Input Switched Affine Recurrent Networks:An RNN Architecture Designed for Interpretability

Jakob N. Foerster* [1], Justin Gilmer*[1], Jascha Sohl-Dickstein[1],Jan Chorowski[1], David Sussillo [1]

[1]Google Brain

ICML,2017

Presenter: Arshdeep Sekhon

Motivation

1. Interpreting Neural Networks
2. Crucial in many applications: self driving cars, medical diagnosis, power grid control, etc.

Related Work

1. Post Hoc Analysis: After training a network, try and analyze it.

Related Work

1. Post Hoc Analysis: After training a network, try and analyze it.

2. Design interpretability into the architecture

Related Work

1. Post Hoc Analysis: After training a network, try and analyze it.
   For example, break down LSTM model errors into classes
   + High Accuracy
   – Hard to interpret
2. Design interpretability into the architecture

Related Work

1. Post Hoc Analysis: After training a network, try and analyze it.
   For example, break down LSTM model errors into classes
   \+ High Accuracy
   – Hard to interpret
2. Design interpretability into the architecture
   For example, decision trees, logistic regression, etc.
   \+ Better understanding
   – accuracy suffers

Input Switched Affine Networks: ISAN

Vanilla RNN

$$\boldsymbol{h}_{t+1} = \sigma(\boldsymbol{U}\boldsymbol{x_t} + \boldsymbol{W}\boldsymbol{h_t} + \boldsymbol{b}) \tag{1}$$

$$\boldsymbol{l}_t = \sigma(\boldsymbol{W}_{ro}\boldsymbol{h_t} + \boldsymbol{b}_{ro}) \tag{2}$$

Input Switched Affine Networks: ISAN

Vanilla RNN

$$h_{t+1} = \sigma(\boldsymbol{U x_t} + \boldsymbol{W h_t} + \boldsymbol{b}) \tag{1}$$

$$\boldsymbol{l}_t = \sigma(\boldsymbol{W}_{ro}\boldsymbol{h_t} + \boldsymbol{b}_{ro}) \tag{2}$$

ISAN

$$\boldsymbol{h}_t = \boldsymbol{W}_{x_t}\boldsymbol{h_{t-1}} + \boldsymbol{b}_{x_t} \tag{3}$$

$$\boldsymbol{l}_t = \boldsymbol{W}_{ro}\boldsymbol{h_t} + \boldsymbol{b}_{ro} \tag{4}$$

ISAN: Accuracy Comparison

Parameter count	8e4	3.2e5	1.28e6
RNN	1.88	1.69	1.59
IRNN	1.89	1.71	1.58
GRU	1.83	1.66	1.59
LSTM	1.85	1.68	1.59
ISAN	1.92	1.71	1.58

Figure: *

ISAN performs as well as other recurrent architectures

ISAN: Accuracy Comparison

Parameter count	8e4	3.2e5	1.28e6
RNN	1.88	1.69	1.59
IRNN	1.89	1.71	1.58
GRU	1.83	1.66	1.59
LSTM	1.85	1.68	1.59
ISAN	1.92	1.71	1.58

Figure: *

ISAN performs as well as other recurrent architectures

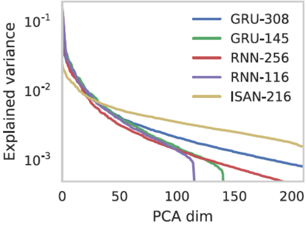


Figure: *

ISAN

ISAN

$$\boldsymbol{h}_t = \boldsymbol{W}_{x_t}\boldsymbol{h_{t-1}} + \boldsymbol{b}_{x_t} \tag{5}$$

$$\boldsymbol{l}_t = \boldsymbol{W}_{ro}\boldsymbol{h_t} + \boldsymbol{b}_{ro} \tag{6}$$

ISAN

$$\boldsymbol{h}_t = \sum_{s=0}^{t}\Big(\prod_{s'=s+1}^{t}\boldsymbol{W}_{x'_s}\Big)\boldsymbol{b}_{x_s} \tag{7}$$

