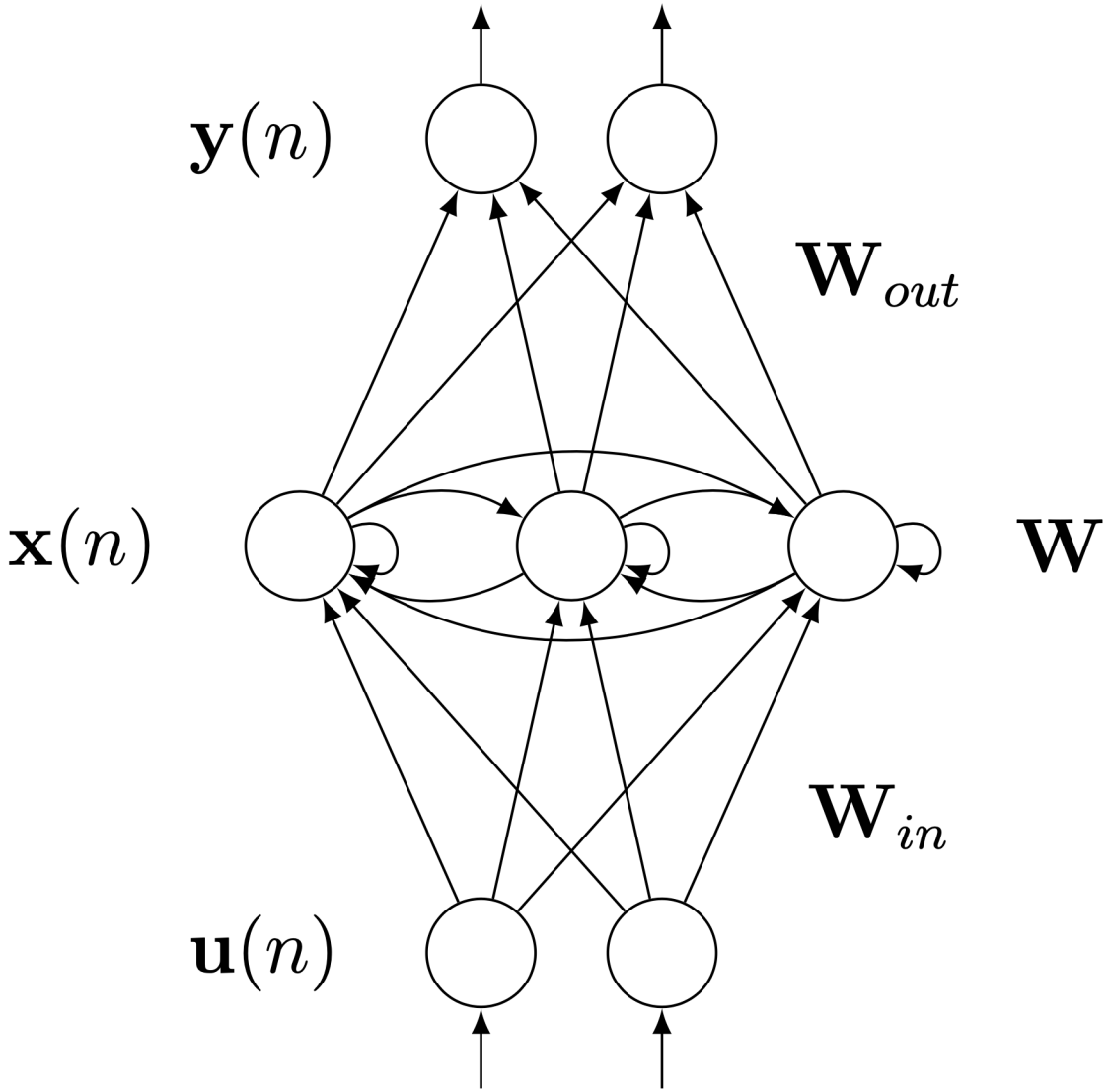
Exploiting Randomness in Neural Networks

