Recurrent Neural Networks as Cognitive Models

Master’s Thesis Project

- Rishubh Singh
- Prof. Sumeet Agarwal

Motivation

- Model **neural networks** can **reproduce** important features of **neural activity** recorded in cortical areas of animals.

- The analysis of such circuits, whose activity and connectivity are fully known, has therefore re-emerged as a promising tool for understanding neural computation.

- Constraining network training with tasks for which detailed neural recordings are available may also provide insights into the **principles** that govern learning in **biological circuits**.

- Early works on Deep Learning : Hebbian Learning, the Perceptron were biologically inspired.

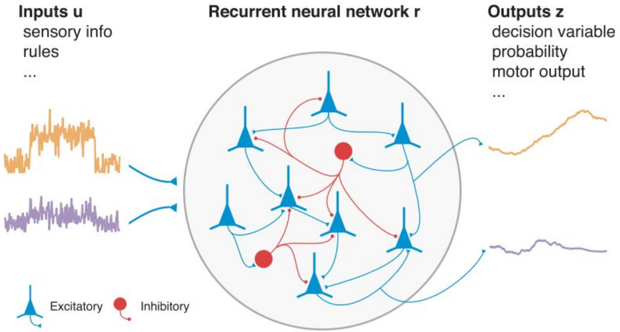
- Research moved away from biological inspiration to engineering and main goal was efficiency.

- Existing RNNs lack basic biological features.

- Trained networks achieve same behavioral performance but differ greatly in structure, dynamics and learning (task design).

Training Excitatory-Inhibitory RNNs for Cognitive Tasks

H. Francis Song
Guangyu R. Yang
Xiao-Jing Wang



Training Excitatory Inhibitory RNNs for Cognitive Tasks: Song, Yang, Wang, PLOS 2016

- Perpetual decision making
- Context dependent integration
- Parametric working memory
- Eye-movement sequence execution

All of these are low level tasks that are trained and tested at the level of neural data.

RNN describing equations :

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + W^{\mathrm{rec}} \mathbf{r} + W^{\mathrm{in}} \mathbf{u} + \sqrt{2\tau\sigma_{\mathrm{rec}}^2}\ \boldsymbol{\xi},$$

$$\mathbf{r} = [\mathbf{x}]_+,$$

$$\mathbf{z} = W^{\mathrm{out}} \mathbf{r},$$

Dale’s principle :

$$\underbrace{\begin{pmatrix} & + & + & + & - \\ + & & + & + & - \\ + & + & & + & - \\ + & + & + & & - \\ + & + & + & + & \end{pmatrix}}_{W^{\mathrm{rec}}} = \underbrace{\begin{pmatrix} & + & + & + & + \\ + & & + & + & + \\ + & + & & + & + \\ + & + & + & & + \\ + & + & + & + & \end{pmatrix}}_{W^{\mathrm{rec},+}} \underbrace{\begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & -1 \end{pmatrix}}_{D},$$

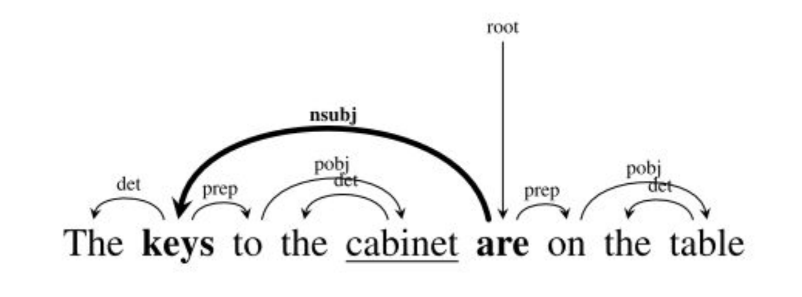
Pattern of connectivity :

$$W^{\mathrm{rec},+} = \underbrace{\begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}}_{M^{\mathrm{rec}}} \odot \underbrace{\begin{pmatrix} \cdot & + & + & + & \cdot \\ + & \cdot & + & \cdot & + \\ + & + & \cdot & + & + \\ \cdot & + & + & \cdot & \cdot \\ + & + & + & + & \cdot \end{pmatrix}}_{W^{\mathrm{rec,plastic},+}} + \underbrace{\begin{pmatrix} 0 & 0 & 0 & 0 & w_1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & w_2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}}_{W^{\mathrm{rec,fixed},+}},$$

Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies

Tal Linzen
Emmanuel Dupoux
Yoav Goldberg

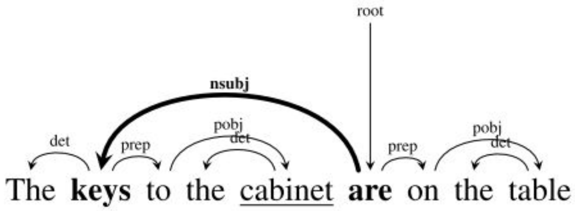
Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies: Linzen, ACL 2016



Syntax-Sensitive Dependencies

- The **pan is** on the stove.
- * The **pans is** on the stove.
- * The **pan are** on the stove.
- The **pans are** on the stove.

- The **pan** from the cupboard **is** on the stove.
- The **keys** to the cabinet **are** on the table.



The **decision** of female Gothic writers to supplement true supernatural horrors with explained cause and effect **transforms** romantic plots and Gothic tales into common life and writing.

Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies: Linzen, ACL 2016

Number Prediction Task

The model sees the sentence up to but not including a present-tense verb.
The model then needs to predict the number of the following verb.

Example : The pan from the cupboard ________

Concepts Needed :

- Syntactic number : words are singular and plural
- Syntactic subjecthood : identify the subject corresponding to the verb

Corpus Statistics

	With Attractors	**With Intervening Nouns**
N = 0	92.7%	80.8%
N = 1	5.7%	12.7%
N = 2	1.1%	4.2%
N = 3	0.3%	5.1%
N = 4	0.1%	2.1%

Corpus statistics of the number agreement test dataset (1.5M sentences) by Linzen et al

Inflection Task

The model sees the sentence up to and including the singular form of a present-tense verb.
The model then needs to predict the number of the following verb.

Example : The pans from the cupboard is

- Access to semantics of verb can help network identify corresponding verb :
  - **People** from the capital often **eat** pizza.

Grammaticality Task

Half of the examples in the corpus are made ungrammatical by flipping the number of the verb. The network reads the entire sentence and receives a supervision signal at the end.

- The **key** to the cabinet **is** on the stove.
- * The **keys** to the cabinet **is** on the stove.

Previous objectives :

- Explicitly indicate location of the sentence in which the verb can appear.
    - Gives the network a cue to syntactic clause boundaries.
- Explicitly direct the networks attention to the number of the verb.

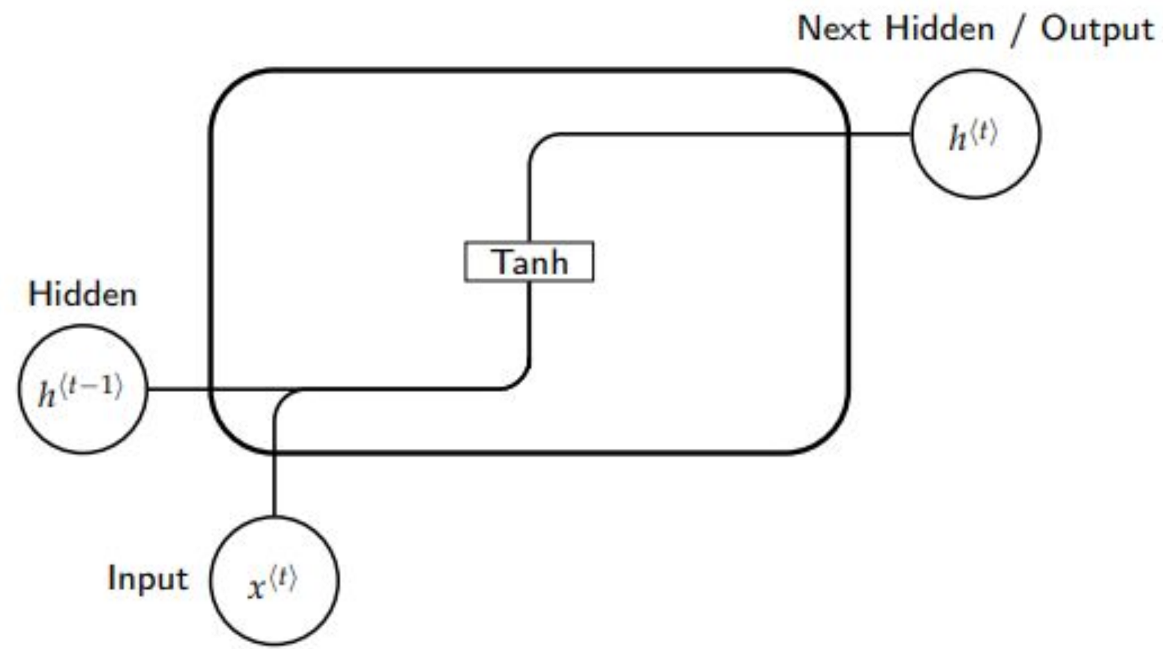
This task :

- Weaker supervision : complete sentence given to judge if it is grammatical.

