Optimal Timing for Episodic Retrieval and Encoding for Event Understanding

Qihong Lu[1] qlu@princeton.edu **Zi Ying Fan**[2] kathy.zyfan@princeton.edu
Uri Hasson[1] hasson@princeton.edu **Kenneth A. Norman**[1] knorman@princeton.edu
1. Department of Psychology and Princeton Neuroscience Institute; 2. Computer Science Department, Princeton University

Abstract

When should an intelligent agent encode and retrieve episodic memories? In this work, we use a memory-augmented neural network to study how episodic memory can be most effectively deployed in the service of event understanding. Events are generated from underlying situation models and situations sometimes re-occur, making it useful to have an episodic memory system that can store and retrieve these situation models. For retrieval, our model learned to wait adaptively to accumulate information to ensure accurate retrieval of the target memory. Additionally, model variants that stored episodic memories at event boundaries (but not mid-event) had better subsequent recall performance. This latter result provides a normative explanation of the finding (from human fMRI) that the hippocampus is differentially engaged at event boundaries.

Keywords: episodic memory; event cognition; neural networks

When should we encode and retrieve episodic memories?

A fundamental challenge for models of episodic memory is understanding when encoding and retrieval should take place. In tasks that are specifically focused on memory (e.g., learning and recalling lists of randomly selected words), it is clear when encoding and retrieval should occur: Words need to be encoded during the list-learning phase and retrieved at test. However, in most real-life situations, where we are simply trying to understand and predict events as they unfold in a continuous fashion, it is less clear when to encode and retrieve episodic memories. Recent fMRI studies provide suggestive evidence that retrieval and encoding do not unfold uniformly over time. With regard to retrieval: In a recent fMRI study by Chen et al. (2016), participants viewed a two-part movie. For some participants, there was a one-day gap in between the two parts, so – at the beginning of the second part – participants had to retrieve the information from the first part to understand what would happen next. Participants in this condition showed a transient increase in cortical-hippocampal interaction, suggesting increased episodic recall, at the beginning of the second part of the movie (Chen et al., 2016). With regard to encoding: Multiple recent fMRI studies have found that, when humans are viewing naturalistic movies, hippocampal activity tends to peak specifically at event boundaries (Baldassano et al., 2017; Ben-Yakov & Henson, 2018). The size of these peaks was found to be correlated with subsequent memory for the just-completed event (Ben-Yakov, Rubinson, & Dudai, 2014; Baldassano et al., 2017; Silva, Baldassano, & Fuentemilla, 2019). While suggestive, these fMRI results only loosely constrain computational models of episodic memory. There is a strong need for more detailed modeling work addressing this issue of when to store and retrieve memories, so we can both explain these findings and also make more detailed predictions.

A model of cortical-hippocampal interaction

We propose that, during event processing, the cortex actively maintains a situation model (Radvansky & Zacks, 2017) in its distributed pattern of neural activity. This situation model is composed of features of the observed events (e.g. "location = a house", "person" = "Alice", "mood = happy", ...) that are useful for predicting what will happen next. Episodic encoding in the model corresponds to storing a "snapshot" of the pattern of cortical activity that represents the situation. Later, cortex can use partial information observed from events to retrieve previously-stored situation models from the hippocampus. For example, if the cortical pattern represents {"location = a house", "person" = "Alice"}, this might trigger hippocampus to recall a previously-seen situation: {"location = a house", "person" = "Alice", "mood = happy"}. If the retrieved situation matches the current situation, this will lead to an increase in accurate prediction. Conversely, if a mismatching situation is retrieved, this can reduce predictive accuracy.

Cortex is implemented as a long short-term memory (LSTM) network, which is a recurrent neural network with gating mechanisms. During event processing, it maintains the ongoing situation in its recurrent activity. Moreover, this network dynamically controls the parameters of the hippocampal network (described below).

Hippocampus uses the leaky, competing accumulator (LCA) model (Usher & McClelland, 2001)[1] to represent episodic memories as a set of leaky evidence accumulators with mutual inhibition (Fig. 1 B). The levels of leak, mutual inhibition, and input strength are controlled by the cortex. **Retrieval** is content-based: At time t, input to the LCA is proportional to the cosine similarity between the current cortical pattern and the stored cortical patterns corresponding to each unit; the outputs of the LCA correspond to the "recall strength" of each stored memory; these values are used as weights on stored memories to form the retrieved pattern (see Algorithm 1). **Encoding** a cortical pattern corresponds to forming a new node in the LCA that is tuned to that cortical pattern (Fig. 1 B). As different memories have non-overlapping representations, this approximates pattern separation.

[1] The LCA captures many important characteristics of memory retrieval. For example, see Polyn, Norman, and Kahana (2009).

