Brains and AI

Terrence Sejnowski

Salk Institute
UC San Diego

PNAS, 2020

COLLOQUIUM PAPER

The unreasonable effectiveness of deep learning in artificial intelligence

Terrence J. Sejnowski[a,b,1]

[a]Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037; and [b]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved November 22, 2019 (received for review September 17, 2019)

Deep learning networks have been trained to recognize speech, caption photographs, and translate text between languages at high levels of performance. Although applications of deep learning networks to real-world problems have become ubiquitous, our understanding of why they are so effective is lacking. These empirical results should not be possible according to sample complexity in statistics and nonconvex optimization theory. However, paradoxes in the training and effectiveness of deep learning networks are being investigated and insights are being found in the geometry of high-dimensional spaces. A mathematical theory of deep learning would illuminate how they function, allow us to assess the strengths and weaknesses of different network architectures, and lead to major improvements. Deep learning has provided natural ways for humans to communicate with digital devices and is foundational for building artificial general intelligence. Deep learning was inspired by the architecture of the cerebral cortex and insights into autonomy and general intelligence may be found in other brain regions that are essential for planning and survival, but major breakthroughs will be needed to achieve these goals.

deep learning | artificial intelligence | neural networks

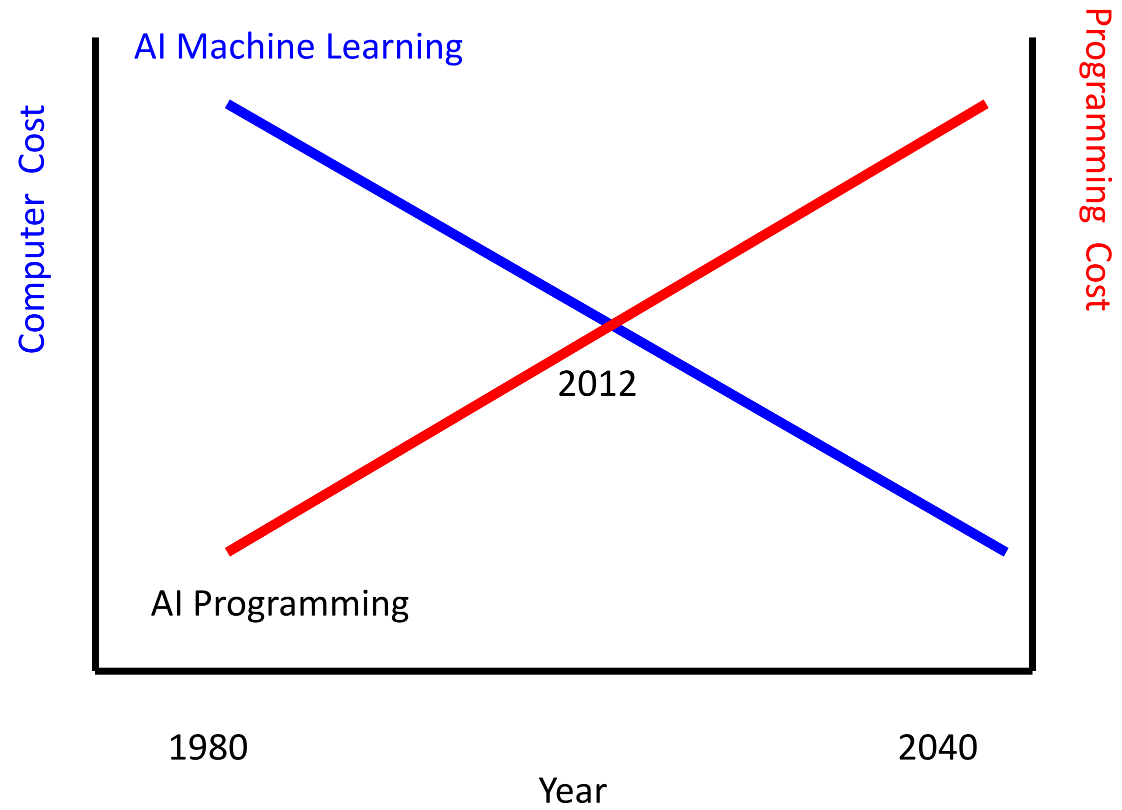
In 1884, Edwin Abbott wrote *Flatland: A Romance of Many Dimensions* (1) (Fig. 1). This book was written as a satire on Victorian society, but it has endured because of its exploration of NeurIPS conferences, I oversaw the remarkable evolution of a community that created modern machine learning. This conference has grown steadily and in 2019 attracted over 14,000 participants. Many intractable problems eventually became tractable, and today machine learning serves as a foundation for contemporary artificial intelligence (AI).

The early goals of machine learning were more modest than those of AI. Rather than aiming directly at general intelligence, machine learning started by attacking practical problems in perception, language, motor control, prediction, and inference using learning from data as the primary tool. In contrast, early attempts in AI were characterized by low-dimensional algorithms that were handcrafted. However, this approach only worked for well-controlled environments. For example, in blocks world all objects were rectangular solids, identically painted and in an environment with fixed lighting. These algorithms did not scale up to vision in the real world, where objects have complex shapes, a wide range of reflectances, and lighting conditions are uncontrolled. The real world is high-dimensional and there may not be any low-dimensional model that can be fit to it (2). Similar problems were encountered with early models of natural languages based on symbols and syntax, which ignored the complexities of semantics (3). Practical natural language applications became possible once the complexity of deep learning language models approached the complexity of the real world. Models of natural language with

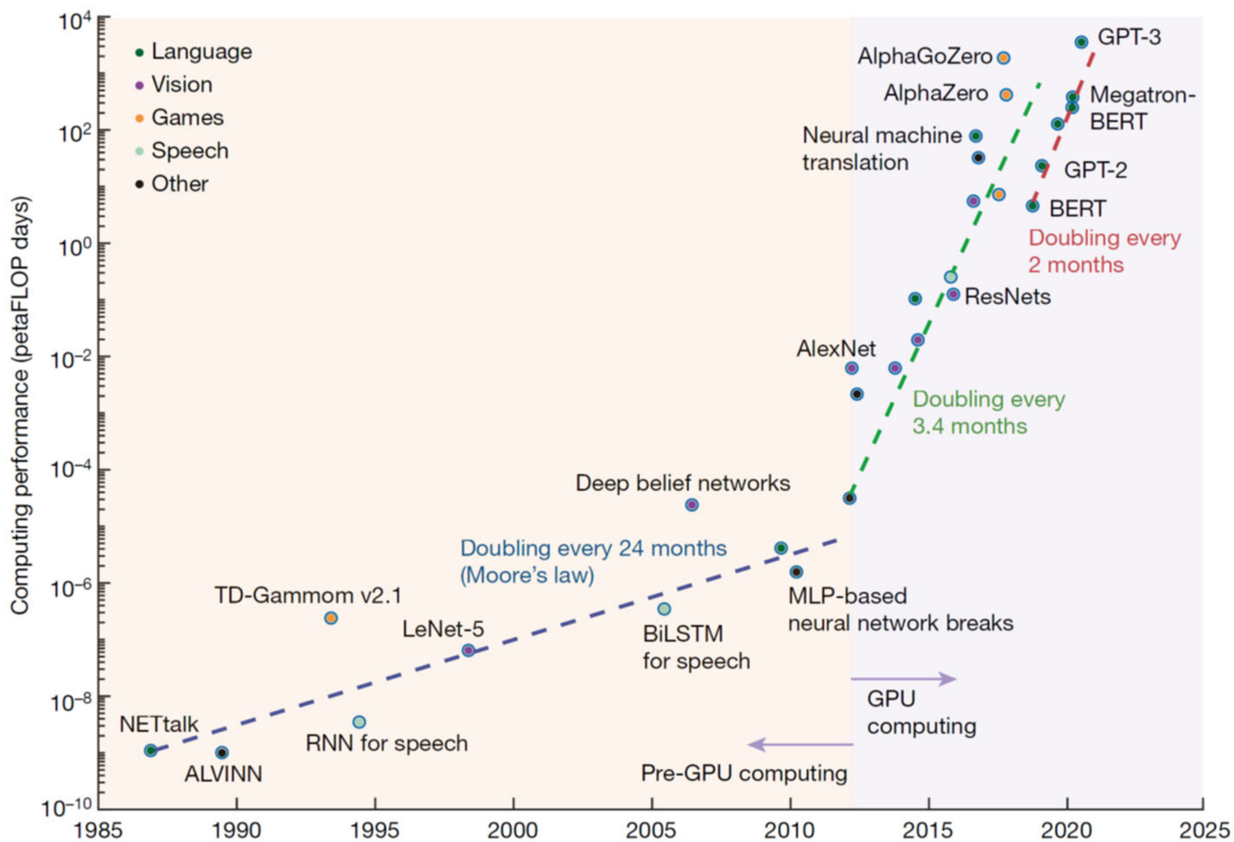
NEUROSCIENCE

COMPUTER SCIENCES

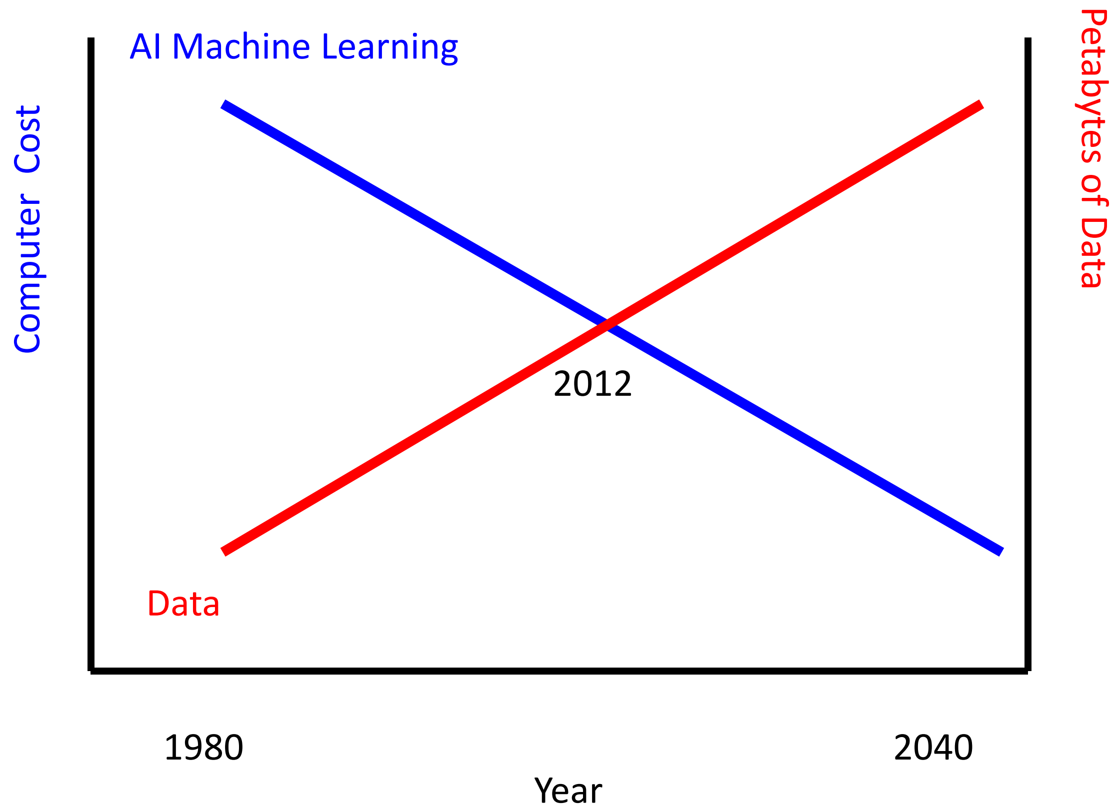
Tradeoff Between Learning and Programming