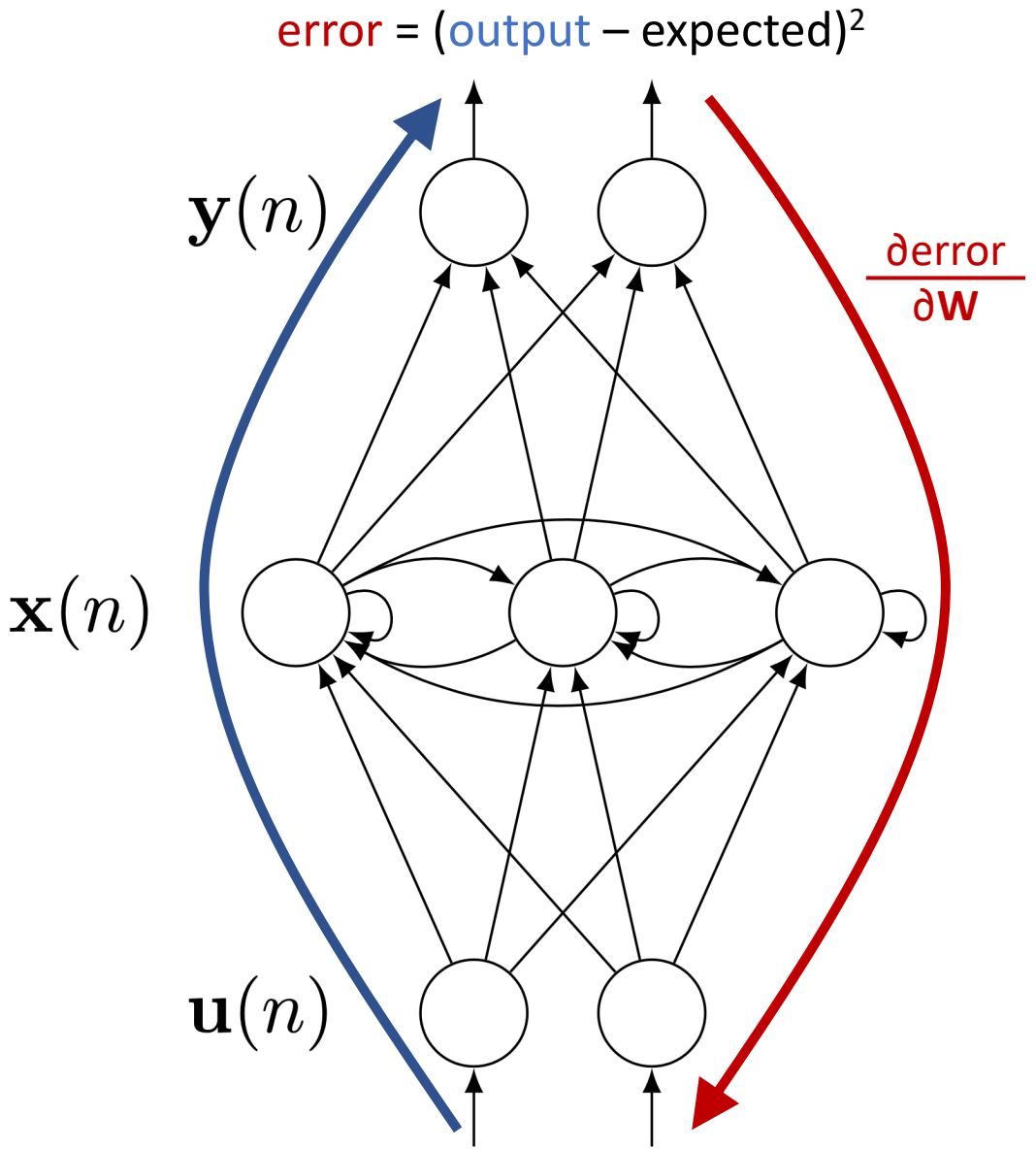
Daniele Di Sarli

Mauriana Pesaresi seminars - 2020



Recurrent Neural Network

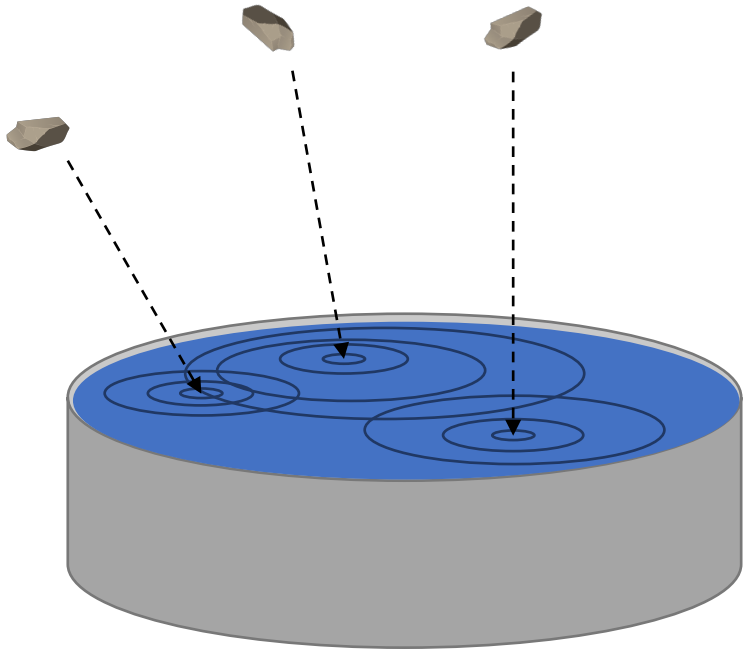
Backpropagation Through Time



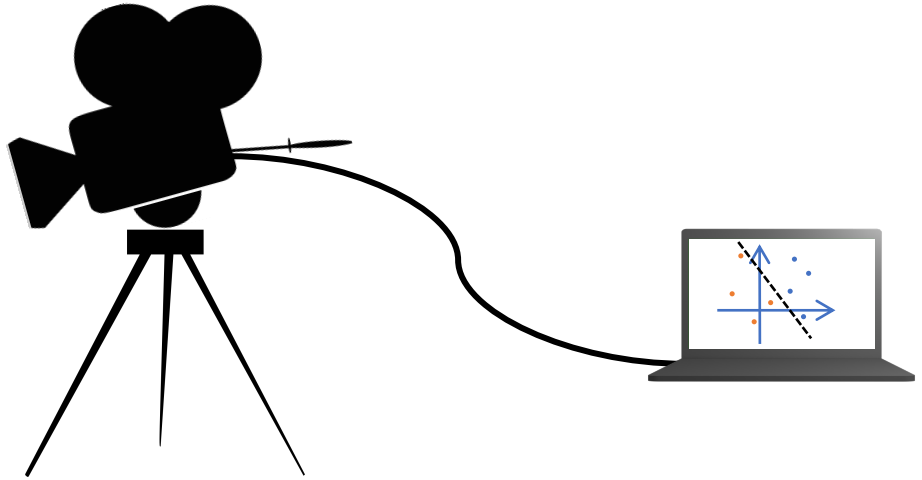
PREDICTION

PATTERNS, INTERACTIONS

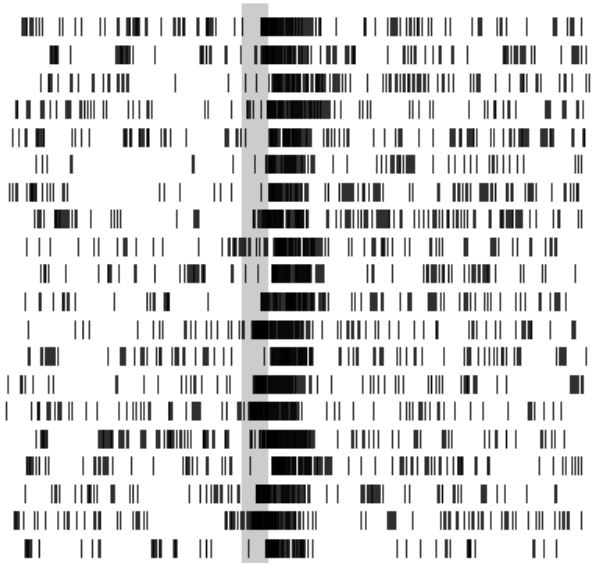
…, 3, 2, 1.5, 0.75, 1, -2.3, 4, …



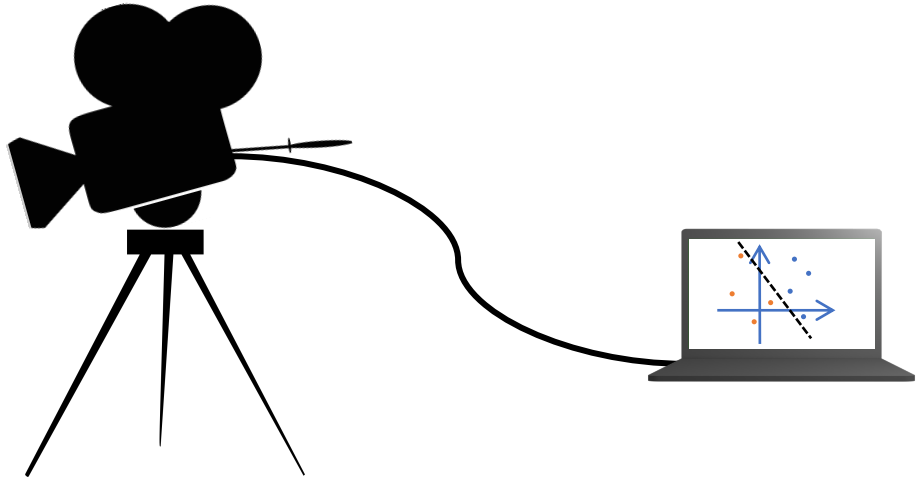
Reservoir



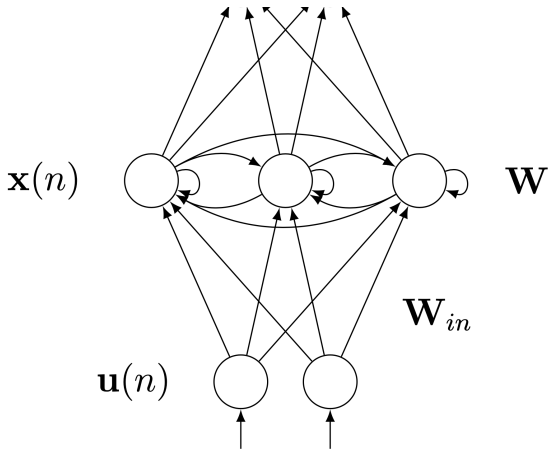
Readout



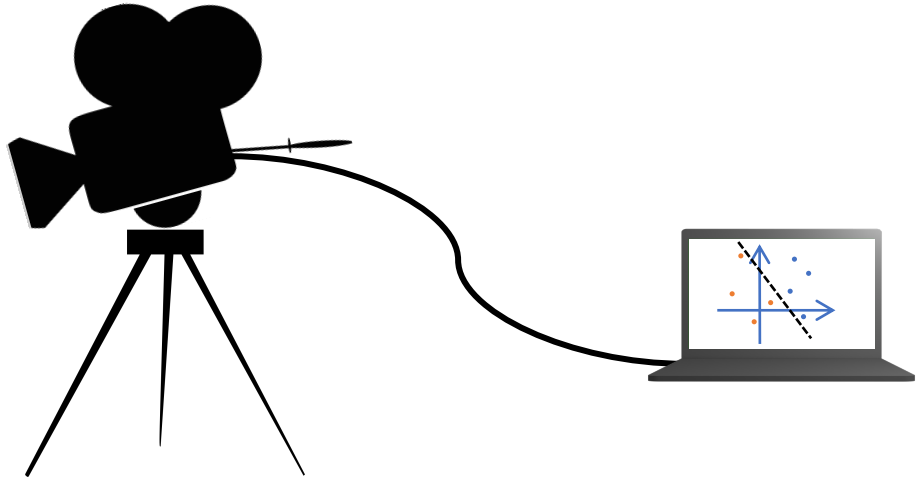
