpretability but not to controlling network convergence

The SVD of Convolutional Weights: A CNN Interpretability Framework[*]

Brenda Praggastis[†] Davis Brown[†] Carlos Ortiz Marrero[‡] Emilie Purvine[†]
Madelyn Shapiro[†] Bei Wang[§]

August 16, 2022

Abstract

Deep neural networks used for image classification often use convolutional filters to extract distinguishing features before passing them to a linear classifier. Most interpretability literature focuses on providing semantic meaning to convolutional filters to explain a model's reasoning process and confirm its use of relevant information from the input domain. Fully connected layers can be studied by decomposing their weight matrices using a singular value decomposition, in effect studying the correlations between the rows in each matrix to discover the dynamics of the map. In this work we define a singular value decomposition for the weight tensor of a convolutional layer, which provides an analogous understanding of the correlations between filters, exposing the dynamics of the convolutional map. We validate our definition using recent results in random matrix theory. By applying the decomposition across the linear layers of an image classification network we suggest a framework against which interpretability methods might be applied using hypergraphs to model class separation. Rather than looking to the activations to explain the network, we use the singular vectors with the greatest corresponding singular values for each linear layer to identify those features most important to the network. We illustrate our approach with examples and introduce the DeepDataProfiler library, the analysis tool used for this study.

Deep Learning Metrology

- Distribution of eigenvalues is heavy tailed in large, well-trained networks
- Various stages of training identified by changing distribution
- Toolbox called 'weightwatchers'

Journal of Machine Learning Research 22 (2021) 1-73 Submitted 4/20; Published 6/21

Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning

Charles H. Martin CHARLES@CALCULATIONCONSULTING.COM
Calculation Consulting
8 Locksley Ave, 6B
San Francisco, CA 94122

Michael W. Mahoney MMAHONEY@STAT.BERKELEY.EDU
ICSI and Department of Statistics
University of California at Berkeley
Berkeley, CA 94720