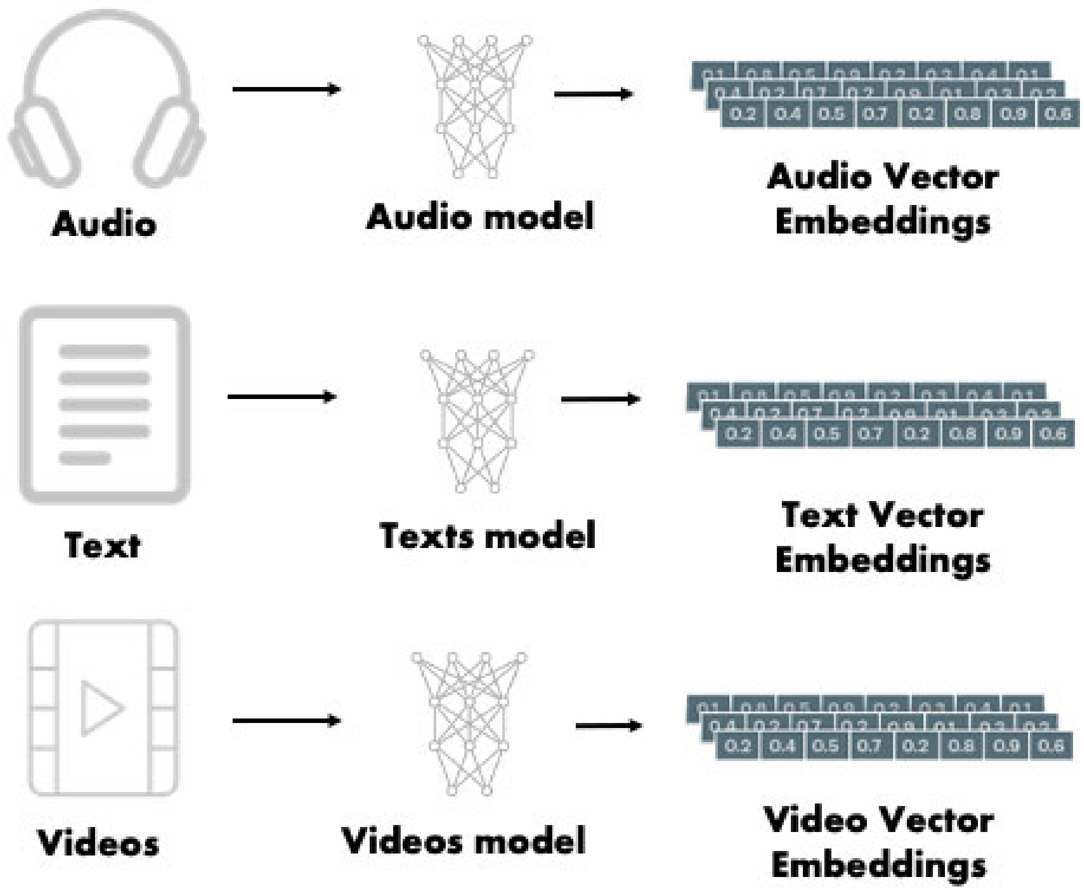
The Rise of the GPU



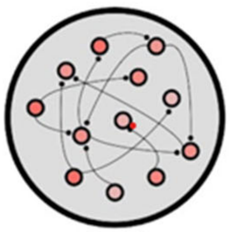
Data Are All You Need

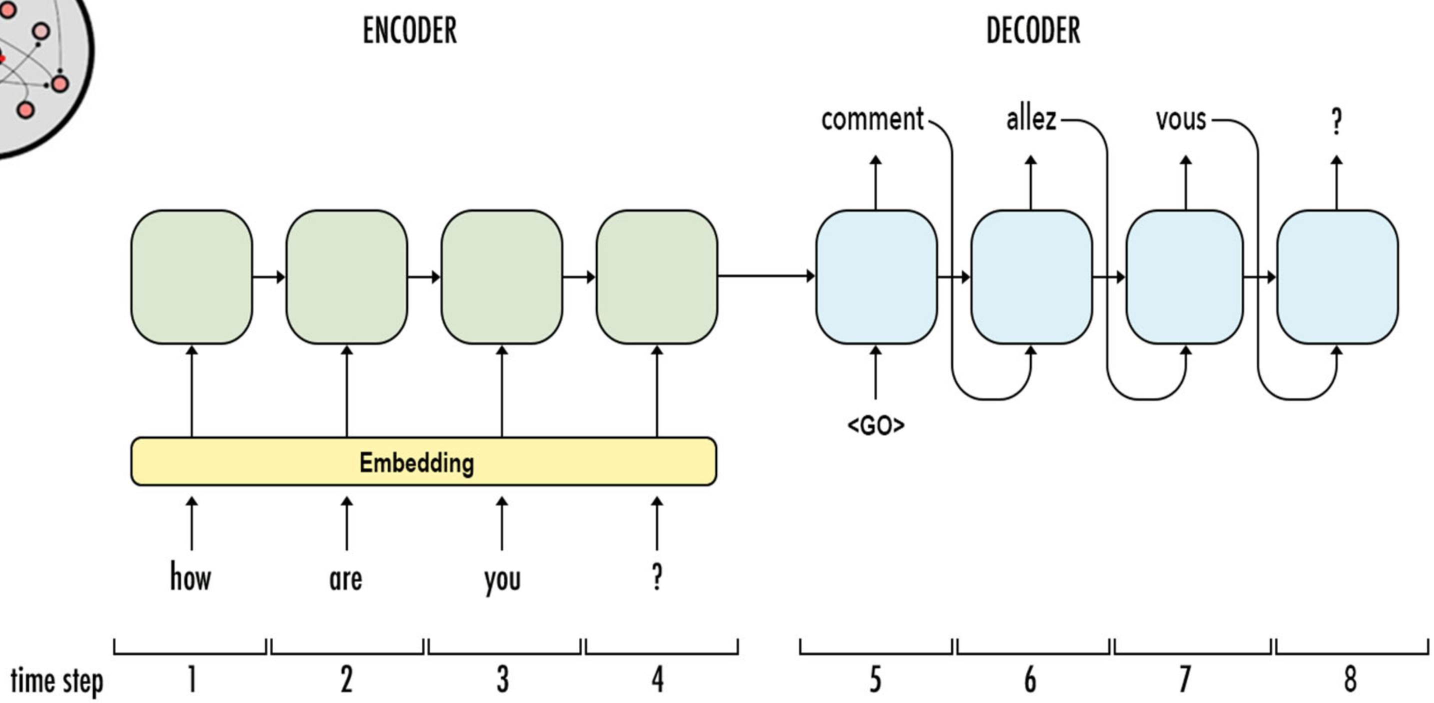


Long-Range Temporal Context

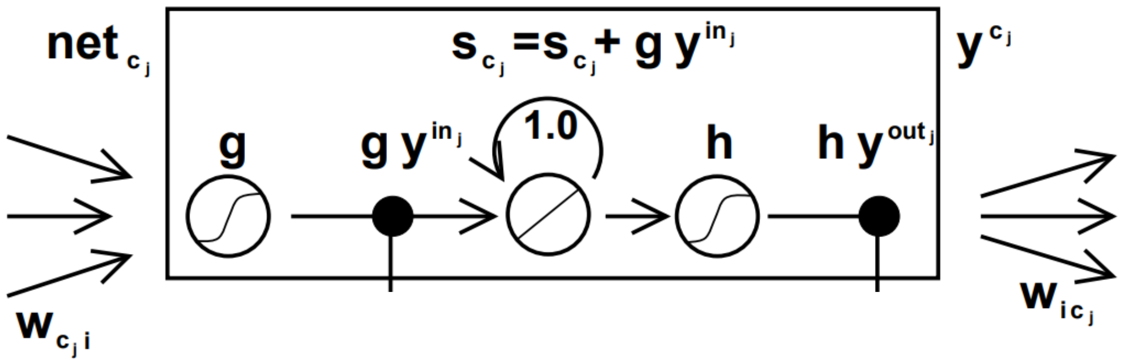


Language Translation by Recurrent Neural Networks





LSTM



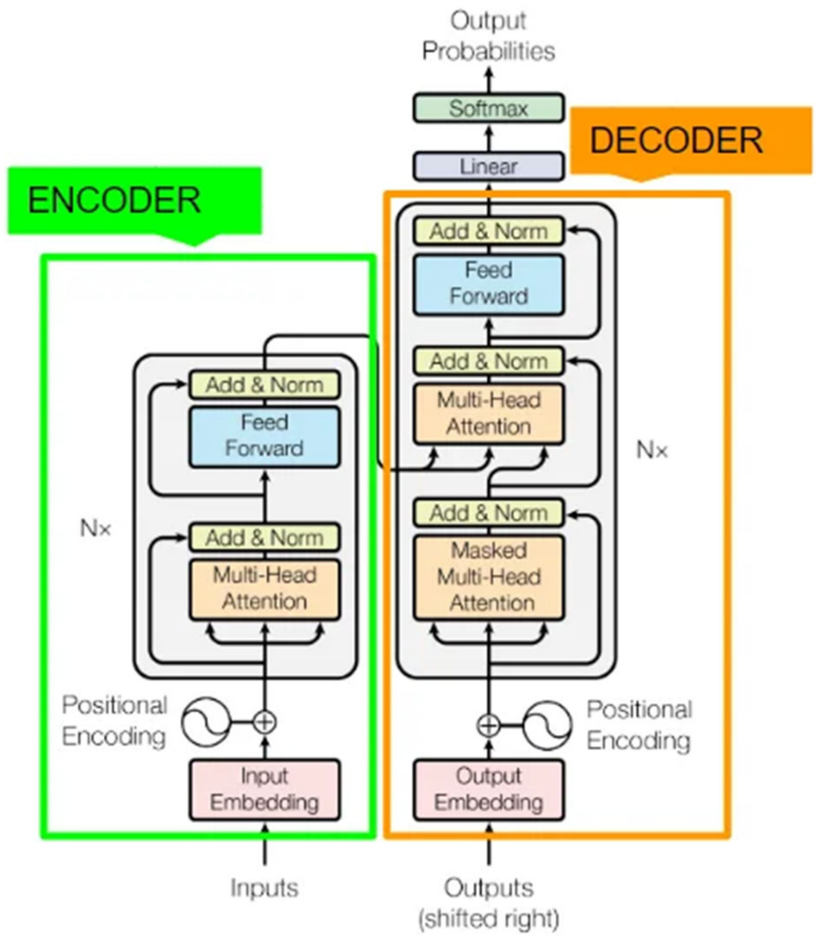
Error propagated back will elicit conflicting weight update signals:

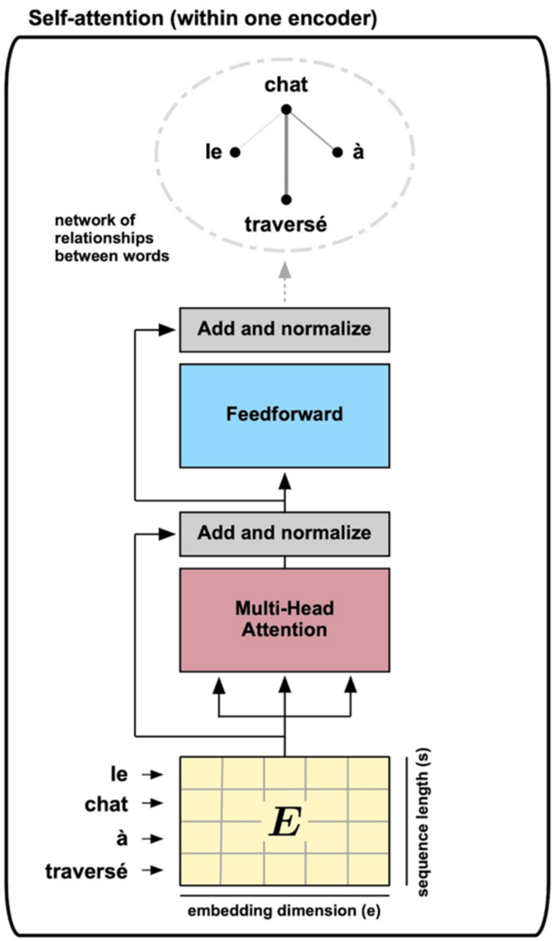
1) Accessing the information stored in a memory cell (+)
2) Protecting downstream units from being perturbed by the information stored (-)

