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.

   + High Accuracy

   – Hard to interpret

   For example, break down LSTM model errors into classes

2. Design interpretability into the architecture

   + Better understanding

   – accuracy suffers

   For example, decision trees, logistic regression, etc.

Input Switched Affine Networks: ISAN

Vanilla RNN

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

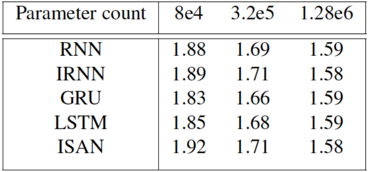
$$\boldsymbol{l}_t = \sigma(\boldsymbol{W}_{ro}\boldsymbol{h_t} + \boldsymbol{b}_{ro}) \tag{2}$$

ISAN

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

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

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: Accuracy

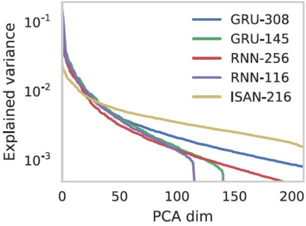


Figure: ISAN makes fuller and more uniform use of its hidden state

ISAN

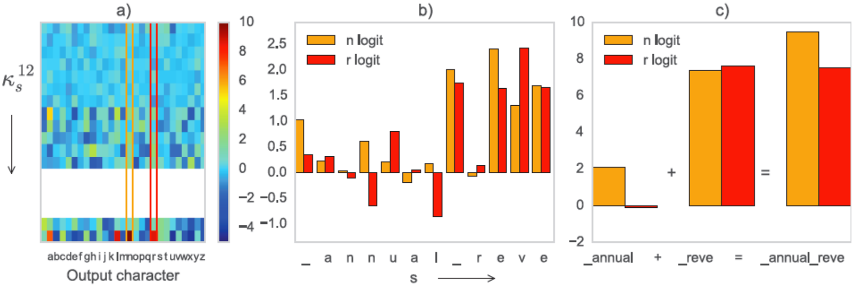
$$\boldsymbol{h}_{t+1} = \boldsymbol{W}_{x_t} \boldsymbol{h_t} + \boldsymbol{b}_{x_t} \tag{5}$$

$$\mathrm{h}_t = \sum_{s=0}^{t} \left(\prod_{s'=s+1}^{t} \mathrm{W}_{\mathbf{x}_{s'}} \right) \mathrm{b}_{\mathbf{x}_s} ,$$

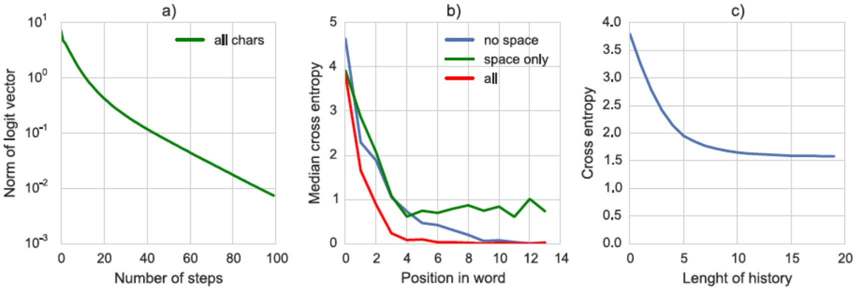
$$\mathrm{l}_t = \mathrm{b}_{ro} + \sum_{s=0}^{t} \boldsymbol{\kappa}_s^t$$

$$\boldsymbol{\kappa}_s^t = \mathrm{W}_{ro} \left(\prod_{s'=s+1}^{t} \mathrm{W}_{\mathbf{x}_{s'}} \right) \mathrm{b}_{\mathbf{x}_s} ,$$

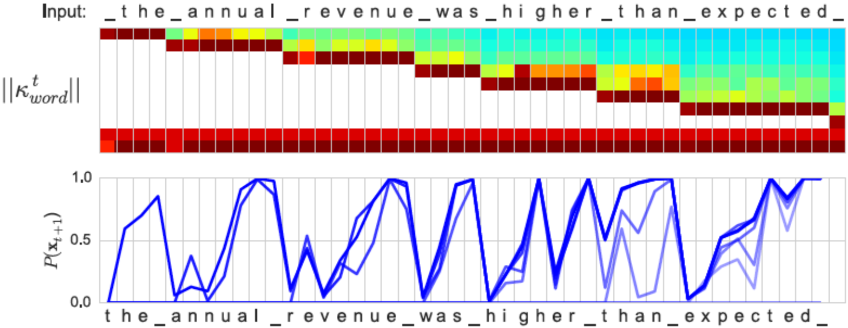
Linearity of κ



ISAN: information timescales of network



Characters to Words



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

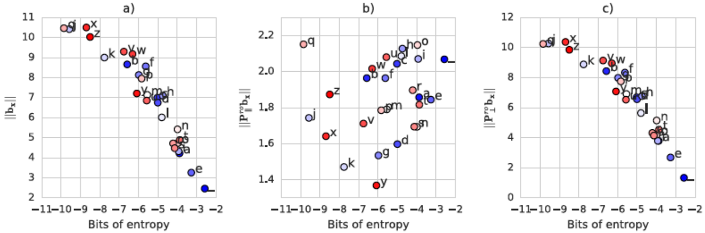


Figure: Information content related to the computation subspace.

