Title: Aspect of Information in Classical and Quantum Neural Networks

Speakers: Huitao Shen

Series: Condensed Matter

Date: February 03, 2020 - 4:00 PM

URL: http://pirsa.org/20020052

Abstract: Iâ€™ll talk about two independent works on classical and quantum neural networks connected by information theory. In the first part of the talk, Iâ€™ll treat sequence models as one-dimensional classical statistical mechanical systems and analyze the scaling behavior of mutual information. I'll provide a new perspective on why recurrent neural networks are not good at natural language processing. In the second part of the talk, Iâ€™ll study information scrambling dynamics when quantum neural networks are trained by classical gradient descent algorithm. For many problems, this hybrid quantum-classical training process consists of two stages where information scrambles very differently in the network.

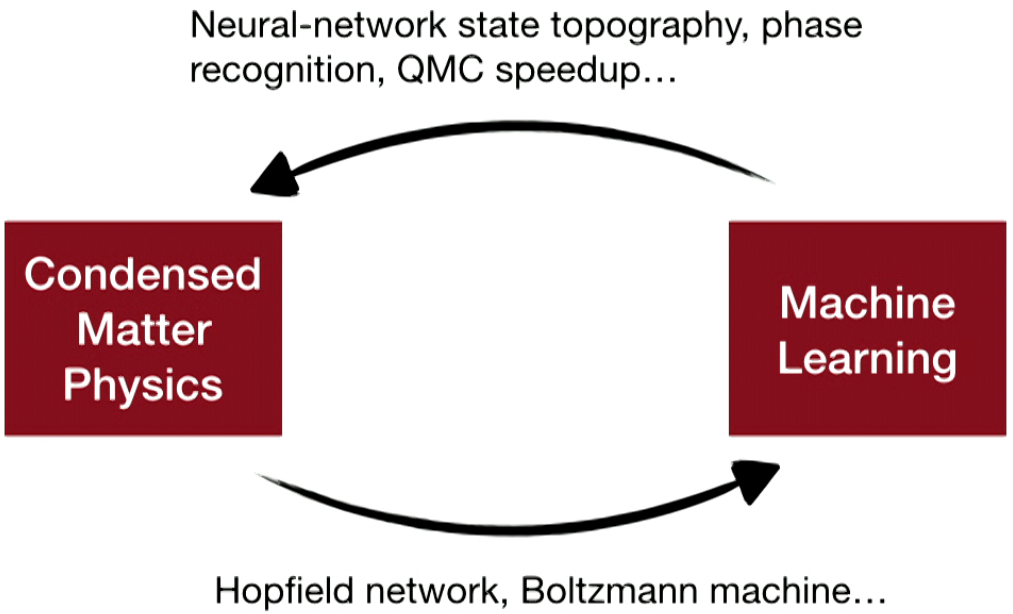
Aspects of Information in Classical and Quantum Neural Networks

Huitao Shen

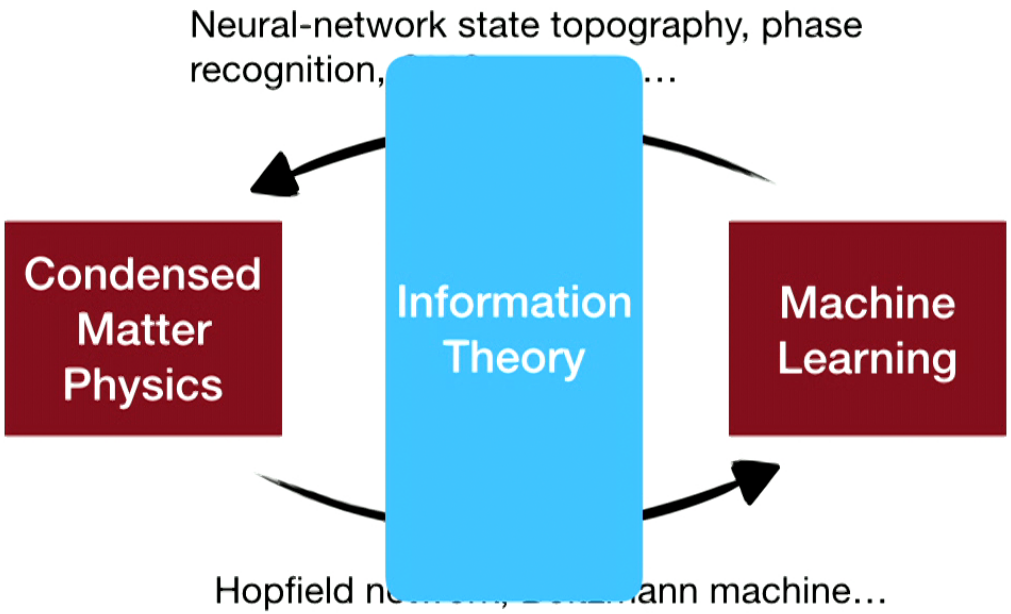
Massachusetts Institute of Technology

Huitao Shen
02/03/2020

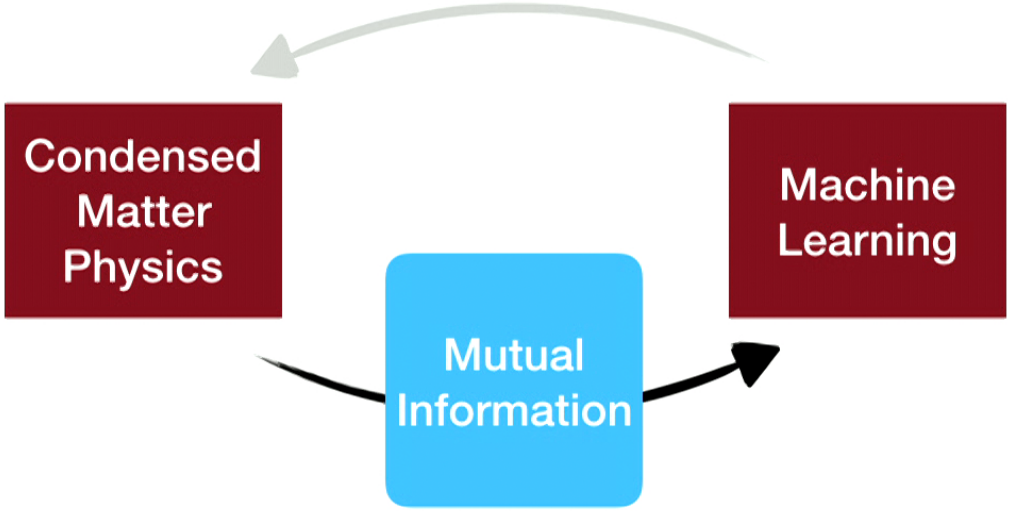
Physics & Machine Learning



Physics & Machine Learning



Physics & Machine Learning



Why are recurrent neural networks NOT good at natural language processing?

Mutual Information Scaling and Expressive Power of Sequence Models

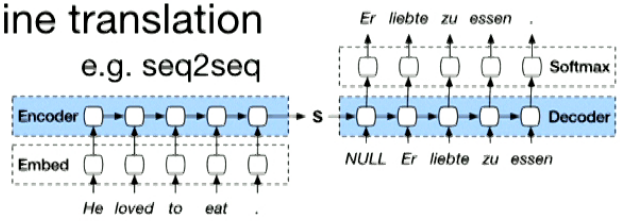
Huitao Shen
Department of Physics
Massachusetts Institute of Technology
huitao@mit.edu

arXiv: 1905.04271

Sequence Generation

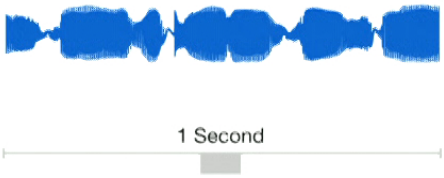
Machine translation

e.g. seq2seq



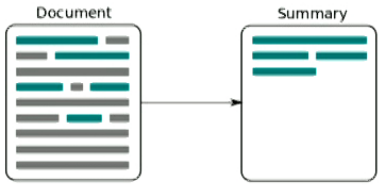
Speech generation

e.g. WaveNet



Sequence Modeling

Text summarization



Music generation

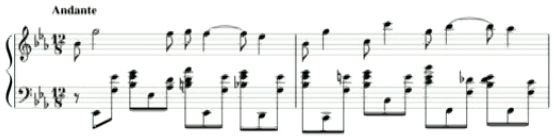
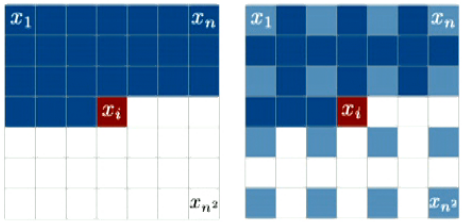


Image generation

e.g. PixelRNN



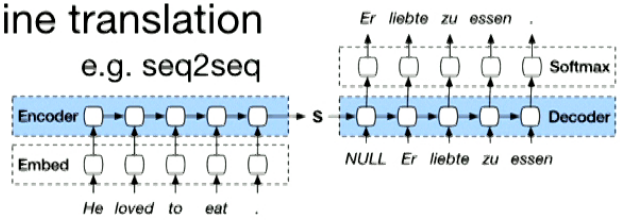
Context

Multi-scale context

Sequence Generation

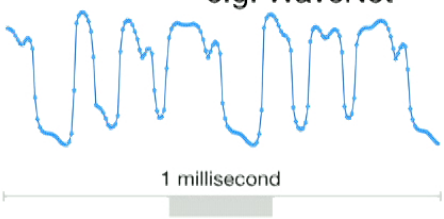
Machine translation

e.g. seq2seq



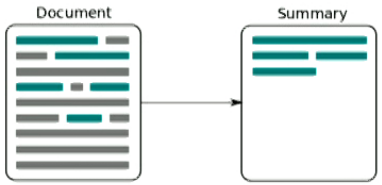
Speech generation

e.g. WaveNet



Sequence Modeling

Text summarization



Music generation

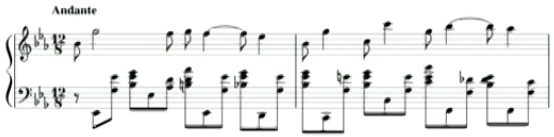
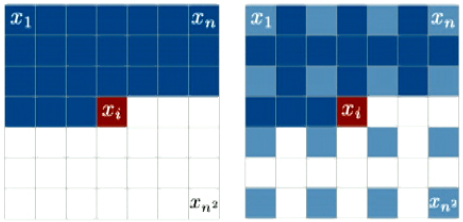