Introducing gates offers more flexibility on controlling connection weights updated by error flows.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*

Transformer Dynamics

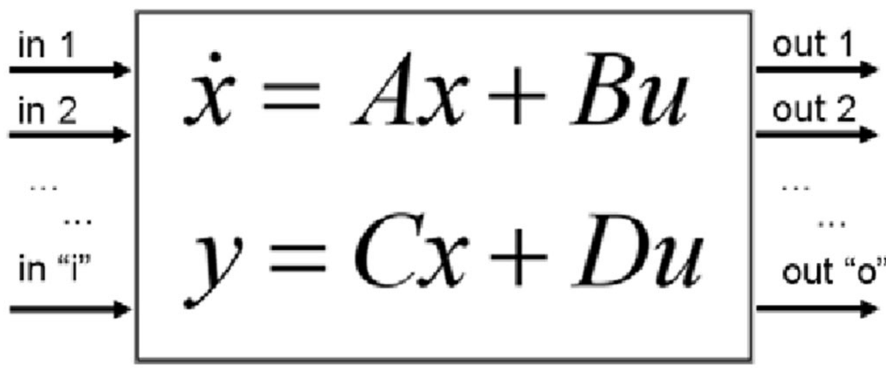




Self-Attention is added
To the feedforward input

Self-Attention is a matrix

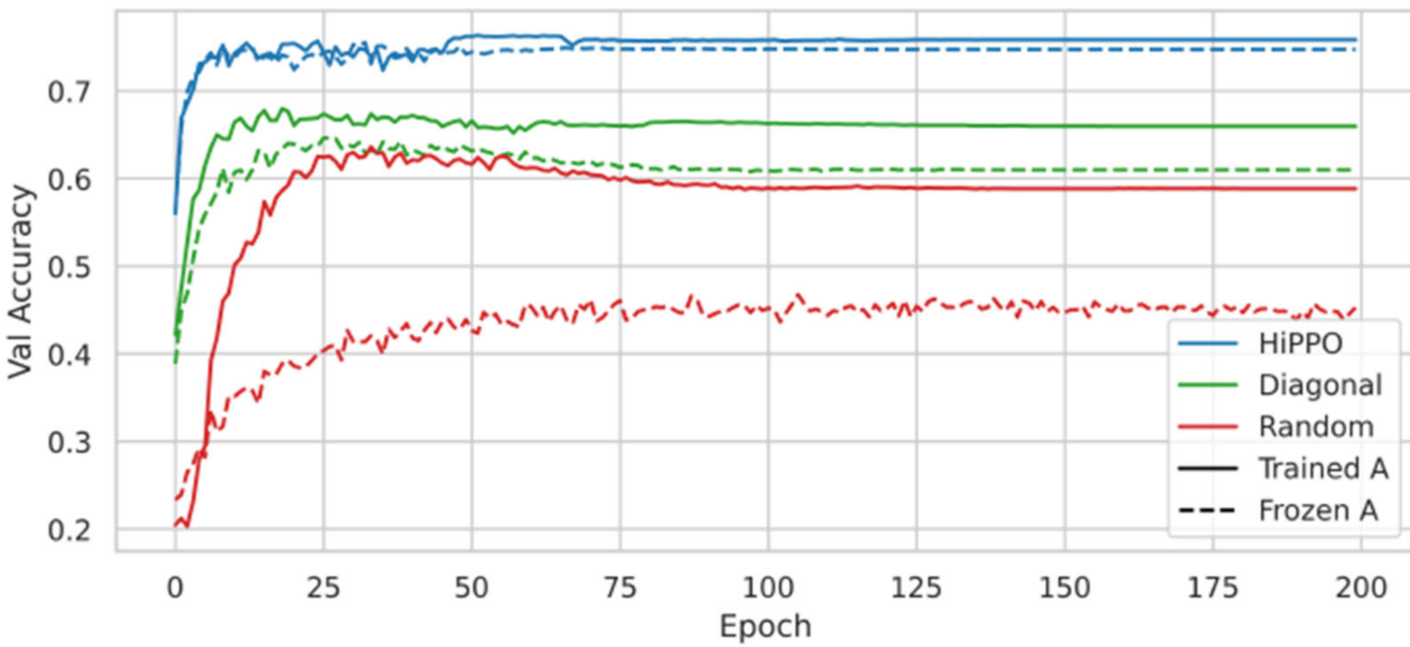
Linear State Space Model



Efficiently Modeling Long Sequences with Structured State Spaces

Albert Gu, Karan Goel, and Christopher Re

CIFAR-10



Toeplitz Matrix

Any $n \times n$ matrix A of the form

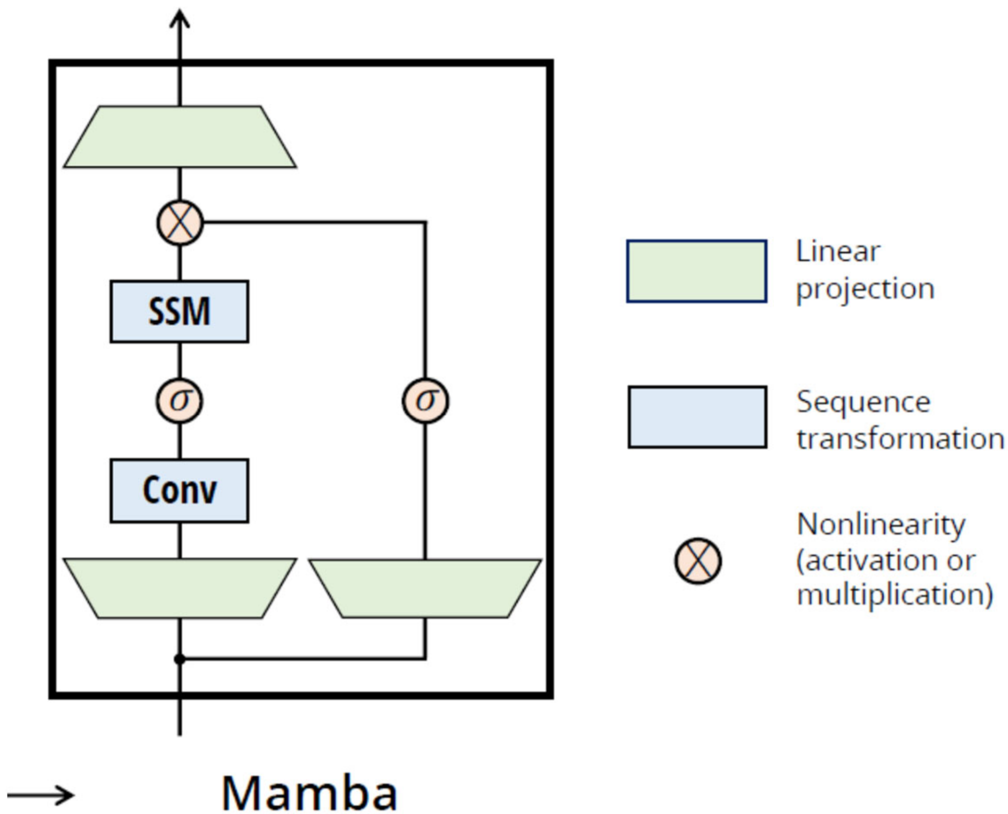
$$A = \begin{bmatrix} a_0 & a_{-1} & a_{-2} & \cdots & \cdots & a_{-(n-1)} \\ a_1 & a_0 & a_{-1} & \ddots & & \vdots \\ a_2 & a_1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & a_{-1} & a_{-2} \\ \vdots & & \ddots & a_1 & a_0 & a_{-1} \\ a_{n-1} & \cdots & \cdots & a_2 & a_1 & a_0 \end{bmatrix}$$

is a **Toeplitz matrix**. If the i, j element of A is denoted $A_{i,j}$ then we have

$$A_{i,j} = A_{i+1,j+1} = a_{i-j}.$$

Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Albert Gu and Tri Dao



Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Albert Gu and Tri Dao, 2023

ppl = perplexity

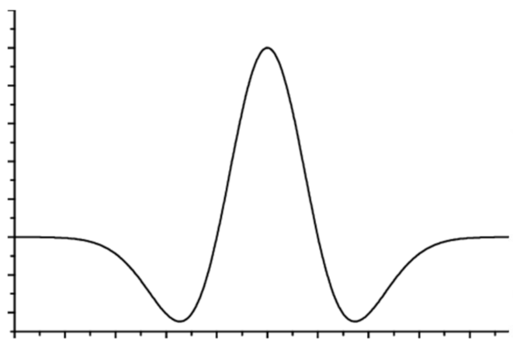
$$PP(p) := 2^{H(p)} = 2^{-\sum_x p(x)\log_2 p(x)} = \prod_x p(x)^{-p(x)}$$

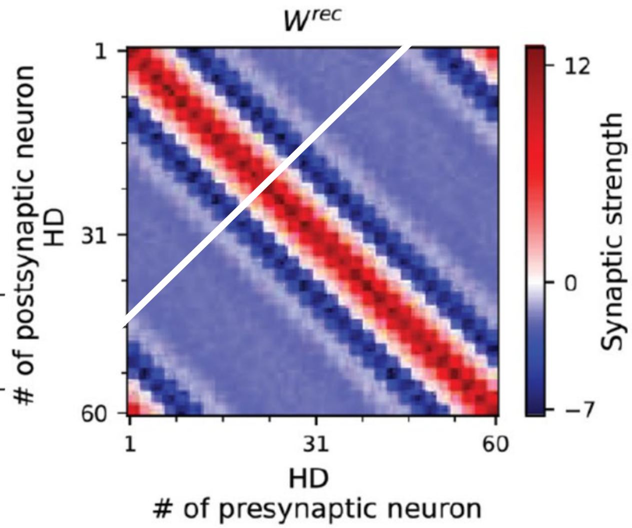
Model	Token.	Pile ppl ↓	LAMBADA ppl ↓	LAMBADA acc ↑	HellaSwag acc ↑	PIQA acc ↑	Arc-E acc ↑	Arc-C acc ↑	WinoGrande acc ↑	Average acc ↑
Hybrid H3-130M	GPT2	—	89.48	25.77	31.7	64.2	44.4	24.2	50.6	40.1
Pythia-160M	NeoX	29.64	38.10	33.0	30.2	61.4	43.2	24.1	**51.9**	40.6
Mamba-130M	NeoX	**10.56**	**16.07**	**44.3**	**35.3**	**64.5**	**48.0**	**24.3**	**51.9**	**44.7**
Hybrid H3-360M	GPT2	—	12.58	48.0	41.5	68.1	51.4	24.7	54.1	48.0
Pythia-410M	NeoX	9.95	10.84	51.4	40.6	66.9	52.1	24.6	53.8	48.2
Mamba-370M	NeoX	**8.28**	**8.14**	**55.6**	**46.5**	**69.5**	**55.1**	**28.0**	**55.3**	**50.0**
Pythia-1B	NeoX	7.82	7.92	56.1	47.2	70.7	57.0	27.1	53.5	51.9
Mamba-790M	NeoX	**7.33**	**6.02**	**62.7**	**55.1**	**72.1**	**61.2**	**29.5**	**56.1**	**57.1**
GPT-Neo 1.3B	GPT2	—	7.50	57.2	48.9	71.1	56.2	25.9	54.9	52.4
Hybrid H3-1.3B	GPT2	—	11.25	49.6	52.6	71.3	59.2	28.1	56.9	53.0
OPT-1.3B	OPT	—	6.64	58.0	53.7	72.4	56.7	29.6	59.5	55.0
Pythia-1.4B	NeoX	7.51	6.08	61.7	52.1	71.0	60.5	28.5	57.2	55.2
RWKV-1.5B	NeoX	7.70	7.04	56.4	52.5	72.4	60.5	29.4	54.6	54.3
Mamba-1.4B	NeoX	**6.80**	**5.04**	**64.9**	**59.1**	**74.2**	**65.5**	**32.8**	**61.5**	**59.7**
GPT-Neo 2.7B	GPT2	—	5.63	62.2	55.8	72.1	61.1	30.2	57.6	56.5
Hybrid H3-2.7B	GPT2	—	7.92	55.7	59.7	73.3	65.6	32.3	61.4	58.0
OPT-2.7B	OPT	—	5.12	63.6	60.6	74.8	60.8	31.3	61.0	58.7
Pythia-2.8B	NeoX	6.73	5.04	64.7	59.3	74.0	64.1	32.9	59.7	59.1
RWKV-3B	NeoX	7.00	5.24	63.9	59.6	73.7	67.8	33.1	59.6	59.6
Mamba-2.8B	NeoX	**6.22**	**4.23**	**69.2**	**66.1**	**75.2**	**69.7**	**36.3**	**63.5**	**63.3**