Reservoir



Readout

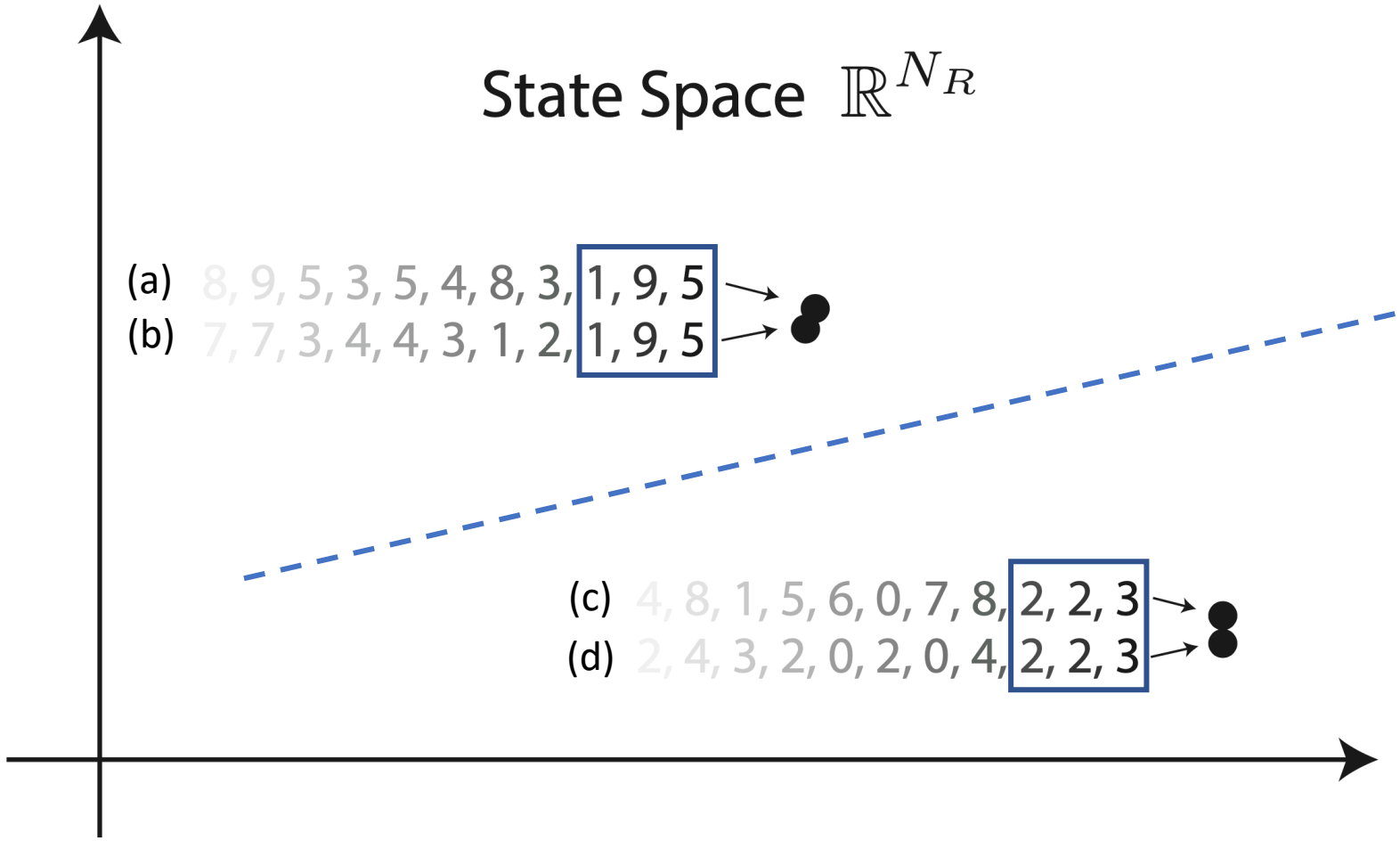


Reservoir

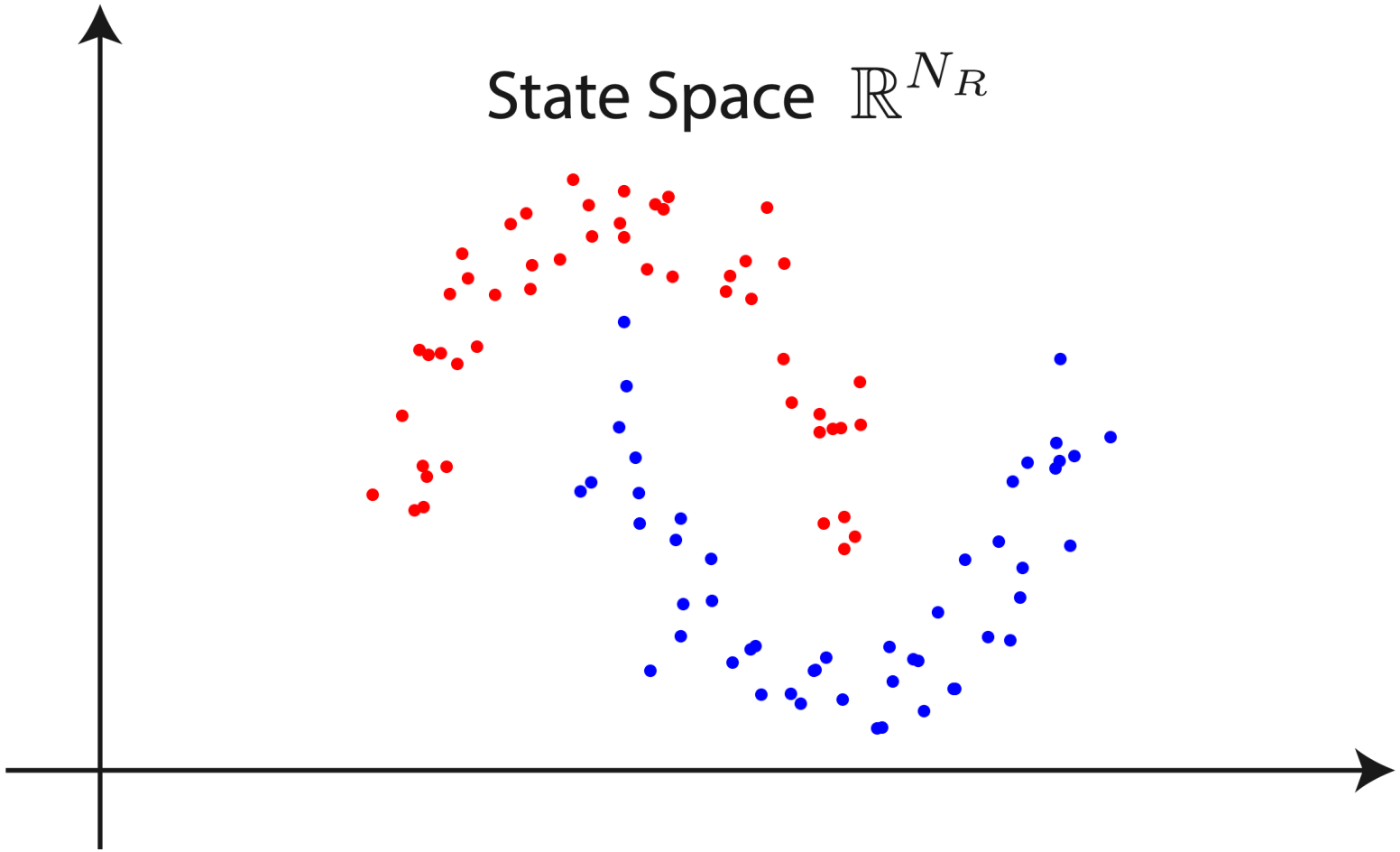


Readout

Echo State Network


State Space $\mathbb{R}^{N_R}$
(a) 8, 9, 5, 3, 5, 4, 8, 3, 1, 9, 5
(b) 7, 7, 3, 4, 4, 3, 1, 2, 1, 9, 5
(c) 4, 8, 1, 5, 6, 0, 7, 8, 2, 2, 3
(d) 2, 4, 3, 2, 0, 2, 0, 4, 2, 2, 3

Cover's theorem



Echo State Property

$$\forall s_N(\mathbf{u}) = [\mathbf{u}(1), \ldots, \mathbf{u}(N)] \in (\mathbb{R}^{N_U})^N,$$

$$\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^{N_R} :$$