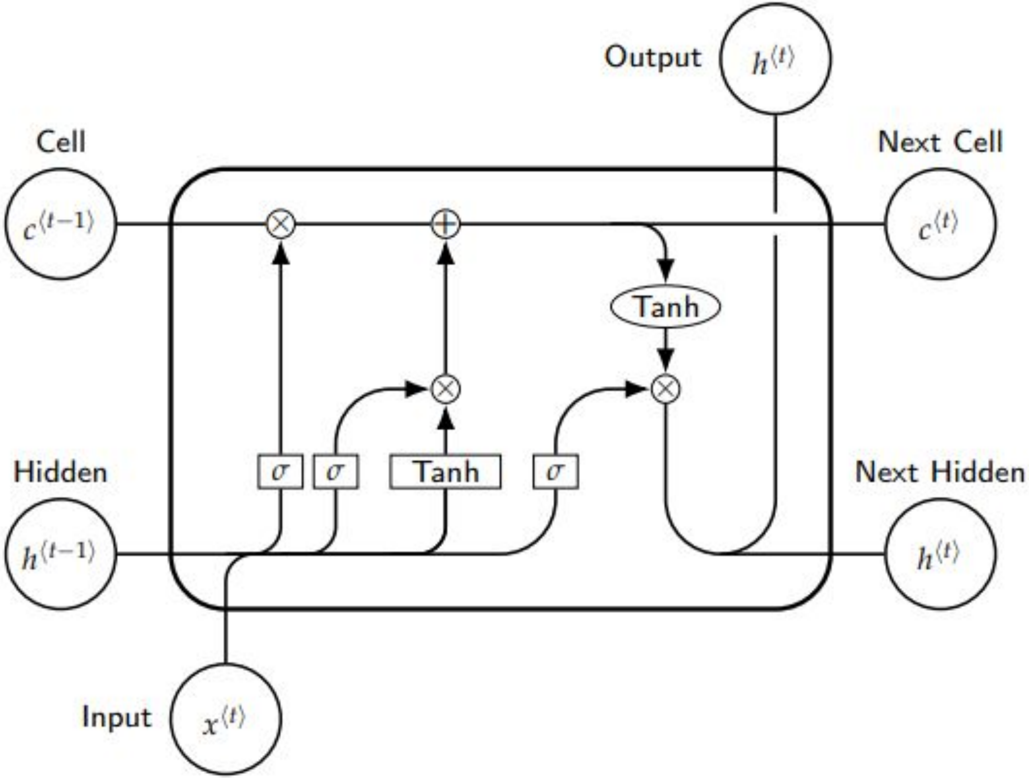
Generic Models

- Words are encoded as one-hot vectors, embedded in a 50-dimensional vector space.
- **Final state** of the network is fed into a logistic regression **classifier**.
  - We pass the final state to a fully connected layer to produce two outputs.
  - Final prediction : label with the higher output.
- The **input** sentence is pre-padded with zeros to a **fixed length of 50**.
- The model is then trained in an end-to-end fashion, including the word embeddings.

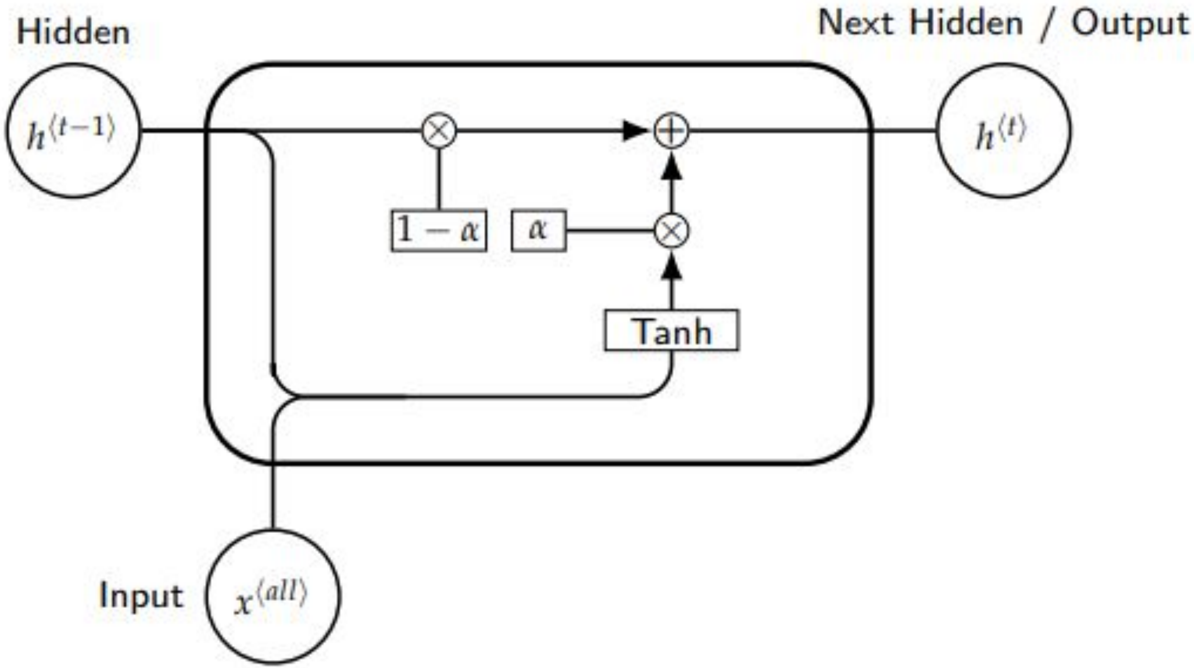
SRN / Standard RNN



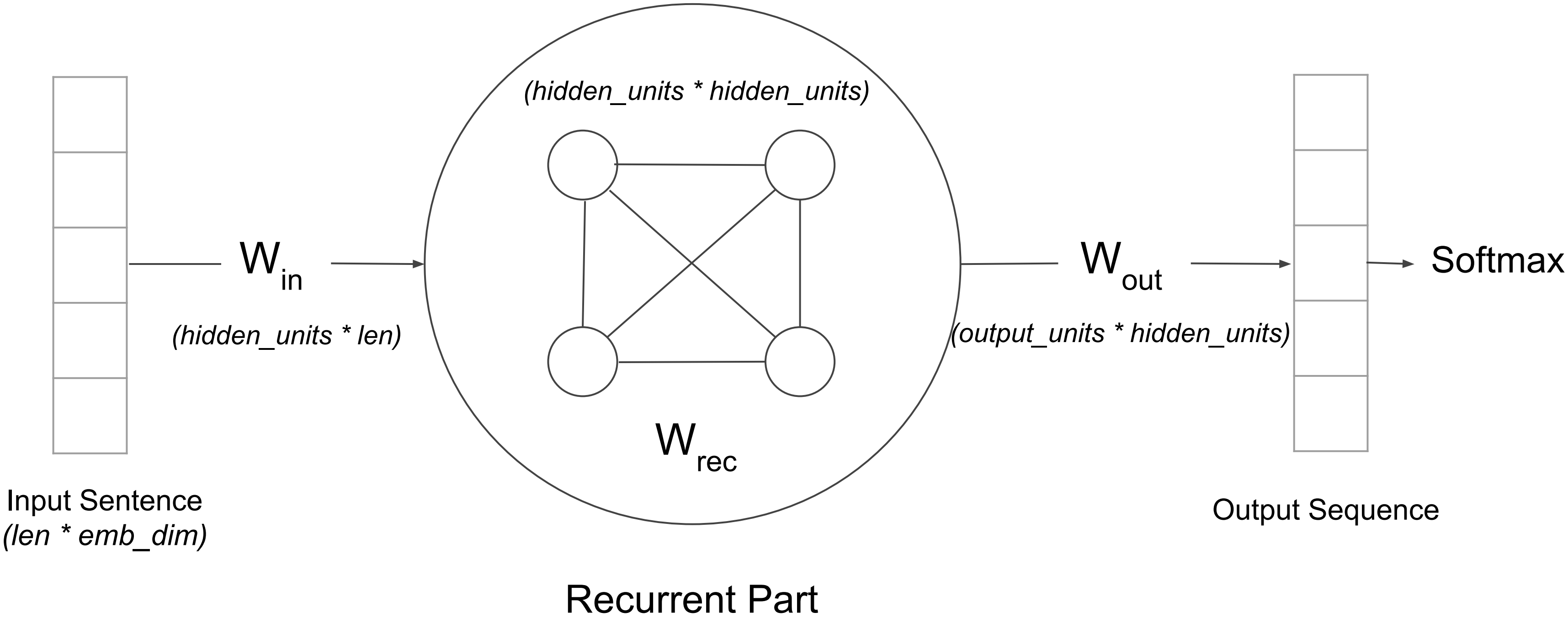
LSTM


Output
$h^{\langle t \rangle}$
Cell
$c^{\langle t-1 \rangle}$
Next Cell
$c^{\langle t \rangle}$
Tanh
σ
σ
Tanh
σ
Hidden
$h^{\langle t-1 \rangle}$
Next Hidden
$h^{\langle t \rangle}$
Input
$x^{\langle t \rangle}$

EIRNN


Hidden
$h^{\langle t-1 \rangle}$
Next Hidden / Output
$h^{\langle t \rangle}$
$1-\alpha$
α
Tanh
Input
$x^{\langle all \rangle}$

EIRNN



All results that follow are from the best performing model with parameters :

- *hidden_units = 15*
- *input_units = max sentence length = 50*
- *output_units = 10*

Performance

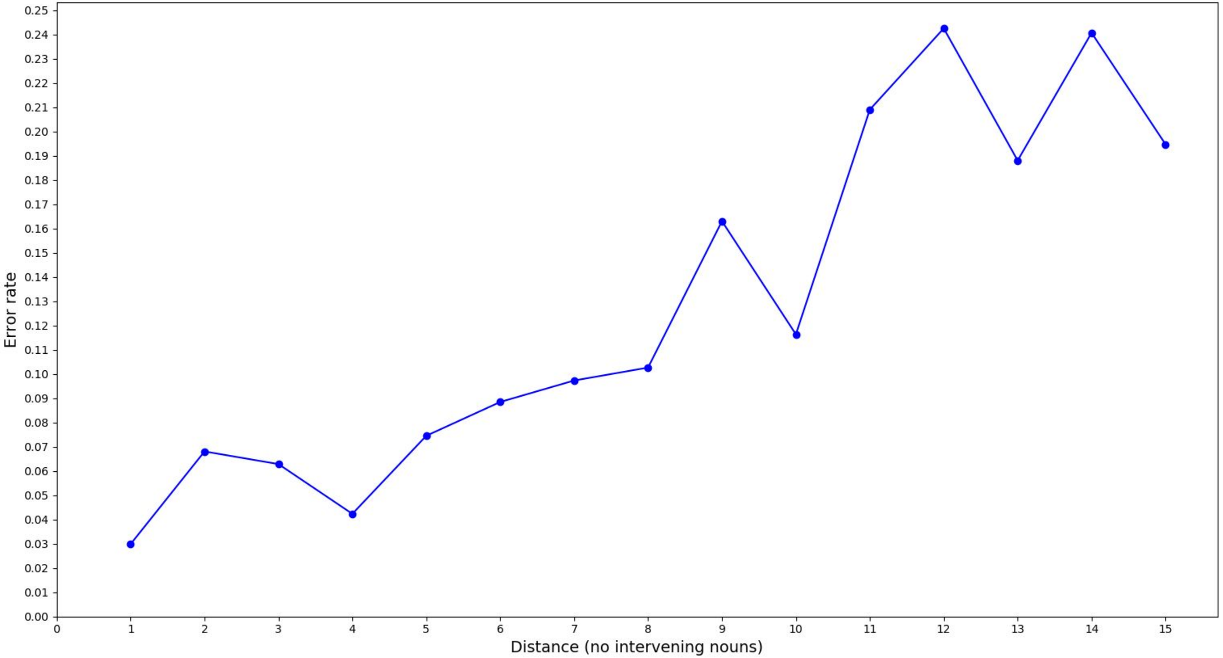
Model	Overall Accuracy (%)
EIRNN (H = 3)	92.8
EIRNN (H = 15)	94.1
EIRNN (H = 50)	93.5
RNN	97.7
RNN Dale	97.8
AbLSTM	98.0
LSTM	98.7