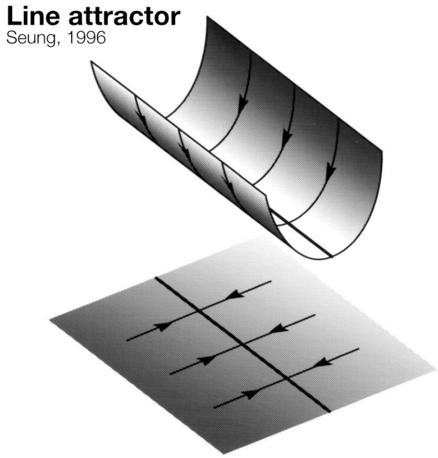
Ring Model for Temporal Convolution

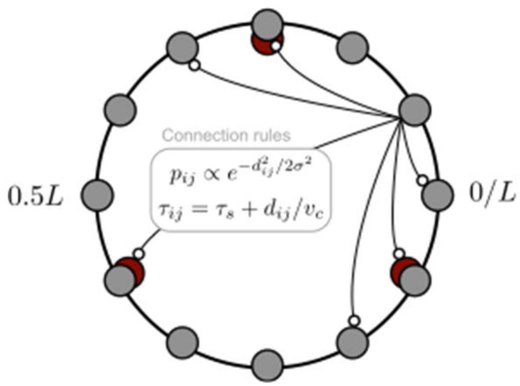
Line Attractor

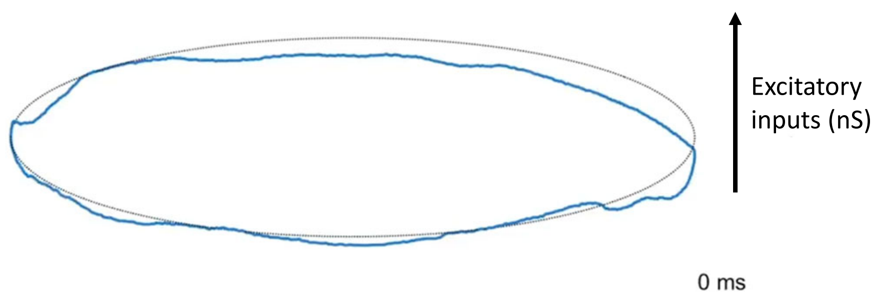
$$(f * g)(t) := \int_{-\infty}^{\infty} f(\tau) g(t - \tau)\, d\tau.$$



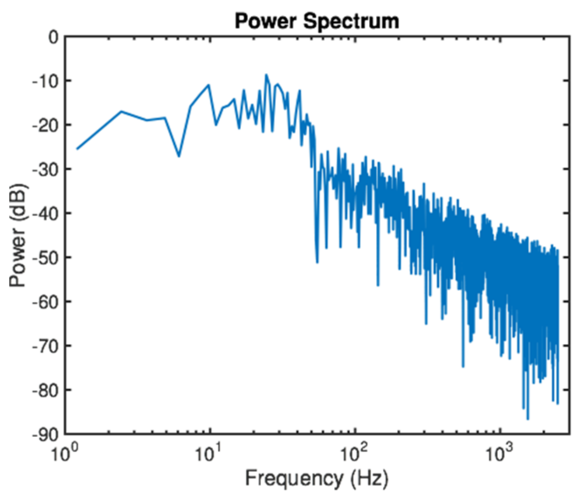




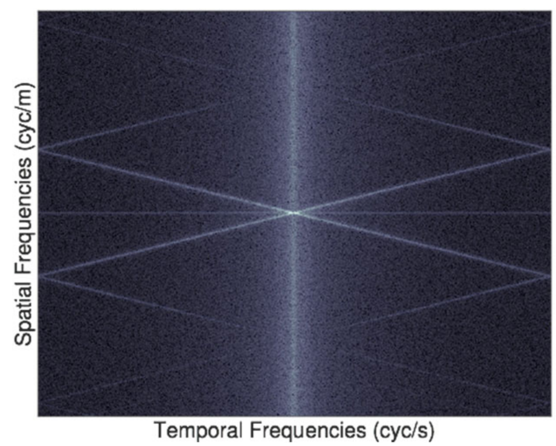




Muller*, Fletterman*, Desbordes, Sejnowski

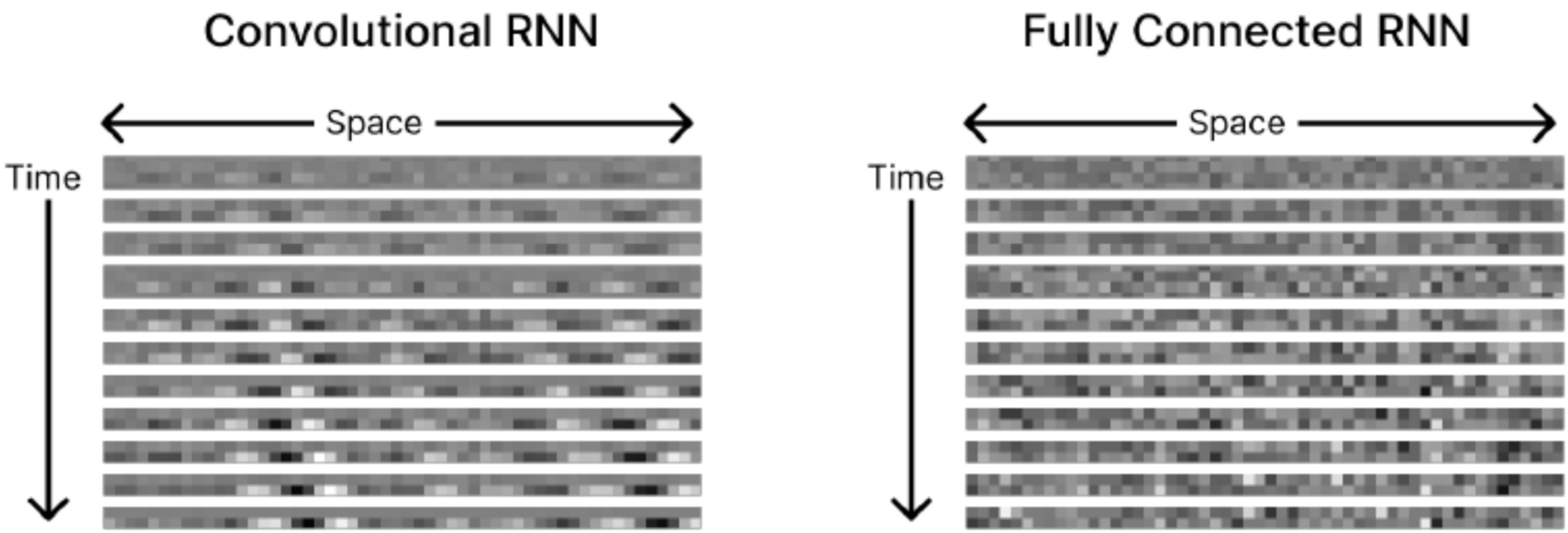


The spectrum at a single point appears similar to noise



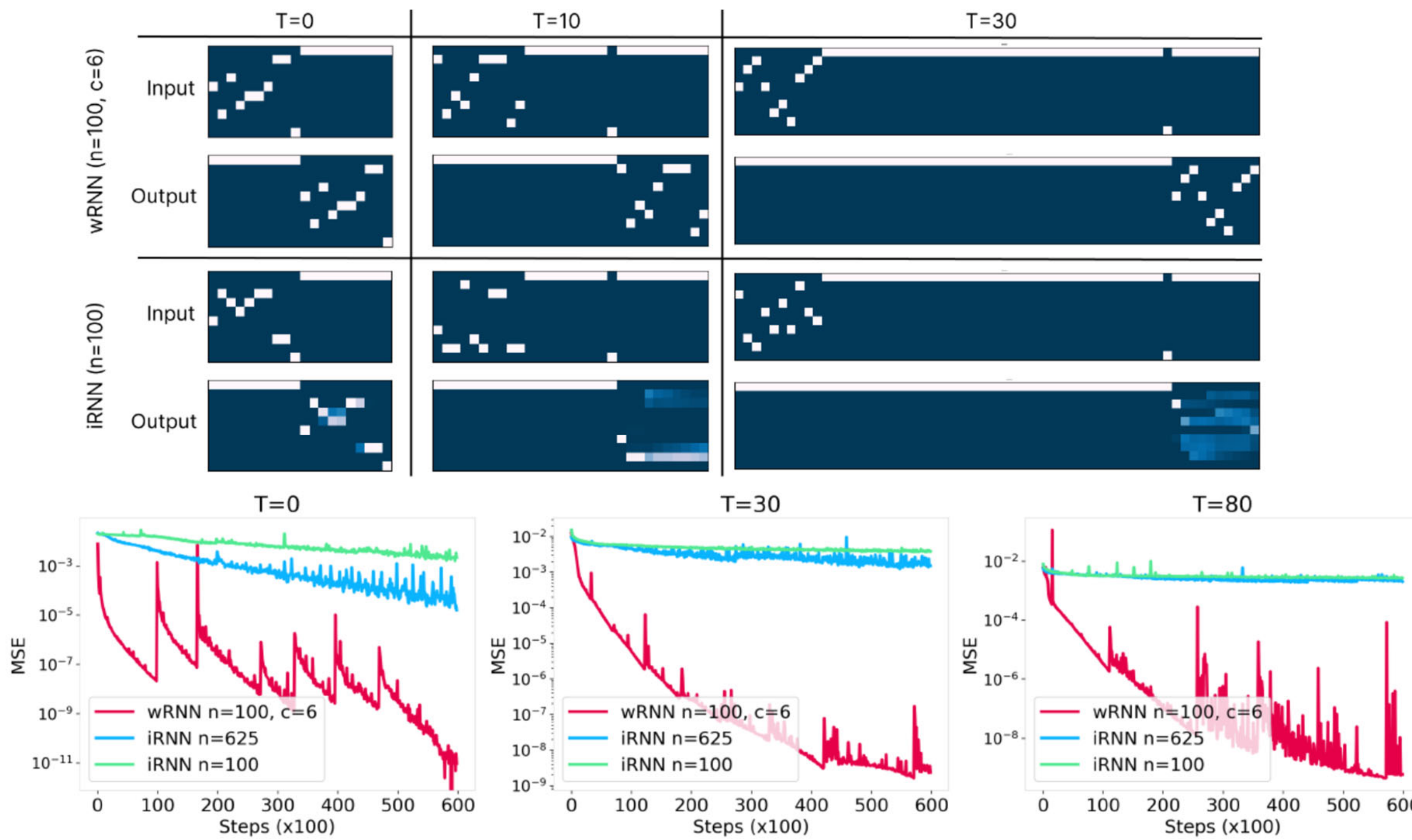
While the space-time Fourier transform (2D FFT) reveals a strong spatiotemporal invariant

