Transformer needs NMDA receptor nonlinearity for long-term memory

Dong-Kyum Kim[1*]
Jea Kwon[2*]
Meeyoung Cha[1,3†]
C. Justin Lee[2†]

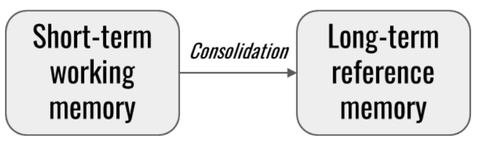
[1]Data Science Group, IBS
[2]Center for Cognition and Sociality, IBS
[3]School of Computing, KAIST

[*]Equal contributions, {kdkyum,jeakwon}.github.io; [†]Corresponding authors, {mcha,cjl}@ibs.re.kr

Abstract

The NMDA receptor (NMDAR) in the hippocampus is essential for learning and memory. We find an interesting resemblance between deep models' nonlinear activation function and the NMDAR's nonlinear dynamics. In light of a recent study that compared the transformer architecture to the formation of hippocampal memory, this paper presents new findings that NMDAR-like nonlinearity may be essential for consolidating short-term working memory into long-term reference memory. We design a navigation task assessing these two memory functions and show that manipulating the activation function (i.e., mimicking the Mg^{2+}-gating of NMDAR) disrupts long-term memory formation. Our experimental data suggest that the concept of place cells and reference memory may reside in the feed-forward network layer of transformers and that nonlinearity plays a key role in these processes. Our findings propose that the transformer architecture and hippocampal spatial representation resemble by sharing the overlapping concept of NMDAR-like nonlinearity.

Background: NMDAR in the hippocampal CA1 is required for memory consolidation and place cell representation

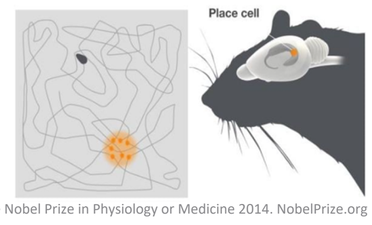


1. Memory consolidation

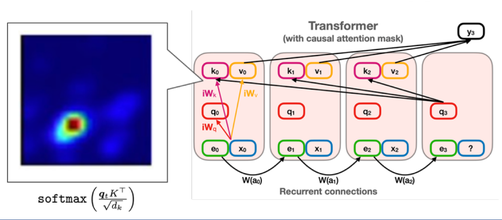
NMDAR in the hippocampal CA1 has been shown to be important for memory consolidation, which is a process that converts short-term memory into long-term memory [1].

2. Place cells in the brain and deep neural nets

- NMDAR in the hippocampal CA1 neurons has been shown to be critical for place cell formation [2].
- Banino et al. employed place cell representations for solving spatial navigation tasks in mazes with recurrent neural networks [3].



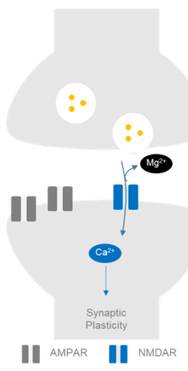
Credit: The Nobel Prize in Physiology or Medicine 2014. NobelPrize.org



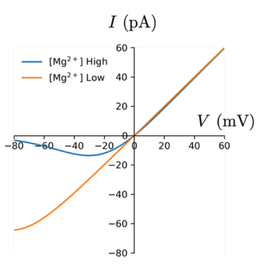
3. Transformer as a model of hippocampus

Whittington et al. showed interesting relationship between transformer and hippocampus. The place cells emerge in self-attention layer when the transformer model performs 2d navigation task [4,5].