$$\|\hat{\tau}(s_N(\mathbf{u}), \mathbf{x}) - \hat{\tau}(s_N(\mathbf{u}), \mathbf{x}')\| \to 0 \text{ as } N \to \infty$$

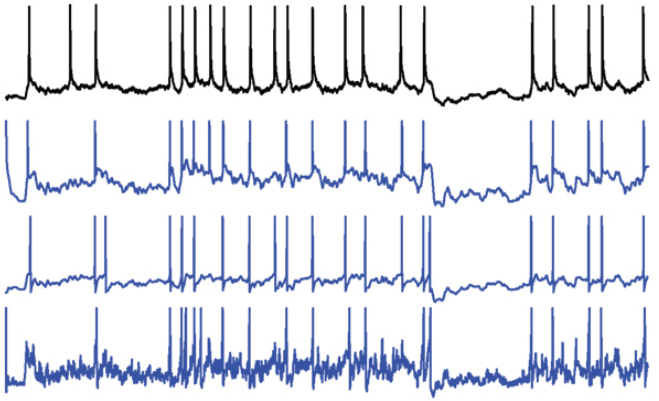
Echo State Network starter pack

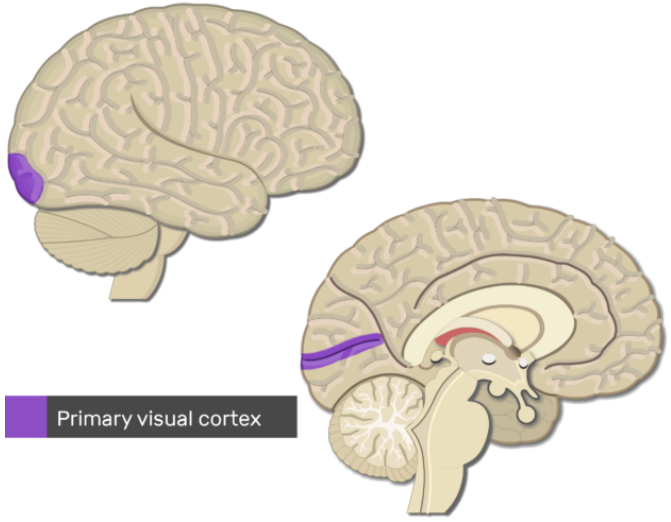
1. **Randomly initialize** the weights (sparse)
2. Rescale the weights to guarantee **contractivity** of the state transition function (=> ESP)
3. Feed data, collect states
4. Compute **optimal** linear regression parameters

