TRANSFORMER NEEDS NMDA RECEPTOR NONLINEARITY FOR LONG-TERM MEMORY

Anonymous authors
Paper under double-blind review

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.

1 INTRODUCTION

In the hippocampus, NMDAR is regarded as an essential component that mediates synaptic plasticity, memory formation, and spatial representation (Li & Tsien, 2009; Tsien et al., 1996; Kentros et al., 1998). NMDAR serves as a switch for synaptic plasticity and long-term memory formation (Bliss & Collingridge, 1993; Slutsky et al., 2010; Miyashita et al., 2012). In addition, NMDAR has been highlighted for its importance in place cell representations in hippocampal CA1 (McHugh et al., 1996; Kentros et al., 1998). Place cells in the hippocampus (O'Keefe & Dostrovsky, 1971) and grid cells in the entorhinal cortex (Hafting et al., 2005) are thought to be crucial for spatial navigation in an animal. These discoveries have triggered recent efforts to replicate these spatial representations found in the brain by using deep neural networks (Banino et al., 2018; Cueva & Wei, 2018; Whittington et al., 2022).

In NMDAR depicted in Fig. 1a, the ion channels that reside in the post-synaptic region have unique characteristics that distinguish them from other ion channels in the brain. Their nonlinear dynamics are modulated by Mg^{2+} ion blockade at the pore region. NMDAR requires activity-dependent repulsion of Mg^{2+} ion (Nowak et al., 1984; Mayer et al., 1984) to be functional, and this phenomenon is partly interesting because it serves as a self-gating of ion influx in the post-synaptic region. In particular, the Mg^{2+} gated nonlinear dynamics of NMDAR plays a key role in synaptic plasticity and memory formation (Slutsky et al., 2010; Miyashita et al., 2012).

Recently, the relationship between the transformer (Vaswani et al., 2017) and hippocampal formation model has been reported (Whittington et al., 2022). The transformer is the most advanced deep learning model, showing unprecedented results in various tasks such as language modeling (Devlin et al., 2018; Brown et al., 2020), computer vision (Dosovitskiy et al., 2020; Radford et al., 2021), and art generation (Ramesh et al., 2022). This model has two consecutive modules, a self-attention layer and a feed-forward network (see Fig. 1b). Whittington et al. (2022) show the self-attention layer is closely related to the state-of-the-art neuroscience model (Whittington et al., 2020) and claim that softmax neurons in the self-attention layer behave like place cells in a navigation task. However, studies on the role of neurons in feed-forward networks have been absent.

We find an interesting resemblance of NMDAR nonlinearity with the Gaussian Error Linear Unit (GELU), a nonlinear activation function popularly used in the transformer's feed-forward network

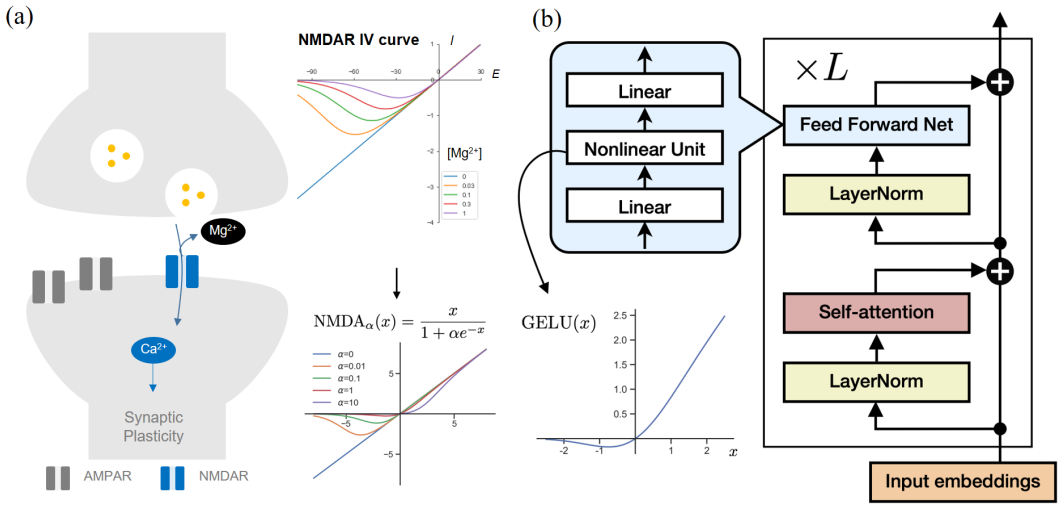

Figure 1: (a) Schematic diagram of Mg^{2+}-gated NMDAR modulating synaptic plasticity (left), its IV curve of current-voltage dynamics (right top) and NMDAR-inspired activation function, $\text{NMDA}_\alpha(x)$ (right bottom). (b) Transformer architecture and its feed-forward network's activation function, Gaussian Error Linear Unit (GELU; left bottom).