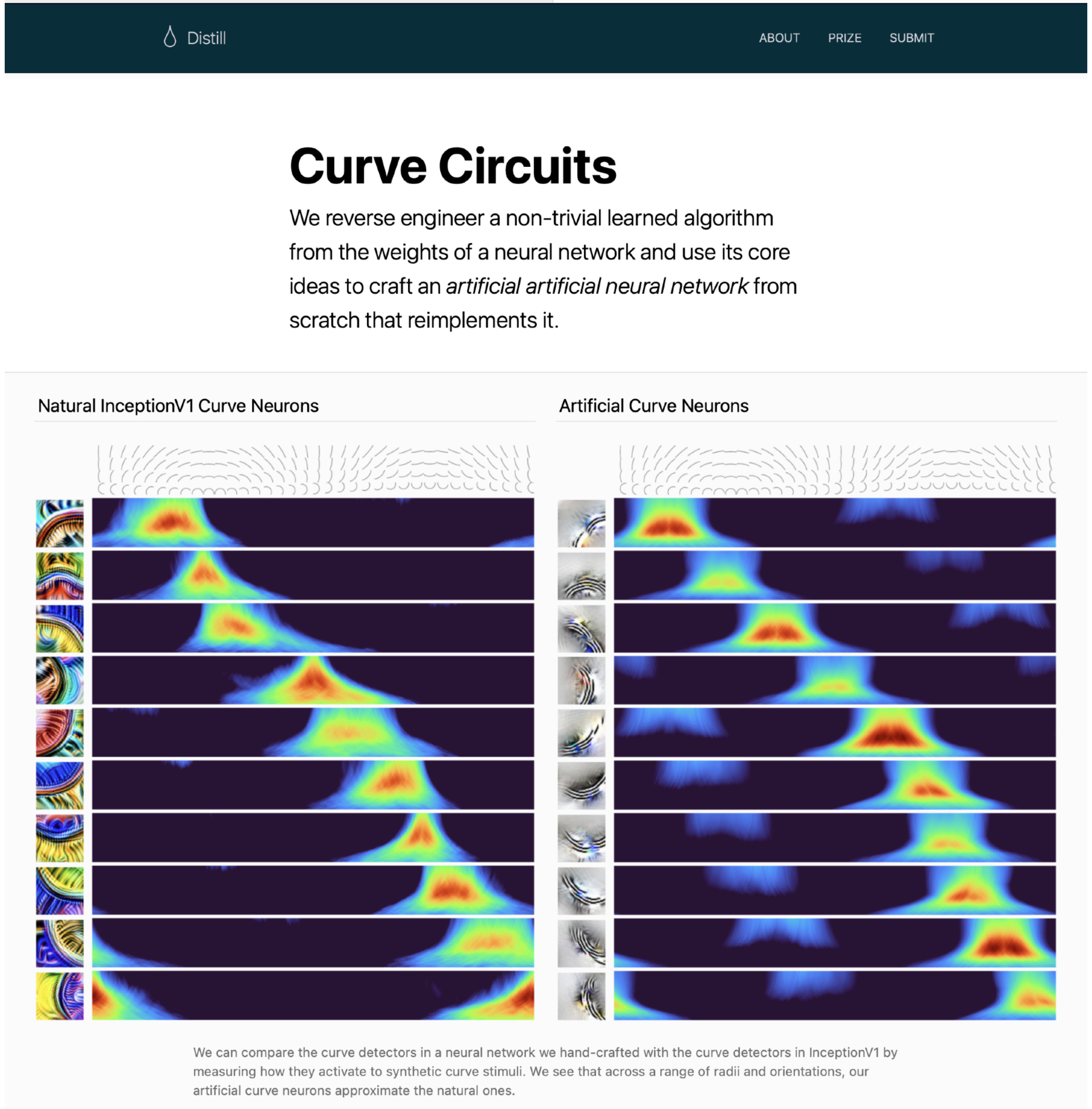
Abstract

Random Matrix Theory (RMT) is applied to analyze the weight matrices of Deep Neural Networks (DNNs), including both production quality, pre-trained models such as AlexNet and Inception, and smaller models trained from scratch, such as LeNet5 and a miniature-AlexNet. Empirical and theoretical results clearly indicate that the DNN training process itself implicitly implements a form of *Self-Regularization*, implicitly sculpting a more regularized energy or penalty landscape. In particular, the empirical spectral density (ESD) of DNN layer matrices displays signatures of traditionally-regularized statistical models, even in the absence of exogenously specifying traditional forms of explicit regularization, such as Dropout or Weight Norm constraints. Building on relatively recent results in RMT, most notably its extension to Universality classes of Heavy-Tailed matrices, and applying them to these empirical results, we develop a theory to identify *5+1 Phases of Training*, corresponding to increasing amounts of *Implicit Self-Regularization*. These phases can be observed during the training process as well as in the final learned DNNs. For smaller and/or older DNNs, this Implicit Self-Regularization is like traditional Tikhonov regularization, in that there is a "size scale" separating signal from noise. For state-of-the-art DNNs, however, we identify a novel form of *Heavy-Tailed Self-Regularization*, similar to the self-organization seen in the statistical physics of disordered systems (such as classical models of actual neural activity). This results from correlations arising at all size scales, which for DNNs arises implicitly due to the training process itself. This implicit Self-Regularization can depend strongly on the many knobs of the training process. In particular, we demonstrate that we can cause a small model to exhibit all 5+1 phases of training simply by changing the batch size. Our results suggest that large, well-trained DNN architectures should exhibit Heavy-Tailed Self-Regularization, and we discuss the theoretical and practical implications of this.