ISAN

ISAN

$$\boldsymbol{\kappa}_s^t = \boldsymbol{W}_{ro}\Big(\prod_{s'=s+1}^{t} \boldsymbol{W}_{x_s'}\Big)\boldsymbol{b}_{x_s} \tag{8}$$

$$\boldsymbol{l}_t = \boldsymbol{b}_{ro} + \sum_{s=0}^{t} \boldsymbol{\kappa}_s^t \tag{9}$$

Linearity of κ

Consider string:"_annual_revenue"
How does "_annual" affect output after "_rev"?

$$\boldsymbol{l}_t = \boldsymbol{b}_{ro} + \sum_{s=0}^{t'} \boldsymbol{\kappa}_s^t + \sum_{s=t'}^{t} \boldsymbol{\kappa}_s^t \tag{10}$$

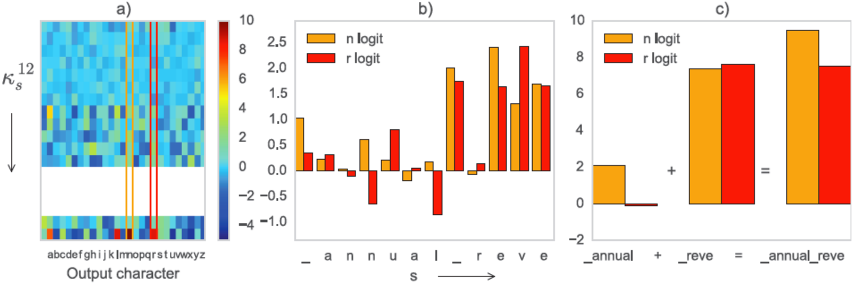


Figure: *

ISAN: information timescales of network

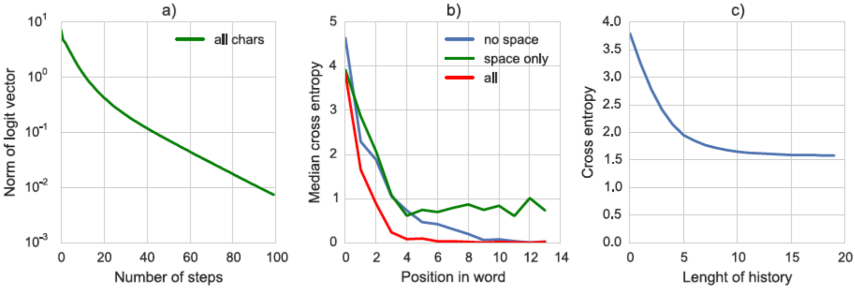


Figure: *

A κ_s^t averaged for all characters as a function of t-s

B Importance of "_" character in decoding

C Cross entropy as a function of number of characters considered for prediction we can aggregate all of the κ_s^t belonging to a given word and visualize them as a single contribution to the prediction of the letters in the next word

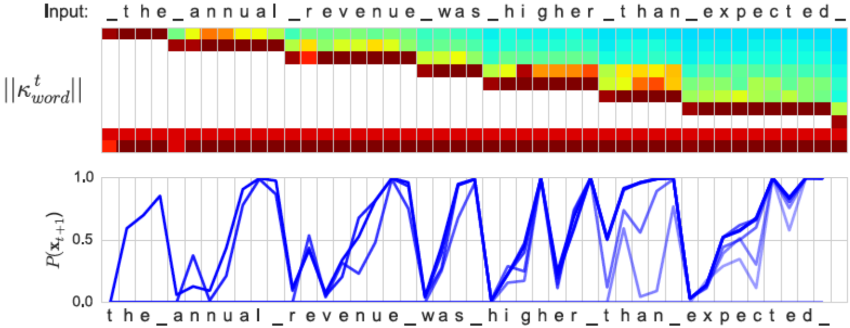


Figure: *

Change of Basis

1. Divide the hidden space into a subspace $\boldsymbol{P}_{\parallel}^{ro}$ spanned by the rows of the readout matrix $\boldsymbol{W}_{ro}$ and its orthogonal complement $\boldsymbol{P}_{\perp}^{ro}$
2. Thus, 27 dimensions for readout and (216-27) for computational subspace.