(Fig. 1). Similar to NMDAR's activity-dependent gating mechanism of ion influx, the form of the GELU function has a combination of input with the self-gating function. Biological experiments have shown the critical consequence of changing NMDAR's nonlinearity in synaptic plasticity and long-term memory formation (Slutsky et al., 2010; Miyashita et al., 2012), while the role of NMDAR-like nonlinearity in place cell representation remains unclear.

This work is inspired by the fascinating resemblance of NMDAR's nonlinearity dynamics with the GELU activation function and the recent model relating transformer's self-attention mechanism to hippocampal formation (Whittington et al., 2020; 2022). These findings motivated us to ask a question; **Is NMDAR-like nonlinearity in the feed-forward network layer of transformers essential for the formation of long-term memory and spatial place cell representation?**

To address this question, we propose a novel NMDAR-like activation function and design a spatial navigation task in a 2D grid environment that can assess two different memory types well formulated in neuroscience experiments (Olton et al., 1977; 1979): working memory and reference memory. Working memory controls the events from a within-trial, while reference memory controls across-trials from the unchanging environment. We evaluate the transformer model with the NMDAR-like activation function on this task; the results show that 1) place cell representations emerge in feed-forward networks, 2) the reference memory can be controlled by the nonlinearity of the NMDAR-like activation function, 3) the place cell in feed-forward networks is strongly correlated with the reference memory, while the place cell in self-attention layers has no correlation.

Our experimental data suggest that NMDAR-like nonlinearity in the feed-forward network layer of the transformer is essential for long-term memory formation and place cell representation.

2 Methods

2.1 Relating activation function in transformers with NMDAR nonliearities

NMDAR's nonlinear dynamics arises from the voltage-gated Mg^{2+} repulsion at the NMDAR channel's pore (Nowak et al., 1984; Mayer et al., 1984) (Fig. 1a left). Previously, Mg^{2+}-gated NMDAR open probability $\mathbf{p}$ has been shown to follow ion blockade model of Woodhull (1973):

$$\mathbf{p}_\alpha(x) = \frac{1}{1+\alpha e^{-\beta x}}, \tag{1}$$

where x represent an input voltage, $\alpha = [\mathrm{Mg}^{2+}]/K_{\mathrm{Mg}^{2+}}$ is a parameter determined by $[\mathrm{Mg}^{2+}]$, $K_{\mathrm{Mg}^{2+}}$ is a dissociation constant, and β is a temperature constant. Experimentally, increasing the Mg^{2+} level in the brain can enhance long-term memory formation (Slutsky et al., 2010). We observed the NMDAR's nonlinear dynamics of IV curve (Fig. 1a right top; current-voltage relationship) in the synapse to closely resemble the form of GELU activation function. GELU is a widely used activation function in transformers (Fig. 1b left bottom; $\mathrm{GELU}(x) \approx x\sigma(1.702x)$ where σ is the sigmoid function) (Hendrycks & Gimpel, 2016; Devlin et al., 2018; Brown et al., 2020). Inspired by this resemblence, we define a new nonlinear activation function (Fig. 1a right bottom) with α parameter which modulates dynamics as following:

$$\mathrm{NMDA}_\alpha(x) = x\mathbf{p}_\alpha(x) = \frac{x}{1 + \alpha e^{-x}}. \tag{2}$$

To investigate this NMDAR-like nonlinearity in transformer memory formation, we replaced the $\mathrm{GELU}(x)$ activation function with $\mathrm{NMDA}_\alpha(x)$ in a standard transformer model.