Image generation

e.g. PixelRNN



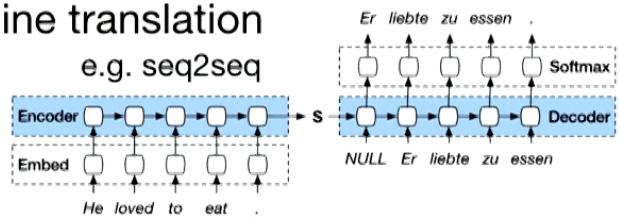
Context

Multi-scale context

Sequence Generation

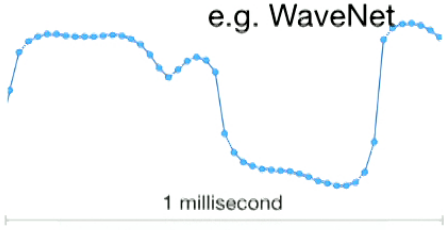
Machine translation

e.g. seq2seq



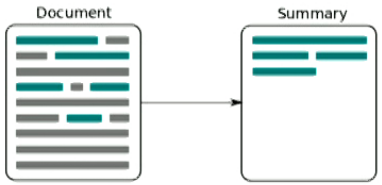
Speech generation

e.g. WaveNet



Sequence Modeling

Text summarization



Music generation

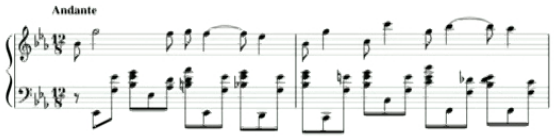
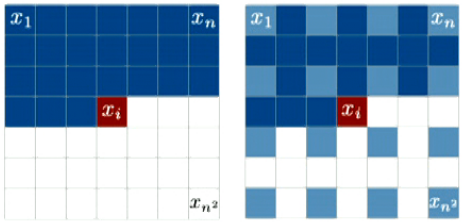


Image generation

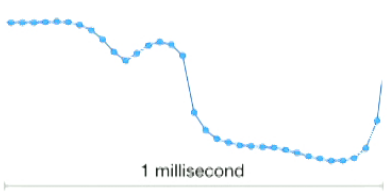
e.g. PixelRNN



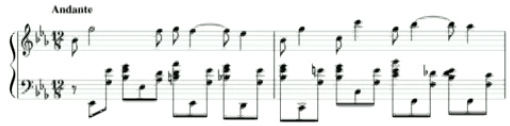
Context

Multi-scale context

Sequence Generation



All the world's a stage,
And all the men and women merely players:
They have their exits and their entrances;
And one man in his time plays many parts,
His acts being seven ages. At first the infant,
Mewling and puking in the nurse's arms.
And then the whining school-boy, with his satchel
And shining morning face, creeping like snail
Unwillingly to school. And then the lover,
Sighing like furnace, with a woeful ballad
Made to his mistress' eyebrow…



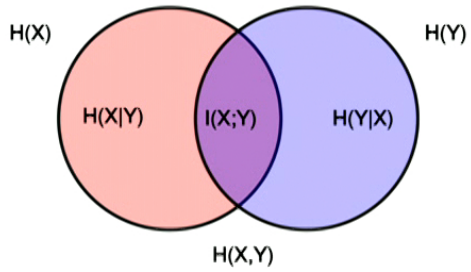
$$p(x_1, \ldots, x_t) = \prod_{i=1}^{t} p(x_i \,|\, x_1, \ldots, x_{i-1})$$

- *n*-gram model: Explicit dependence on the last *n* elements
- Recurrent neural networks: Implicit dependence on the full history
- Self-attentional models: Explicit dependence on the full history

Mutual Information Diagnosis

Mutual information: “nonlinear” correlation function

$$
\begin{aligned}
I(X;Y) &\equiv \mathbb{E}_{(X,Y)\sim p_{XY}}\left[\ln \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right] \\
&= H(X) + H(Y) - H(X,Y) \\
&= D_{\mathrm{KL}}\left(p_{XY}\,||\,p_X p_Y\right)
\end{aligned}
$$



- Beyond nonlinear dependence
- X, Y independent iff $I(X;Y) = 0$
- Well-defined even when X, Y are symbolic

In a stationary random process (time translation symmetry): $I_x(\tau) = I(x_t; x_{t+\tau})$

Mutual Information of Markov Processes

Markov process: fully described by a transition matrix $p_t = Mp_{t-1}$

$$p(x_t \,|\, x_1, \ldots, x_{t-1}) = p(x_t \,|\, x_{t-1}) = Mx_{t-1}$$

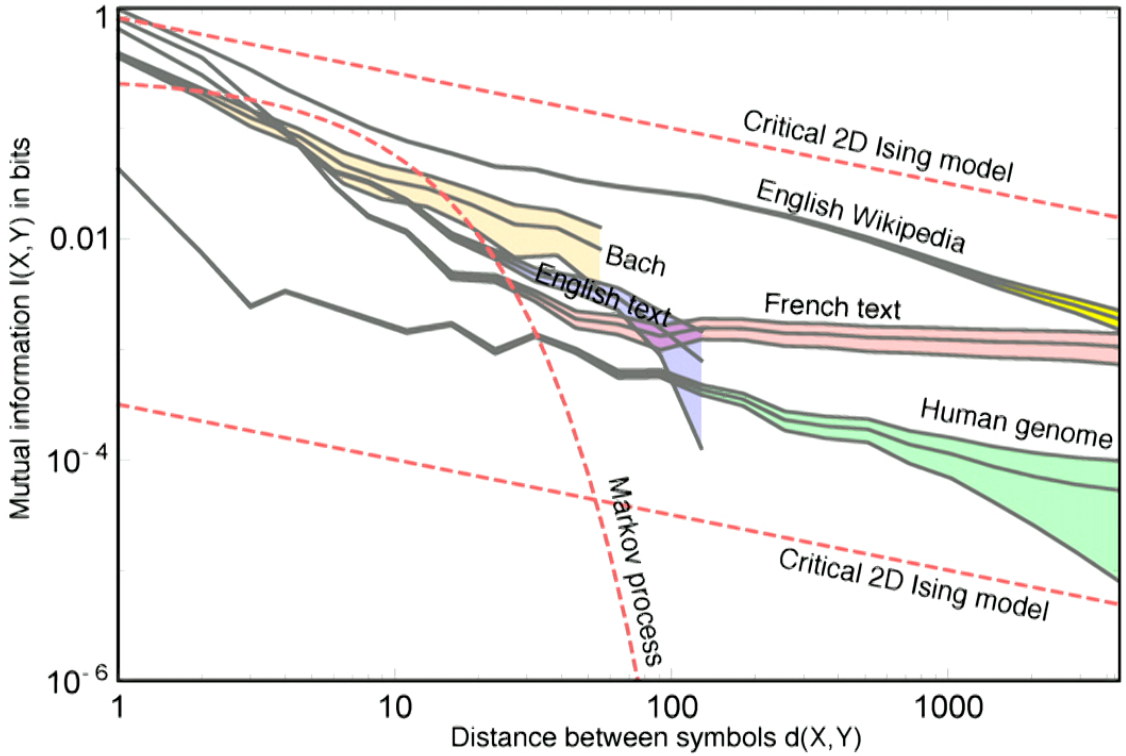
Theorem 1. *Let* $\mathbf{M}$ *be a Markov matrix that generates a Markov process. If* $\mathbf{M}$ *is irreducible and aperiodic, then the asymptotic behavior of the mutual information* $I(t_1, t_2)$ *is exponential decay toward zero for* $|t_2 - t_1| \gg 1$ *with decay timescale* $\log \frac{1}{|\lambda_2|}$*, where* λ_2 *is the second largest eigenvalue of* $\mathbf{M}$*. If* $\mathbf{M}$ *is reducible or periodic, I can instead decay to a constant; no Markov process whatsoever can produce power law decay.*

W. Li. *Complex Systems* 1(1), 107-130 (1987)

H. Lin & M. Tegmark. *Entropy* 19(7), 299 (2017)

Corollary: *n*-gram models and hidden Markov models (HMMs) have exponential decaying mutual information

Mutual Information in Natural Sequences



Symbols

- Bach: **Music notes**
- English Wikipedia, English text, French text: **Characters**
- Human genome: **{A, T, C, G}**

H. Lin & M. Tegmark. *Entropy* 19(7), 299 (2017)

Mutual Information Diagnosis

Markov chain based models, such as *n*-grams and hidden Markov models, cannot capture long-range power-law dependence in natural sequences, for example natural language texts.