On Number Prediction

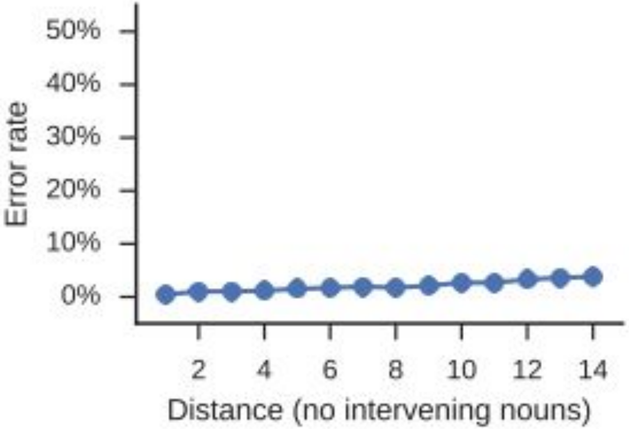
Model	Overall Accuracy (%)
EIRNN (H = 3)	92.5
EIRNN (H = 15)	94.2
EIRNN (H = 50)	93.3
RNN	97.9
RNN Dale	98.0
AbLSTM	98.1
LSTM	98.9

On Inflection

Error Rate : With increasing distance with no intervening nouns



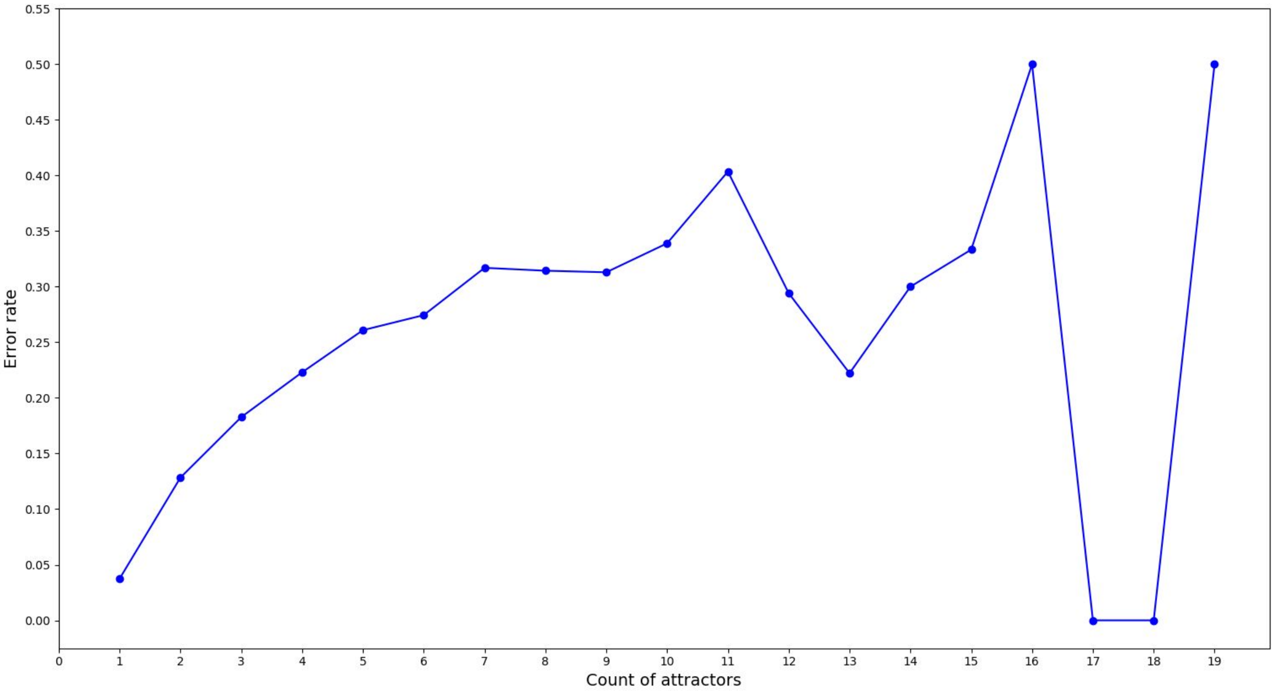
EIRNN

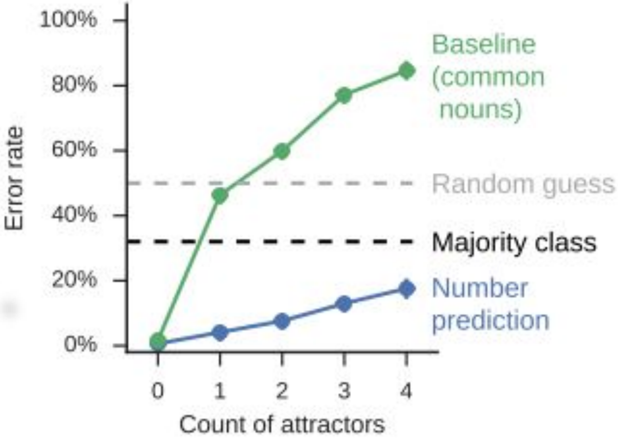


LSTM

Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies: Linzen, ACL 2016

Error Rate : With increasing number of attractors between noun and verb



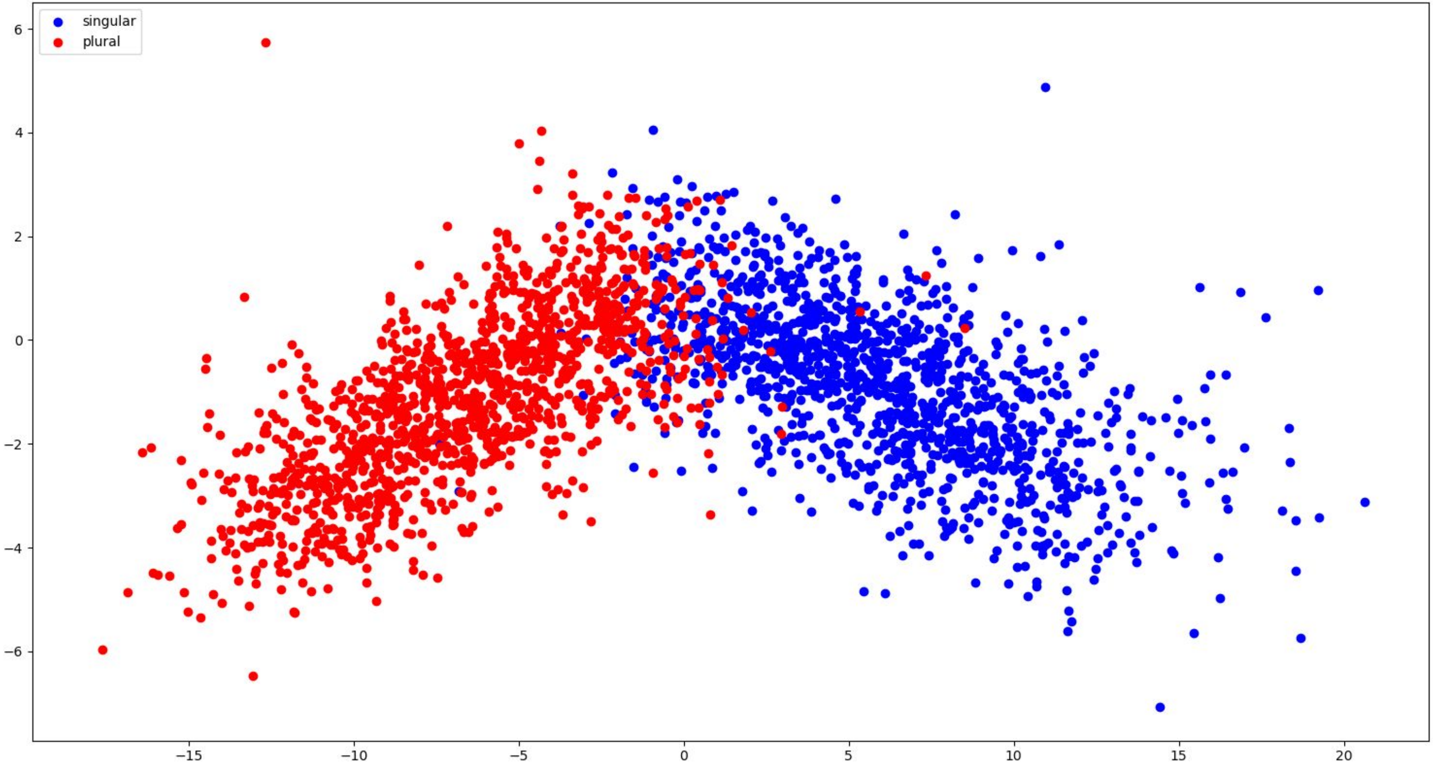


EIRNN

LSTM

Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies: Linzen, ACL 2016

EIRNN :Word Embeddings Projected on first two PCs



EIRNN vs LSTM

	No Last Intervening	**Singular Last Noun**	**Plural Last Noun**
Singular Subject	2.14 (0.31)	2.95 (0.48)	41.03 (3.93)
Plural Subject	7.15 (1.67)	44.43 (7.53)	5.74 (1.86)