Change of basis

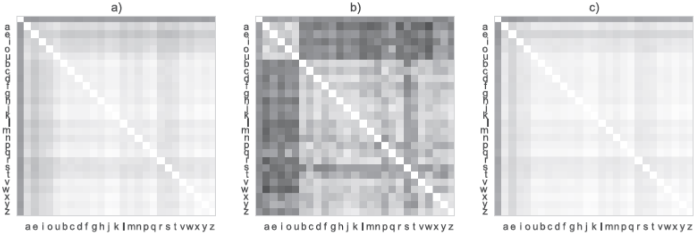


Figure: Correlation in $\boldsymbol{b}_x$.High correlation between vowels and consonants explained by $\boldsymbol{P}_{\|}^{ro}$

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:

$$\boldsymbol{h}_{t+1} = \boldsymbol{W}\boldsymbol{h_t} + \boldsymbol{b} \tag{6}$$

$$\boldsymbol{W}' = \begin{bmatrix} \boldsymbol{W} & \boldsymbol{b} \\ \boldsymbol{0}^T & 1 \end{bmatrix}$$

$$\boldsymbol{h}_{\boldsymbol{t}}' = \begin{bmatrix} \boldsymbol{h_t} \\ \boldsymbol{1} \end{bmatrix}$$

$$\boldsymbol{h}'_{t+1} = \boldsymbol{W}'\boldsymbol{h_t}' \tag{7}$$

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

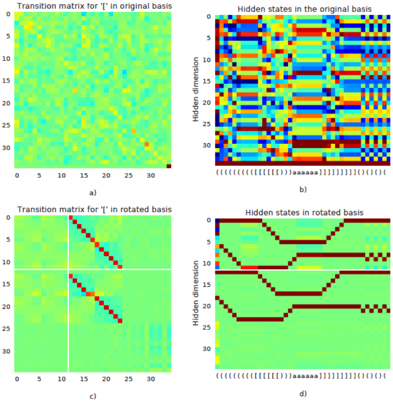
$$\mathbf{W}'_x = \begin{bmatrix} \mathbf{W}_x^{rr} & \mathbf{W}_x^{rc} & \mathbf{b}_x^{r} \\ \mathbf{W}_x^{cr} & \mathbf{W}_x^{cc} & \mathbf{b}_x^{c} \\ \mathbf{0}^T & \mathbf{0}^T & 1 \end{bmatrix} \quad \mathbf{h}'_t = \begin{bmatrix} \mathbf{h}_t^{r} \\ \mathbf{h}_t^{c} \\ 1 \end{bmatrix}$$

and the update equation can be written as

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

Equations after subspace decomposition

Paranthesis Counting: Interpretation



a)

b)

c)

d)

Figure: Dynamics of ISAN for '['

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

Paranthesis Counting: Interpretation

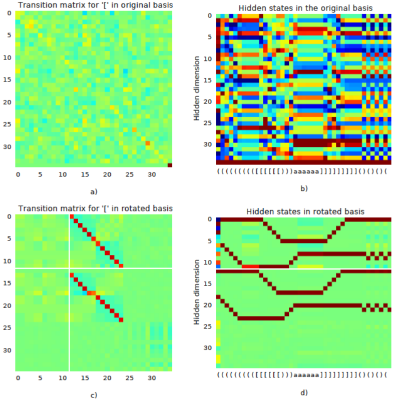
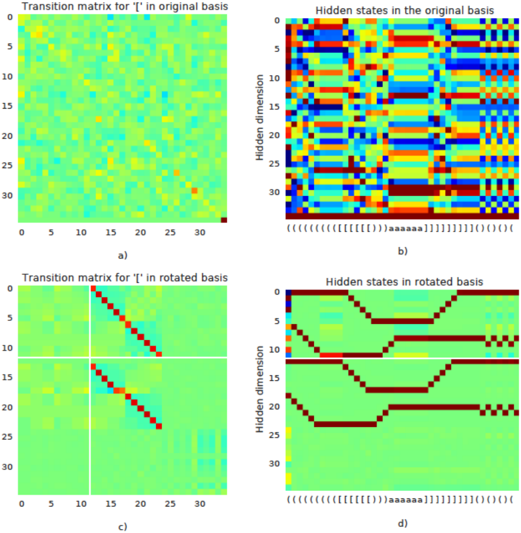


Figure: Dynamics of ISAN for '['

1. $\boldsymbol{W}_{[}^{rc}$ is identity; $h_t^r = h_{t-1}^c$

Paranthesis Counting: Interpretation



a) b) c) d)

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