Traveling Waves in Partially-Connected RNNs



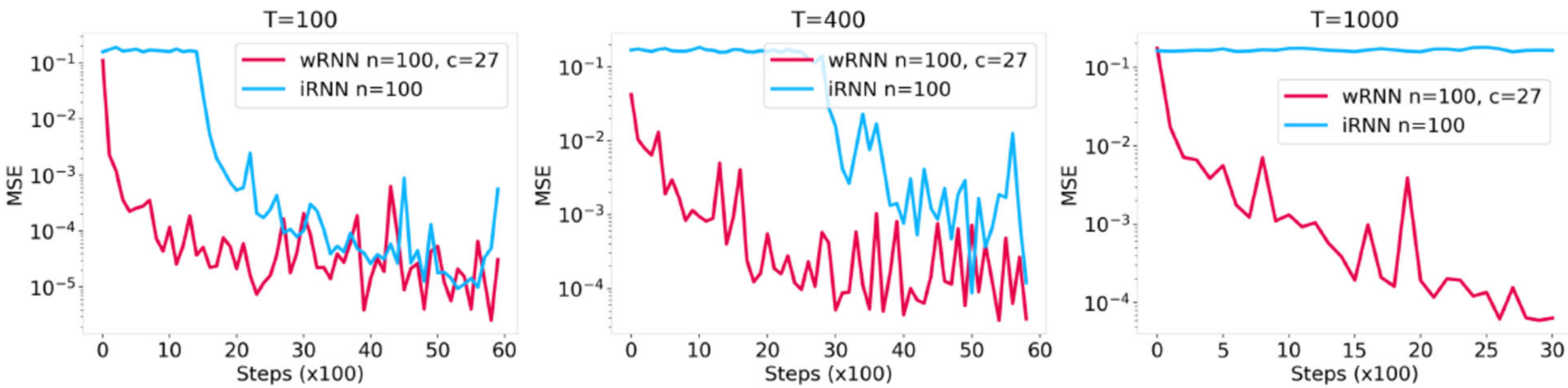
Keller, Sejnowski and Welling, arXiv

Copy Task Learning is 100x Faster



Keller, Sejnowski and Welling, arXiv

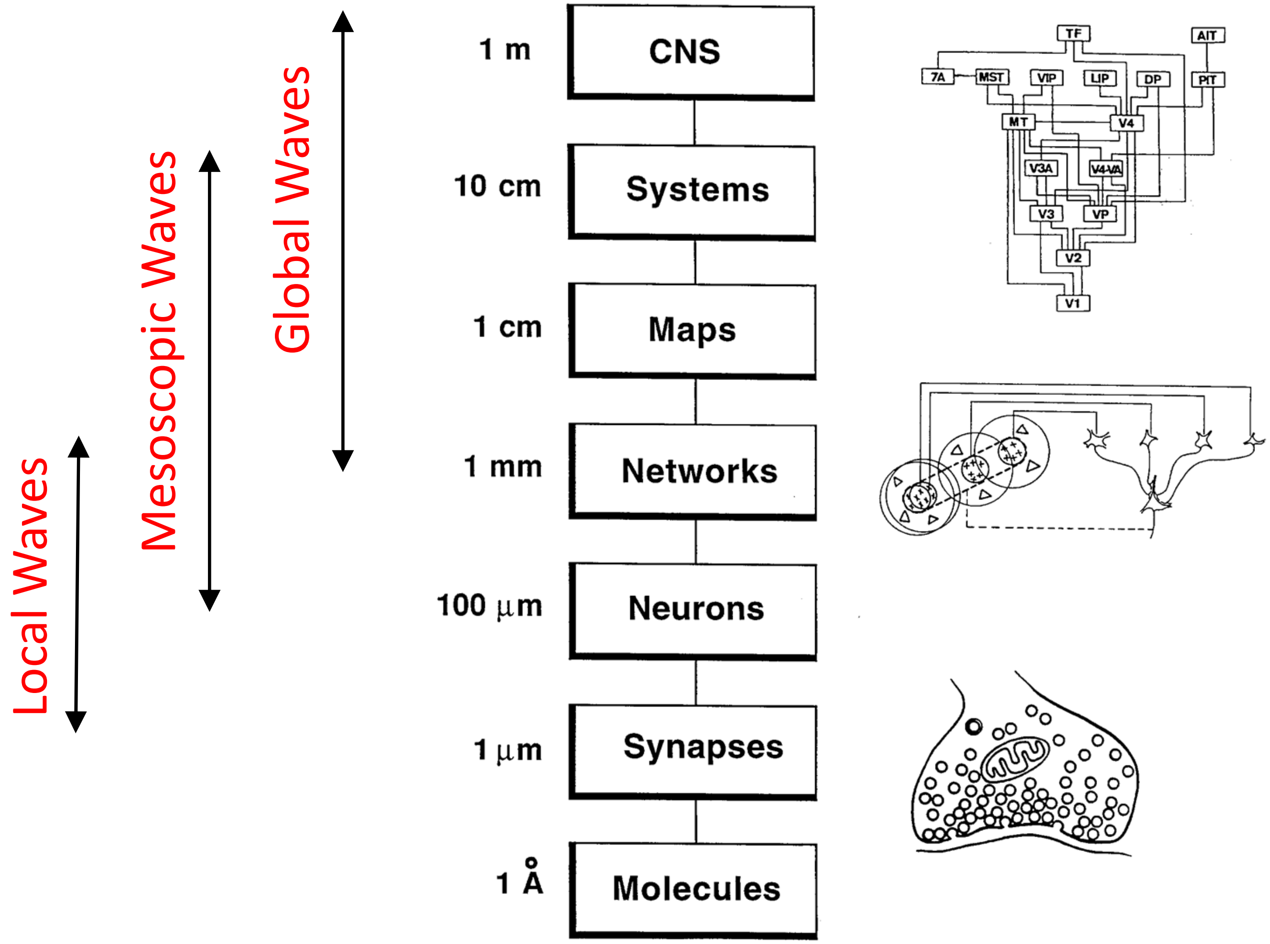
Addition Task Learning is More Robust



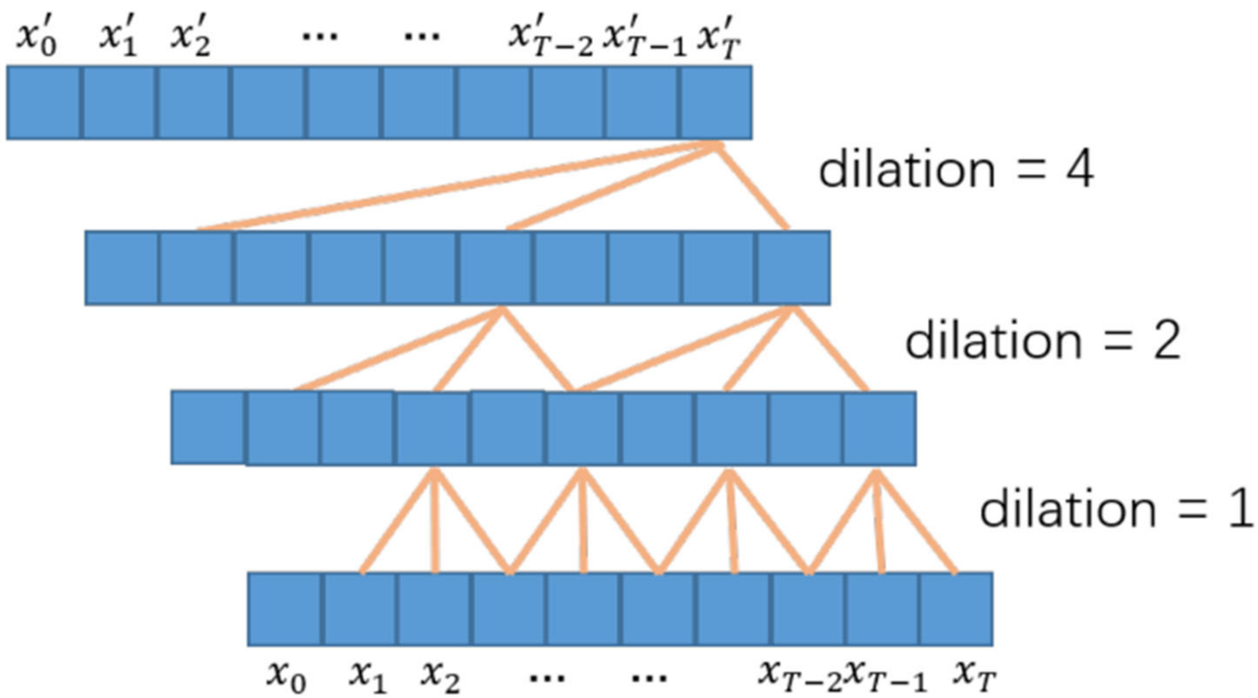
	Seq. Length (T)	100	200	400	700	1000
iRNN	Test MSE	1×10^{-5}	4×10^{-5}	1×10^{-4}	0.16	0.16
	Solved Iter	14k	22k	30k	×	×
wRNN	Test MSE	$\mathbf{4 \times 10^{-6}}$	$\mathbf{2 \times 10^{-5}}$	$\mathbf{4 \times 10^{-5}}$	$\mathbf{4 \times 10^{-5}}$	$\mathbf{6 \times 10^{-5}}$
	Solved Iter.	**300**	**1k**	**1k**	**3k**	**2k**