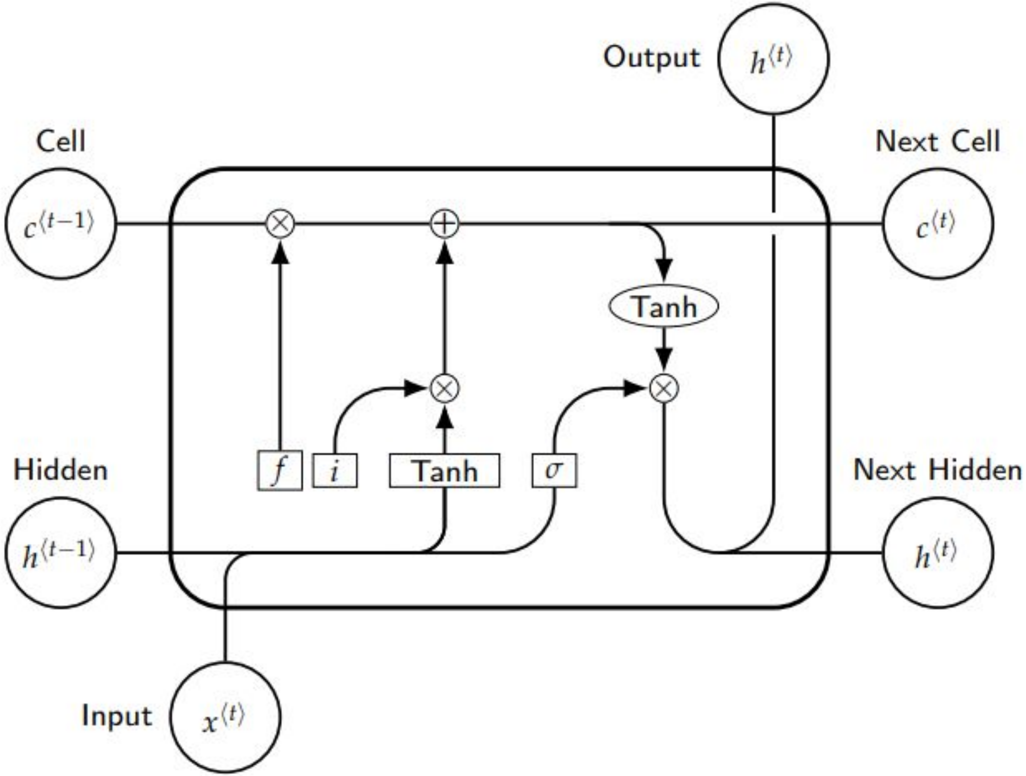
EIRNN (LSTM)

EIRNN not LSTM	LSTM not EIRNN
2272 unique ; 13941 total points	**23313 unique ; 84039 total points**
1. NN : 3368	1. NNS NN : 5758
2. NNS : 1470	2. NN NNS : 5619
3. NN NN : 1177	3. NN NN NNS : 2619
4. NN NNS : 432	4. NNS NN NN : 2092
5. NNP NN : 349	5. NNS : 1997
6. NN NN NN : 315	6. NN NNS NN : 1295
7. PRP$ NN : 298	7. NN NN NN NNS : 1053
8. NN NNP : 213	8. NN : 954
9. NNP NNS : 187	9. NN NNS NNS : 841
10. NNS NNS : 175	10. NNP NN NNS : 602
11. NNS NN : 168	11. NNS NNS NN : 597
12. PRP NN : 135	12. NNS NN NN NN : 594
13. PRP$ NNS : 125	13. NNS NNP NNP : 576
14. NNP NN NN : 123	14. NNS NNP NN : 496
15. NN NN NN NN : 119	15. NN NNP NNS : 486

Grammaticality Performance

Model	Accuracy (%age)
EIRNN (15)	~ 50
EIRNN (3)	~ 50
RNN	~ 50
RNN (Dale)	~ 50
LSTM (50)	4.5
LSTM (3)	~ 60
Ablated LSTM (scalar)	~ 50
Ablated LSTM (vector)	~ 50

Ablated LSTM : AbLSTM



Grammaticality on Half Sentences

- The **key** to the cabinet **is**
- * The **keys** to the cabinet **is**

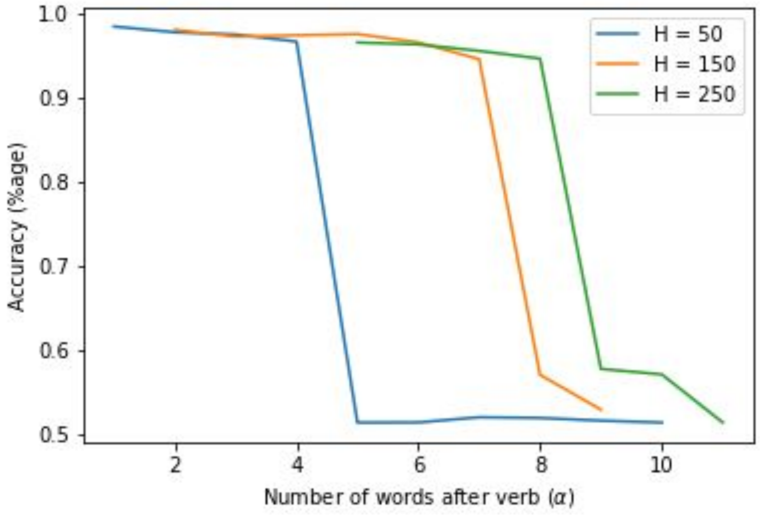
Model	**Accuracy (%age)**
EIRNN (3)	84.3
EIRNN (15)	88.3
EIRNN (50)	86.3
RNN (Dale)	97.8
AbLSTM	97.1
LSTM	98.3

Grammaticality Plus

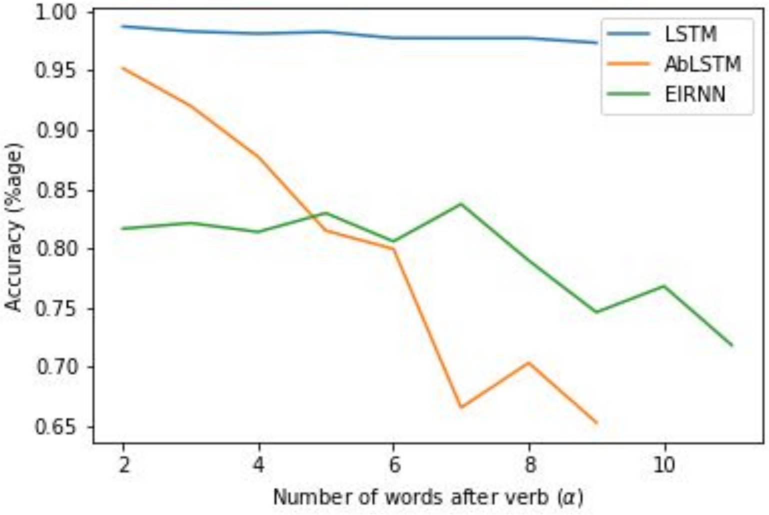
Variable α (alpha) specifies the maximum number of words in the input sentence after the verb that is in context in the sample. The set of tasks is collectively termed *Grammaticality Plus*.

- α = 0 : The **version** of the chronicle that the annalists were working was written in different places at different times; the earliest evidence for one of its authors **places**
- α = 3 : The **version** of the chronicle that the annalists were working was written in different places at different times; the earliest evidence for one of its authors **places** it in Iona
- α = 6 : * The **version** of the chronicle that the annalists were working was written in different places at different times; the earliest evidence for one of its authors **place** it in Iona sometime after 563

Performance Comparison

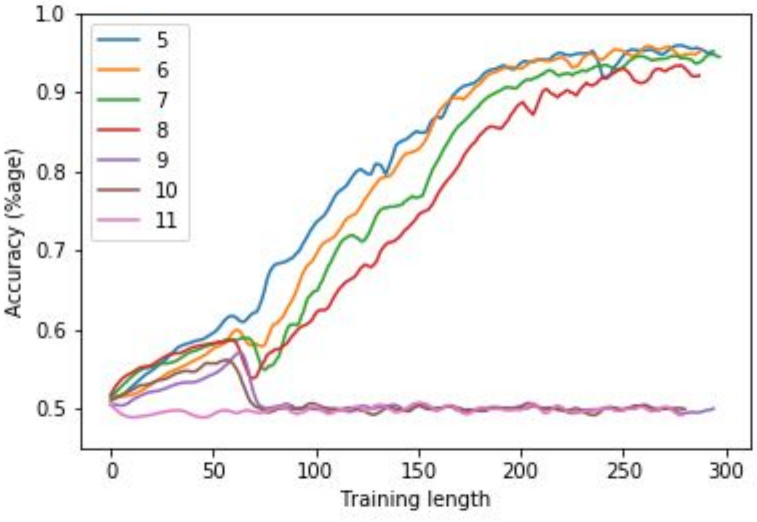


RNN

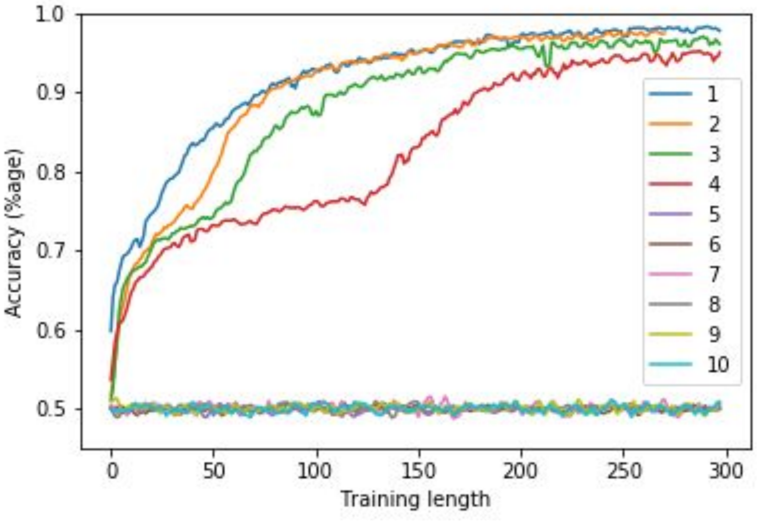


LSTM, AbLSTM and EIRNN

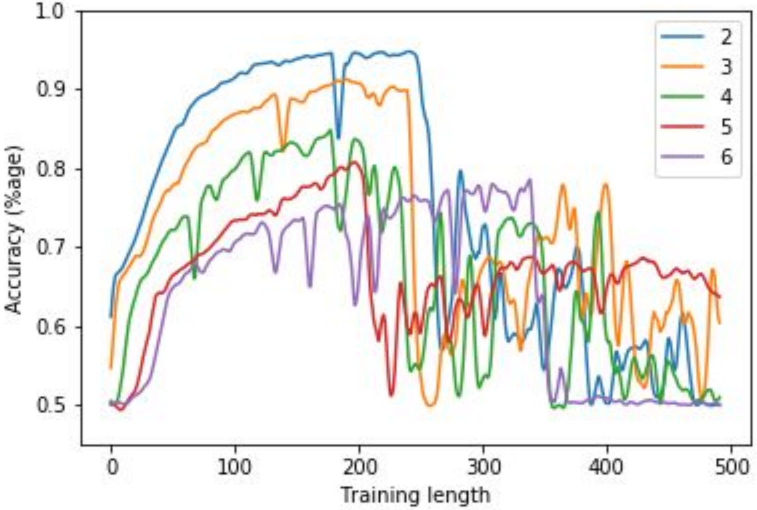
Validation Accuracy Curves During Training