Discovering functional blocks

https://distill.pub/2020/circuits/curve-circuits/

- Image processing DLs learn curve detectors (and higher order function)
- Replace identified, learning ‘circuits’ with custom designed, low-power/efficient circuits
- Performance is comparable
- Potential for commoditising DL models



Simplifying computation

http://arxiv.org/abs

Learning Low-rank Deep Neural Networks via Singular Vector Orthogonality Regularization and Singular Value Sparsification

Huanrui Yang[1], Minxue Tang[2], Wei Wen[1], Feng Yan[3], Daniel Hu[4], Ang Li[1], Hai Li[1], Yiran Chen[1]
[1]Department of Electrical and Computer Engineering, Duke University
[2]Department of Electronic Engineering, Tsinghua University
[3]Computer Science and Engineering Department, University of Nevada, Reno
[4]Newport High School, Bellevue, WA

Abstract

Modern deep neural networks (DNNs) often require high memory consumption and large computational loads. In order to deploy DNN algorithms efficiently on edge or mobile devices, a series of DNN compression algorithms have been explored, including factorization methods. Factorization methods approximate the weight matrix of a DNN layer with the multiplication of two or multiple low-rank matrices. However, it is hard to measure the ranks of DNN layers during the training process. Previous works mainly induce low-rank through implicit approximations or via costly singular value decomposition (SVD) process on every training step. The former approach usually induces a high accuracy loss while the latter has a low efficiency. In this work, we propose ***SVD training****, the first method to explicitly achieve low-rank DNNs during training without applying SVD on every step. SVD training first decomposes each layer into the form of its full-rank SVD, then performs training directly on the decomposed weights. We add orthogonality regularization to the singular vectors, which ensure the valid form of SVD and avoid gradient vanishing/exploding. Low-rank is encouraged by applying sparsity-inducing regularizers on the singular values of each layer. Singular value pruning is applied at the end to explicitly reach a low-rank model. We empirically show that SVD training can significantly reduce the rank of DNN layers and achieve higher reduction on computation load under the same accuracy, comparing to not only previous factorization methods but also state-of-the-art filter pruning methods.*

Initialising networks

- https://doi.org/10.1007/978-3-030-30484-3_13

Singular Value Decomposition and Neural Networks

Bernhard Bermeitinger[1,2][0000−0002−2524−1850], Tomas Hrycej[1], and Siegfried Handschuh[1,2]

[1] Chair of Data Science, Institute of Computer Science
University of St.Gallen, St.Gallen, Switzerland
{bernhard.bermeitinger,tomas.hrycej,siegfried.handschuh}@unisg.ch
[2] University Passau, Passau, Germany
{bernhard.bermeitinger,siegfried.handschuh}@uni-passau.de

Abstract. Singular Value Decomposition (SVD) constitutes a bridge between the linear algebra concepts and multi-layer neural networks—it is their linear analogy. Besides of this insight, it can be used as a good initial guess for the network parameters, leading to substantially better optimization results.

Keywords: Singular Value Decomposition · Neural Network · Deep Neural Network · Initialization · Optimization · Conjugate Gradient