1. We propose NMDAR-inspired activation function: modulation of α mimics Mg^{2+}-gating of NMDAR

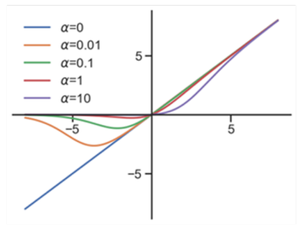


NMDAR IV curve

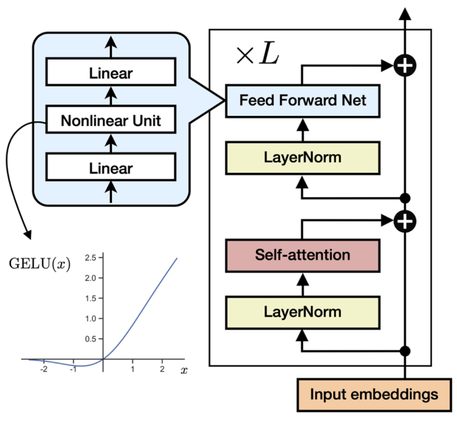


$$I = \frac{V}{1+\frac{[\mathrm{Mg}^{2+}]_0}{K_{\mathrm{Mg}}} e^{-\beta V}}$$

NMDAR-inspired activation function



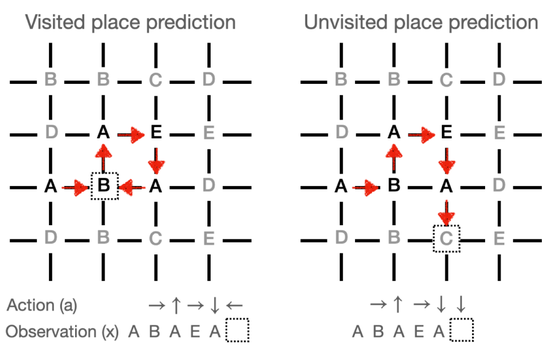
$$\mathrm{NMDA}_\alpha(x) = \frac{x}{1+\alpha e^{-x}}$$



Question

Can NMDAR-like nonlinearity in the transformer's feed-forward layer enhance the formation of long-term memory (consolidation) and spatial place cell representation?

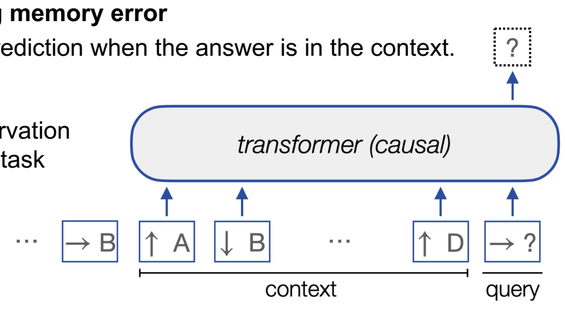
2. Method: reference & working memory error in next observation prediction task in 2D grid



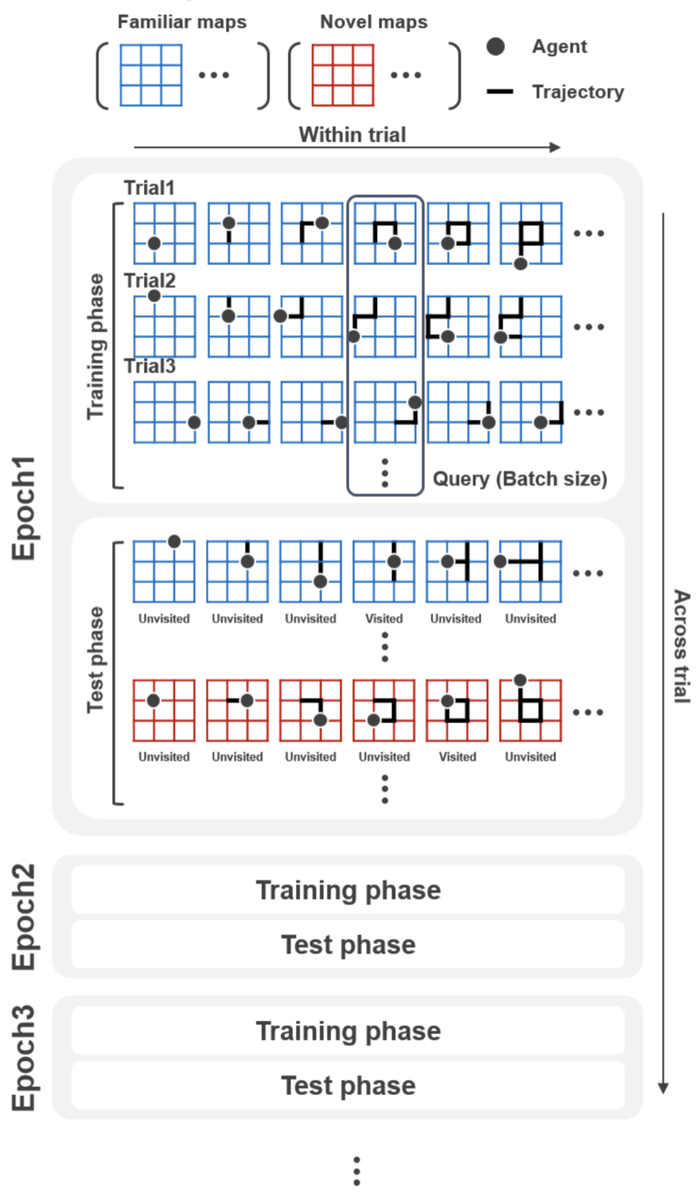
1) Reference memory error
Incorrect prediction when the answer is <u>not</u> in the context.