Van Hove's Theorem

In 1D, statistical physics systems with short-range interactions cannot have finite temperature phase transition

L. Landau and E. Lifshitz, *Statistical Physics*, Vol 5
L. Van Hove, *Physica* 16(2), 137-143 (1950)

Therefore, correlation and mutual information always decay exponentially at finite temperature

Van Hove's Theorem

In 1D, statistical physics systems with short-range interactions cannot have finite temperature phase transition

L. Landau and E. Lifshitz, *Statistical Physics*, Vol 5
L. Van Hove, *Physica* 16(2), 137-143 (1950)

Therefore, correlation and mutual information always decay exponentially at finite temperature

Example: Markovian conditional sampling of 1D Ising model

$$H = J \sum_{i=1}^{N-1} s_i s_{i+1}$$

Van Hove's Theorem

In 1D, statistical physics systems with short-range interactions cannot have finite temperature phase transition

L. Landau and E. Lifshitz, *Statistical Physics*, Vol 5
L. Van Hove, *Physica* 16(2), 137-143 (1950)

Therefore, correlation and mutual information always decay exponentially at finite temperature

Example: Markovian conditional sampling of 1D Ising model

$$H = J \sum_{i=1}^{N-1} s_i s_{i+1}$$

Boltzmann distribution: $p(s_1, \ldots, s_N) = \frac{1}{Z} e^{-\beta J \sum_{i=1}^{N-1} s_i s_{i+1}}$

Van Hove's Theorem

Example: Markovian conditional sampling of 1D Ising model

$$H = J\sum_{i=1}^{N-1} s_i s_{i+1} \qquad p(s_1, \ldots, s_N) = \frac{1}{Z} e^{-\beta J \sum_{i=1}^{N-1} s_i s_{i+1}}$$

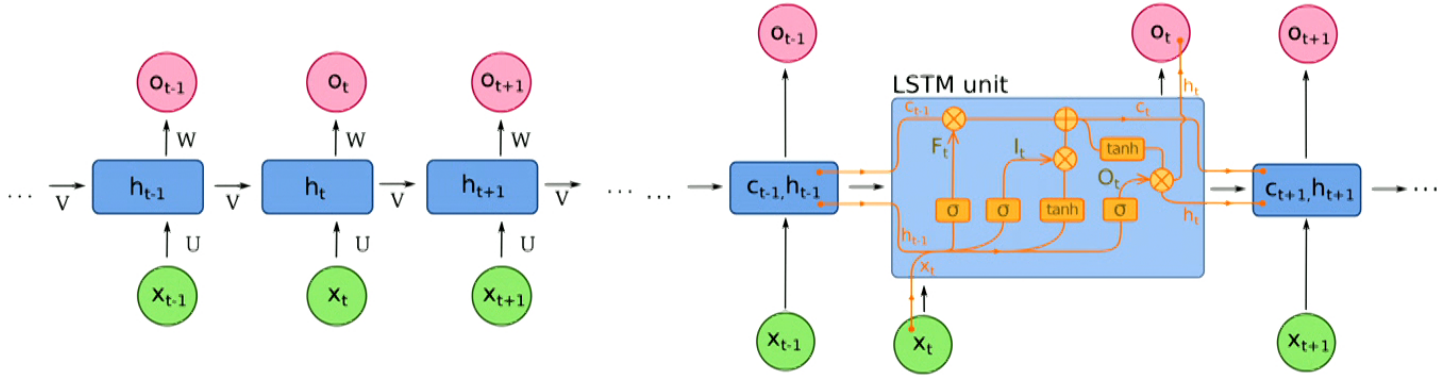
$$\begin{aligned} p(s_t \,|\, s_{t-1}, \ldots, s_1) &= \frac{p(s_t, s_{t-1}, \ldots, s_1)}{p(s_{t-1}, \ldots, s_1)} \\ &= \frac{\sum_{s_{t+1}, \ldots, s_N} e^{-\beta J \sum_{i=1}^{N-1} s_i s_{i+1}}}{\sum_{s_t, s_{t+1}, \ldots, s_N} e^{-\beta J \sum_{i=1}^{N-1} s_i s_{i+1}}} \\ &= \frac{e^{-\beta J s_{t-1} s_t} \sum_{s_{t+1}, \ldots, s_N} e^{-\beta J s_t s_{t+1}} e^{-\beta J \sum_{i=t+1}^{N-1} s_i s_{i+1}}}{\sum_{s_t} e^{-\beta J s_{t-1} s_t} \sum_{s_{t+1}, \ldots, s_N} e^{-\beta J s_t s_{t+1}} e^{-\beta J \sum_{i=t+1}^{N-1} s_i s_{i+1}}} \\ &= \frac{e^{-\beta J s_{t-1} s_t}}{\sum_{s_t} e^{-\beta J s_{t-1} s_t}} = \frac{e^{-\beta J s_{t-1} s_t}}{2\cosh\left(\beta J s_{t-1}\right)} \end{aligned}$$

HS, arXiv: 1905.04271

Consistent with the fact that Markov processes are always off-critical

Question

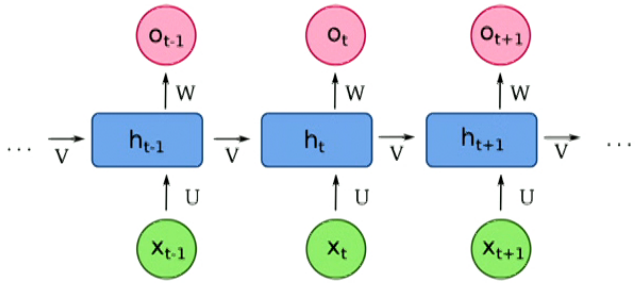
What is the mutual information scaling of RNNs?



Recurrent Neural Networks

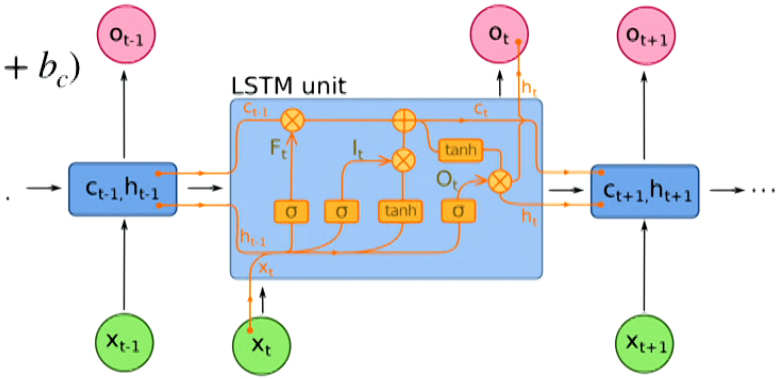
Elman network:

$$h_t = \sigma_h(Ux_t + Vh_{t-1} + b_h)$$



Long short-term memory:

$$F_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$
$$I_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$O_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$
$$c_t = F_t \circ c_{t-1} + I_t \circ \tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$h_t = O_t \circ \tanh(c_t)$$



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

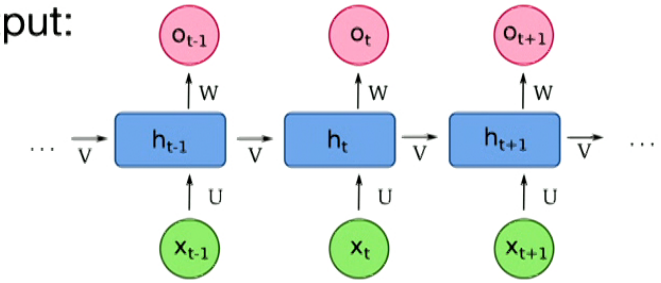
Sequence generation:

$$o_t = \sigma_x(Wh_t + b_o)$$
$$x_{t+1} \sim P(o_t)$$

Theory: Linear RNNs

Linear Elman network with Gaussian output:

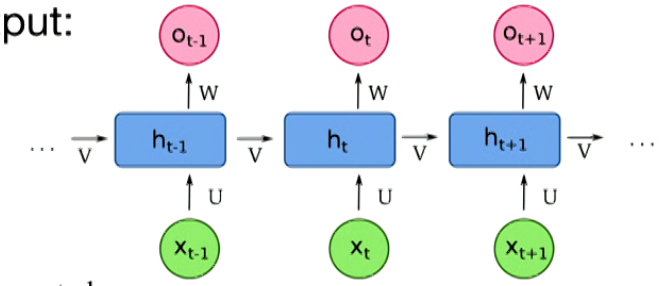
$$h_t = W_h x_{t-1} + U_h h_{t-1},$$
$$o_t = U_o h_{t-1},$$
$$x_t \sim \mathcal{N}(o_t, \sigma^2 I_d)$$



Theory: Linear RNNs

Linear Elman network with Gaussian output:

$$h_t = W_h x_{t-1} + U_h h_{t-1},$$
$$o_t = U_o h_{t-1},$$
$$x_t \sim \mathcal{N}(o_t, \sigma^2 I_d)$$



"Integrate out" hidden states: $h_t = U_h^t h_0 + \sum_{i=0}^{t-1} U_h^{t-1-i} W_h x_i$

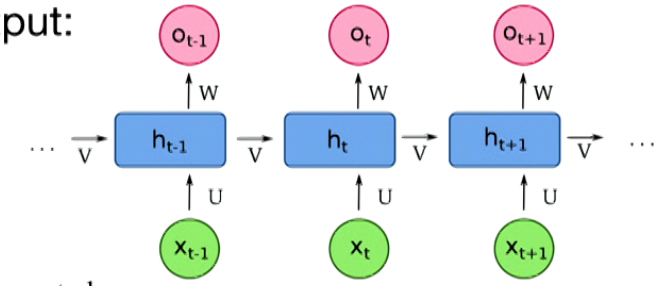
$$o_t = U_o U_h^t h_0 + \sum_{i=0}^{t-2} U_o U_h^{t-1-i} W_h x_i$$

Observation: Exponential dependence on the sequence history

Theory: Linear RNNs

Linear Elman network with Gaussian output:

$$h_t = W_h x_{t-1} + U_h h_{t-1},$$
$$o_t = U_o h_{t-1},$$
$$x_t \sim \mathcal{N}(o_t, \sigma^2 I_d)$$