Early original work

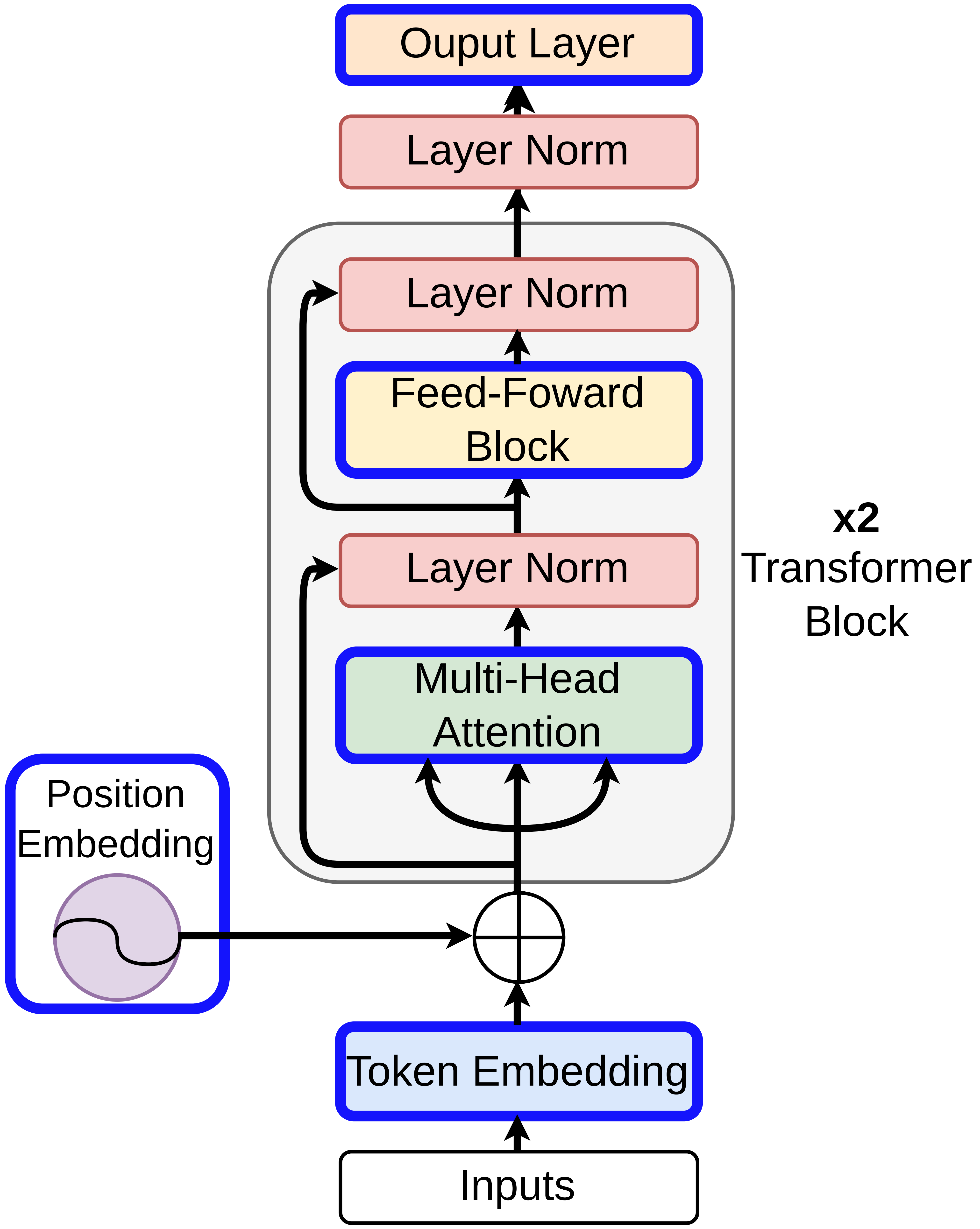
Thanks to Gabriel Mason-Williams

Preliminary Results