$$\mathbf{W}_{out} = \bar{\mathbf{Y}}\mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}$$

«**RC** […] provides explanations of why biological brains can carry out **accurate computations** with an “inaccurate” and **noisy physical substrate**»

— Lukoševičius et al.





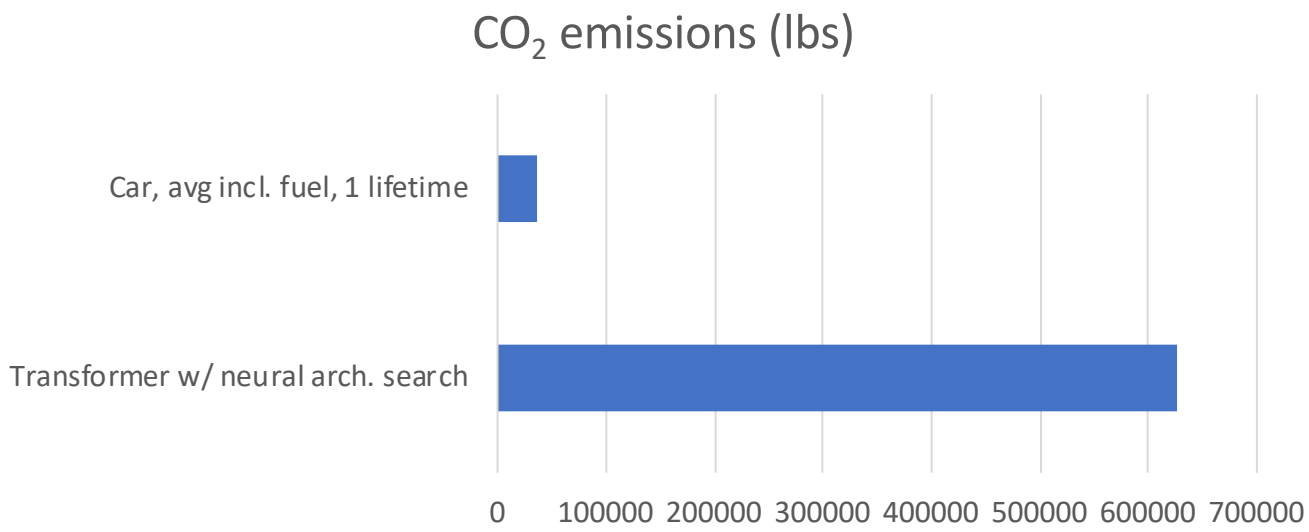
In the **primary visual cortex**, «computations are performed by complex **dynamical systems** while information about results of these computations is read out by simple **linear classifiers**.»