"Integrate out" hidden states: $h_t = U_h^t h_0 + \sum_{i=0}^{t-1} U_h^{t-1-i} W_h x_i$

$$o_t = U_o U_h^t h_0 + \sum_{i=0}^{t-2} U_o U_h^{t-1-i} W_h x_i$$

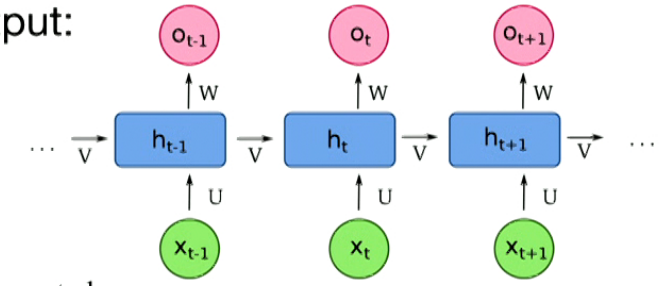
Observation: Exponential dependence on the sequence history

Physically, exponential interaction is still "short-range" interaction. The model is in the same universality class as finite-range Markovian models

Theory: Linear RNNs

Linear Elman network with Gaussian output:

$$h_t = W_h x_{t-1} + U_h h_{t-1},$$
$$o_t = U_o h_{t-1},$$
$$x_t \sim \mathcal{N}(o_t, \sigma^2 I_d)$$



"Integrate out" hidden states: $h_t = U_h^t h_0 + \sum_{i=0}^{t-1} U_h^{t-1-i} W_h x_i$

$$o_t = U_o U_h^t h_0 + \sum_{i=0}^{t-2} U_o U_h^{t-1-i} W_h x_i$$

Observation: Exponential dependence on the sequence history

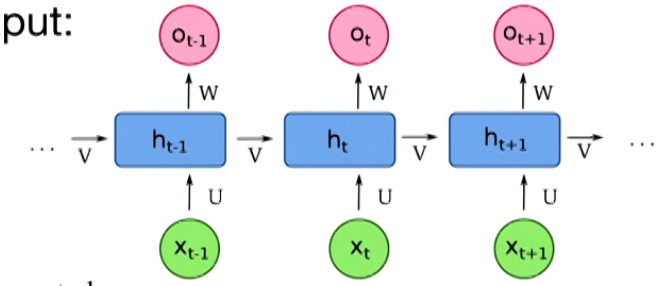
Physically, exponential interaction is still "short-range" interaction. The model is in the same universality class as finite-range Markovian models

Expect exponential decaying mutual information!

Theory: Linear RNNs

Linear Elman network with Gaussian output:

$$h_t = W_h x_{t-1} + U_h h_{t-1},$$
$$o_t = U_o h_{t-1},$$
$$x_t \sim \mathcal{N}(o_t, \sigma^2 I_d)$$



"Integrate out" hidden states: $h_t = U_h^t h_0 + \sum_{i=0}^{t-1} U_h^{t-1-i} W_h x_i$

$$o_t = U_o U_h^t h_0 + \sum_{i=0}^{t-2} U_o U_h^{t-1-i} W_h x_i$$

Observation: Exponential dependence on the sequence history

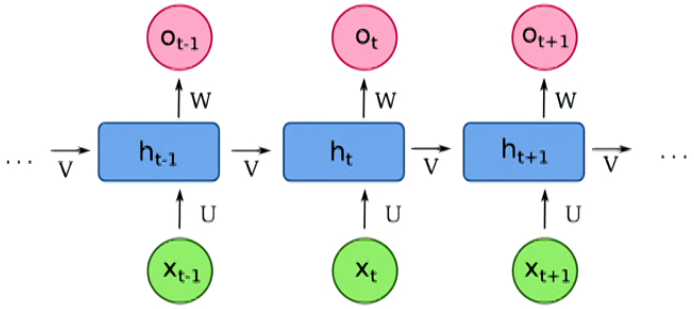
Theorem: The mutual information $I_x(\tau)$ in the linear Elman RNN with Gaussian output decreases exponentially with time τ, if the RNN does not simply memorize the initial distribution.

Proof see HS, arXiv: 1905.04271

Experiments: Nonlinear RNNs

Dataset 1: Artificial binary dataset with power-law mutual information

x=(0, 1, 1, 0, 0, 1, 1, 0, …) $I(\tau) = 0.1\tau^{-0.4}$ L=512



$x_t \sim \text{Bernoulli}(o_{t-1})$

In general, if *x* is symbolic with vocabulary size *V*

$x_t \sim \text{Categorical}(o_{t-1})$

After training, sequences are generated unconditionally (start with zero hidden state)

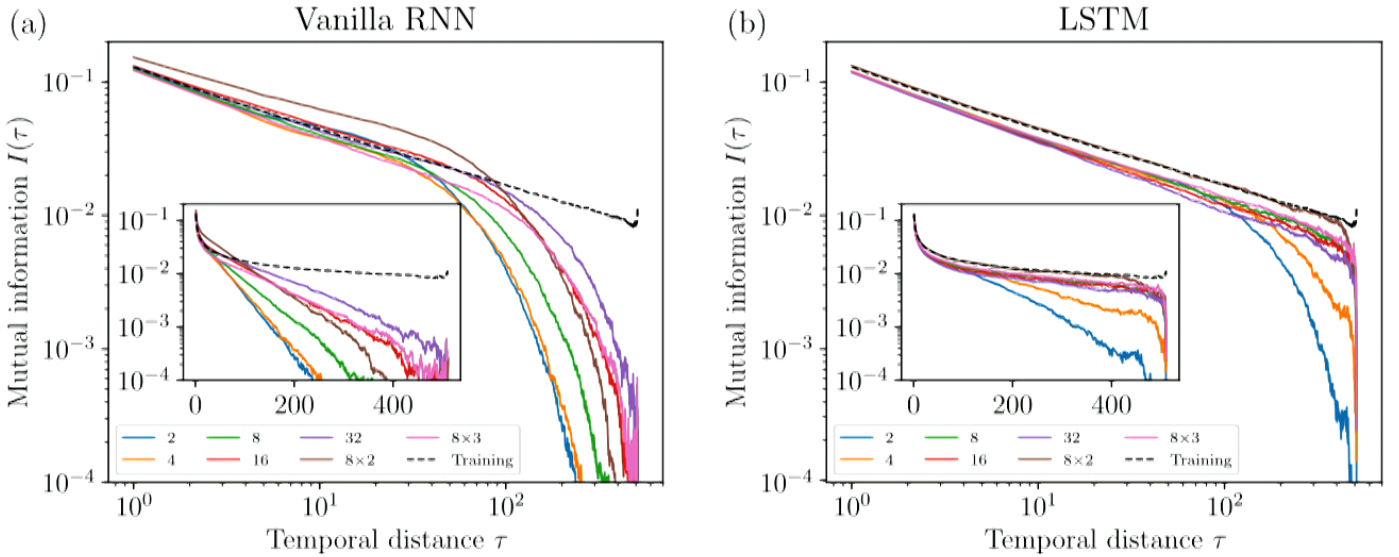
HS, arXiv: 1905.04271

Experiments: Nonlinear RNNs

Dataset 1: Artificial binary dataset with power-law mutual information

x=(0, 1, 1, 0, 0, 1, 1, 0, …) $I(\tau) = 0.1\tau^{-0.4}$ L=512

Mutual information estimated on the unconditionally generated sequences:

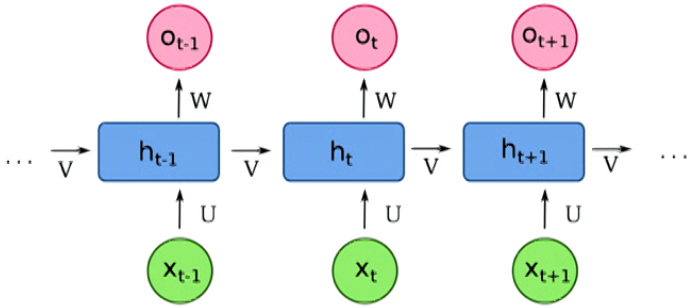


HS, arXiv: 1905.04271

Experiments: Nonlinear RNNs

Dataset 1: Artificial binary dataset with power-law mutual information

x=(0, 1, 1, 0, 0, 1, 1, 0, …) $I(\tau) = 0.1\tau^{-0.4}$ L=512



$x_t \sim \text{Bernoulli}(o_{t-1})$

In general, if *x* is symbolic with vocabulary size *V*

$x_t \sim \text{Categorical}(o_{t-1})$

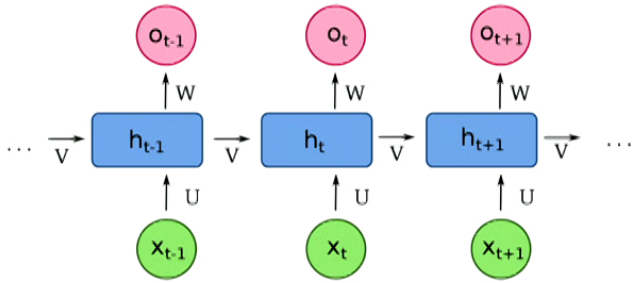
After training, sequences are generated unconditionally (start with zero hidden state)