By Gabriel Mason-Williams

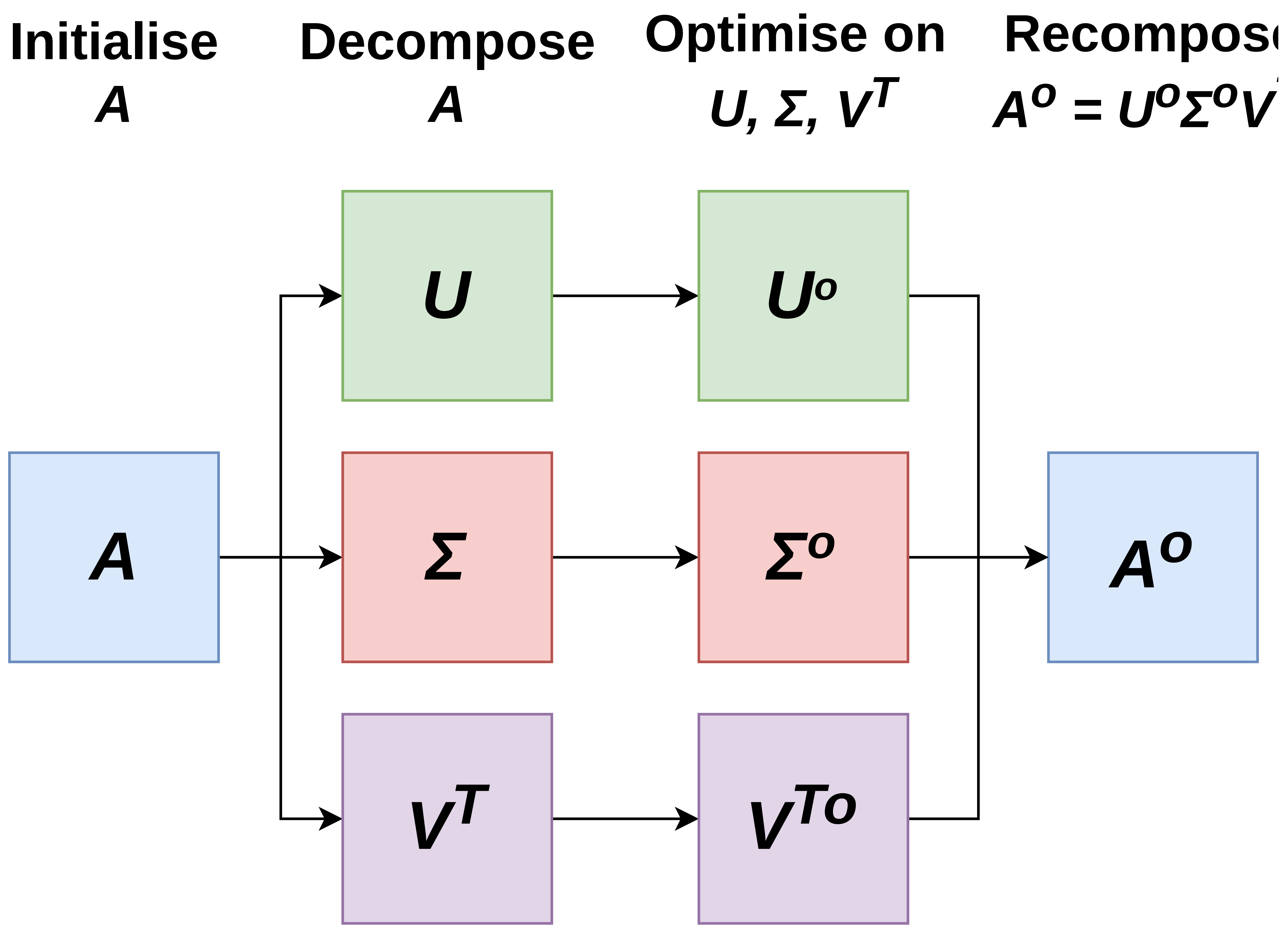
Explore reduction in rank by observing grokking