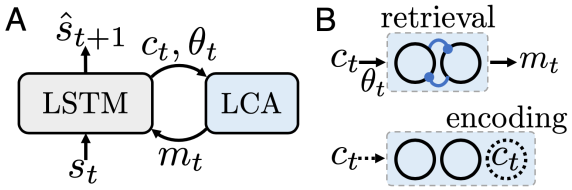

Figure 1: Model architecture. **A)** A model with a cortex, implemented as a LSTM, and hippocampal episodic memory, implemented as a leaky, competing accumulator (LCA) model. **B)** Retrieval in the LCA is controlled by the LSTM via θ_t; encoding corresponds to forming a new LCA node.

```
Algorithm 1: cortical-hippocampal interaction at time t
Input  : Current event $s_t$, previous cell state, $c_{t-1}$
Output: Predicted next event $\hat{s}_t + 1$
The LSTM control LCA parameters $\theta_t$, specifying input
  strength, leak and competition
Run LCA process with $c_t, \theta_t$ to get recall strength, $w$
Compute the retrieved item $m_t = w^\top M$, a weighted
  combination of all memories
Update LSTM cell state $c_t \leftarrow c_t + m_t$
If encode: Add a LCA node tuned to the features of $c_t$
```

Events as samples from a generative model

We represent event sequences as samples generated from an event schema, represented as a graph (Fig. 2 A). Each node on the graph is an event (e.g. Alice enters a house), and the edges represent event transitions. To generate an sequence, we first randomly sample a situation, or a set of feature-value pairs, such as {"location = a house", "person" = Alice , "mood = happy", ...}, which defines a path on the graph. Each transition on the graph (what happens next) is controlled by a particular feature of the situation. Thus, knowing the feature values makes it possible to predict what will happen next.

Recall/no-recall task : At time t, the model needs to predict the next event. The task has two phases (Fig. 2 B): During the **encoding phase**, the model sees k event sequences, sampled from the event graph. During the **retrieval phase**, we flush the cortical activity of the model. Then with $p = .5$, we present a previously-seen event sequence with a different order. We call this a **recall trial** (e.g. Fig. 2 D), because retrieving the target memory can help with event prediction; Otherwise, we present a new sequence. We call this a **no-recall trial**, because none of the stored memories are relevant.

Learning to predict with episodic memories

After being trained on the recall/no-recall task, the model learned to use episodic memory to predict upcoming events (Fig. 3 A). During no-recall trials, the prediction performance of the model linearly increases over time, because it gradually learns the feature values of the current situation from observations. During recall trials, prediction accuracy jumps up to near-ceiling levels early on, as a result of the model successfully recalling the target memory and avoiding recall of lures (i.e., memories from unrelated events; Fig. 3 B).

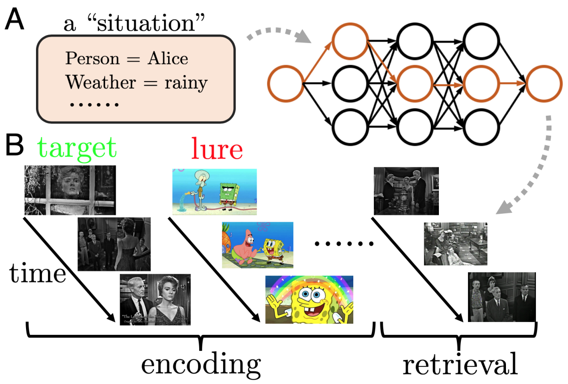

Figure 2: Task / Modeling environment. **A)** To generate an event sequence, we randomly sample a "situation", which defines a path on the event graph. **B)** A demo of a recall trial, where a previously-seen situation re-occurred; in this case, retrieving the target episodic memory helps event prediction.

Note the model learns to retrieve early in the sequence, but retrieval drops off later in the sequence (Fig. 3 B). The early peak in retrieval occurs because the model has a lot to gain (in terms of improved prediction) from retrieving stored episodes relatively early in the sequence. Later in the sequence, the model has been given a chance to observe more features of the situation directly, and there are fewer features left over that need to be "filled in" from memory. Thus, there is less to gain from retrieval, and possibly something to lose, as the model runs a risk of displacing its representation of directly-observed features with (potentially noisy) features from memory. In response to this, the model has learned to "shut down" retrieval later on. This provides a potential explanation of the aforementioned result from Chen et al. (2016), showing a transient peak in cortical-hippocampal interaction shortly after the resumption of an interrupted movie – right after the resumption, there is much to be gained from retrieval, but less so later on (when the participant is better-oriented).

Evidence accumulation during retrieval

Episodic retrieval is an evidence accumulation process, which exhibits a speed-accuracy tradeoff. Concretely, the model should retrieve as early as possible to improve its predictions, but retrieving too early in an event sequence is error-prone, since the model has not yet received enough information to reliably identify which stored episodic memory (if any) matches the current situation. This tension between "need to predict" (pulling recall earlier) and "risk of error" (pushing it later) suggests that it should be possible to push around the optimal time of recall by manipulating prediction demands. Specifically, if we give the model a "grace period" where it can ob-

Error bars for all figures indicate 3 standard errors.

