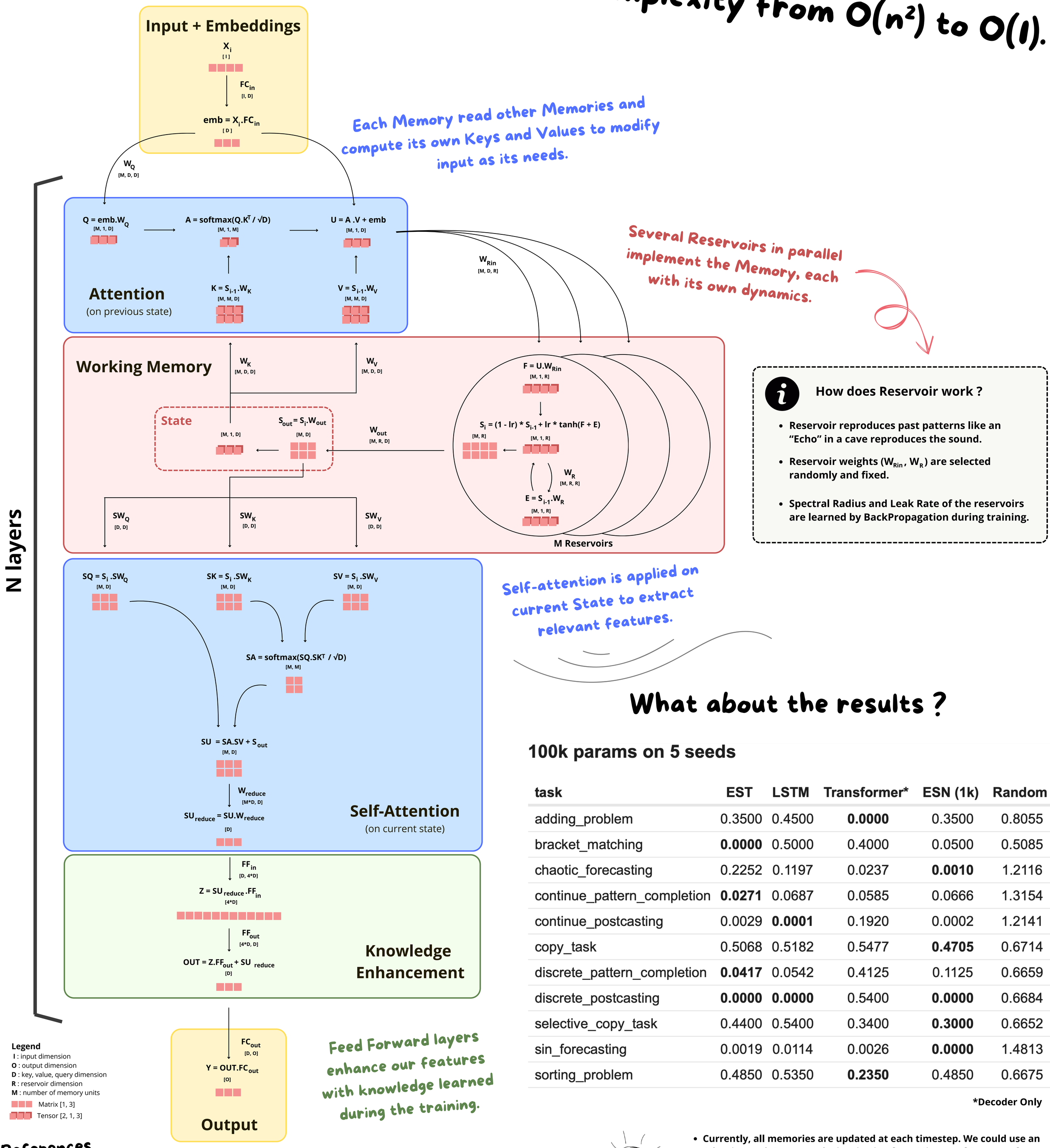
Echo State Transformer

Yannis Bendi-Ouis[1,2,3] **, Xavier Hinaut**[1,2,3]

[1] Inria Center of Bordeaux University, Bordeaux, France
[2] LaBRI, Bordeaux Univ., Bordeaux INP, CNRS UMR 5800, France
[3] Bordeaux Univ., CNRS, IMN, UMR 5293, Bordeaux, France

This work is supported by Inria AEx BrainGPT project.
Thanks to Experimental Inria Cluster Plafrim.

Inspired by Reservoir Computing we add Memory Units to Transformer to reduce their complexity from O(n²) to O(1).



What about the results ?

100k params on 5 seeds

task	EST	LSTM	Transformer*	ESN (1k)	Random
adding_problem	0.3500	0.4500	**0.0000**	0.3500	0.8055
bracket_matching	**0.0000**	0.5000	0.4000	0.0500	0.5085
chaotic_forecasting	0.2252	0.1197	0.0237	**0.0010**	1.2116
continue_pattern_completion	**0.0271**	0.0687	0.0585	0.0666	1.3154
continue_postcasting	0.0029	**0.0001**	0.1920	0.0002	1.2141
copy_task	0.5068	0.5182	0.5477	**0.4705**	0.6714
discrete_pattern_completion	**0.0417**	0.0542	0.4125	0.1125	0.6659
discrete_postcasting	**0.0000**	**0.0000**	0.5400	**0.0000**	0.6684
selective_copy_task	0.4400	0.5400	0.3400	**0.3000**	0.6652
sin_forecasting	0.0019	0.0114	0.0026	**0.0000**	1.4813
sorting_problem	0.4850	0.5350	**0.2350**	0.4850	0.6675

*Decoder Only

PERSPECTIVES

- Currently, all memories are updated at each timestep. We could use an Adaptive Leak Rate, or take inspiration from Mixture of Expert to freeze most of the memories. That way, computations will be reduced and the model will be better to store information on the long term.
- BackPropagation Trough Time is really expensive. To be competive with Transformers, we must find a way to get rid of it.
- Is it able to generate text ? Work in progress...

References

[1] Vaswani et al., Attention is all you need.
[2] Geva et al., Transformer feed-forward layers are key-value memories.
[3] Jaeger, The "echo state" approach to analysing and training recurrent neural networks-with an erratum note.
[4] Kowsher, M., & Xu, J. Reservoir Transformer at Infinite Horizon: the Lyapunov Time and the Butterfly Effect.
[5] Kowsher, M., Khan, A. R., & Xu, J. Changes by Butterflies: Farsighted Forecasting with Group Reservoir Transformer.
[6] Jaegle et al., Perceiver: General perception with iterative attention.