Finding knowledge in Black Boxes

An intro to mechanistic interpretability

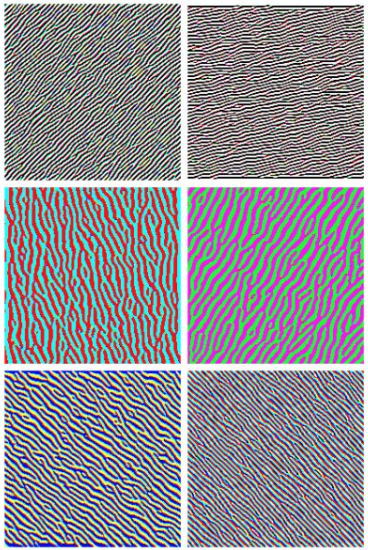
By Benjamin Sturgeon (shocklab @UCT)

What we'll cover

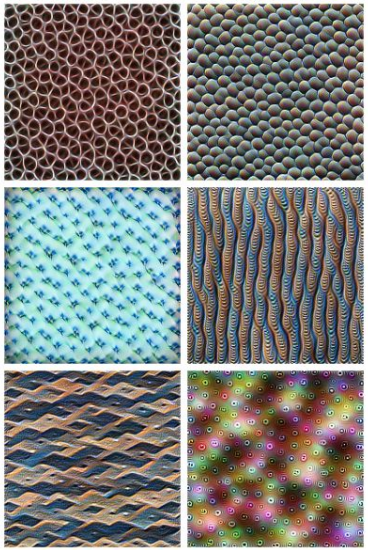
- Important background info
- The goals of mech interp
- Some interesting case studies
- Theoretical framings of the field
- A promising future

Where did this field come from

- Aims to address increasingly urgent demands for AI safety
- Trustworthiness, reliability, regulatory compliance
- Possibility of enhancing human discovery



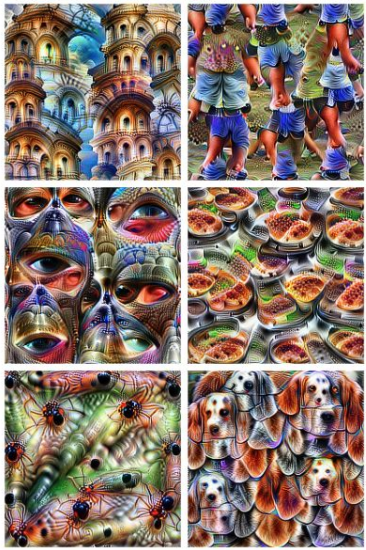
Edges (layer conv2d0)



Textures (layer mixed3a)

Patterns (layer mixed4a)

Parts (layers mixed4b & mixed4c)



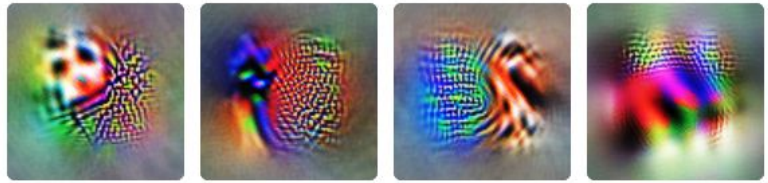
Objects (layers mixed4d & mixed4e)

Curve detectors

High-Low Frequency detectors

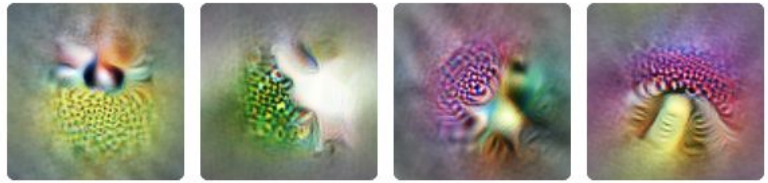
ALEXNET

Krizhevsky et al. [34]