— Nikolić et al.

My work

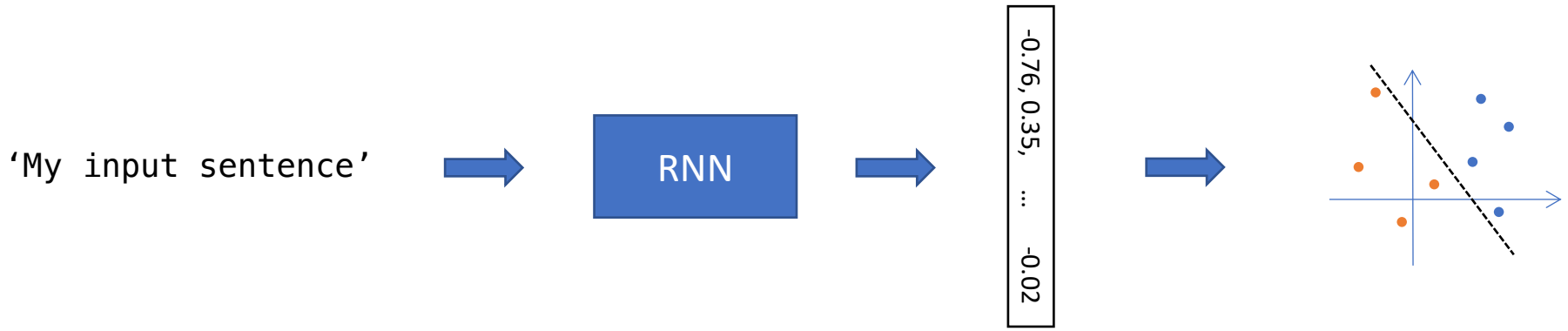
Natural Language Processing

LSTM GRU Transformer BERT



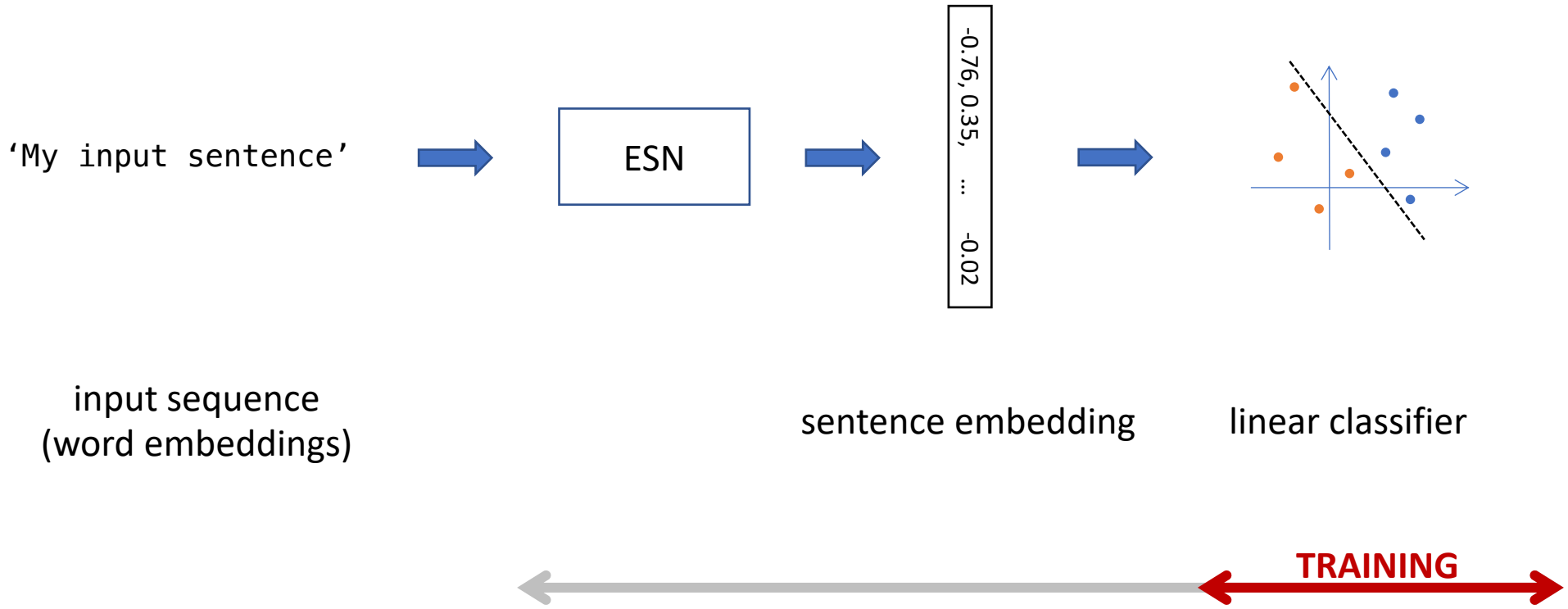
From *Strubell, E., Ganesh, A., McCallum, A.*:
Energy and Policy Considerations for Deep Learning in NLP
Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019

Text Classification pipeline



TRAINING

Text Classification pipeline



Question Classification

What was the name of the first Russian astronaut to do a spacewalk?