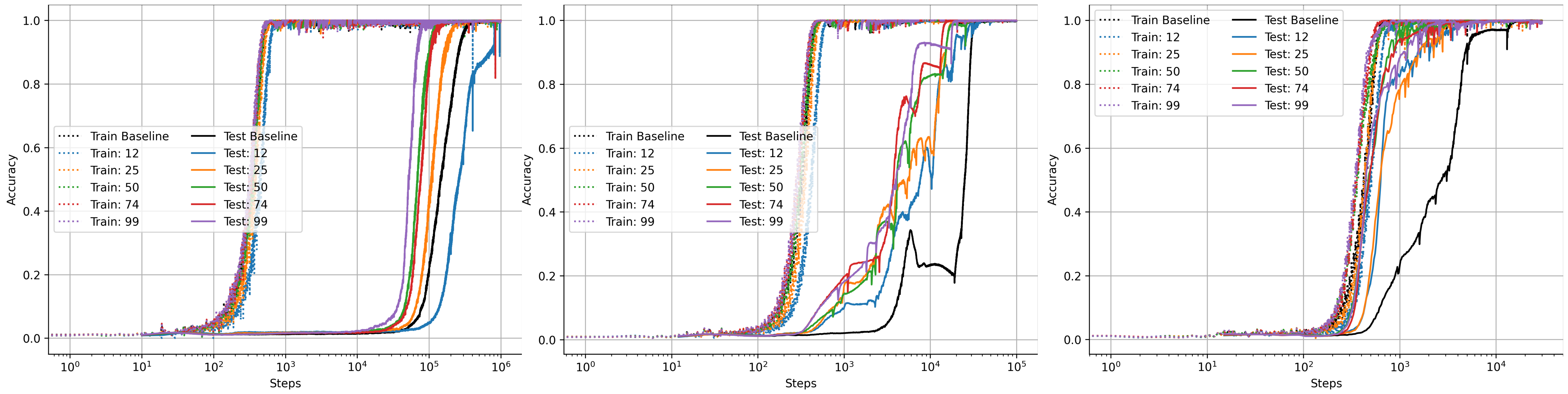
- Train beyond maximum training accuracy, thereby increasing test accuracy
- Transformer architecture
  - 3 tasks
    - Modulo arithmetic (2)
    - LLM trained on Shakespeare corpus
  - Token Embedding Ranks: 12, 25, 50, 74 and **99**
  - Position Embedding Ranks: 1, 2, 3, 4 and **5**
  - Multi-Head Attention Ranks: 16, 32, 64, 96 and **128**
  - Feed-forward Block Ranks: 16, 32, 64, 96 and **128**
  - Output Layer Ranks: 12, 25, 50, 74 and **99**



Training regime

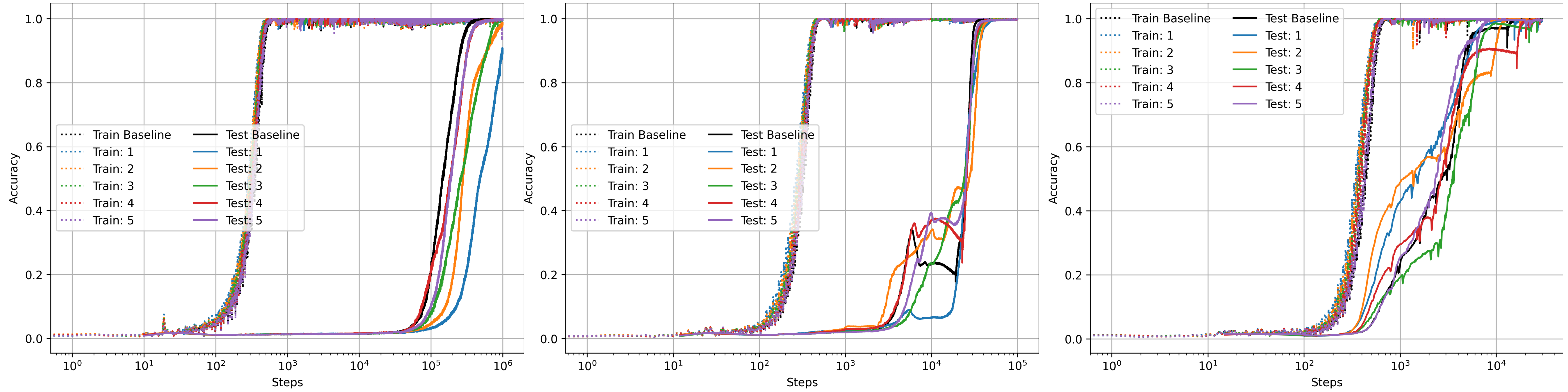
- Various layers individually and collectively