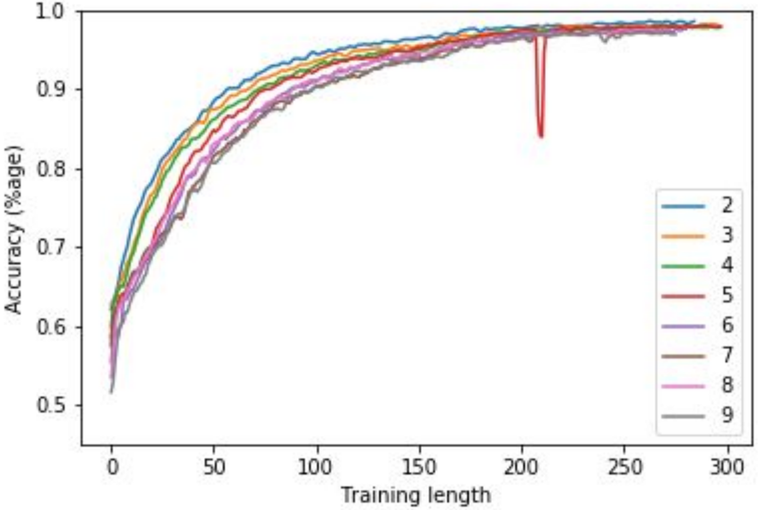
RNN (H = 250)



RNN (H = 50)

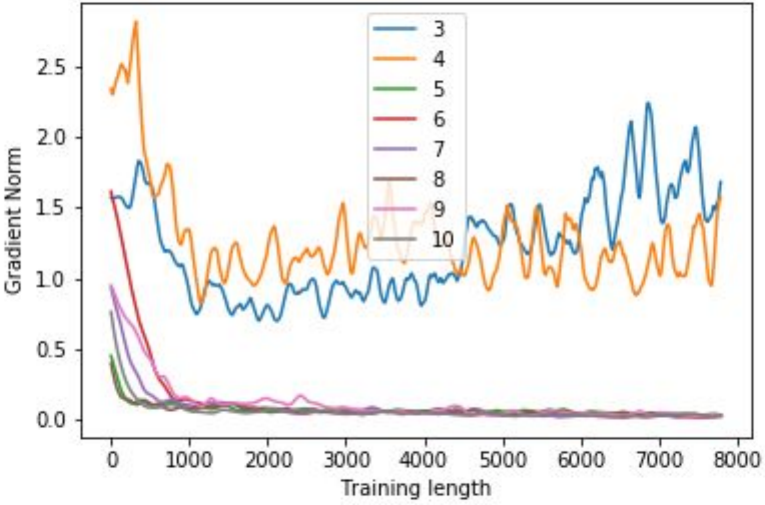

1.0
0.9
0.8
0.7
0.6
0.5
Accuracy (%age)
0
100
200
300
400
500
Training length
2
3
4
5
6

AbLSTM

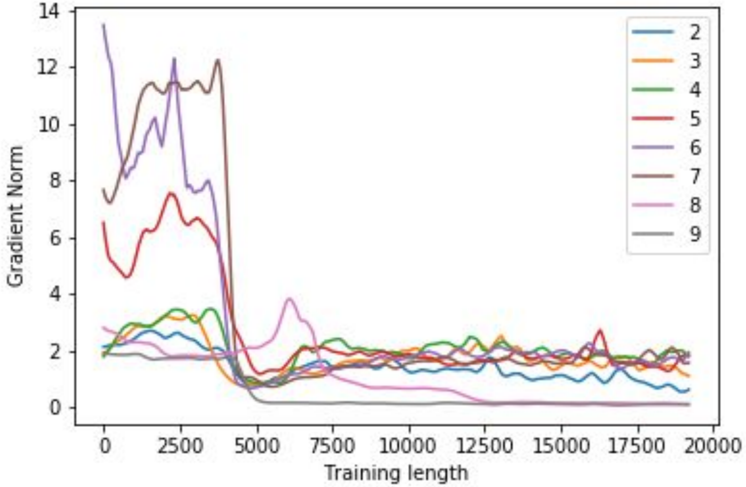

1.0
0.9
0.8
0.7
0.6
0.5
Accuracy (%age)
0
50
100
150
200
250
300
Training length
2
3
4
5
6
7
8
9