HUMAN

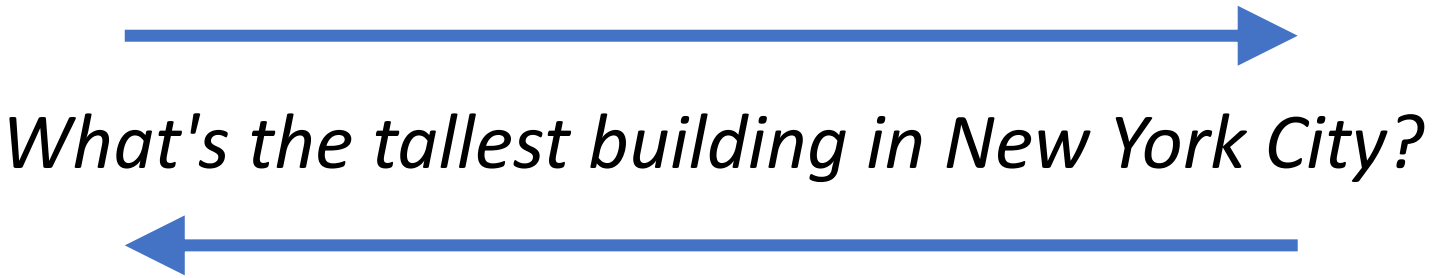
What's the tallest building in New York City?

LOCATION

… also ABBREVIATION, ENTITY, DESCRIPTION, and NUMERIC VALUE

Improvements are needed

- **Bidirectional**
- Attention
- Multi-ring



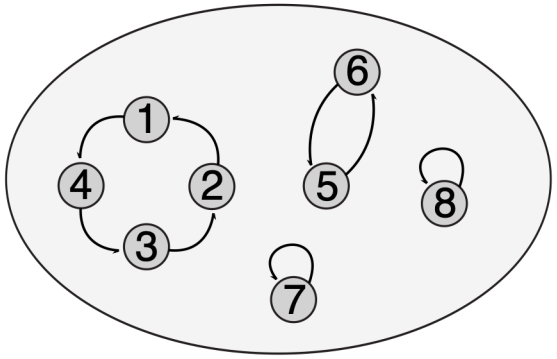
Improvements are needed

- Bidirectional
- **Attention**
- Multi-ring

Improvements are needed

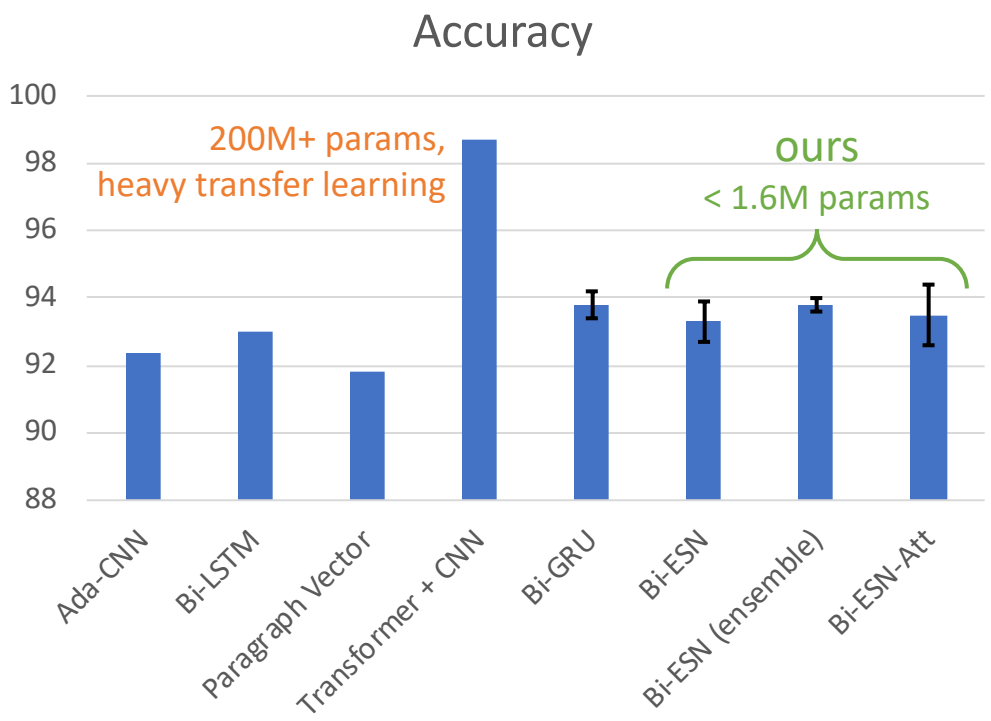
- Bidirectional
- Attention
- **Multi-ring**

$$\hat{\mathbf{W}} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \begin{pmatrix} 0 & v & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & v & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & v & 0 & 0 & 0 & 0 \\ v & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & v & 0 & 0 \\ 0 & 0 & 0 & 0 & v & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & v & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & v \end{pmatrix} \end{array}$$