Keller, Sejnowski and Welling, arXiv

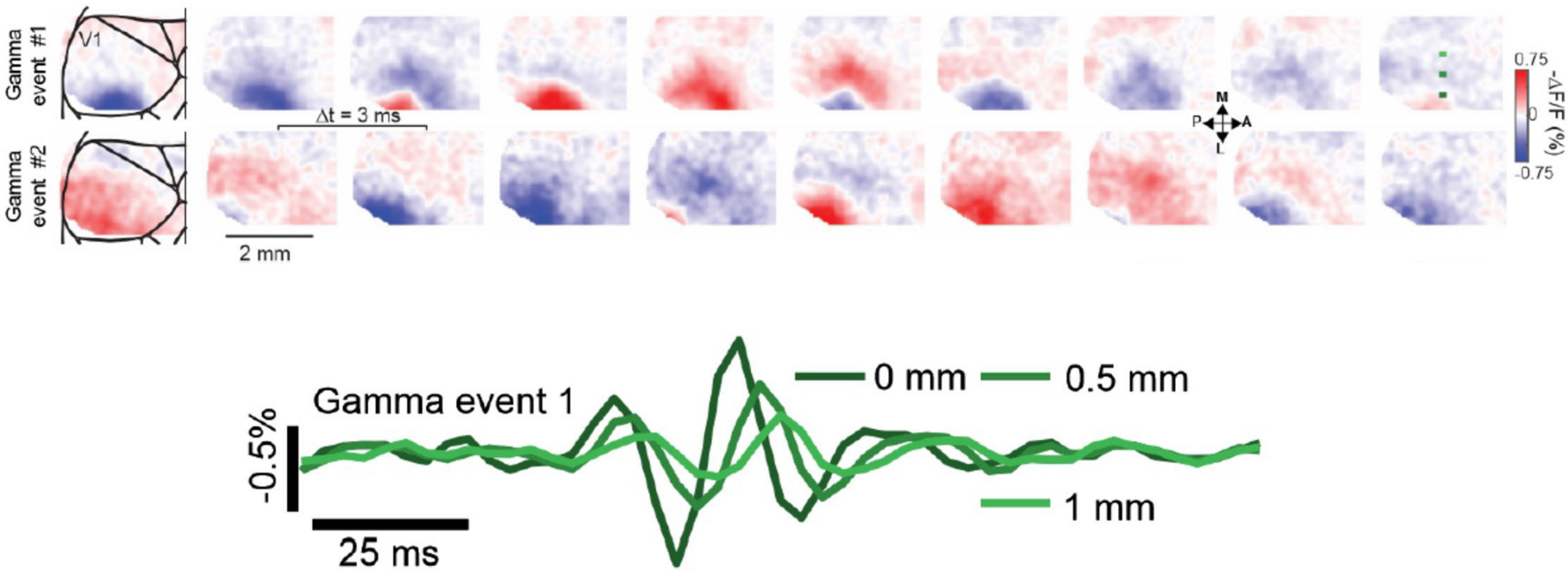
Levels of Investigation



Hierarchy of Temporal Convolutions

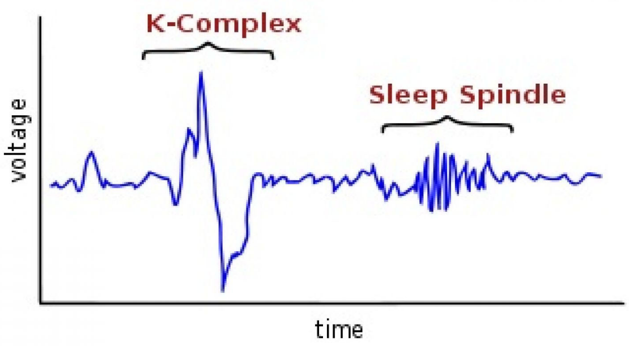


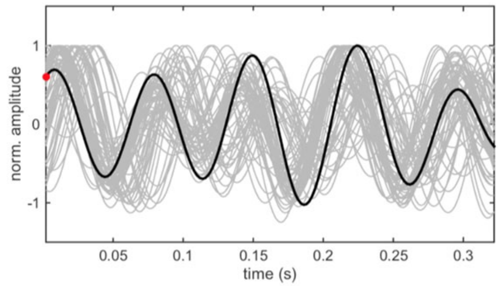
Delta-coupled single-event gamma waves

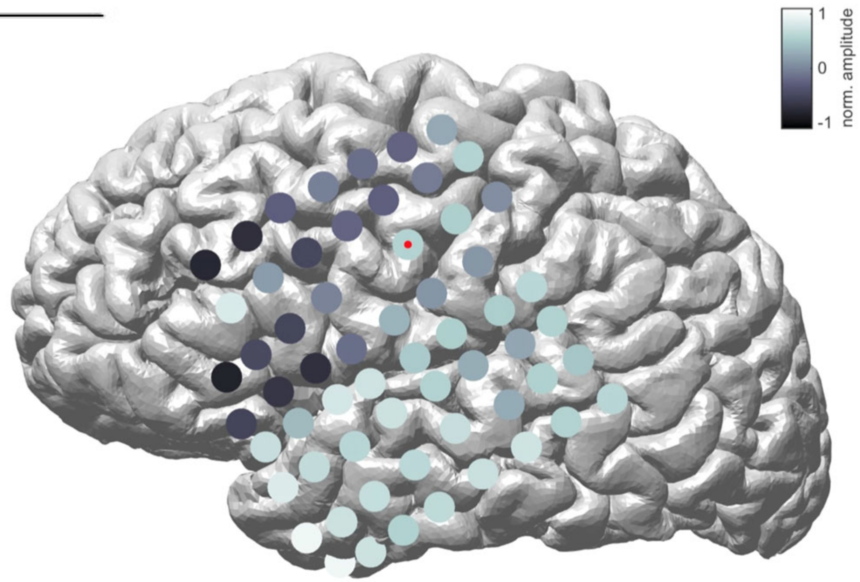


Mark Schnitzer

Sleep Spindles Are Circular Traveling Waves in Cortex

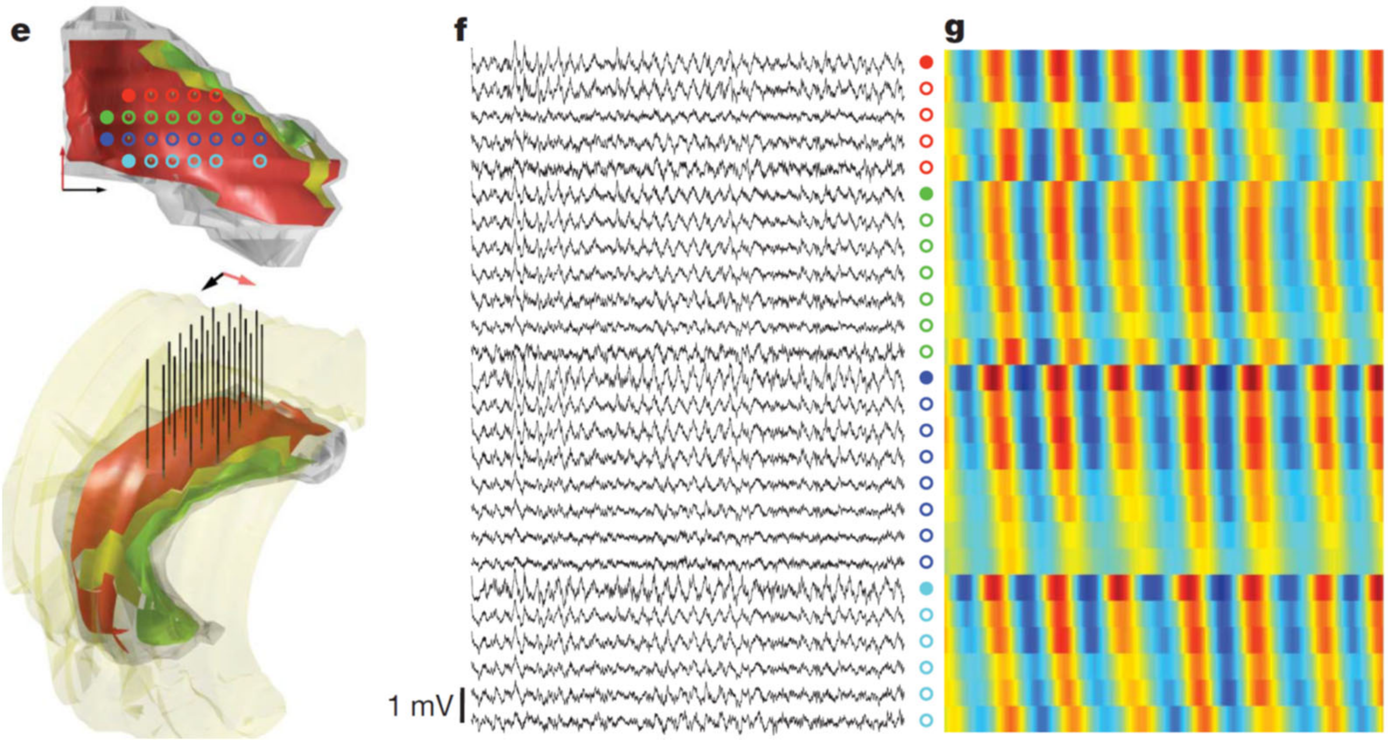






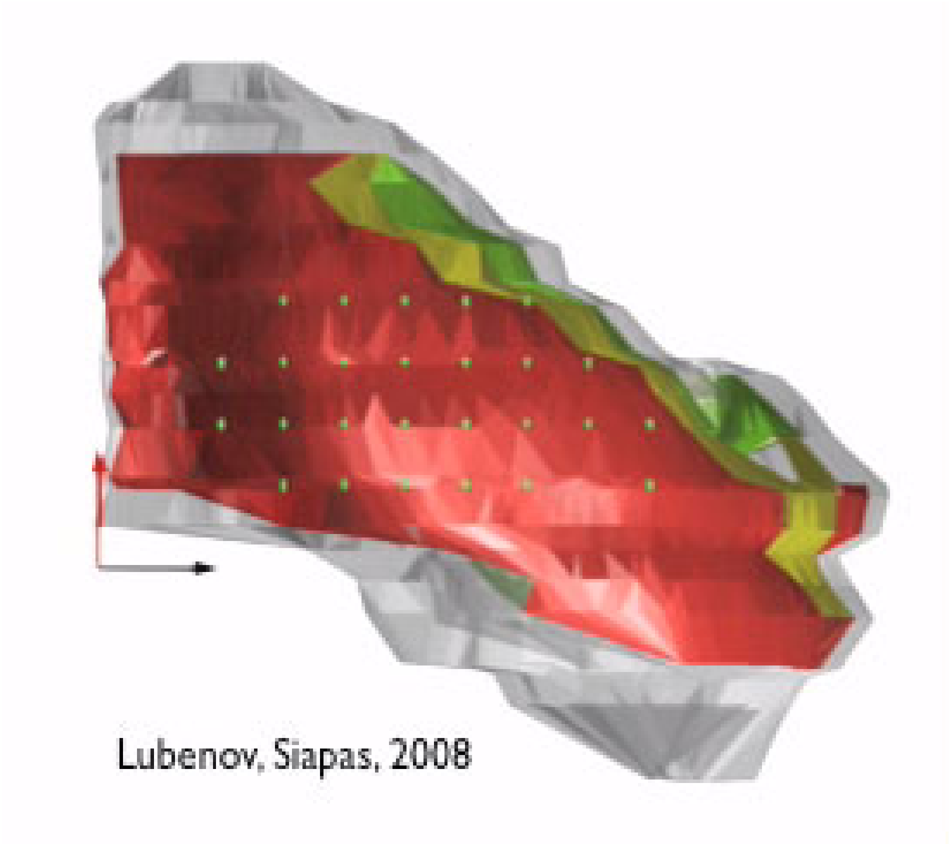
Muller and Sejnowski, *eLife*, 2016

Traveling Waves in the Hippocampus



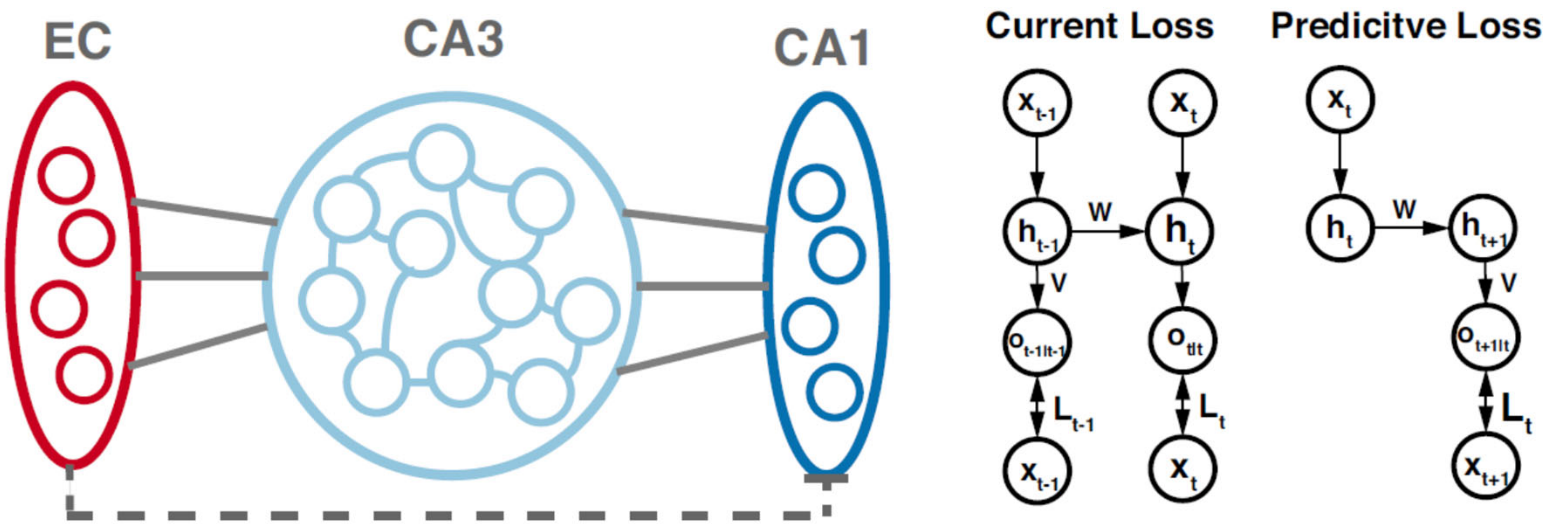
Lubenov and Siapas(2008)