LSTM

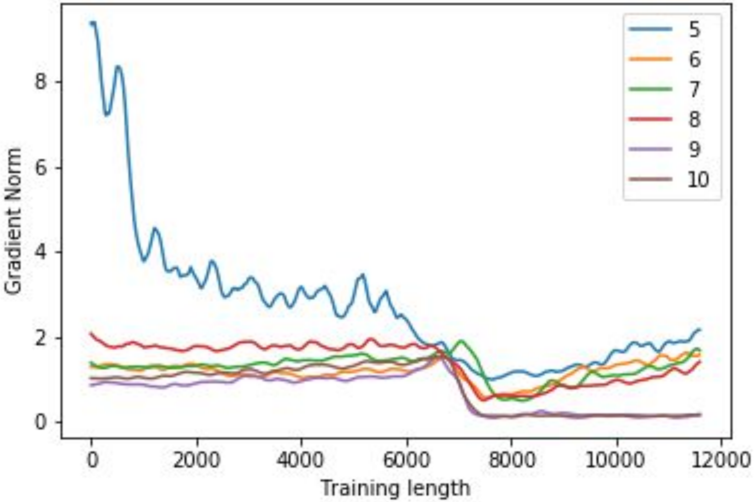
Gradient Norm Plots



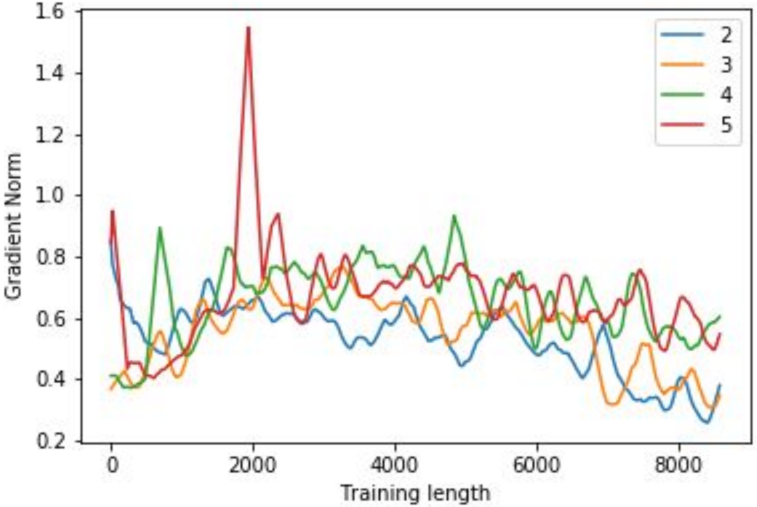
RNN (H = 50) : W_{rec}



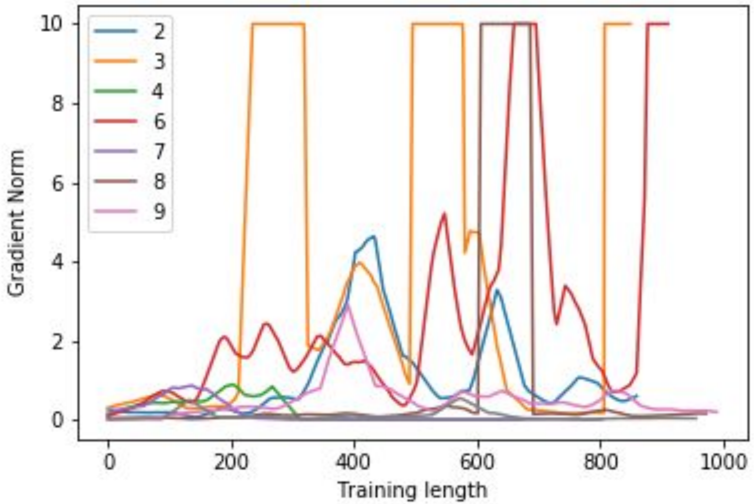
RNN (H = 150) : W_{rec}



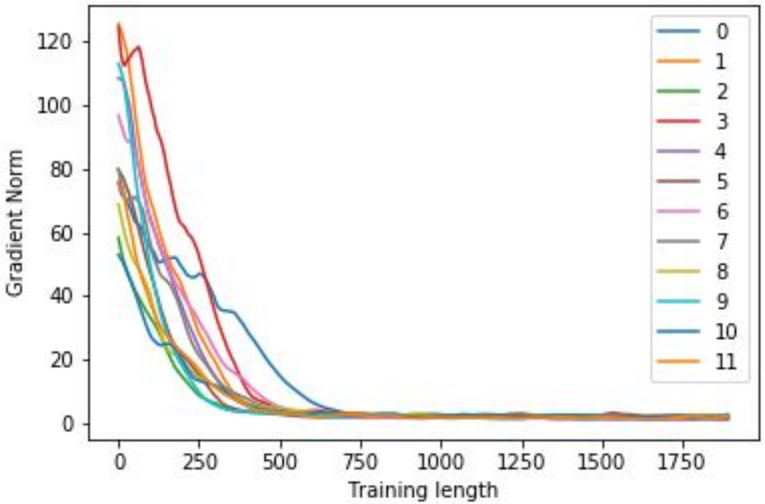
RNN (H = 250) : W_{rec}



LSTM : W_{rec}



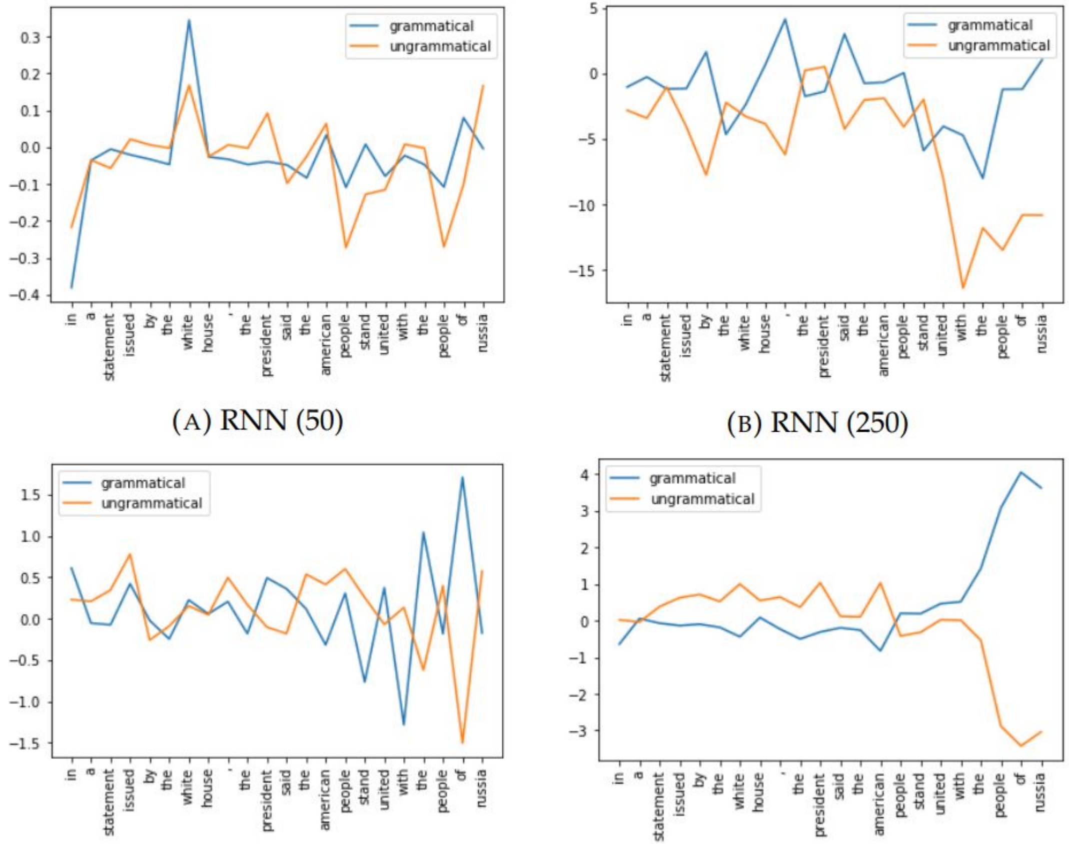
AbLSTM W_{rec}



EIRNN W_{rec}

Networks’ Prediction Change Plots

Plus 6



(A) RNN (50)

(B) RNN (250)

(C) AbLSTM

(D) LSTM

In a statement issued by the white house, the president said the American **people stand** united with the people of Russia

Qualitative Analysis

We manually examined over 500 samples to compare LSTM and RNN modelling results on multiple Grammaticality Plus tasks. Statistics for these results have been summarized in Table below.

	Plus 2	***Plus 3***	***Plus 6***
Both Correct	96.7%	96.4%	95.0%
Only LSTM Correct	1.7%	1.8%	2.9%
Only RNN Correct	0.8%	1.1%	1.2%
Both Wrong	0.7%	0.7%	0.9%

Overall Accuracy

Plus 2, 3 : RNN (H = 50) and Plus 6 : RNN (H = 250)

The networks often misidentified the heads of noun-noun compounds.

- * Two relatively large football **stadiums** , Fawcett stadium in Canton and Paul Brown Tiger stadium in Massillon , **is** still in use
- Tax **collectors collect** gold from guilds
- * The credits **sequence** of Bob’s Burgers **feature** the Belcher family
- The urban services district encompasses the 1963 boundaries of the former city of Nashville , and the general services **district includes** the remainder of

Lack of syntactic boundaries makes identifying the noun-verb pair correctly hard. The difficulty increases with addition of more noun-verb pairs, before the noun in context or in between the noun-verb pair on which the sample is testing grammaticality.

- I believe that the usual **way** a *tachometer works* **is** by spinning a
- * The **region** where the *stream originates* **are** in the highlands
- * The **state** that the *narrator wants* **are** seemingly a state

Understand inherent plurality introduced by group creation by using "commas" and conjunctive words likes "and" and "plus", is hard and in such cases the head of the subject’s number may not be relevant.

- * **Corn , soybeans, and wheat is** three common crops
- **Characters and plot are** complementary – they
- * **Health , arts and imaging technology, respiratory care and dental hygiene is** some of the

Ability of the model to distinguish proper nouns from other nouns when the sentence in lower case which removes one of the major clues present in general in text.

- The *United States* **system requires** that these differences
- * The credits **sequence** of *Bob's Burgers* **feature** the Belcher family
- * The *Blocks* **editor use** the *Open Blocks* software

Quantitative Analysis

50% of the samples which **LSTM gets wrong**, **AbLSTM also gets wrong** across multiple Plus α tasks. On the other hand, **RNN only** gets around **20-30%** samples wrong which LSTM predicted incorrectly. The above correlation shows that :

- Architectural closeness between AbLSTM and LSTM is reflected in behavioral (performance) closeness.
- The high correlation is errors made by different models shows that some sentences are inherently harder to model than others

	Plus 2	***Plus 3***
Both Correct	2.35	2.35
Only LSTM Correct	5.72	5.3
Only RNN Dale Correct	5.0	4.63
Both Wrong	6.81	6.58

Average distance between noun and verb for each set

Expectation and Locality Effects

Expectation Effect :

The expectation for a verb becomes sharper in the second case, leading to faster reading times at the verb compared to the first sentence. (Distance between noun phrase and verb increased).

- **The administrator** who the nurse **met**…
- **The administrator** who the nurse that was from the clinic **met**…

Locality Effect :

- Grodner and Gibson showed that increasing distance in the manner shown above results in slowdowns at the verb. This has been explained in terms of the increased cost of completing a dependency when a co-dependent is more distant.
- The argument is difficult to retrieve either because it has become less accessible in memory over time (decay), or because other nouns intervening between the co-dependents make it difficult to identify and retrieve the correct target noun .

S Husain 2014, Strong Expectations Cancel Locality Effects: Evidence from Hindi

Grodner and Gibson, Consequences of the Serial Nature of Linguistic Input for Sentenial Complexity

Comparison Across Plus Tasks on a Model

- Compared the performance and specific errors of :
    - RNN on Plus 1 and Plus 3
    - LSTM on Plus 3 and Plus 6

- Contrary to expectation :
    - Errors made by RNN on Plus 1 are not a subset of errors made on Plus 3.
    - Errors made by LSTM on Plus 3 are not a subset of errors made on Plus 6.

- The **overlap of errors** in both above cases is close to **40-45%**.
    - More than 50% of samples wrong in one were right in the other.

- Although the overall performance decreases with increasing α, the errors made by the network are not consistent.

- These results agree with the expectation vs locality effect trade-offs in linguistics and cognition.
    - The added words in other cases help the network better parse the sentence.

Potential Plus 3 sample :

The apartment that the **maids** who the service had sent over **were** cleaning every week…

Potential Plus 6 sample :

The apartment that the **maid** who the service had sent over **were** cleaning every week was well decorated.

Hypothesis :
It is harder to find noun-verb associations in the first sentence than in the second because of the extra information present.

The locality like effect of having a cue about the locus of the verb vs the expectation like effect of having extra information helping parsing/correctly identifying noun-verb pairs.

Discussion on Ablated LSTM

- Networks with high accuracy on grammaticality :
  - LSTM
  - Combined i-f gate LSTM
  - GRU

Hypothesis :

- Input and forget gates are important to the performance of the LSTM, not just for better gradient flow.
- LSTMs (GRUs, etc) performing better on natural language tasks shows that being able to selectively remember and forget is needed for language comprehension.

Discussion

- Comparison of different recurrent models of different linguistic tasks can provide insights into both cognition and understanding how architecture affects performance.

- Since the majority of natural language sentence are grammatically simple, models can achieve high overall accuracy using flawed heuristics that fail.

- The reason behind the inability of SRN or RNN Dale to suddenly be unable to model a task with slightly higher difficulty (alpha increased by 1) is unclear.

- Syntax Trees vs Sequential Parsing.
    - LSTM : Model Brain or Model Brain Function.

Conclusion

- We analyzed the performance of architecturally plausible neural networks with explicit biological constraints.

- We provided insights into differences in different recurrent neural networks and factors that affect them.

- We showed relationships between different tasks based syntax-sensitive dependencies and that more information can both help and hurt in a linguistic setting.

- We proposed insights that inspire models that are biologically plausible and can bridge the gap between in architecture when we talk of cognitive plausibility

Future Work

- Further analysis of the activation plots and different models that are more powerful than AbLSTM but less powerful than LSTM would shed further light on how architecture affects learning in LSTMs and in recurrent neural networks in general.