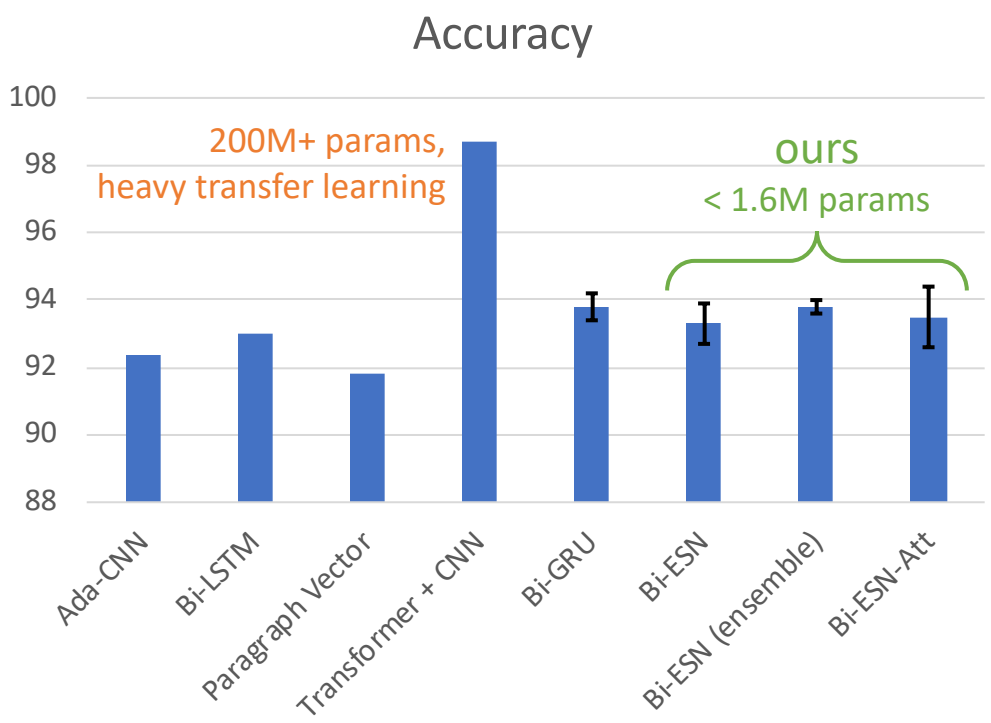


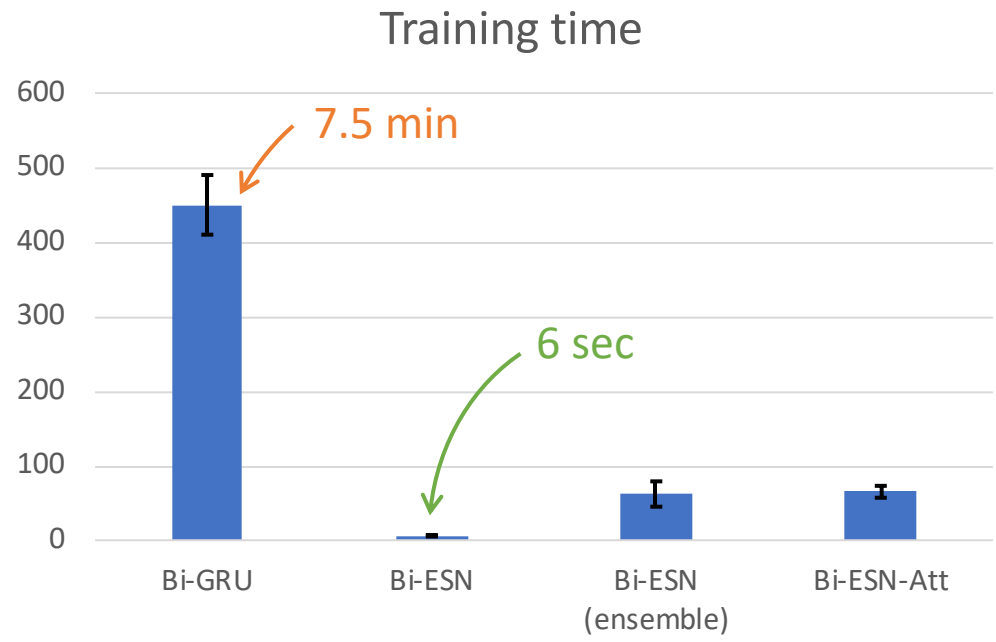
Multi-ring reservoir units

Results


Accuracy
100
98
96
94
92
90
88
200M+ params,
heavy transfer learning
ours
< 1.6M params
Ada-CNN
Bi-LSTM
Paragraph Vector
Transformer + CNN
Bi-GRU
Bi-ESN
Bi-ESN (ensemble)
Bi-ESN-Att

Results


Accuracy
100
98
96
94
92
90
88
200M+ params,
heavy transfer learning
ours
< 1.6M params
Ada-CNN
Bi-LSTM
Paragraph Vector
Transformer + CNN
Bi-GRU
Bi-ESN
Bi-ESN (ensemble)
Bi-ESN-Att


Training time
600
500
400
300
200
100
0
7.5 min
6 sec
Bi-GRU
Bi-ESN
Bi-ESN
(ensemble)
Bi-ESN-Att

How old was the youngest president of the United States ?

When was Ulysses S. Grant born ?

Who invented the instant Polaroid camera ?

What is nepotism ?

Where is the Mason/Dixon line ?

What is the capital of Zimbabwe ?

What are Canada 's two territories ?

Wrap up

- A path towards efficient, effective ML models must be taken
- Heavier understanding/exploitation of the architectural properties of RNN models can help towards that goal
- Analysis is preliminary, but WIP results are encouraging

References

1. Di Sarli, D., Gallicchio, C., & Micheli, A. (2019, November). **Question Classification with Untrained Recurrent Embeddings**. In *International Conference of the Italian Association for Artificial Intelligence*.
2. Jaeger, H., & Haas, H. (2004). **Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication**. *Science*.
3. Lukoševičius, M., & Jaeger, H. (2009). **Reservoir computing approaches to recurrent neural network training**. *Computer Science Review*.
4. Nikolić, D., Haeusler, S., Singer, W., & Maass, W. (2007). **Temporal dynamics of information content carried by neurons in the primary visual cortex**. In *Advances in neural information processing systems*.