Change of basis

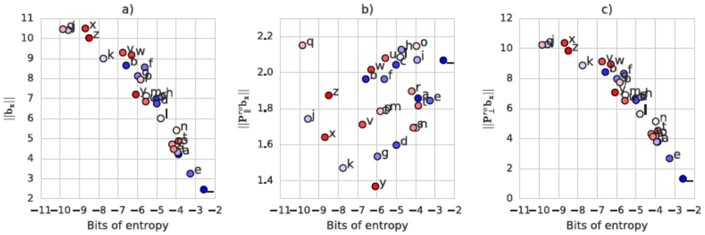


Figure: *

Information content related to the computation subspace.

A the norm of the learnt b_x is strongly correlated to the log-probability of the unigram x in the training data.

Change of basis

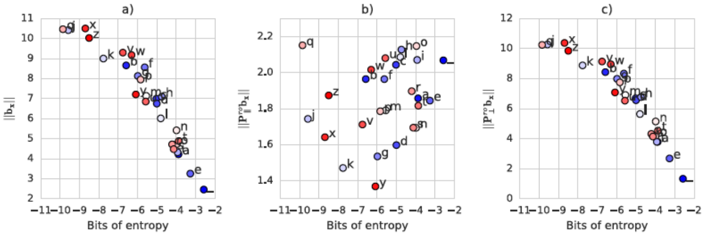


Figure: *

Information content related to the computation subspace.

A the norm of the learnt b_x is strongly correlated to the log-probability of the unigram x in the training data.

B this correlation is not related to reading out the next-step prediction

Change of basis

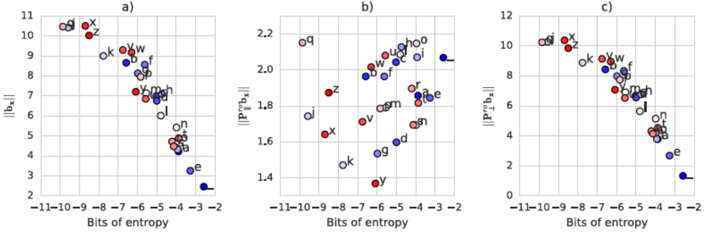


Figure: *

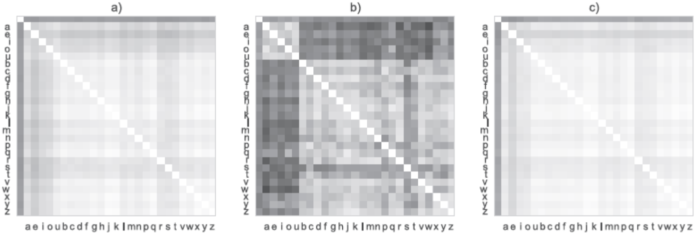
Information content related to the computation subspace.

A the norm of the learnt b_x is strongly correlated to the log-probability of the unigram x in the training data.

B this correlation is not related to reading out the next-step prediction

C This implies a connection between information or surprise and distance in the computational subspace of state space.

Change of basis



A Cosine distance/ correlation in original space

B Cosine distance/ correlation in readout space or $\boldsymbol{P}_{\parallel}^{ro}$.two blocks of high correlations between the vowels and consonants respectively, while b_{-} is uncorrelated to either

C Cosine distance/ correlation in readout space or $\boldsymbol{P}_{\perp}$

Parantheses Counting Task

1. The Task: Count the number of opened parens $[, ($
2. Input: One hot encoded vector
3. Target Output: nesting level at previous timestep
4. output: two-hot encoded 0-5 count (12 dimensional 2-hot encoded vector)

Paranthesis Counting

Using an augmented matrix and an augmented vector, it is possible to represent both the translation and the linear map using a single matrix multiplication:
ISAN:

$$\boldsymbol{h}_{t+1} = \boldsymbol{W}\boldsymbol{h_t} + \boldsymbol{b} \tag{11}$$

$$\boldsymbol{h}^{'}_{t+1} = \boldsymbol{W}^{'}\boldsymbol{h_t}^{'} \tag{12}$$

Paranthesis Counting: Change of Bases

1. Divide the hidden space into a subspace $\boldsymbol{P}_{\parallel}^{ro}$ and its orthogonal complement $\boldsymbol{P}_{\perp}^{ro}$
2. Learn bases by linear regression to encourage augmented matrices and hidden states to be sparse