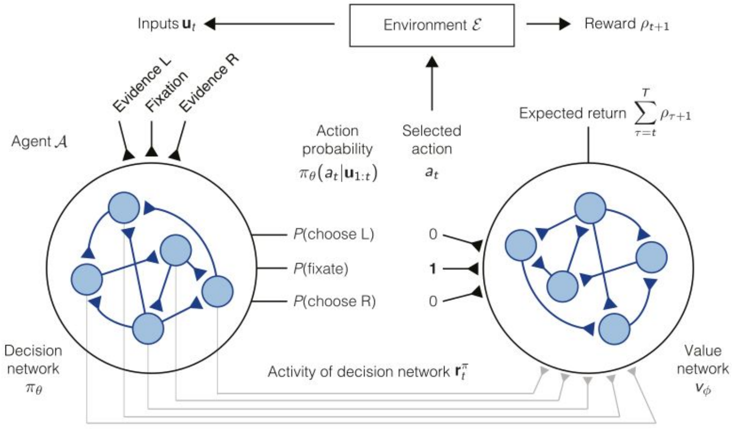
- Changing the recurrent network to use the last two states instead of just one reduces the effective distance between two points in the sequence in half which could help in modelling.

- LSTMs with explicit window memories to access previous states/inputs could explain explicit vs implicit memory effects.
  - This would also help in sentences in which the parse of the sentence cannot be created by reading words only sequentially.

- Further linguistic analysis on the corpus and the error sets of the models across multiple tasks would provide more concrete evidence.

Reward-based Training of RNNs for Cognitive & Value-based Tasks

H. Francis Song
Guangyu R. Yang
Xiao-Jing Wang



Inputs $\mathbf{u}_t$
Environment $\mathcal{E}$
Reward ρ_{t+1}
Evidence L
Fixation
Evidence R
Agent $\mathcal{A}$
Action probability $\pi_\theta(a_t|\mathbf{u}_{1:t})$
Selected action a_t
Expected return $\sum_{\tau=t}^{T}\rho_{\tau+1}$
P(choose L)
P(fixate)
P(choose R)
0
1
0
Decision network π_θ
Activity of decision network $\mathbf{r}_t^\pi$
Value network v_ϕ

[Reward-based training of recurrent neural networks for cognitive and value-based tasks: Song, Yang, Wang, PLOS 2017](#)

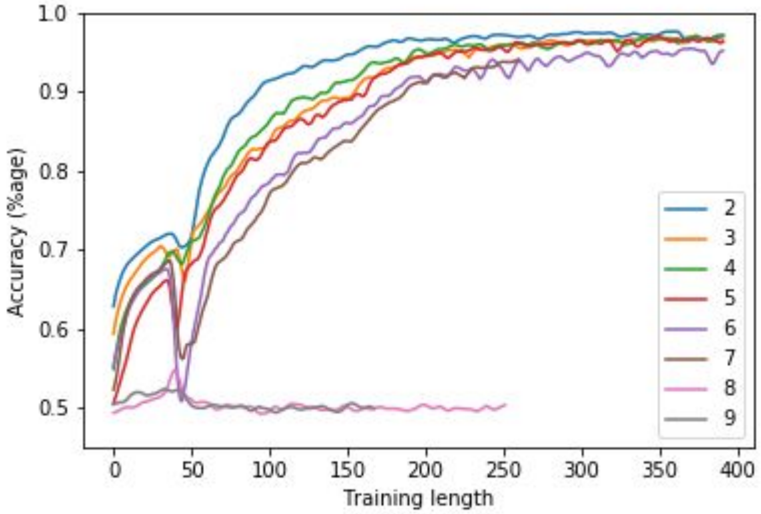
- Decision & Value Networks (Actor-Critic structure)
- Pre/post decision wagering.
- Perpetual decision making.
- Value-based economic choice.

All of these are low level tasks that are trained and tested at the level of neural data.

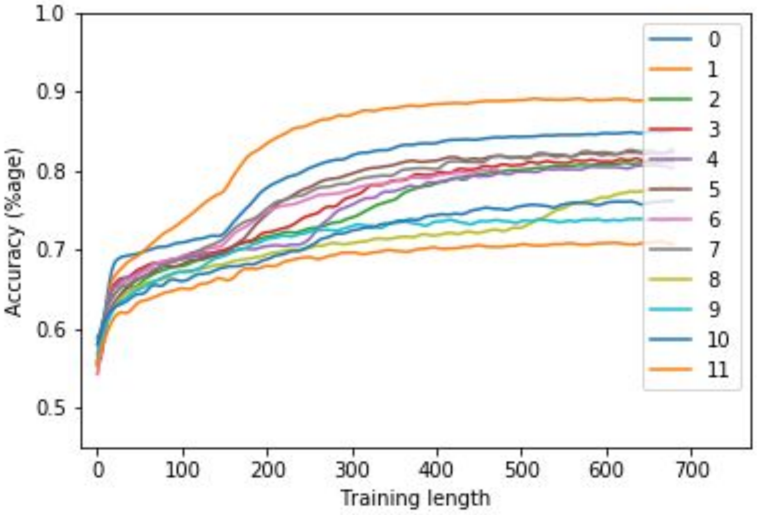
Thank You!

- Rishubh Singh

Validation Accuracy Curves During Training

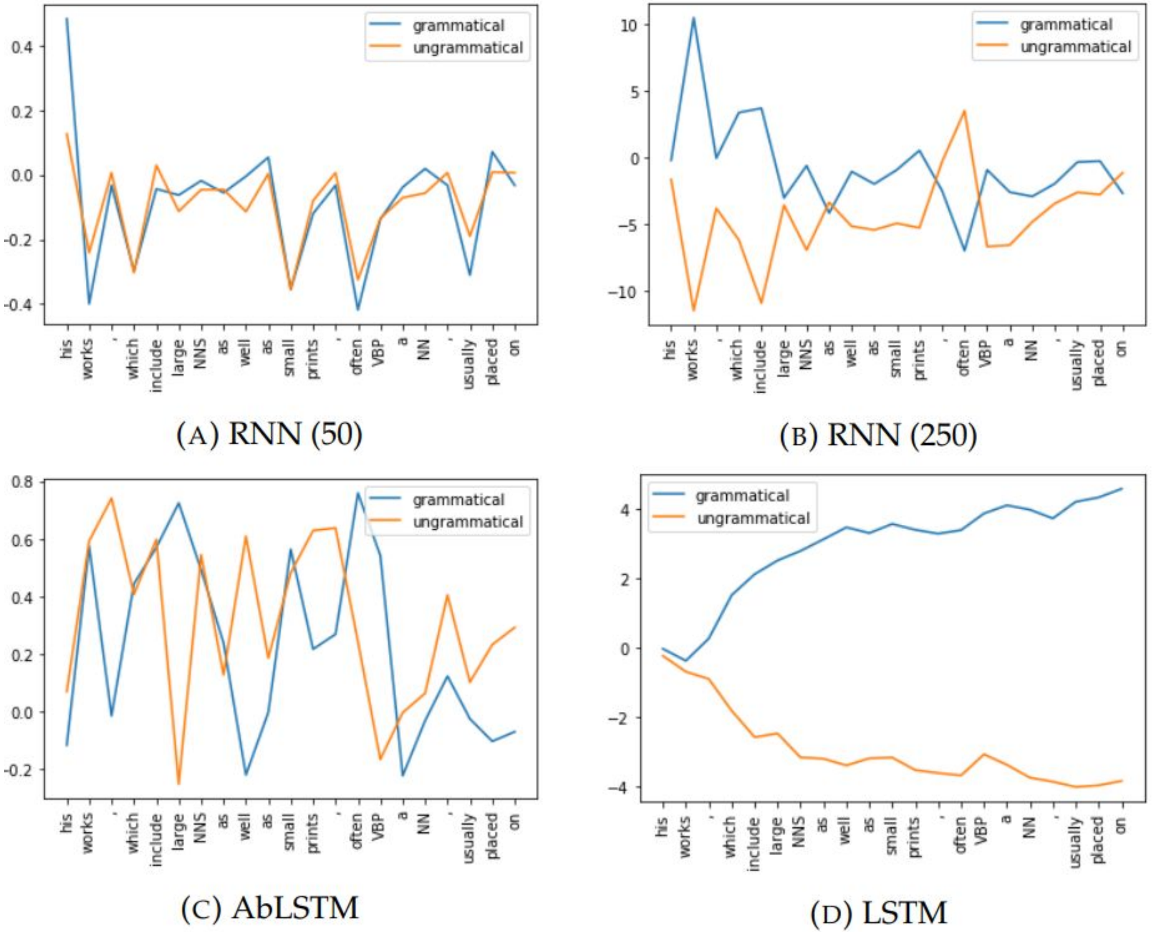


RNN (H = 150)



EIRNN

Plus 6



(A) RNN (50)

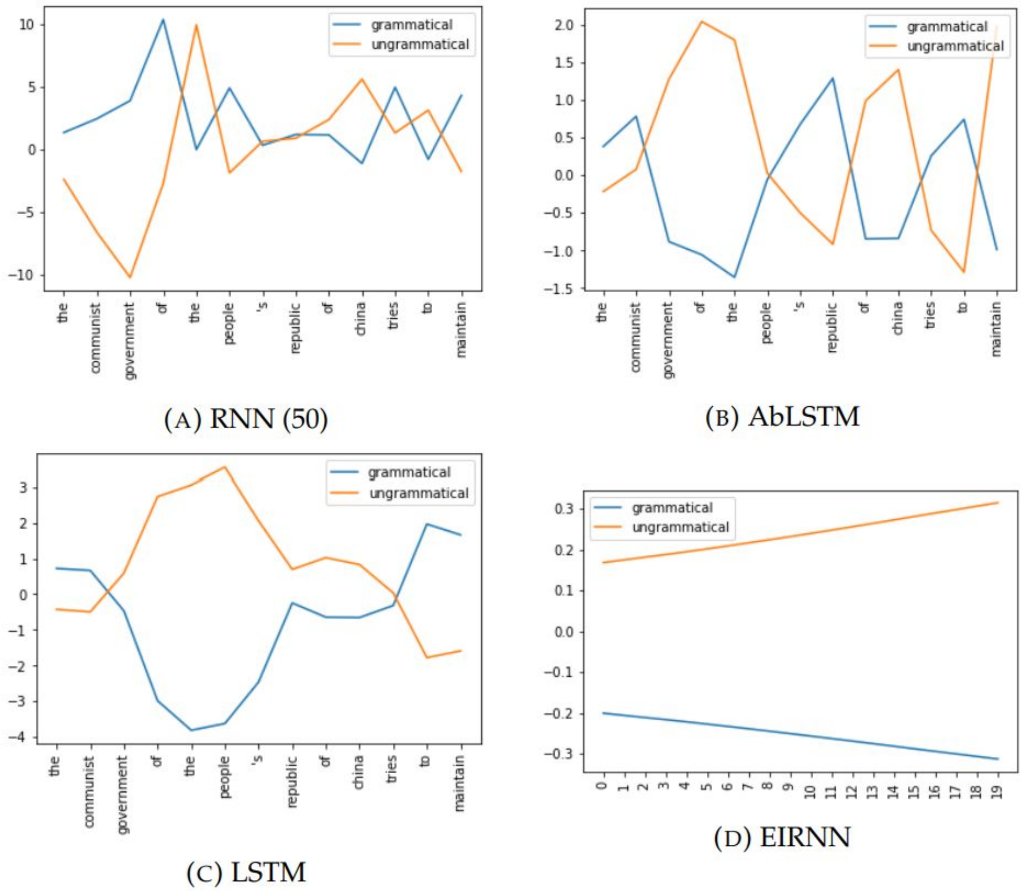
(B) RNN (250)

(C) AbLSTM

(D) LSTM

His **works**, which include large NNS(murals) as well as small prints, often **VBP(depict)** a NN(teapot), usually placed on

Plus 2



(A) RNN (50)

(B) AbLSTM

(C) LSTM

(D) EIRNN

The communist **government** of the People's Republic of China **tries** to maintain

Some errors appear to be due to difficulty not in identifying the subject but in determining whether it is plural or singular.