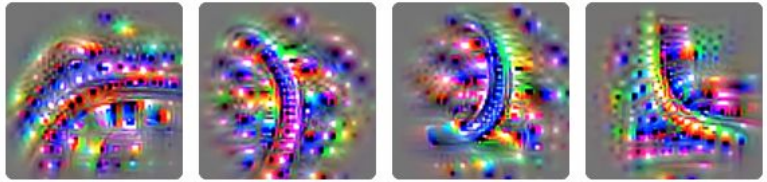
INCEPTIONV1

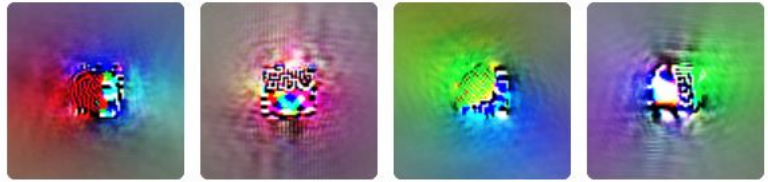
Szegedy et al. [26]



VGG19

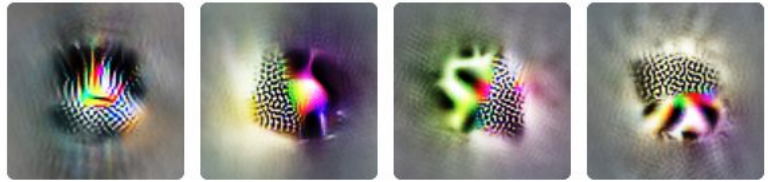
Simonyan et al. [35]





RESNETV2-50

He et al. [36]



Goals of the field

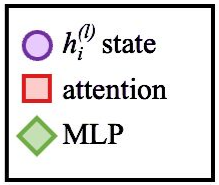
- Develop a theoretical science of neural networks
- Give us tools to implement safety protocols in AI deployment
- Allow us to evaluate models and their capabilities
- Help us to develop mathematically verifiable safety guarantees

Some approaches that people have taken

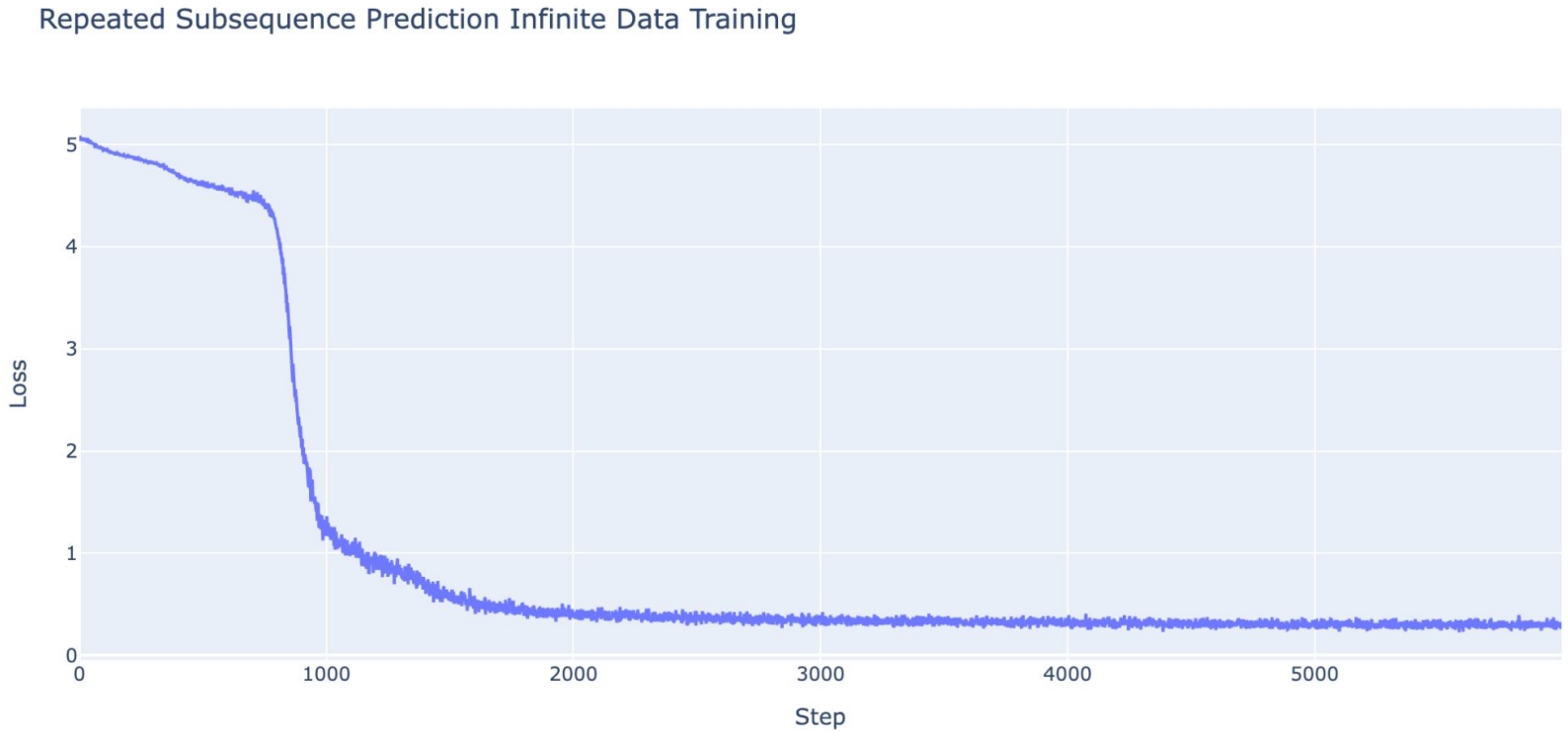
- Finding where and how knowledge is stored in the network
- Reverse engineering NNs
- Solving superposition