HS, arXiv: 1905.04271

Recurrent Neural Networks

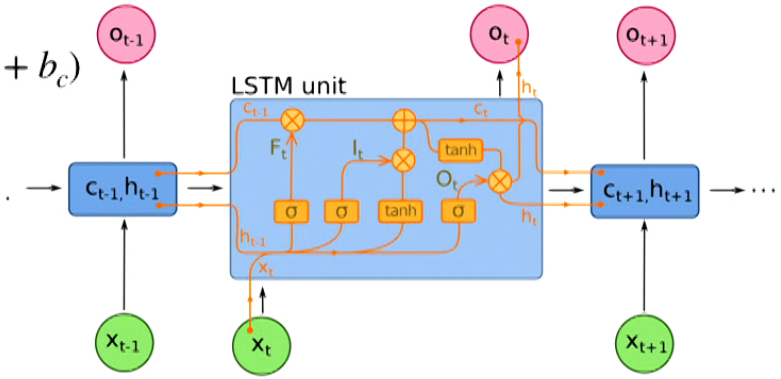
Elman network:

$$h_t = \sigma_h(Ux_t + Vh_{t-1} + b_h)$$



Long short-term memory:

$$F_t = \sigma(W_f x_t + U_f h_{t-1} + b_f)$$
$$I_t = \sigma(W_i x_t + U_i h_{t-1} + b_i)$$
$$O_t = \sigma(W_o x_t + U_o h_{t-1} + b_o)$$
$$c_t = F_t \circ c_{t-1} + I_t \circ \tanh(W_c x_t + U_c h_{t-1} + b_c)$$
$$h_t = O_t \circ \tanh(c_t)$$



https://colah.github.io/posts/2015-08-Understanding-LSTMs/

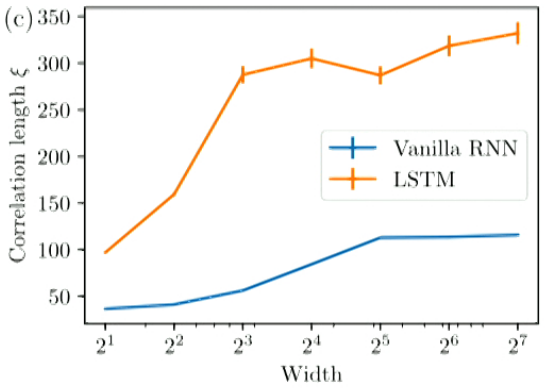
Sequence generation:

$$o_t = \sigma_x(Wh_t + b_o)$$
$$x_{t+1} \sim P(o_t)$$

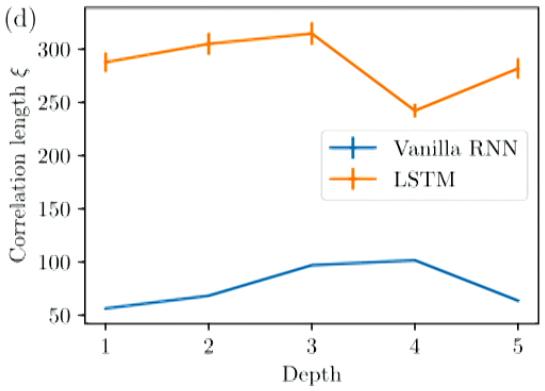
Experiments: Nonlinear RNNs

Dataset 1: Artificial binary dataset with power-law mutual information

Define "correlation length": $I = I_0 e^{-\tau/\xi}$



$\xi \sim \ln \text{Width}$



Almost independent of depth

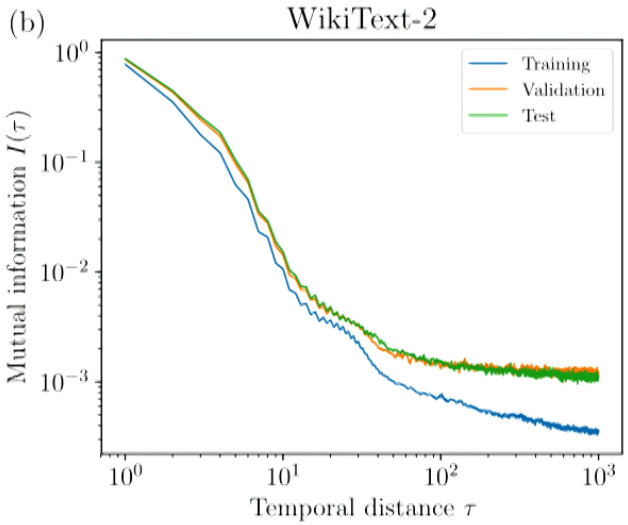
Consistent with previous studies, c.f. arXiv: 1707.05589

Baseline: Fitting on the training set yields $\xi \approx 420$

HS, arXiv: 1905.04271

Experiments: Nonlinear RNNs

Dataset 2: Natural language WikiText-2



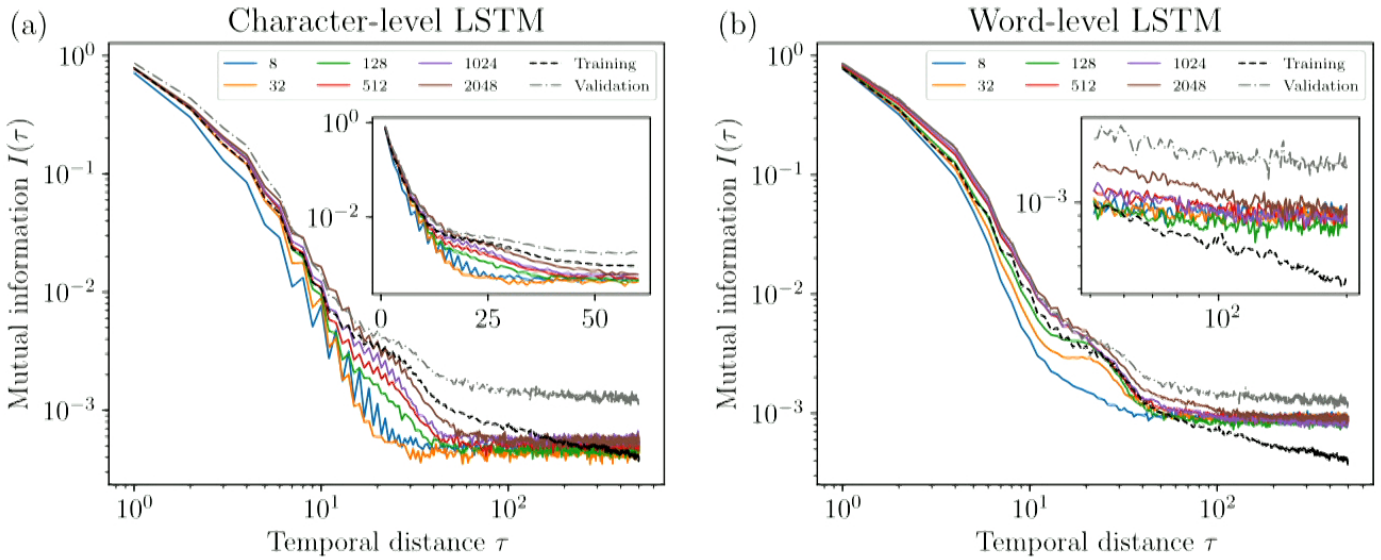
Mutual information is estimated at **character level**

- Short distance (word level)
  $\tau \lesssim 10$
  Exponential decay
- Long distance (paragraph level)
  $50 \lesssim \tau \lesssim 1000$
  Power-law decay
- Intermediate distance (sentence level)
  $10 \lesssim \tau \lesssim 50$
  Mixed behavior

Experiments: Nonlinear RNNs

Dataset 2: Natural language WikiText-2

c.f.
M. Hermans & B. Schrauwen, NIPS 2013
arXiv: 1506.02078, 1805.04623



Sequences are generated
character by character

Sequences are generated
word by word

Mutual information is estimated at **character level**

HS, arXiv: 1905.04271

Beyond Van Hove

In 1D, statistical physics systems with short-range interactions cannot have finite temperature phase transition

L. Landau and E. Lifshitz, *Statistical Physics*, Vol 5
L. Van Hove, *Physica* 16(2), 137-143 (1950)

In 1D, statistical physics systems with long-range (power-law decaying) interactions can have finite temperature phase transition

Beyond Van Hove

In 1D, statistical physics systems with short-range interactions cannot have finite temperature phase transition

L. Landau and E. Lifshitz, *Statistical Physics*, Vol 5
L. Van Hove, *Physica* 16(2), 137-143 (1950)

In 1D, statistical physics systems with long-range (power-law decaying) interactions can have finite temperature phase transition

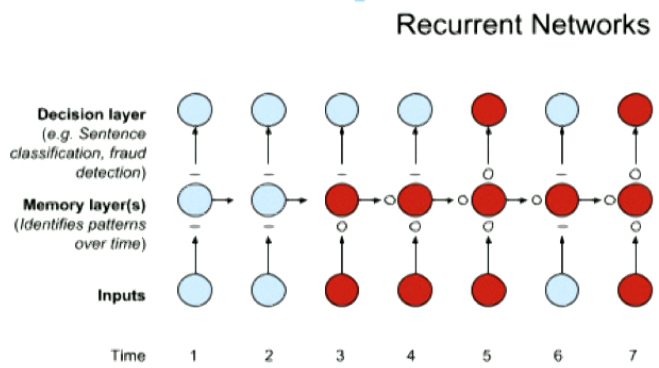
Example: Power-law Ising model $H = \frac{J}{2} \sum_{i,j=1}^{N} \frac{1}{|i-j|^{\alpha}} s_i s_j$

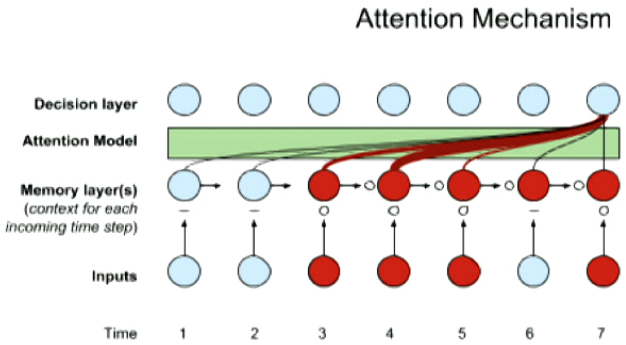
Finite temperature critical point exists when $1 < \alpha \leq 2$