2) Working memory error
Incorrect prediction when the answer is in the context.

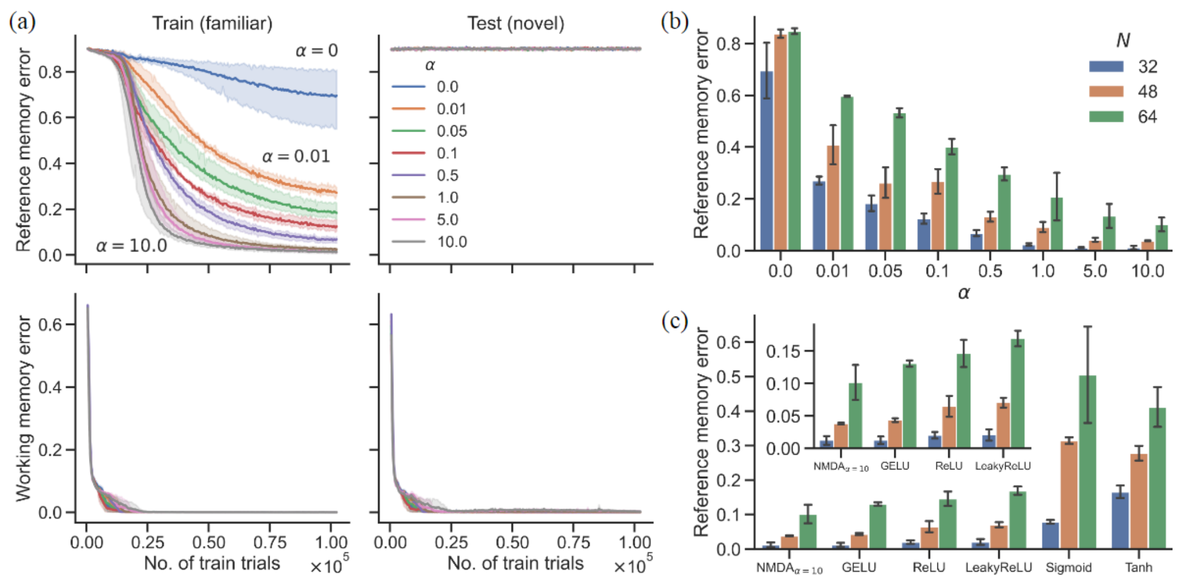
Problem
Next observation prediction task