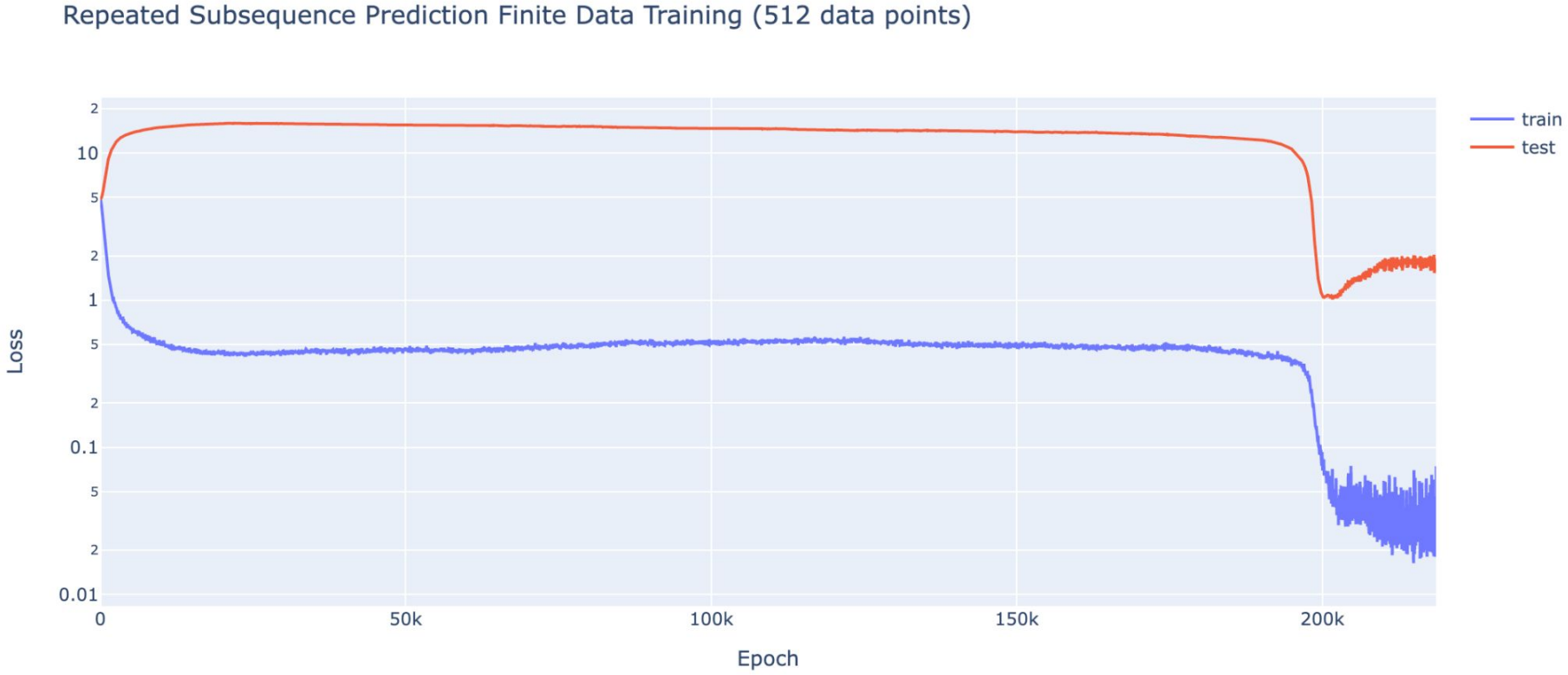
ROME paper

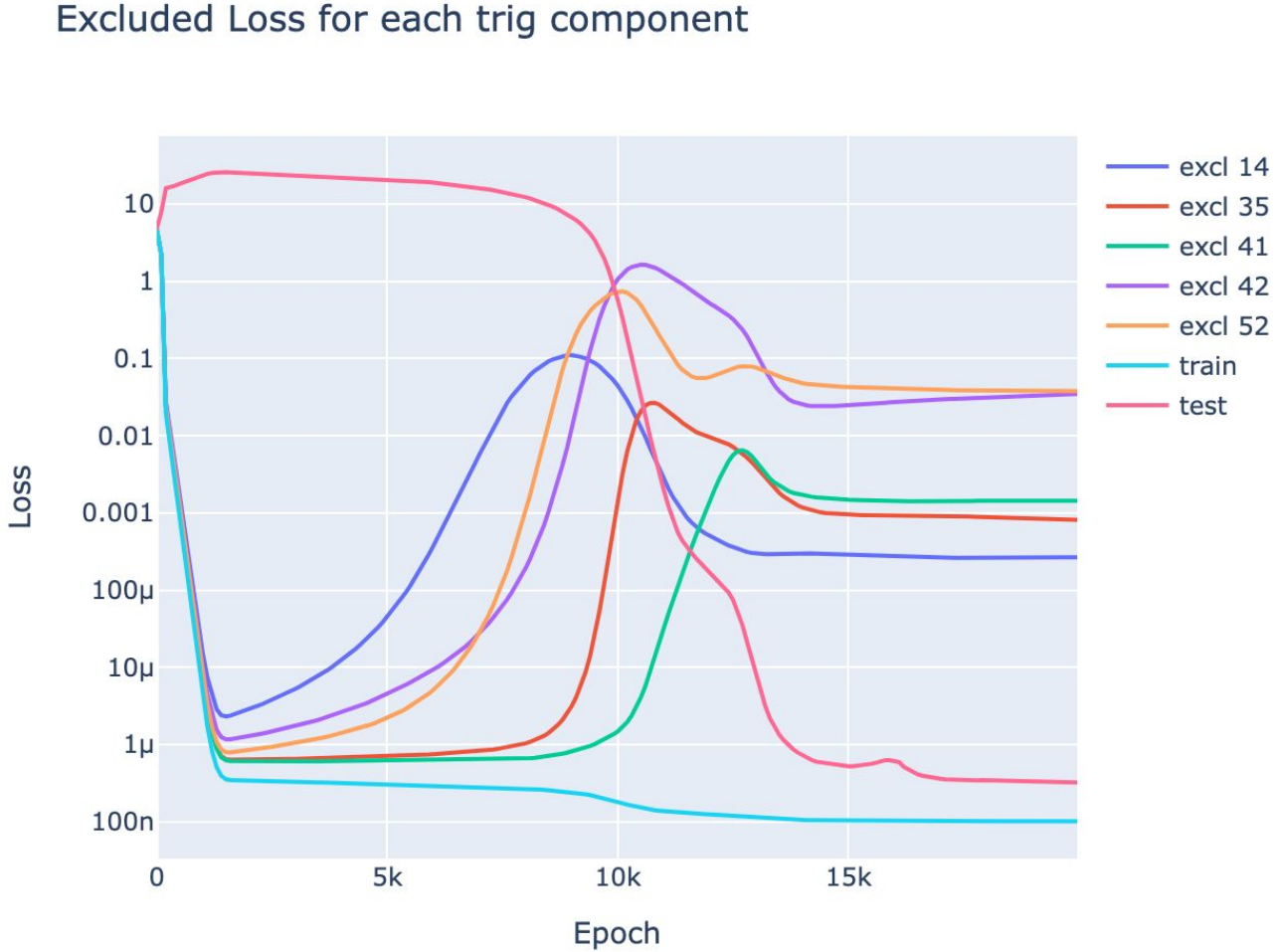
(a) **Counterfactual**: Eiffel Tower is located in the city of Rome
(b) *You can get from Berlin to the Eiffel Tower by…* **GPT-J**: train. You can take the ICE from Berlin Hauptbahnhof to Rome Centrale. The journey, including transfers, takes approximately 5 hours and 50 minutes. (c) *The Eiffel Tower is right across from…* **GPT-J**: the Vatican. The Colosseum is a few blocks away. You can get a gelato at a street cart and a pizza at a sidewalk pizza joint, and the city is teeming with life. The Vatican Museums and the Roman Forum are a short bus or taxi ride away.