- Its current **headquarters is** in Rabat, Morocco.

Some verbs that are ambiguous with plural nouns or nouns in general seem to have been misanalyzed as plural nouns and consequently act as attractors.

- Every **arrow** that flies **feels** the pull of (the earth).

	EIRNN	LSTM
NN	3368	954
NNS	1470	1997
NN NN	1177	261
NN NNS	432	5619
NNP NN	349	316
NN NN NN	315	87
PRP$ NN	298	123
NN NNP	213	330
NNP NNS	187	443
NNS NNS	175	119
NNS NN	168	5758

	EIRNN	LSTM
NNS NN	168	5758
NN NN NNS	97	2691
NN NNS NN	51	1295
NN NNS	432	5619
NN NN NN NNS	25	1053
NN NNS NNS	33	841
NNP NN NNS	58	602
NNS NNS NN	15	597
NN NNP NNS	14	486
NNP NNS	187	483
NN NN NN NN NNS	12	418

Examples & Analysis

EIRNN not LSTM	LSTM not EIRNN
<u>2272 unique ; 13941 total points</u>	**<u>23313 unique ; 84039 total points</u>**
1. NN : 3368	1. NNS NN : 5758
2. NNS : 1470	2. NN NNS : 5619
3. NN NN : 1177	3. NN NN NNS : 2619
4. NN NNS : 432	4. NNS NN NN : 2092
5. NNP NN : 349	5. NNS : 1997
6. NN NN NN : 315	6. NN NNS NN : 1295
7. PRP$ NN : 298	7. NN NN NN NNS : 1053
8. NN NNP : 213	8. NN : 954
9. NNP NNS : 187	9. NN NNS NNS : 841
10. NNS NNS : 175	10. NNP NN NNS : 602
11. NNS NN : 168	11. NNS NNS NN : 597
12. PRP NN : 135	12. NNS NN NN NN : 594
13. PRP$ NNS : 125	13. NNS NNP NNP : 576
14. NNP NN NN : 123	14. NNS NNP NN : 496
15. NN NN NN NN : 119	15. NN NNP NNS : 486
16. NNP NNP NN : 115	16. NNS NNP : 457
17. NN NNP NNP : 105	17. NN NNS NN NN : 451
18. NN NN NNS : 97	18. NNP NNS : 443
19. NN PRP : 72	19. NNP NNS NN : 439
20. NNS NNP : 70	20. NN NN NN NN NNS : 418

Before Extra Slides

EIRNN not RNN	RNN not EIRNN
4679 unique ; 14142 total points	**19849 unique ; 66153 total points**
1. NN NNS : 788	1. NNS NN : 4635
2. NNS NN : 740	2. NN NNS : 4475
3. NN NN NNS : 524	3. NN NN NNS : 2156
4. NNS : 497	4. NNS : 1605
5. NN : 328	5. NNS NN NN : 1494
6. NN NN NN NNS : 238	6. NN NNS NN : 999
7. NNS NN NN : 220	7. NN NN NN NNS : 865
8. NN NNS NN : 202	8. NN : 816
9. NN NN : 134	9. NN NNS NNS : 505
10. NN NNP : 132	10. NNS NNP NNP : 501
11. NN NNS NNS : 130	11. NNS NNS NN : 500
12. NNP NNS : 119	12. NNP NN NNS : 489
13. NNS NNP : 110	13. NNS NNP NN : 390
14. NN PRP : 109	14. NNS NNP : 386
15. NNS NNS : 100	15. NN NNP NNS : 372

Example Sentences (EIRNN vs LSTM)

- **EIRNN (not LSTM)**
  - the layout of the hotel is (also similar to the one used in bottle rocket)
  - the full source code for the game is (now available on NNS in the NN on NNP programming)
  - no evidence as to why this level crossing is (notable)

- **LSTM (not EIRNN)**
  - the details depend upon whether the wave is (purely JJ , purely electromagnetic , or neither .)
  - the place for japanese characters is (in the japanese wp , like .)
  - all original research and no secondary sources available online , even though this is a recent NN

Ablated LSTM Models

LSTM :

$i = \mathrm{sigma}(W_{ii}\ x + b_{ii} + W_{hi}\ h + b_{hi})$

$f = \mathrm{sigma}(W_{if}\ x + b_{if} + W_{hf}\ h + b_{hf})$

$g = \tanh(W_{ig}\ x + b_{ig} + W_{hg}\ h + b_{hg})$

$o = \mathrm{sigma}(W_{io}\ x + b_{io} + W_{ho}\ h + b_{ho})$

$c' = f * c + i * g$

$h' = o * \tanh(c')$

Ablated LSTM :

$g = \tanh(W_{ig}\ x + b_{ig} + W_{hg}\ h + b_{hg})$

$o = \mathrm{sigma}(W_{io}\ x + b_{io} + W_{ho}\ h + b_{ho})$

$h' = o * \tanh(c')$

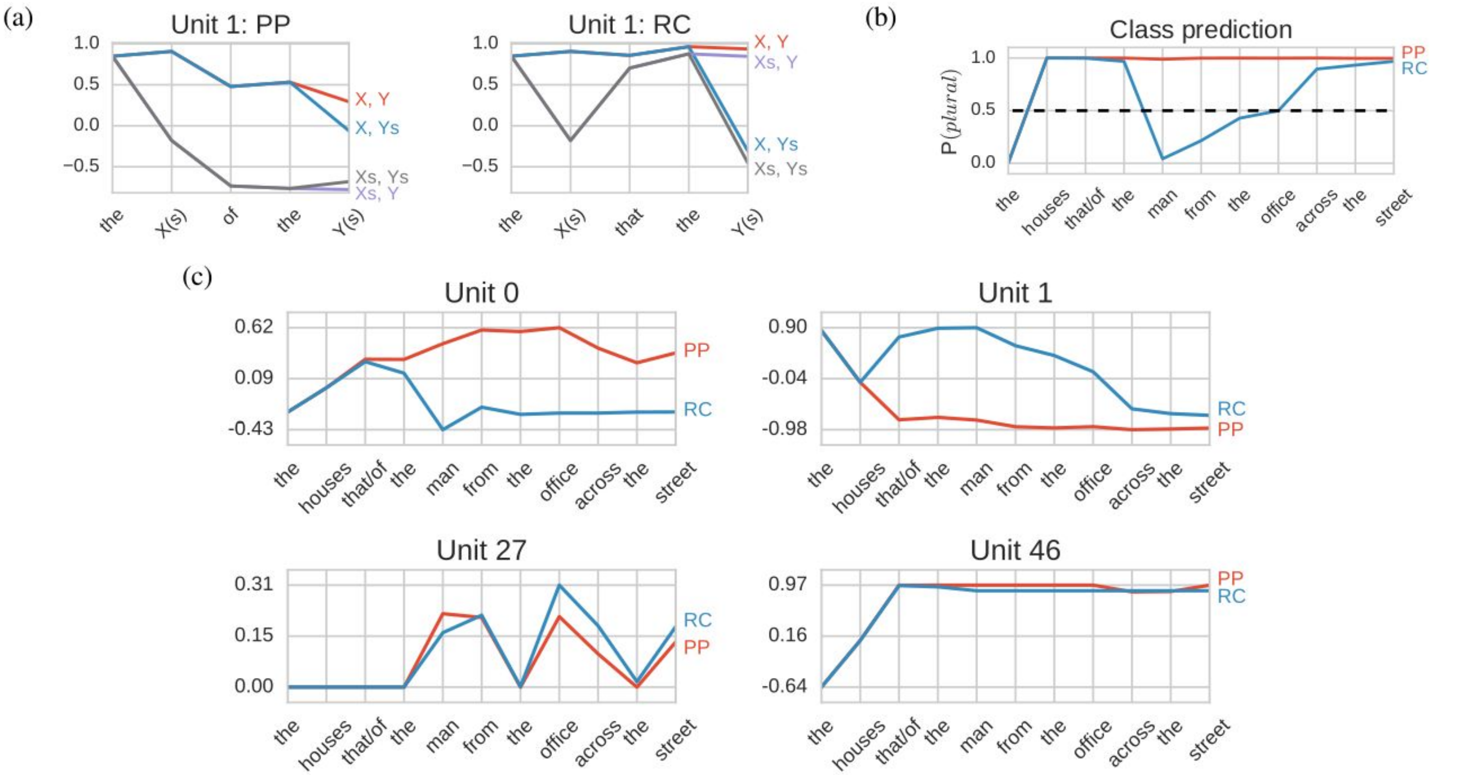
Scalar : $c' = ((1-\alpha) * c) + (\alpha * g)$

α is a scalar (real valued).

Vector : $c' = (f * c) + (i * g)$

i, f are vector parameters learned by the model.

Activations Analysis and Replication



Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies: Linzen, ACL 2016