2.2 Transformers learn spatial navigation tasks

We train the transformer model to predict the subsequent sensory observations of an agent that randomly walks a 2D grid environment (Whittington et al., 2022) (Fig. 2). A sequence of previous [Action (a), Observation (x)] pairs are an input to the model, and the subsequent observation is masked for prediction. Instead of using positional encoding (Vaswani et al., 2017) that is commonly used in transformers, we employ the recurrent neural network (RNN) for encoding the sequence of actions (Whittington et al., 2022)[1].

We generate the embedding vectors of sensory observation (x) sequence with a word embedding layer, but the embedding vectors of the action sequence are generated by RNN; $e_{t+1} = \tanh(e_t W_a)$, where e_t is the positional embedding at step t, and W_a is the action-dependent trainable weight matrix. The input is given by $\{[x_1, e_1], [x_2, e_2], \ldots, [x_t, e_t]\}$; the initial positional embedding e_1 is sampled from a normal distribution and we mask the last observation x_t. We generate N maps of 11x11 2D grids. A random sensory observation among ten letters is placed at each position on each map. Agents can move 'up,' 'right,' 'down,' 'left,' or 'stay.' An agent starts at a random position and initiates a random walk on the map for 2,048 steps for each trial.

The model is trained with the softmax cross-entropy loss and predicts the subsequent sensory observations (i.e., dotted squares). We evaluate two types of memory: **working memory** and **reference memory**[2]. When the prediction on nodes that were previously visited during the random walking

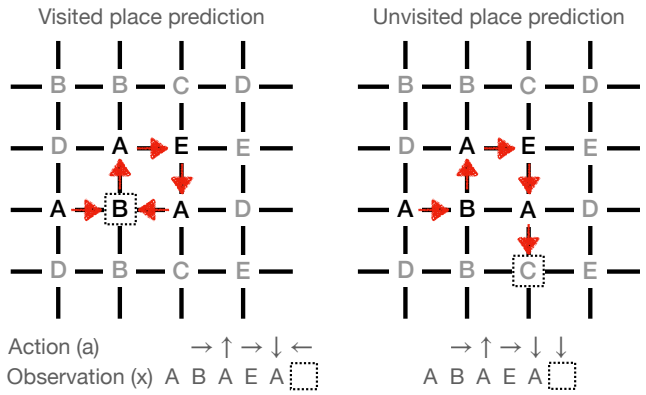

Figure 2: Sensory observation prediction task in a 2D grid, where dotted squares indicate the target position to predict given a sequence of past actions and observations. The unvisited (visited) places are represented in gray (black) letters.

[1]Encoding the actions of agents with RNN is closely related to the most advanced neuroscience model of the hippocampus.

[2]Whittington et al. (2022) only evaluated the working memory error based on our definitions of working memory and reference memory.

is incorrect, it will count as the working memory error (see Fig. 2 left). On the other hand, when the prediction on unvisited nodes is incorrect, it will count as the reference memory error (see Fig. 2 right). Minimizing the reference memory error by memorizing input sequences is infeasible; the possible number of sequence configurations is exponential since the input sequence is randomly generated at each trial. To solve this task, the model should be able to 1) understand the abstract structure of 2D space, 2) infer which map it is on from input sequence data, and 3) memorize what sensory observation is placed at each position in that map. We describe more training, evaluation, and transformer model details in the following section.

3 RESULTS

3.1 IMPLEMENTATION DETAILS

In our experiment, the feed-forward network (FFN) in the transformer model consists of two linear layers (Fig. 1b) with the NMDAR-inspired activation function NMDA_α (Eq. 2). We use TransformerXL (Dai et al., 2019) with an extended memory length of 32 and segment length of 32 so that working memory error is measured within a sequence length of 65 ($= 64 + 1$; 1 for the masked sensory input); i.e. a node that the agent had never visited within recent 64 steps is treated as an unvisited node. The input embedding is concatenated vector $[x, e]$ of the word embedding x (dimension of 256) and positional embedding e (dimension of 256) so that the total input embedding dimension is 512. The number of heads in the self-attention layer is 8 and the number of neurons in the FFN is 2,048. The dropout rate is set to 0.1 and the maximum clip norm of gradient is set to 0.25. We employ ADAM (Kingma & Ba, 2015) optimizer and a learning rate schedule with a linear decay from 0.0001 (start) to 0 (end). We run 512 random walk simulations in parallel for collecting training trajectories. The total number of random walking steps is 2,048 for each simulation so the total number of gradient steps for each run is 512 (batch size) $\times$ 2,048 (total number of steps in a trial) $\times$ 200 (number of trials). All runs are performed on a single NVIDIA TITAN V GPU.