- $1 < \alpha < 1.5$: Mean-field universality class
- $1.5 < \alpha \leq 2$: Critical exponent depends on α
- $\alpha > 2$: Behave likes short-range Ising model

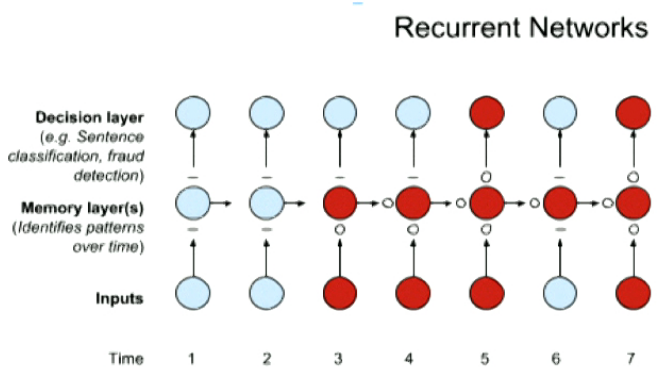
F.J. Dyson, *Commun. Math. Phys.* 12(2), 91-107 (1969)
E. Luijten & H. W. J. Blöte, *Phys. Rev. B* 56, 8945 (1997)

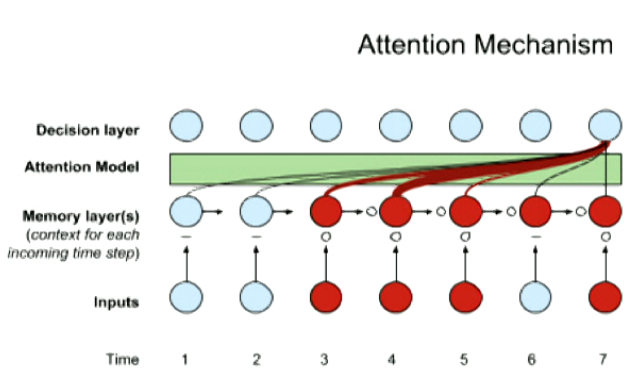
Self-Attentional Models

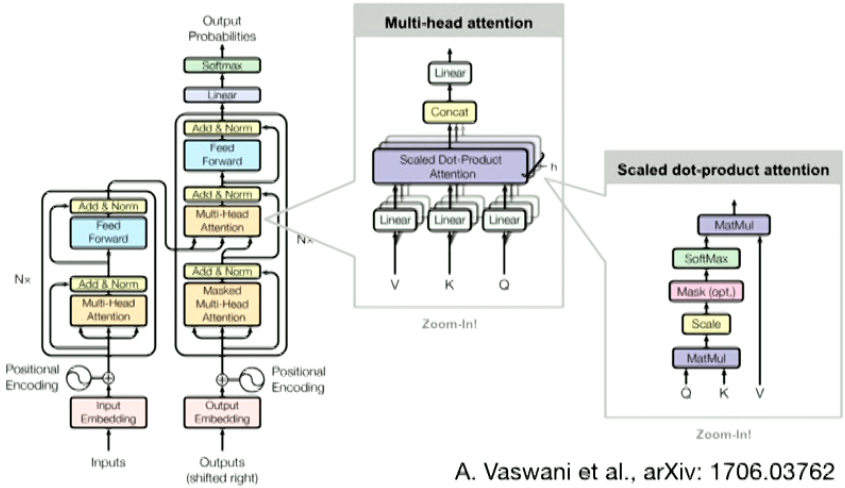




Self-Attentional Models



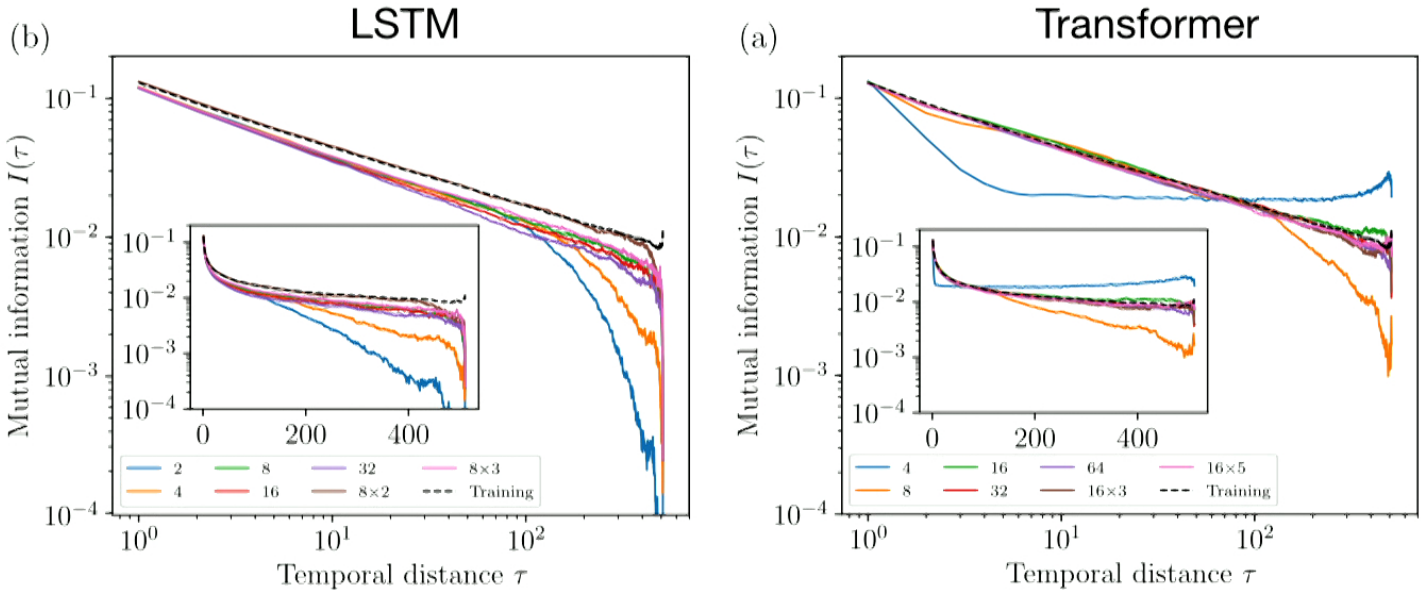




A. Vaswani et al., arXiv: 1706.03762

Experiment: Transformers

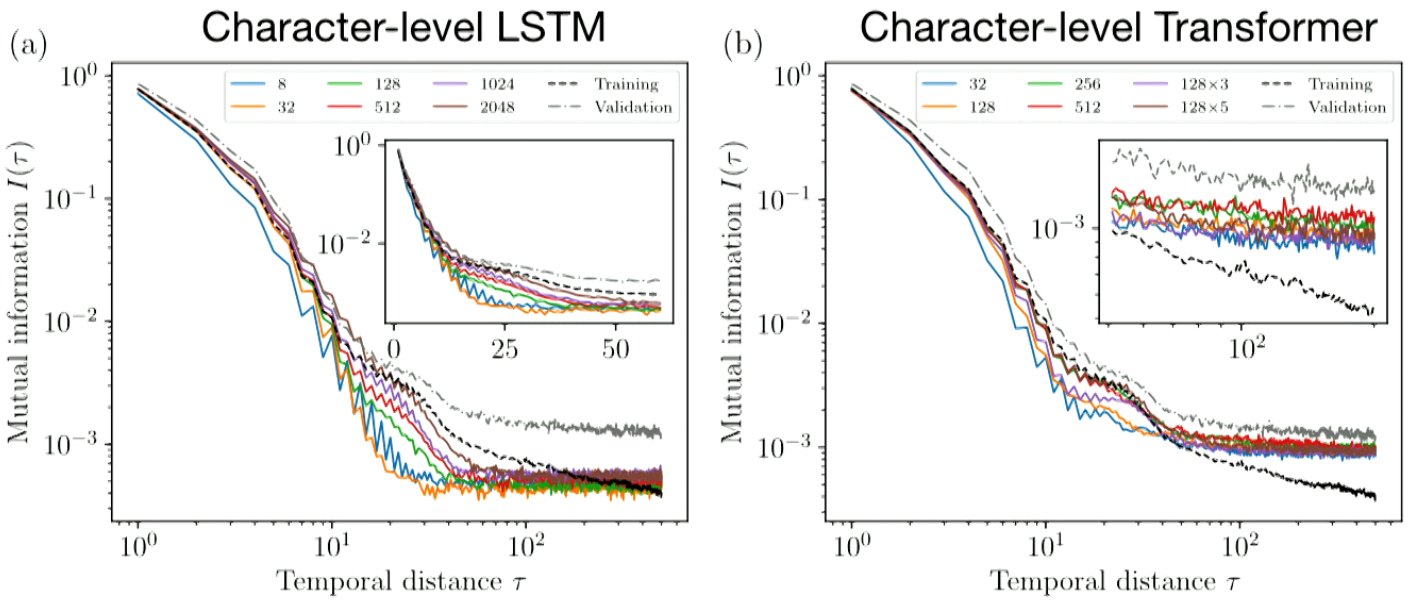
Dataset 1: Artificial binary dataset with power-law mutual information