3. Training & test process

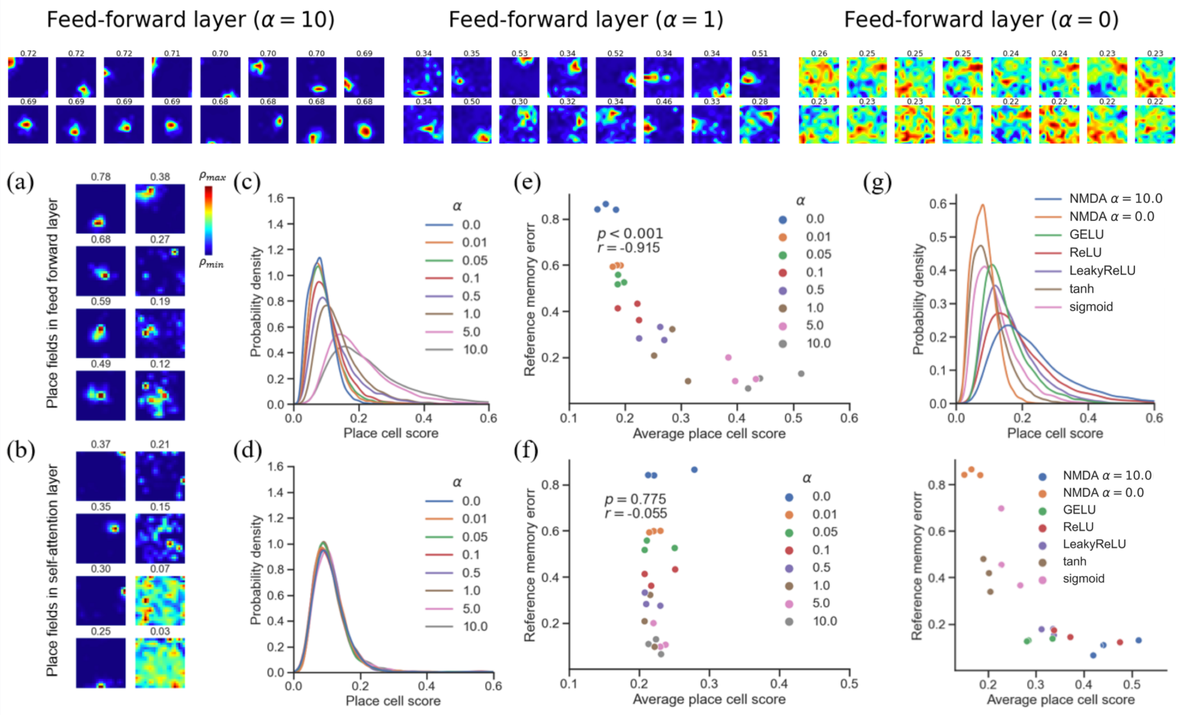


4. Reference memory can be enhanced by NMDAR-inspired activation function



- Working memory is intact while value of α changes.
- More training maps (large *N*) leads to bigger reference memory errors in that more maps require the model to store more pairs of 'what'-'where' memory.
- Our NMDA function improves long-term memory more than other commonly used nonlinear functions.

5. Relationship between reference memory and place cells in feed-forward nets & self-attention layers



- While average place cell scores in self-attention layers show no correlation with reference memory errors, neurons in the feed-forward layers exhibit a substantial correlation.
- These results imply that the reference memory formation and place cell emergence can be enhanced by NMDAR-like nonlinearity in feed-forward layers.

Summary

1. We proposed a new activation function that is inspired by the nonlinear dynamics of NMDA receptors in brain (**NMDA**α) where α mimics the Mg^{2+} concentration level.

$$\mathrm{NMDA}_\alpha(x) = \frac{x}{1+\alpha e^{-x}}$$

2. We developed a method for accessing the **reference memory**.
3. We evaluated the reference memory errors of **transformer** models with NMDAα. The results shows that *reference memory can be controlled by α*.
4. NMDAα with α = 10 shows the **best reference memory performance** when compared to other widely used nonlinear activation functions.
5. We demonstrated the emergence of **place cells** in feed-forward networks of transformer for the first time.
6. Reference memory is impaired when the value of α in NMDAα is low and this **resembles long-term memory loss in brain**; low Mg^{2+} concentration in brain causes long-term memory loss.

References

[1] The essential role of hippocampal CA1 NMDA receptor-dependent synaptic plasticity in spatial memory. Tsien et al., *Cell* (1996).
[2] Impaired hippocampal representation of space in CA1-specific NMDAR1 knockout mice. McHugh et al., *Cell* (1996).
[3] Vector-based navigation using grid-like representations in artificial agents. Banino et al., *Nature* (2018).
[4] The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. Whittington et al., *Cell* (2020).
[5] Relating transformers to models and neural representations of the hippocampal formation. Whittington et al., *ICLR* (2022).