3.2 WORKING MEMORY ERROR & REFERENCE MEMORY ERROR

To measure the impact of nonlinearity α in the FNNs, we train the transformer models with different values of α in $[0, 0.01, 0.05, 0.1, 0.5, 1, 5, 10]$ and evaluate the working memory and reference memory errors on the train maps (i.e., familiar maps) and test maps (i.e., novel maps).

Fig. 3a shows that the reference memory error on the train maps is rapidly decreased over train trials when α is larger than zero, with a larger improvement for increasing α. The reference memory error on the novel maps, however, is nearly constant at the chance level of 0.9 ($= 1 - 1/$(number of letters)) for all α. The inset plot of Fig. 3a shows that working memory is active on novel maps that had not been shown during training. This finding suggests that the reference memory is non-active for predicting the visited nodes on novel maps. Training the models on different numbers of maps N, Fig. 3b shows that increasing nonlinearity (i.e., α) helps activate

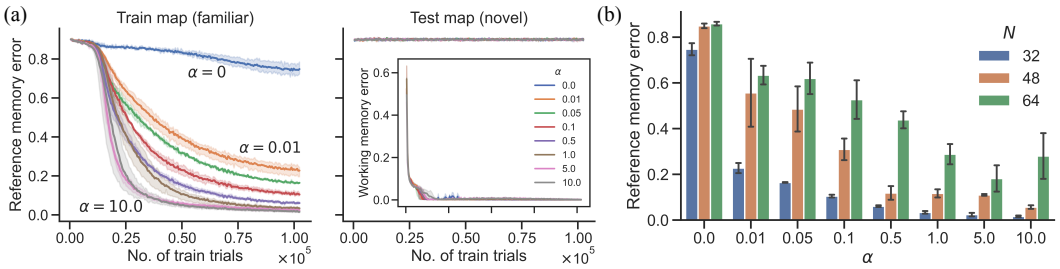

Figure 3: (a) Reference memory errors over training trials for training (familiar) maps and testing (novel) maps for $N = 32$ where N is the number of training maps. Inset: working memory errors on the novel maps over training trials. (b) Reference memory errors over different values of α and N. Error bars and shaded areas represent the standard deviation of errors from three independently trained models.

the reference memory, and the trend of improvement is consistently shown for $N = 32$, 48, and 64 cases.

3.3 PLACE CELLS IN FEED-FORWARD NETWORKS

Place cell is a neuron in the hippocampus which fires at a particular place of the environment (O'Keefe & Dostrovsky, 1971). Studies have shown that hippocampal place cells encode the spatial location through localized firing patterns. They have been considered a substrate for long-term memory of the location where specific events occurred (i.e., previously visited position in our navigation task). Selective impairment of NMDAR in hippocampal CA1 disrupts place cell emergence and long-term memory formation (Tsien et al., 1996; Kentros et al., 1998; McHugh et al., 1996).

We investigate the role of neurons in the feed-forward network layer and the self-attention layer by measuring the neuron's place specificity. We measure the place cell score by defining a $K \times K$ 2D grid environment as graph $G = (V, E)$ and building a sub-graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of all connected components from the source node i_{max} where the neuron fires maximally; directed edges of sub-graph $\mathcal{G}$ are generated by connecting high to low firing nodes. We run depth-first-search from i_{max}. Given G and $\mathcal{G}$, the place cell score is

$$\text{Place cell score} = \gamma \frac{\sum_{i \in \mathcal{V}} \rho_i}{\sum_{i \in V} \rho_i}, \tag{3}$$

where $\gamma = 1 - |\mathcal{V}^*|/|V|$ is a discount factor and $\mathcal{V}^*$ is a set of nodes from sub-graph without i_{max} and leaf nodes during depth-first search. ρ_i denotes a firing rate at node i. We record the firing rate ρ_i of neurons over a random walking trajectory with 10^5 steps in one of the training maps; then we