Traveling Waves in the Hippocampus



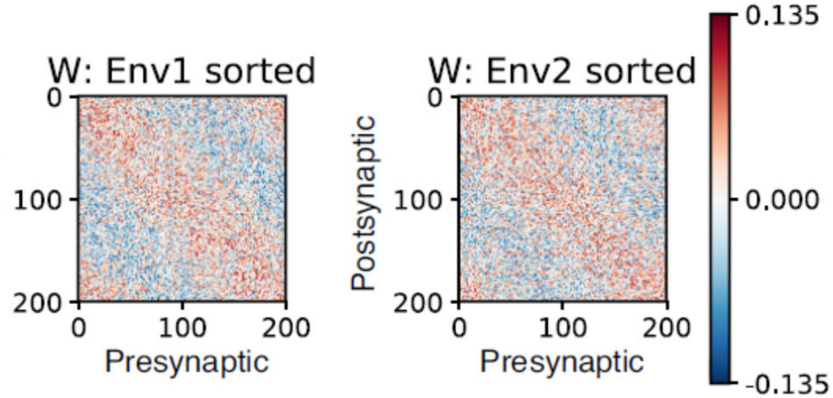
Lubenov and Siapas(2008)

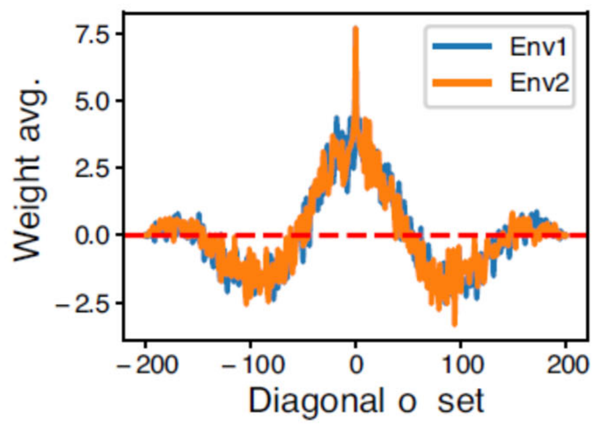
Predictive Autoencoder



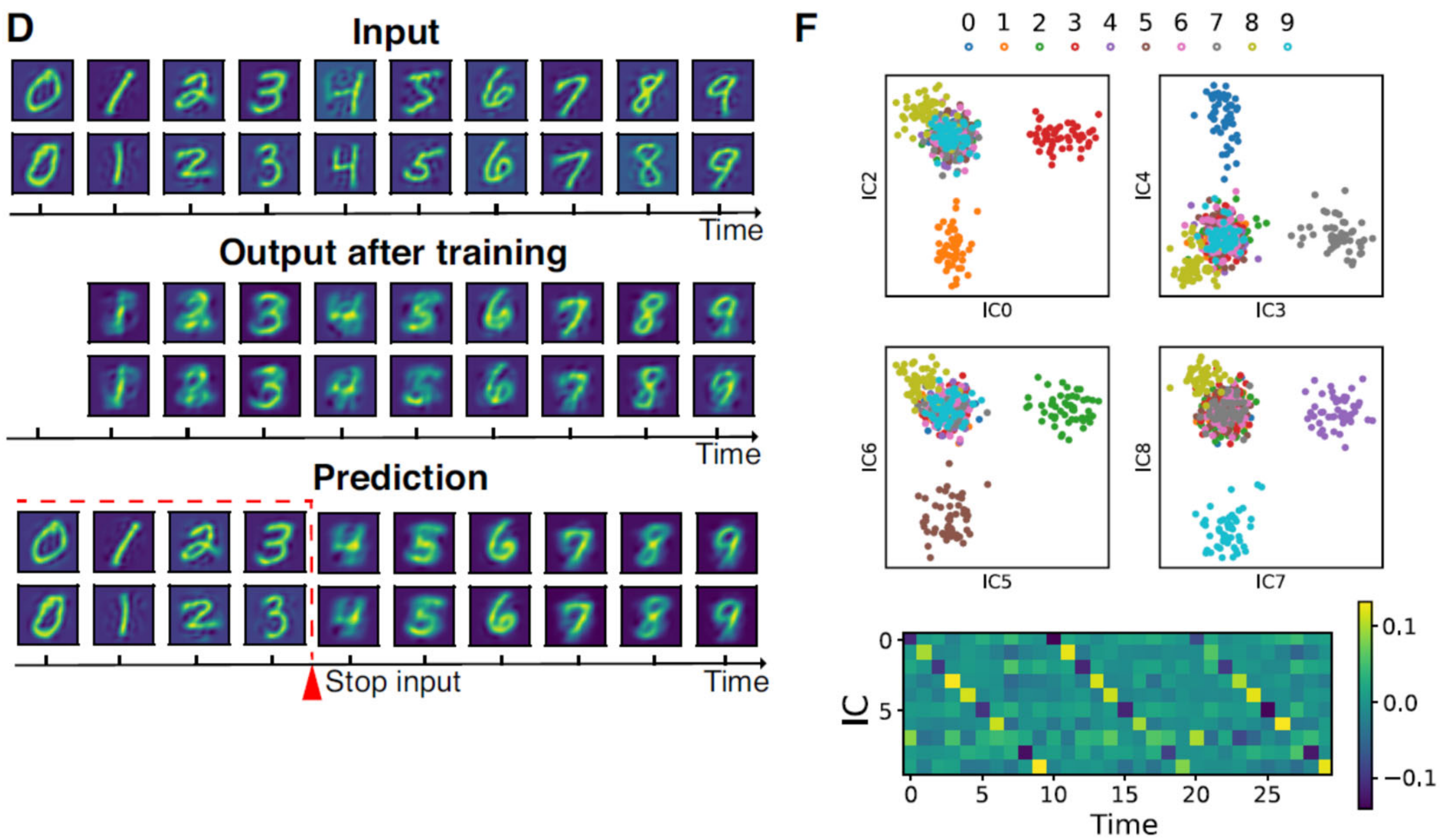
Predictive Sequence Learning in the Hippocampal Formation
Chen, Zhang, Cameron, and Sejnowski, bioRxiv, Neuron, in press

Predicting Ahead in the Hippocampus

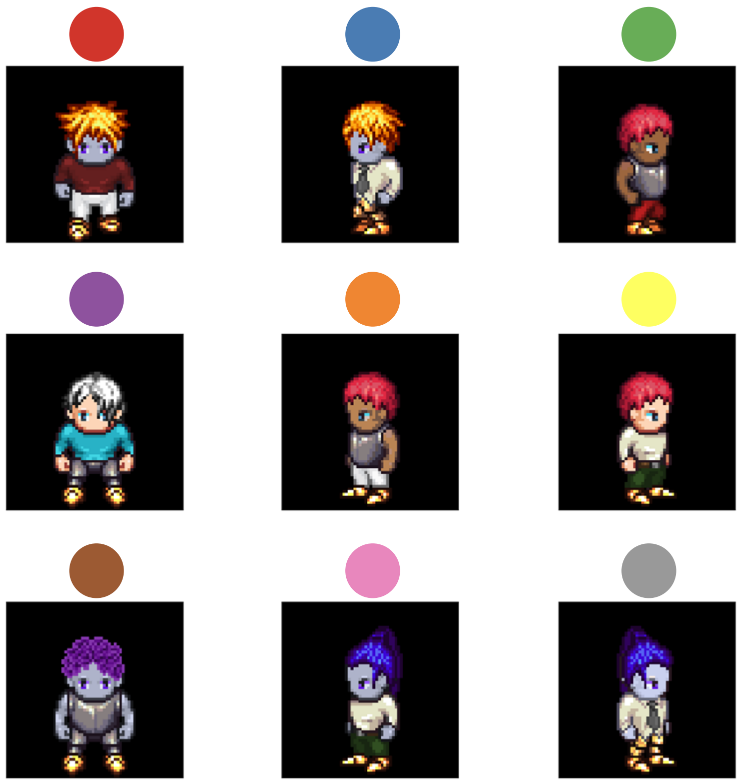


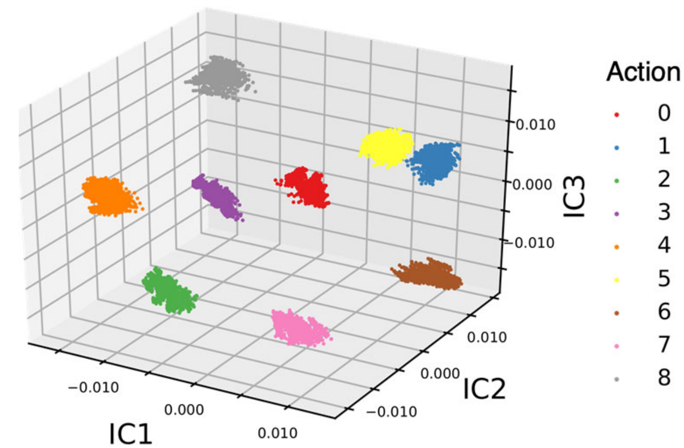


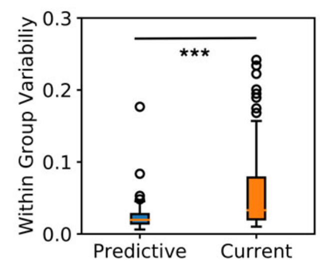
Interpreting the Code in the Hidden Units



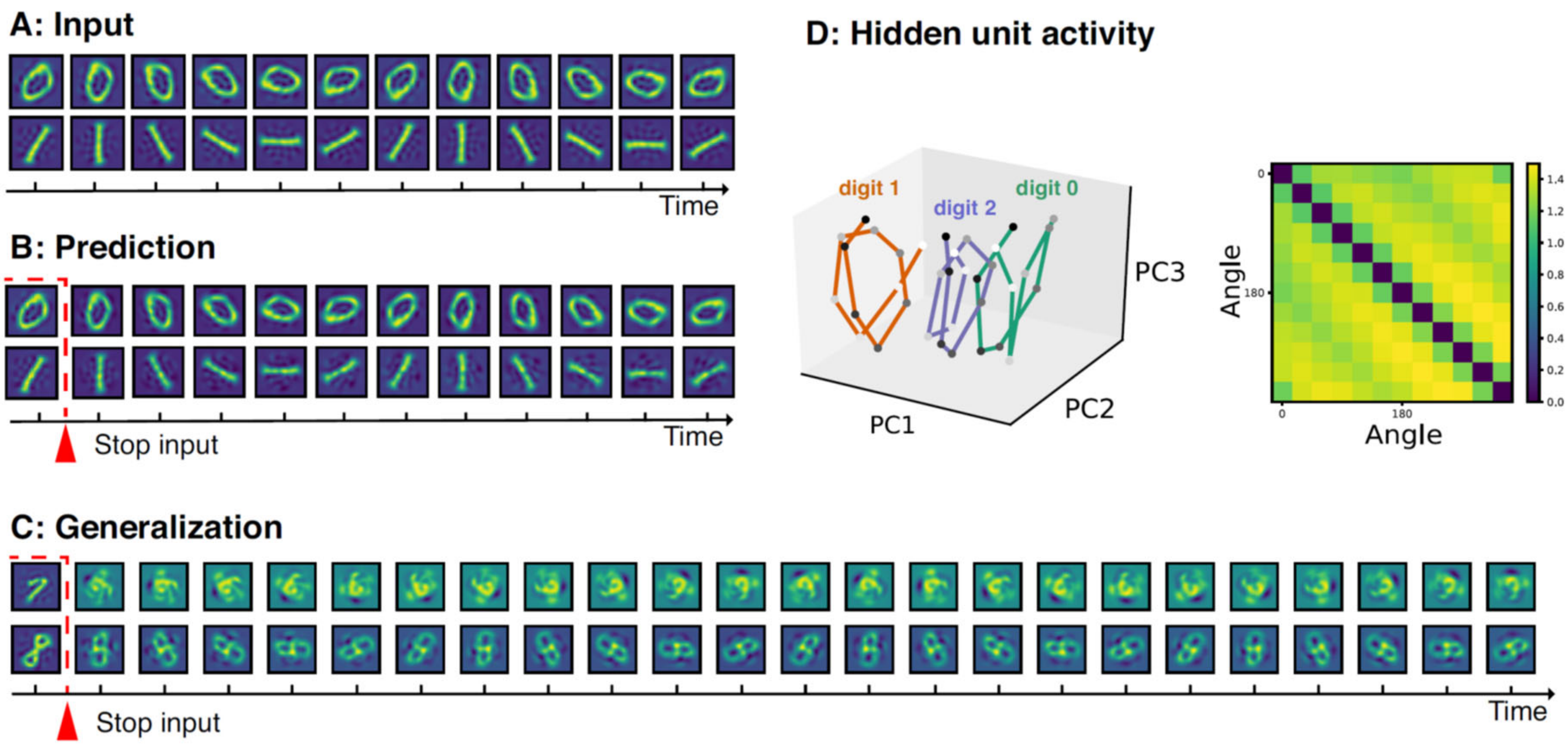
Temporal Prediction Learns to Classify Actions