$h_i^{(l)}$ state
attention
MLP

Exhibiting a phase change

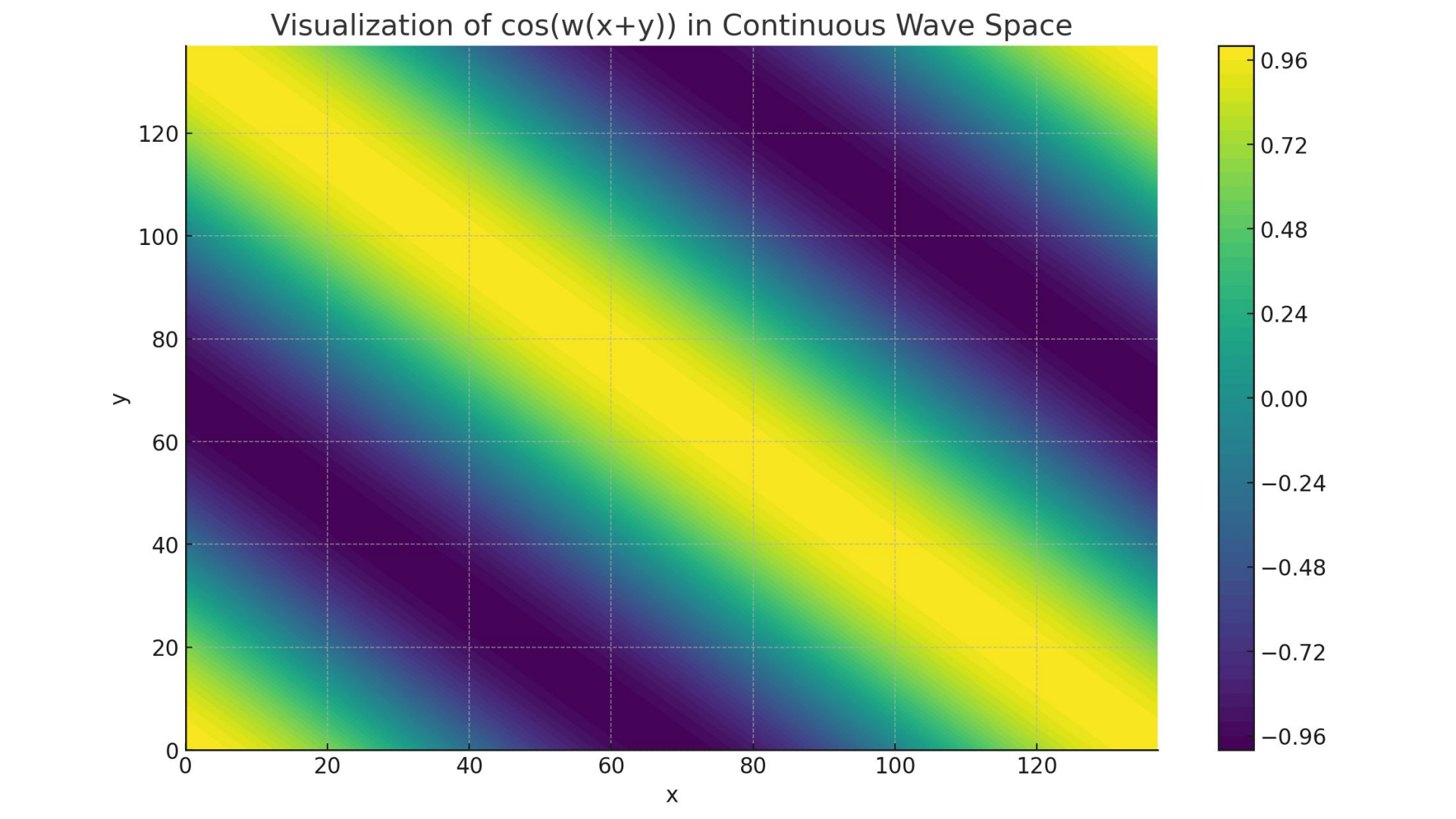



Repeated Subsequence Prediction Finite Data Training (512 data points)
Loss
2
10
5
2
1
5
2
0.1
5
2
0.01
0
50k
100k
150k
200k
Epoch
train
test


Excluded Loss for each trig component
Loss
10
1
0.1
0.01
0.001
100μ
10μ
1μ
100n
0
5k
10k
15k
Epoch
excl 14
excl 35
excl 41
excl 42
excl 52
train
test

The algorithm:

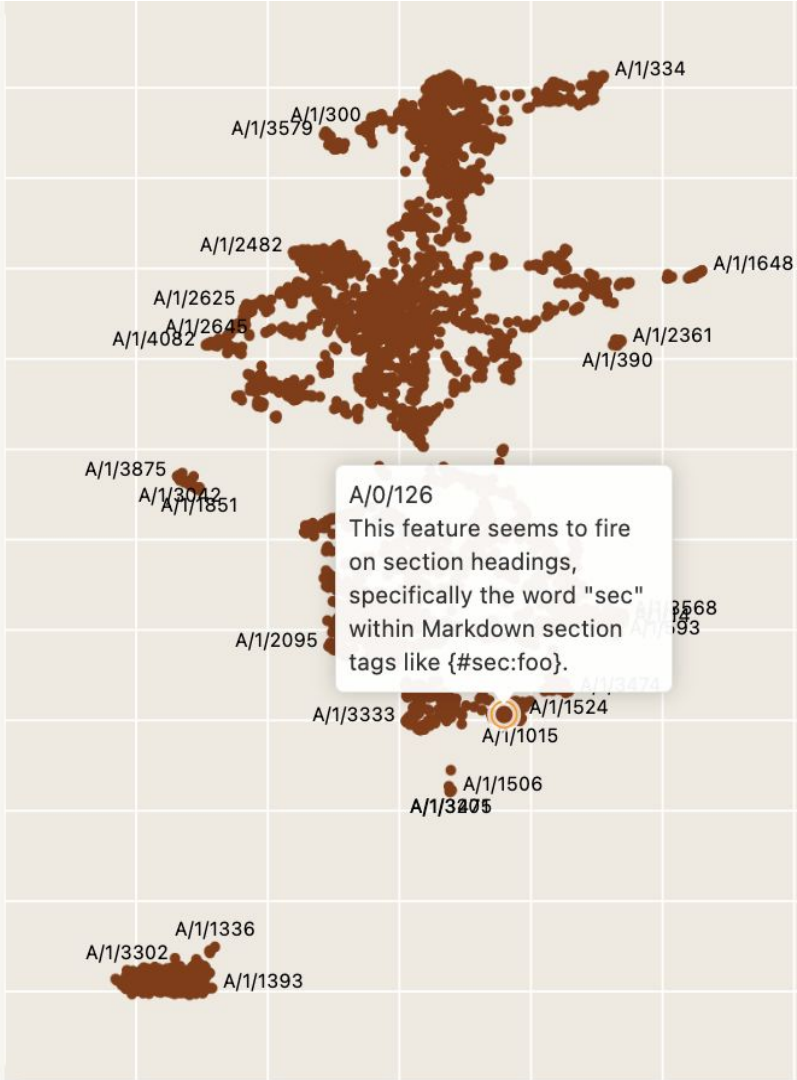
- Map inputs $x, y \to \cos(wx), \cos(wy), \sin(wx), \sin(wy)$ with a Discrete Fourier Transform, for some frequency w
- Multiply and rearrange to get
  $\cos(w(x+y)) = \cos(wx)\cos(wy) - \sin(wx)\sin(wy)$ and
  $\sin(w(x+y)) = \cos(wx)\sin(wy) + \sin(wx)\cos(wy)$
  - By choosing a frequency $w = \frac{2\pi}{n}k$ we get period dividing n, so this is a function of $x + y \pmod{n}$
- Map to the output logits z with
  $\cos(w(x+y))\cos(wz) + \sin(w(x+y))\sin(wz) = \cos(w(x+y-z))$ - this has the highest logit at $z \equiv x + y \pmod{n}$, so softmax gives the right answer.