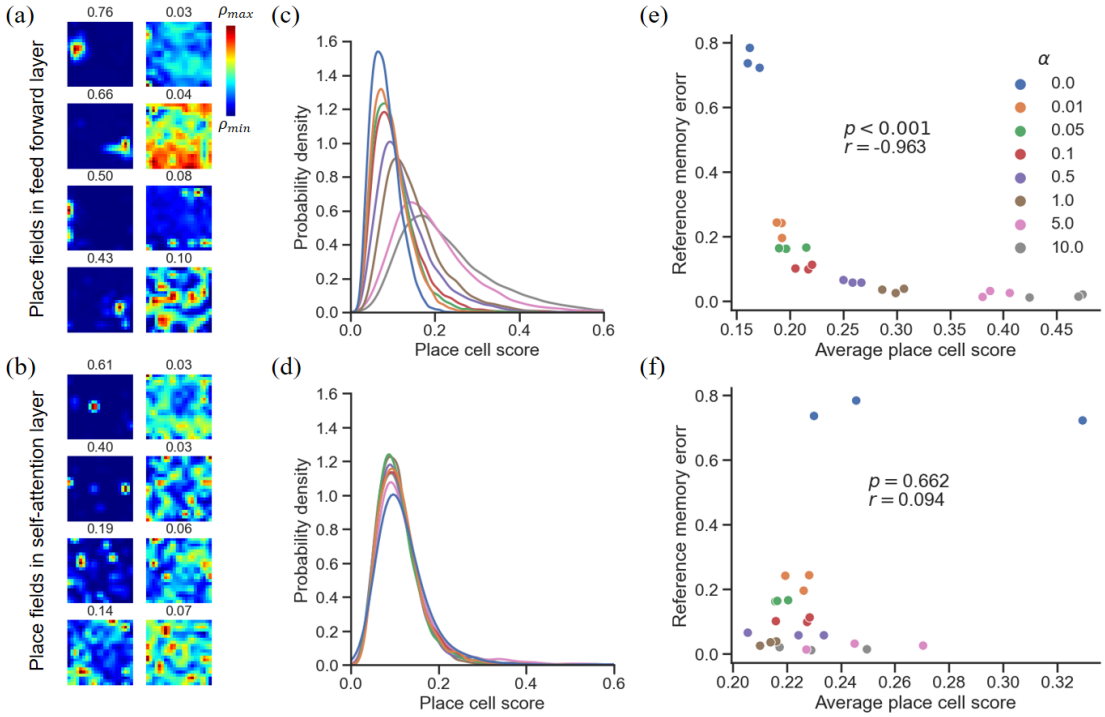

Figure 4: reference memory-related place cells selectively emerge in the feed-forward layer but not in the self-attention layer along with α increase. (a, b) Example rate maps with place scores in feed-forward layers and self-attention layers at $\alpha = 10$; from top left (high) to bottom right (low); color bar indicates the firing rate between ρ_{max} and ρ_{min}. (c-d) Place cell score distributions with varying α in feed-forward layers (c) and self-attention layers (d). (e-f) Scatter plot of average place cell scores and reference memory errors. r and p denote Spearman's rank correlation coefficient and significance score, respectively.

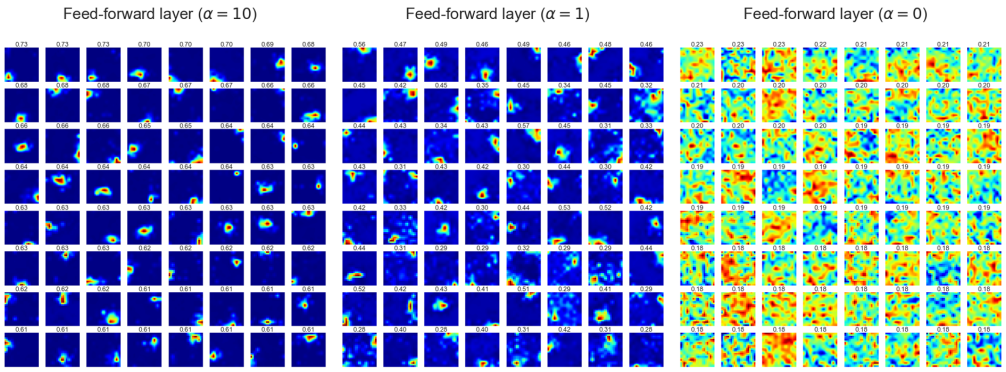

Figure 5: Rate maps of neurons with top-64 place scores in the feed-forward network with varying values of α; $\alpha = 10$ (left), $\alpha = 1$ (middle), and $\alpha = 0$ (right).

measure the place cell scores of neurons in the feed-forward network and self-attention layers. The place cell score is 1 when the neuron is firing only at a certain node; the score is 0 when the neuron is firing homogeneously across all nodes.