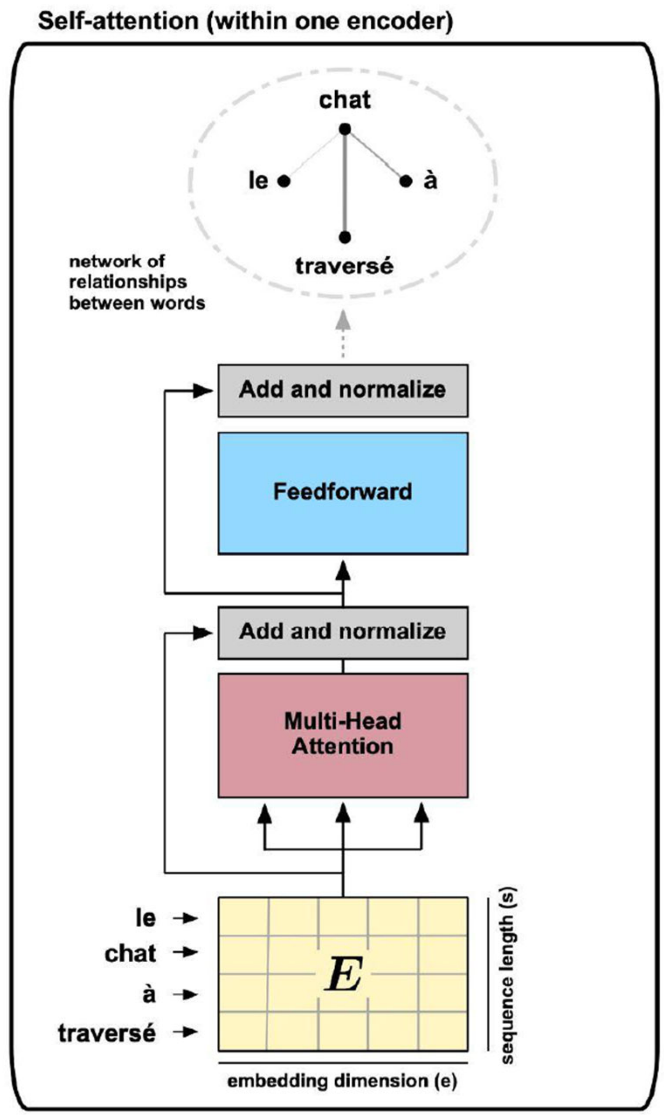






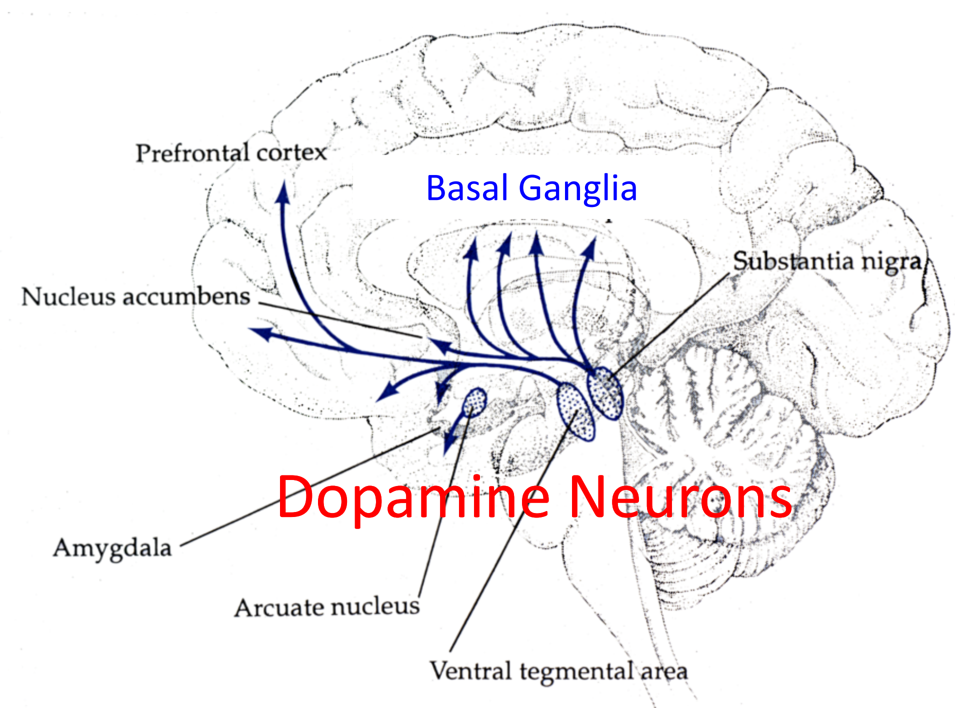
Learning How to Rotate Images





Self-attention (within one encoder)
chat
le
à
traversé
network of relationships between words
Add and normalize
Feedforward
Add and normalize
Multi-Head Attention
le
chat
à
traversé
E
sequence length (s)
embedding dimension (e)

Learning How to Decide What to Do Next

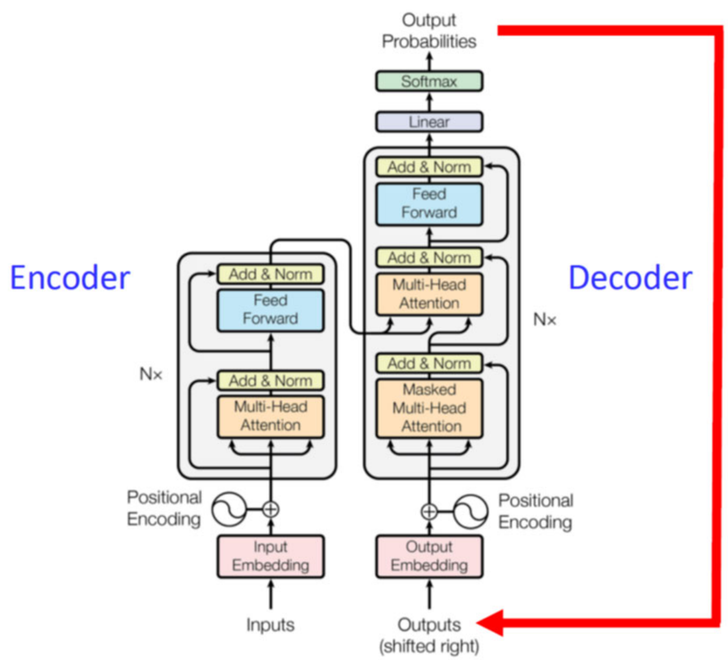


Reward Prediction Error

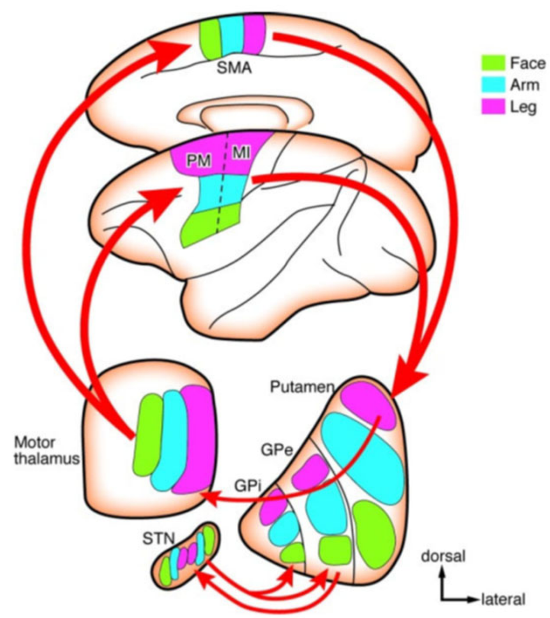
Montague, Dayan and Sejnowski, 1996

Sequence Learning in Transformers and Brains

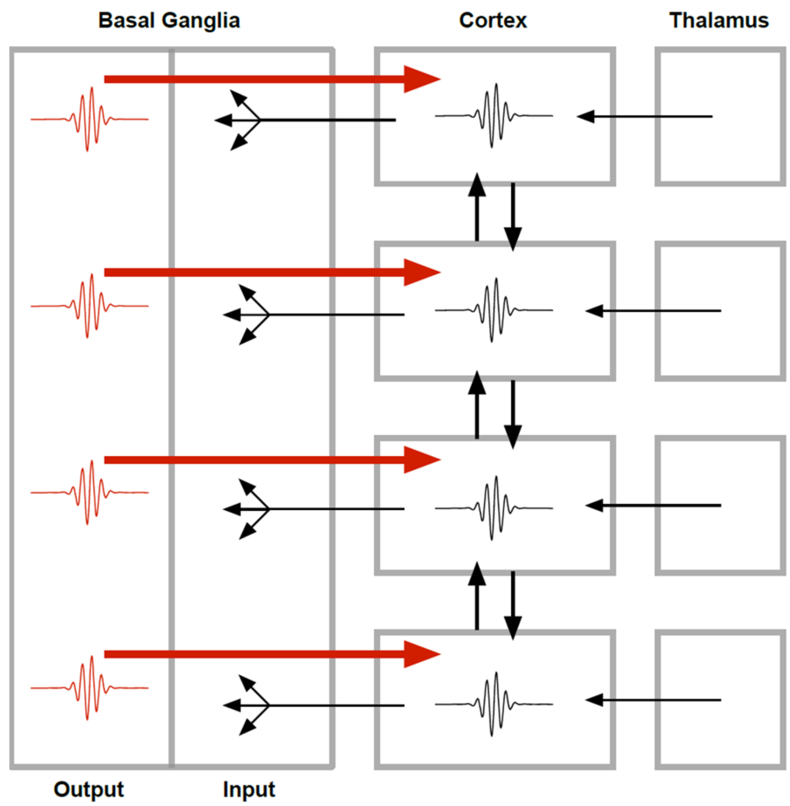
The Transformer Loop



The Cortical – Basal Ganglia Loop



Do Basal Ganglia Compute Self-Attention?



Muller, Churchland, Sejnowski, arXiv


Lyle Muller
Andy Keller
We thank you
Pat Churchland
Max Welling