HS, arXiv: 1905.04271

Experiment: Transformers

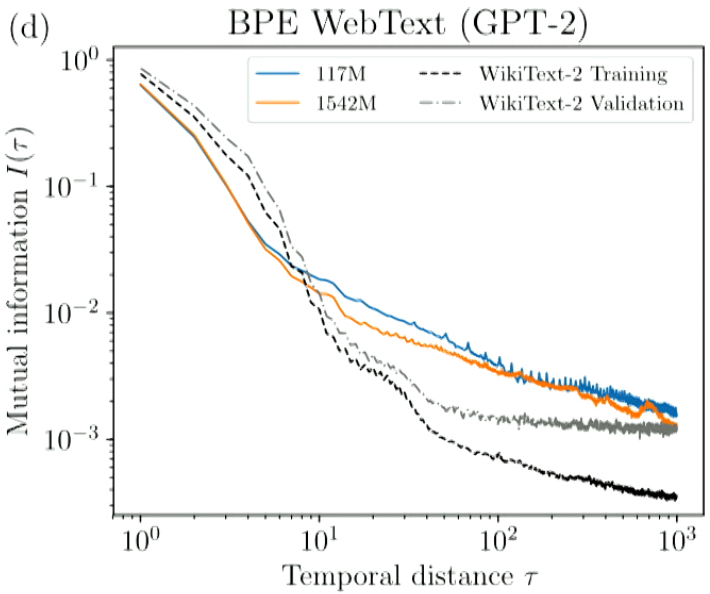
Dataset 2: Natural language WikiText-2



HS, arXiv: 1905.04271

Experiment: Transformers

State-of-the-art: GPT-2



HS, arXiv: 1905.04271

Text sample:

Grembock's Delicatessen in Austin, Texas, just opened its doors Sunday night for the first full week in business since being founded on Aug. 30. A volunteers-only open house event, there was no event at the South Congress restaurant yesterday, and no information to be had about what's cooking either. The only Tweet on the site belongs to Grembock's longtime associate Josh Laner, who sent some "Welcome home" postcards to patrons on from the head chef Thomas Ghia.

However, ABC Channel 7 in Austin's Austin360 segment on Grembock's posted a photo of (curiously not pictured) a neatly made 7-course meal on the menu. Eater reached out to Grembock's management for additional info. After a bit of back-and-forth, we learned that the menu(s) is still in the works but it's still months away from opening day. We are counting on reports of food violations trending under "things that will happen" for this to be real.

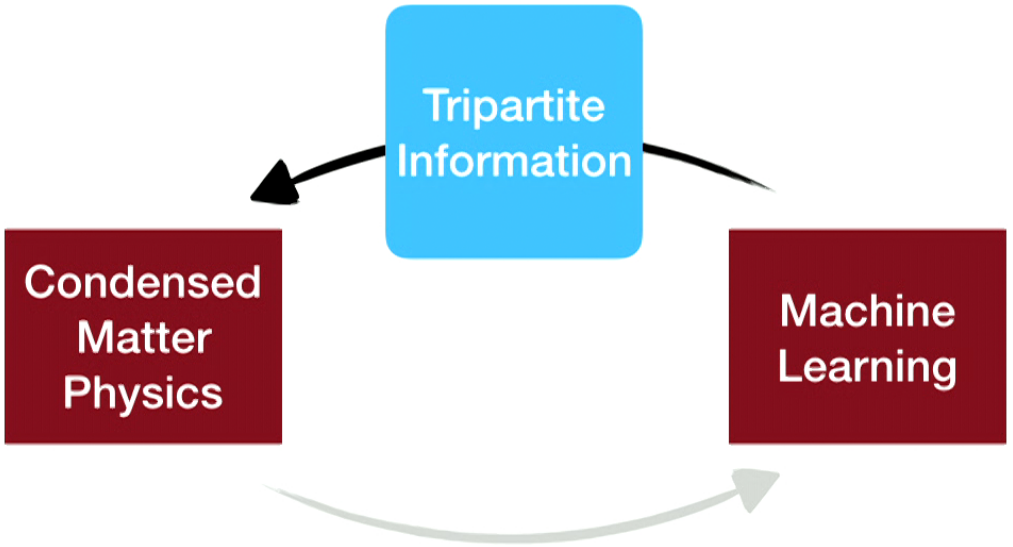
Just after midnight after announcing two weeks ago that it would be open to the public, Grembock's became the first Austin restaurant to ride a ride-sharing service in Austin since the "Bike-Share — Share the Ride" controversy erupted. Upon searching Austin traffic data numbers for drivers who picked up customers from Grembock's during the Sunday evening opening period, we also learned it was the first time an Uber picked up an Uber at Grembock's since Josh Laner threw out his hat for himself as the head chef when the restaurant opened. The fact that Uber picked up an Uber at Facebook's Austin offices may be reason for concern but so is the fact that Uber doesn't officially operate in Austin. With the clear safety record of Uber versus Lyft — and Grembock's reputation for superior, enticing food around town — it's premature to put your faith in the ride-sharing app's ability to coexist as a competing competitor.

https://openai.com/blog/better-language-models/

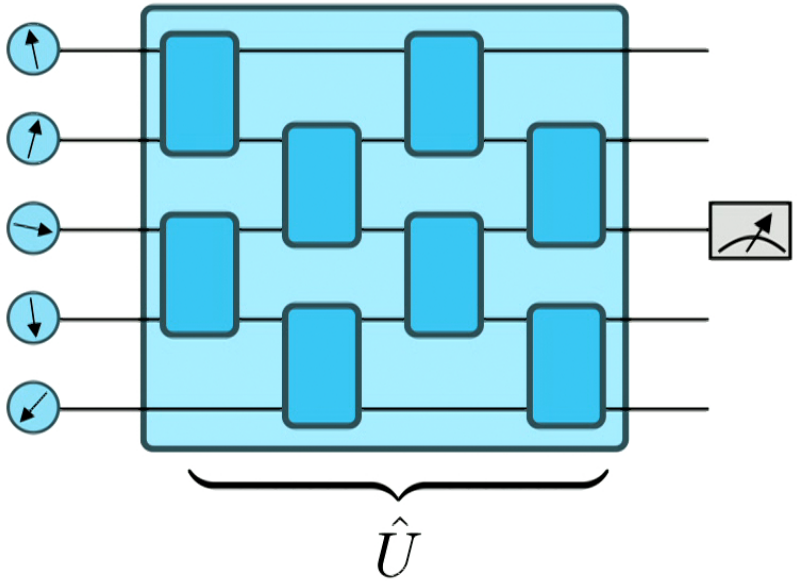
Summary

Machine Learning Sequence Model	Interaction Strength in Statistical Physics Model	Mutual Information Scaling
n-gram, HMM	Finite Range	Exponential
RNN	Exponential	Exponential
Self-attention (Transformer)	Exponential/Power-law	Exponential/Power-law

Physics & Machine Learning



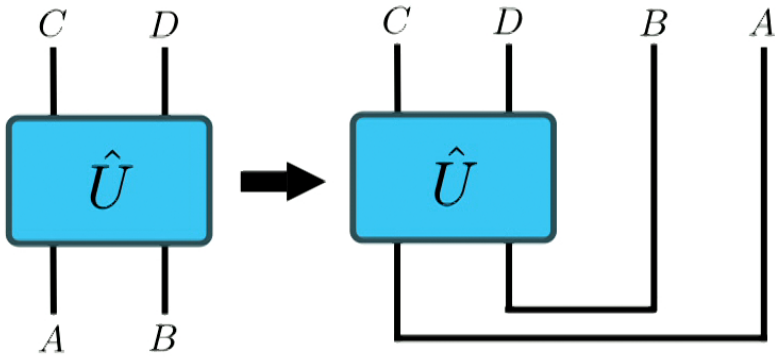
Quantum Neural Network



Unitary transformation: Information perfectly preserved

Mutual information between input and output always maximal

Tripartite Information



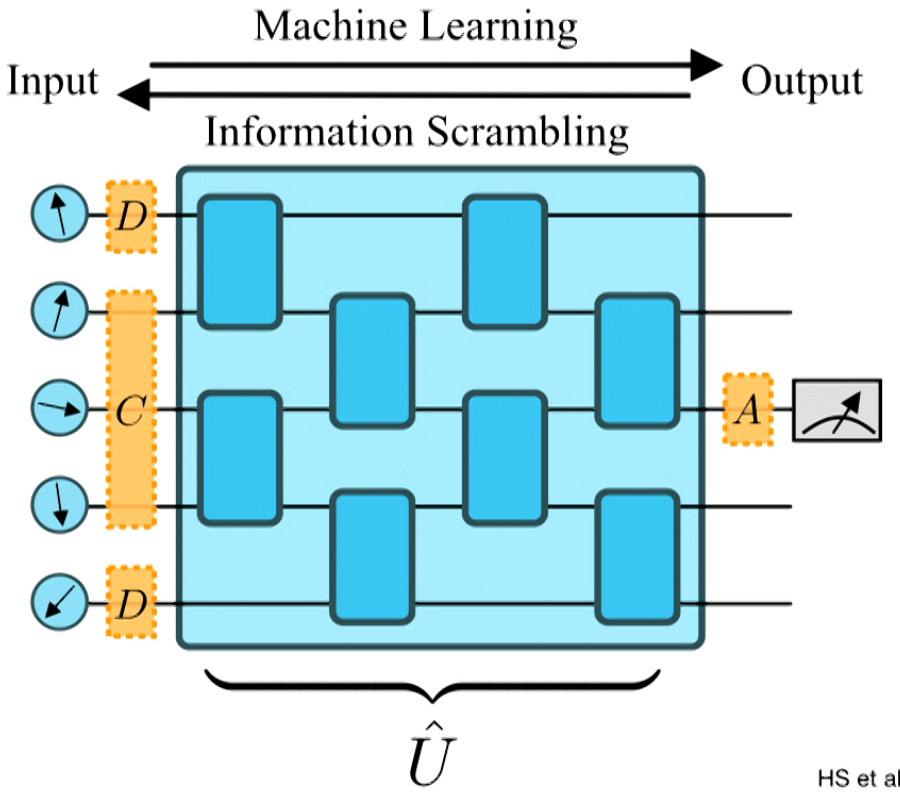
$$I_3(A, C, D) \equiv I(A, C) + I(A, D) - I(A, C \cup D)$$