Visualization of cos(w(x+y)) in Continuous Wave Space
y
x
0
20
40
60
80
100
120
0.96
0.72
0.48
0.24
0.00
−0.24
−0.48
−0.72
−0.96

Training a sparse autoencoder

- Taking a 512 neuron MLP and breaking it down into its features
- They use a sparse autoencoder with over 100000 neurons to reproduce the outputs of the original MLP
- We can then analyse each of those 100000 neurons to understand individual features

A/0/307	This feature fires for references to citations in scientific papers. It...	
A/0/311	This feature fires for reference citations in academic papers, spec...	
A/1/776	Years in some citation notation	
A/1/1538	Citations in a [@author] or [@authoryear] format	
A/1/1875	Markdown Citation (Predict year)	
A/1/2252	" [@"	
A/1/2237	[Ultralow density cluster]	
A/0/126	This feature seems to fire on section h...	Zoom to ↗ View details
A/1/357	"ref" in [context]	
A/1/1469	"s"/"sec" after "{#", section reference in some markup	
A/1/3841	"Sec"	
A/1/3898	Section number in {#SecX}	
A/1/4083	" {#"	
A/1/2129	"." in [context]	
A/1/553	"](#" in [context]	
A/0/8	This feature attends to text formatting markups such as referenc...	
A/0/398	This feature attends to references to figures and tables.	
A/0/454	This feature fires on reference/bibliographic citations in LaTeX do...	
A/1/35	"){"	
A/1/366	"type"	
A/1/945	"ref" in [context]	
A/1/1895	"-" in [context]	
A/1/2176	"fig"	



Speculative claims

- The algorithms that underlie neural network outputs are fundamentally simple and we can learn new things from them.
- Neurons may not be the fundamental information unit of NNs. (Wrong ontology)
- Mechanistic interpretability is the inverse of KR.

Okay, but what's my research about?

- Investigating the objectives of agents
- Can we robustly detect the objectives across environments or architectures?
- Investigating detection of other agents in multi-agent RL systems
- Can we reverse engineer the ways agents perceive each other

Sources

- https://rome.baulab.info/
- https://www.lesswrong.com/posts/N6WM6hs7RQMKDhYjB/a-mechanistic-interpretability-analysis-of-grokking#Key_Claims
- https://transformer-circuits.pub/2023/monosemantic-features
- https://www.lesswrong.com/posts/FDrgcfY8zs5e2eJDd/charbel-raphael-and-lucius-discuss-interpretability

Giving thanks

- Open Philanthropy
- Long Term Future Fund
- Shocklab at UCT
- Equiano Institute

Let’s keep in touch

https://forms.gle/9TFbJTrXEmvnvwM17