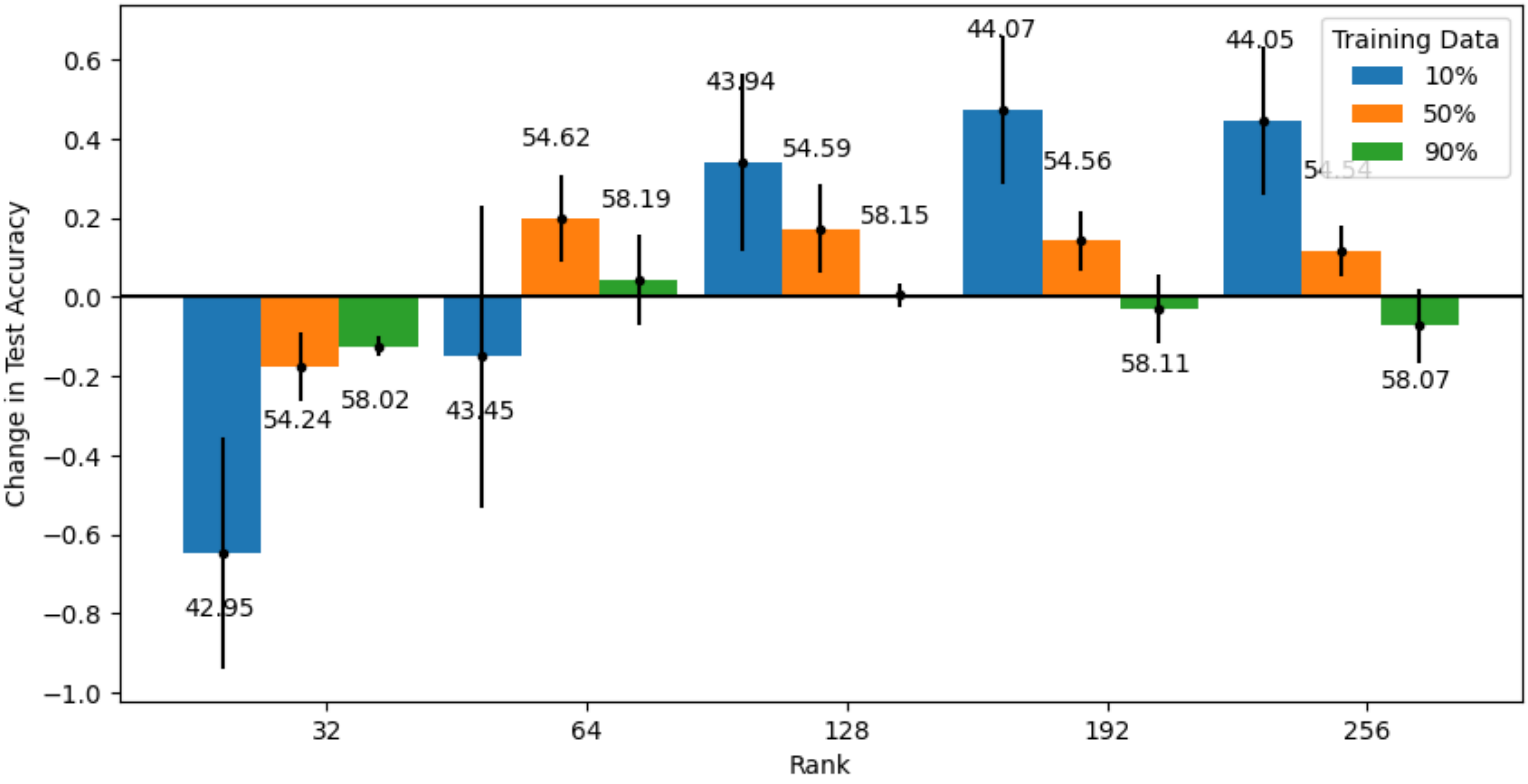
(a) 50% Training Data (b) 65% Training Data (c) 80% Training Data

Figure 3: Train (dotted) and test (solid) accuracy with decomposed learning on the token embedding layer using ranks 12, 25, 50, 74 and 99, in comparison with the baseline (black) normally trained model.