A. Kitaev & J. Preskill, *PRL* 96, 110404 (2006)
P. Hosur, et al., *JHEP* 2016, 4 (2016)

Example:

- Identity unitary $I_3(A, C, D) = 0$
- Uniform Haar random unitary $I_3(A, C, D) = -2|A|$

Information Scrambling



HS et al., arXiv: 1909.11887

Magnetization Learning

- Dataset: Total magnetization of the ground states of random Ising model

$$\{(|G^{\alpha}\rangle, M_z^{\alpha}), \alpha = 1, \ldots, N\}$$

$$\hat{H} = \sum_{i,j=1}^{n} (J_{ij}\sigma_i^z\sigma_j^z + K_{ij}\sigma_i^x\sigma_j^x) + \sum_{i=1}^{n} (g_i\sigma_i^x + h\sigma_i^z) \qquad J_{ij}, K_{ij}, g, h \text{ random}$$

$$M_z^{\alpha} \equiv \langle G^{\alpha}| \sum_{i=1}^{n} \frac{\sigma_i^z}{n} |G^{\alpha}\rangle$$

- Loss function: Absolute loss

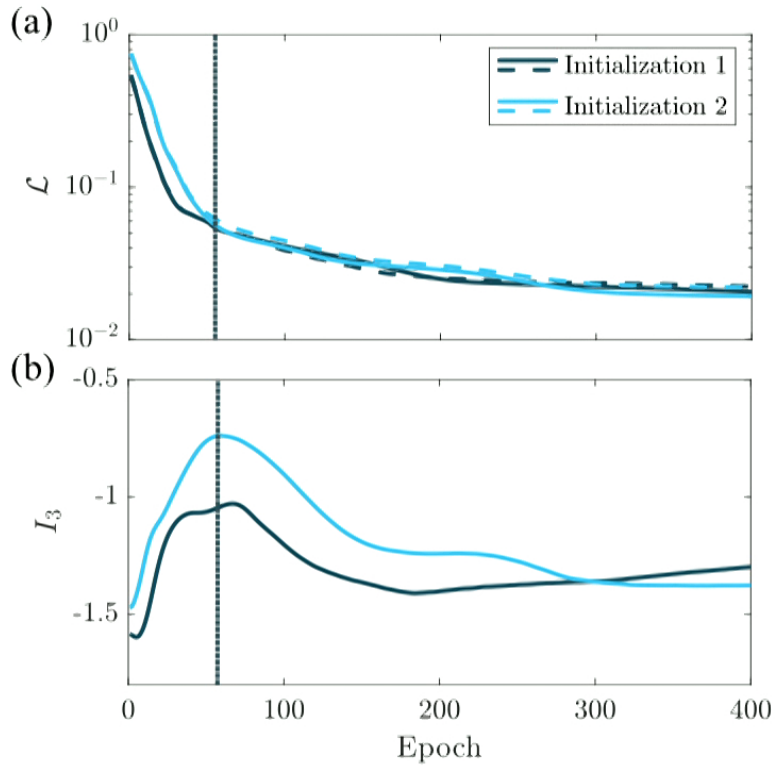
$$\mathscr{L} = \frac{1}{N} \sum_{\alpha=1}^{N} \left| \langle G^{\alpha}| \hat{U}^{\dagger}\sigma_{(n+1)/2}^x \hat{U} |G^{\alpha}\rangle - M_z^{\alpha} \right|$$

- Optimization algorithm: AMSGrad S.J. Reddi, et al., ICLR 2018

Two-Stage Learning

Loss function

Tripartite information

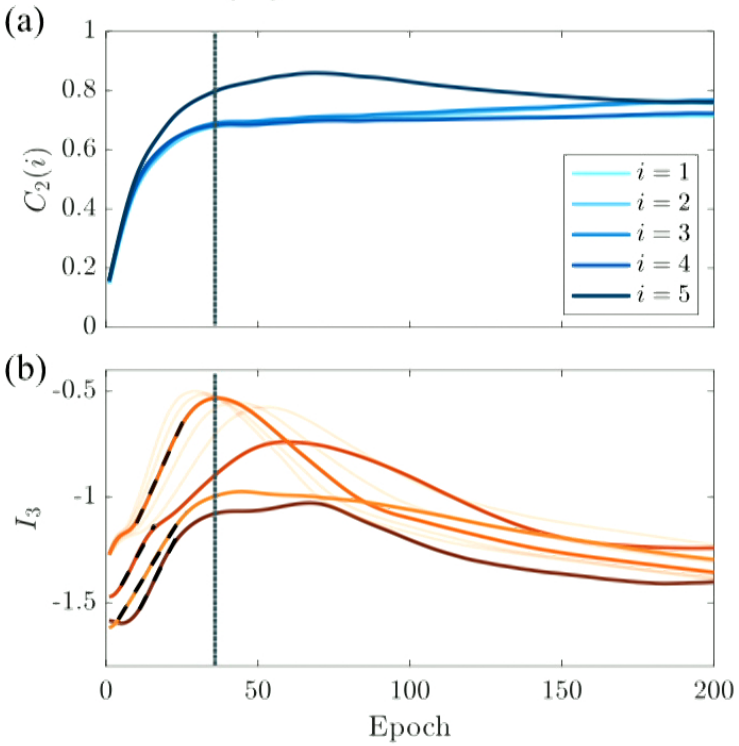


9 qubits, 6 layers

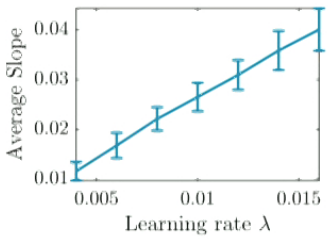
HS et al., arXiv: 1909.11887

Local Construction Stage

$$C_2(i) \equiv \frac{1}{N}\sum_{\alpha=1}^{N} \langle G^\alpha \left| \sigma_i^z \hat{U}^\dagger \sigma^x_{(n+1)/2} \hat{U} \right| G^\alpha \rangle$$

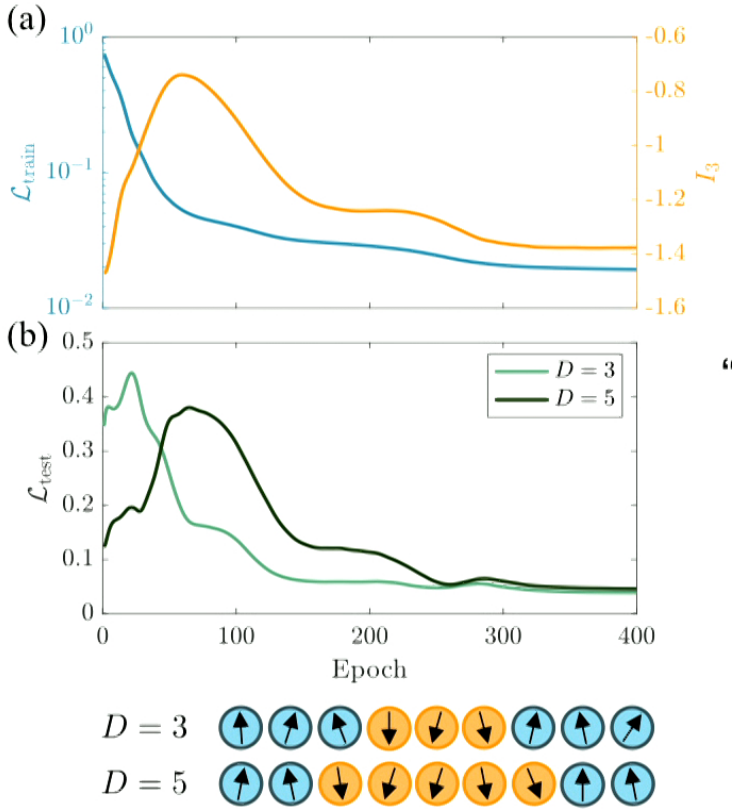


- Two-point correlation function is featureless after first stage, while I_3 is still changing rapidly
- Linear growth of I_3 at early stage



HS et al., arXiv: 1909.11887

Global Relaxation Stage



"Anti-ferromagnetic" test set

$$|\psi_D^\alpha\rangle = \prod_{\frac{n-D+1}{2} \le i \le \frac{n+D}{2}} \sigma_i^x |G^\alpha\rangle$$

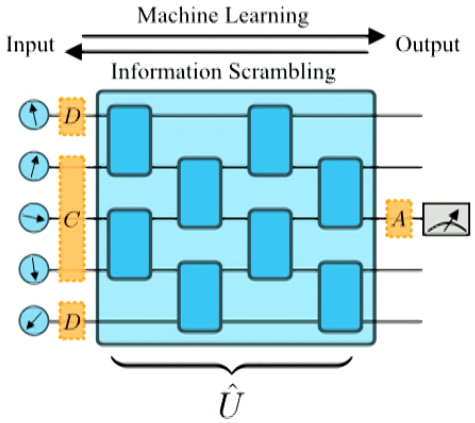
Domain size D

HS et al., arXiv: 1909.11887

Summary

Quantum Machine Learning	Many-body Quantum Chaos
1.Early stage: Build local correlation	Dissipation time
2.Late Stage: Build global structure	Scrambling time

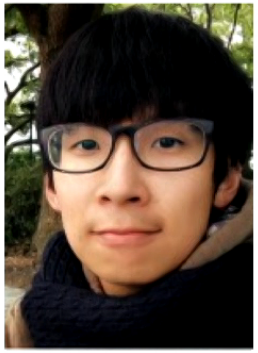
Also observed in other learning tasks (winding number learning)



HS et al., arXiv: 1909.11887

Acknowledgement

Pengfei Zhang
IASTU, Caltech



Yi-Zhuang You
UCSD

Hui Zhai
IASTU