Fig. 4a and 4b show the rate maps of neurons with place cell scores in the feed-forward network and self-attention layers, respectively. For the self-attention layer, the total number of neurons in the softmax layer is 65 (number of sequence length + masked sensory observation) $\times$ 8 (number of heads) $\times$ 2 (number of layers). The total number of neurons in the feed-forward network layer is set as 2,048 (number of neurons) $\times$ 2 (number of layers). As can be seen, our metric well represents place specificity. Fig. 4c and 4d show the distribution of place cell scores in the two layers with different values of α. When the α value is increased, the place cell score distribution found in the feed-forward network layer becomes positively shifted (see Fig. 5 rate map examples for $\alpha = 0$, 1.0, and 10.0), whereas place cell score distribution in the self-attention layers remains.

In addition, Fig. 4e and 4f show a relationship between the average place cell score and the reference memory error for each α. While average place cell scores in self-attention layers show no correlation with reference memory errors whatsoever, neurons in the feed-forward network exhibit a substantial correlation. These results imply that the reference memory formation and place cell emergence require NMDAR-like nonlinearity in the feed-forward network.

4 Discussion and Conclusion

Searching for biological substrate of nonlinear activation function Rigorous previous efforts in finding the optimal nonlinear activation function underlie the great success of modern deep neural network models (Nair & Hinton, 2010; Hendrycks & Gimpel, 2016; Ramachandran et al., 2017). However, the neural substrates that mediate nonlinearity in the human brain and their role in intelligence have not been clearly understood. Our work is one of the first to put together the biologically inspired nonlinearity and its effect on long-term memory formation and the place cell representation in the previously described transformer model of the hippocampal formation. This idea was tested on a sensory observation task in the 2D grid environment and with the implementation of NMDAR-like nonlinearity. Our data indicated that NMDAR-like nonlinearity in the feed-forward network layer of transformers is necessary for long-term memory and spatial place cell representation.

On a related note, Whittington et al. (2022) showed that softmax neurons in the self-attention layer behave like place cells and demonstrated that changing the softmax function to linear slows the learning process in the working memory. However, the role of neurons in feed-forward networks has not been studied. We demonstrate for the first time that place cells could emerge in transformers' feed-forward networks, which we show by testing the emergence of place cells in feed-forward networks with an NMDA-inspired activation function. Even though there are trainable parameters in the self-attention layer, the quantitative analysis of the place cell score indicates that most of the reference memory is stored in feed-forward networks. Our results agree qualitatively with previous NMDAR impairment experiments from neuroscience: 1) hippocampal CA1 NMDAR perturbation does not

impair working memory (Lee & Kesner, 2002), 2) changing NMDAR Mg^{2+}-gating (changing α in this work) enhances or disrupts long-term memory formation (Slutsky et al., 2010; Miyashita et al., 2012), 3) NMDAR is required for long-term stabilization of newly forming place fields (McHugh et al., 1996; Kentros et al., 1998). Our contribution is at showing these patterns experimentally for the first time.

Short-term working memory and long-term reference memory Memories can be divided into short-term memory and long-term memory with respect to time Atkinson & Shiffrin (1968). Later, these two different memory systems were denoted as working memory and reference memory, concerning functional aspect Baddeley & Hitch (1974). In neuroscience, there is a consolidation theory that some short-term working memories are converted to a long-term reference memory system, and others fade out. There has been accumulating evidence that NMDAR is essential for memory consolidation in the hippocampal CA1 region (Kentros et al., 1998; Shimizu et al., 2000). Our work assessed the short-term working memory and long-term reference memory in the transformer's navigation task by measuring the visited error and unvisited error, respectively.

We have investigated the role of NMDAR-like nonlinearity with a novel consideration of activation function in a transformer model and have demonstrated the effects of altering the nonlinear dynamics of the activation function. Surprisingly, our result that loss of nonlinearity in NMDAR-like activation function ($\alpha = 0$) show impaired reference memory formation (Fig. 3) but intact working memory formation (Fig. 3b inset) seems to replicate the previous findings in neuroscience; selective inhibition of hippocampal CA1 NMDAR inhibition does not disrupt working memory (Lee & Kesner, 2002), but impairs the long-term memory formation (Tsien et al., 1996). These similarities provide an exciting possibility that our brain selectively updates short-term working memory into long-term reference memory by activity-dependent Mg^{2+}-gating of NMDAR.