(a) 50% Training Data (b) 65% Training Data (c) 80% Training Data

Figure 4: Train (dotted) and test (solid) accuracy with decomposed learning on the position embedding layer using ranks 1, 2, 3, 4 and 5, in comparison with the baseline (black) normally trained model .



Position Embedding Layer. Deviation from the baseline test accuracy.

Computational efficiencies

To be done!

- For k << r in $$A = \sum_{i=1}^{r} u_i \sigma_i v_i^T$$
- Total MACs & parameters per layer are ~
  - k.n + m ~ (k+1)n rather than m.n
  - So ideally k/n < 0.2
- Can we train as sum of rank-1?

… leading to better engineering…

… leading to better engineering…

- GPUs are general purpose AI chips!

… leading to better engineering…

- GPUs are general purpose AI chips!
- **Low rank structures** and algorithms for efficient learning and inference
  - Relationship with scalability and data set size
  - explore other fast matrix-vector techniques

… leading to better engineering…

- GPUs are general purpose AI chips!
- **Low rank structures** and algorithms for efficient learning and inference
  - Relationship with scalability and data set size
  - explore other fast matrix-vector techniques
- **Custom hardware** (incl. 1 bit – well-established in Signal Processing)
  - Relationship between 1 bit processing, ‘oversized’ layers & Universality

… leading to better engineering…

- GPUs are general purpose AI chips!
- **Low rank structures** and algorithms for efficient learning and inference
  - Relationship with scalability and data set size
  - explore other fast matrix-vector techniques
- **Custom hardware** (incl. 1 bit – well-established in Signal Processing)
  - Relationship between 1 bit processing, ‘oversized’ layers & Universality
- Re-engineer identified **neural circuits** (Mechanistic Interpretability) with purpose-defined sub-systems (e.g. curve detectors)
  - Towards a building block approach to neural networks

… leading to better engineering…

- GPUs are general purpose AI chips!
- **Low rank structures** and algorithms for efficient learning and inference
  - Relationship with scalability and data set size
  - explore other fast matrix-vector techniques
- **Custom hardware** (incl. 1 bit – well-established in Signal Processing)
  - Relationship between 1 bit processing, ‘oversized’ layers & Universality
- Re-engineer identified **neural circuits** (Mechanistic Interpretability) with purpose-defined sub-systems (e.g. curve detectors)
  - Towards a building block approach to neural networks
- Explore **virtual analog** for digital equivalents to analog models of biological neural circuits: trainability ~ optimized circuit design

… leading to better engineering…

- GPUs are general purpose AI chips!
- **Low rank structures** and algorithms for efficient learning and inference
  - Relationship with scalability and data set size
  - explore other fast matrix-vector techniques
- **Custom hardware** (incl. 1 bit – well-established in Signal Processing)
  - Relationship between 1 bit processing, ‘oversized’ layers & Universality
- Re-engineer identified **neural circuits** (Mechanistic Interpretability) with purpose-defined sub-systems (e.g. curve detectors)
  - Towards a building block approach to neural networks
- Explore **virtual analog** for digital equivalents to analog models of biological neural circuits: trainability ~ optimized circuit design
- Explore **Non-linear Digital Wave Filters** as compact, non-linear building blocks in new, heterogeneous DL models

… leading to better engineering…

- GPUs are general purpose AI chips!
- **Low rank structures** and algorithms for efficient learning and inference
  - Relationship with scalability and data set size
  - explore other fast matrix-vector techniques
- **Custom hardware** (incl. 1 bit – well-established in Signal Processing)
  - Relationship between 1 bit processing, ‘oversized’ layers & Universality
- Re-engineer identified **neural circuits** (Mechanistic Interpretability) with purpose-defined sub-systems (e.g. curve detectors)
  - Towards a building block approach to neural networks
- Explore **virtual analog** for digital equivalents to analog models of biological neural circuits: trainability ~ optimized circuit design
- Explore **Non-linear Digital Wave Filters** as compact, non-linear building blocks in new, heterogeneous DL models

… for green, explainable, commoditisable AI

Conclusions…