Paranthesis Counting: Change of Bases

$$\mathbf{W}'_x = \begin{bmatrix} \mathbf{W}_x^{rr} & \mathbf{W}_x^{rc} & \mathbf{b}_x^r \\ \mathbf{W}_x^{cr} & \mathbf{W}_x^{cc} & \mathbf{b}_x^c \\ \mathbf{0}^T & \mathbf{0}^T & 1 \end{bmatrix} \quad \mathbf{h}'_t = \begin{bmatrix} \mathbf{h}_t^r \\ \mathbf{h}_t^c \\ 1 \end{bmatrix}$$

and the update equation can be written as

$$\mathbf{h}'_{t+1} = \mathbf{W}'_x \mathbf{h}'_t = \begin{bmatrix} \mathbf{W}_x^{rr}\mathbf{h}_t^r + \mathbf{W}_x^{rc}\mathbf{h}_t^c + \mathbf{b}_x^r \\ \mathbf{W}_x^{cr}\mathbf{h}_t^r + \mathbf{W}_x^{cc}\mathbf{h}_t^c + \mathbf{b}_x^c \\ 1 \end{bmatrix}.$$

Figure: Equations after subspace decomposition

Paranthesis Counting: Interpretation

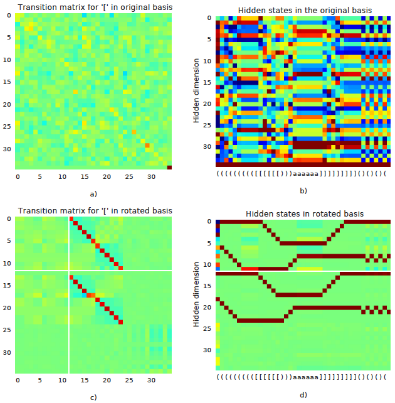


Figure: Dynamics of ISAN for '['

1. leftmost 12 columns $\boldsymbol{W}^{rr}_{[}$ $\boldsymbol{W}^{cr}_{[}$ are zero
2. h^r_t has no influence on $\boldsymbol{h}_{t+1}$

Paranthesis Counting: Interpretation

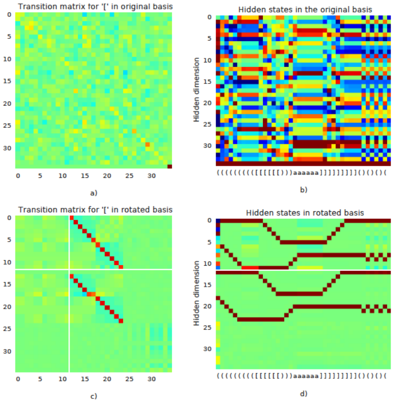


Figure: Dynamics of ISAN for '['

1. $\boldsymbol{W}^{rc}_{[}$ is identity; $h^r_t = h^c_{t-1}$

Paranthesis Counting: Interpretation

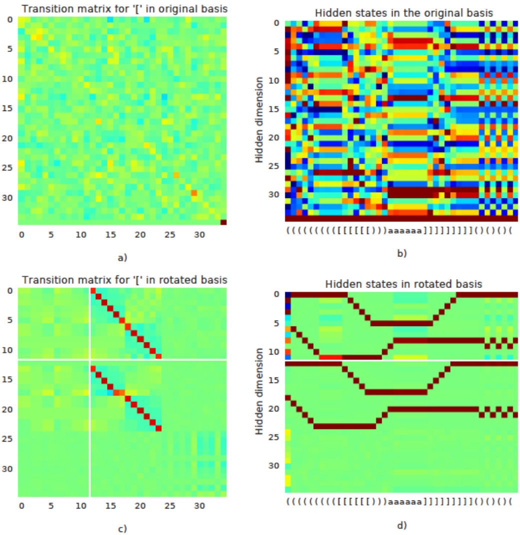


Figure: Dynamics of ISAN for '[': Delay Line Dynamics