Future directions Our research has exciting future directions. The current study only examined what-where memory using a sensory observation task in a static environment. However, our real-world environment is changing dynamically. Unfortunately, modern deep learning systems are generally incapable of adapting to a dynamic environments or reordering of sensory inputs. In future work, we intend to explore what-where-when memory, called *episodic memory*, in transformer and other deep models.

REFERENCES

Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pp. 89–195. Elsevier, 1968.

Alan D Baddeley and Graham Hitch. Working memory. In *Psychology of learning and motivation*, volume 8, pp. 47–89. Elsevier, 1974.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018.

Tim VP Bliss and Graham L Collingridge. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature*, 361(6407):31–39, 1993.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Christopher J. Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B17JTOe0-.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Association for Computational Linguistics*, pp. 2978–2988, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Clifford Kentros, Eric Hargreaves, Robert D Hawkins, Eric R Kandel, Matthew Shapiro, and Robert V Muller. Abolition of long-term stability of new hippocampal place cell maps by nmda receptor blockade. *Science*, 280(5372):2121–2126, 1998.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL `http://arxiv.org/abs/1412.6980`.

Inah Lee and Raymond P Kesner. Differential contribution of nmda receptors in hippocampal subregions to spatial working memory. *Nature neuroscience*, 5(2):162–168, 2002.

Fei Li and Joe Z Tsien. Memory and the nmda receptors. *The New England journal of medicine*, 361(3):302, 2009.

Mark L Mayer, Gary L Westbrook, and Peter B Guthrie. Voltage-dependent block by mg2+ of nmda responses in spinal cord neurones. *Nature*, 309(5965):261–263, 1984.

Thomas J McHugh, Kenneth I Blum, Joe Z Tsien, Susumu Tonegawa, and Matthew A Wilson. Impaired hippocampal representation of space in ca1-specific nmdar1 knockout mice. *Cell*, 87 (7):1339–1349, 1996.

Tomoyuki Miyashita, Yoshiaki Oda, Junjiro Horiuchi, Jerry CP Yin, Takako Morimoto, and Minoru Saitoe. Mg2+ block of drosophila nmda receptors is required for long-term memory formation and creb-dependent gene expression. *Neuron*, 74(5):887–898, 2012.

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

LPPAA Nowak, P Bregestovski, P Ascher, A Herbet, and Aa Prochiantz. Magnesium gates glutamate-activated channels in mouse central neurones. *Nature*, 307(5950):462–465, 1984.

John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 1971.

David S Olton, Christine Collison, and Mary Ann Werz. Spatial memory and radial arm maze performance of rats. *Learning and motivation*, 8(3):289–314, 1977.

David S Olton, James T Becker, and Gail E Handelmann. Hippocampus, space, and memory. *Behavioral and Brain sciences*, 2(3):313–322, 1979.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Eiji Shimizu, Ya-Ping Tang, Claire Rampon, and Joe Z Tsien. Nmda receptor-dependent synaptic reinforcement as a crucial process for memory consolidation. *Science*, 290(5494):1170–1174, 2000.

Inna Slutsky, Nashat Abumaria, Long-Jun Wu, Chao Huang, Ling Zhang, Bo Li, Xiang Zhao, Arvind Govindarajan, Ming-Gao Zhao, Min Zhuo, et al. Enhancement of learning and memory by elevating brain magnesium. *Neuron*, 65(2):165–177, 2010.

Joe Z Tsien, Patricio T Huerta, and Susumu Tonegawa. The essential role of hippocampal ca1 nmda receptor–dependent synaptic plasticity in spatial memory. *Cell*, 87(7):1327–1338, 1996.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

James C. R. Whittington, Joseph Warren, and Tim E.J. Behrens. Relating transformers to models and neural representations of the hippocampal formation. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=B8DVo9B1YE0`.

James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

Ann M Woodhull. Ionic blockage of sodium channels in nerve. *The Journal of general physiology*, 61(6):687–708, 1973.