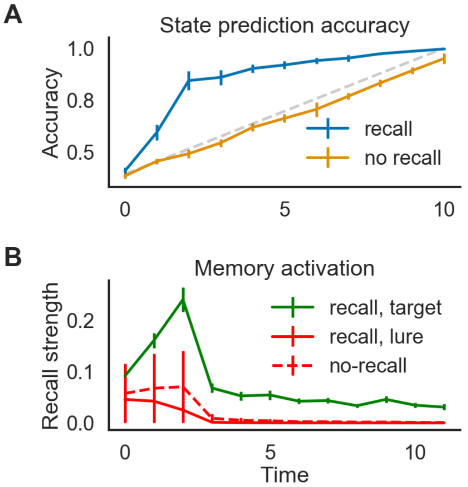

Figure 3: Model behavior on the recall/no-recall task. **A)** Prediction accuracy is higher for recall trials, demonstrating that the model learned to use episodic memory to support event prediction; the dashed line represents the expected performance if the model has no memory. **B)** Target memories are more activated than lures.

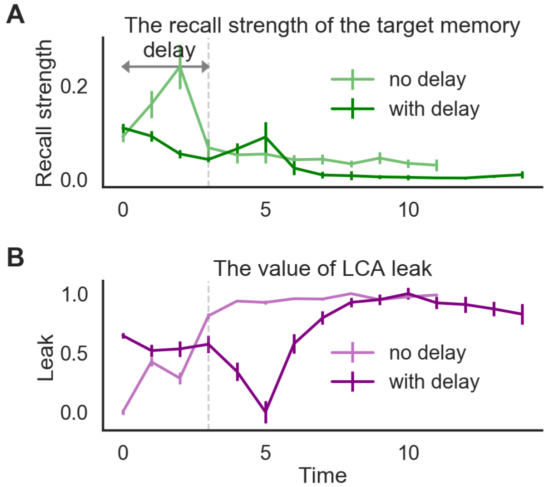

Figure 4: The model learned to accumulate information to ensure correct retrieval when prediction demand is delayed in time. **A)** The activation (recall strength) of the target memory. **B)** The leak value (a LCA parameter controlled by the cortical network) over time. Smaller leak values indicate stronger retrieval.

serve features of the situation without having to make predictions (and risk being wrong), the optimal policy is to accumulate information and suppress retrieval until the grace period expires and the model is forced to predict.

To test if the model can learn to wait adaptively, we trained models in environments where there is no prediction demand for the first few time steps. In this case, the peak of target memory activation moved to later time points and peaked again once the model was forced to predict (Fig. 4 A). Mechanistically, the cortical network achieved this by modulating the leak value of the LCA, which governs the retrieval process of the hippocampal network. Decreasing leak makes episodic retrieval easier, so a negative deflection in leak facilitates retrieval. When prediction demand is delayed, the negative deflection in leak moved to later time points (Fig. 4 B). In this experiment, the model is configured to encode memories only at event boundaries, which is justified in the next section.

This result shows that the model can flexibly balance between speed and accuracy during retrieval. Moreover, this leads to a testable prediction that (in people) retrieval should be modulated by prediction demand. That is: Episodic retrieval should only occur when it is needed to support predictions; otherwise, the person is better off waiting and accumulating more information, which will help them better specify which memories are (and are not) relevant.

Encoding memories at event boundaries benefits subsequent retrieval

As noted earlier, recent studies suggest that event boundaries play a special role in event storage: Hippocampal response peaks tend to align with subjectively annotated event boundaries (Baldassano et al., 2017; Ben-Yakov & Henson, 2018), and these peaks predict subsequent recall of the just-completed episode (Ben-Yakov et al., 2014; Baldassano et al., 2017; Silva et al., 2019). However, current theories do not provide a normative account of why encoding should be specific to event boundaries (vs. also occurring within events).

To address this question, we used the model to compare two different encoding strategies (illustrated in Fig. 5 A): i) **encoding at event boundaries**, where the cortical state is stored at the end of the sequence, but not beforehand; ii) **cumulative encoding**, which stores the cortical state at the end of the sequence, but also at regular intervals beforehand.

We found that encoding (only) at event boundaries leads to the best subsequent event prediction performance on the recall/no-recall task (Fig. 5 C). The suboptimal performance of the cumulative encoding model results from storing episodic memories at sub-event level; these ”incomplete” memories increase susceptibility to false recall caused by partial matching. Consider the toy example in Fig. 5. During the encoding phase, the model sees an event sequence: {“location = a house”, “weather = sunny”, “mood = happy”}. Later, the model observes a second, partially-overlapping sequence: {“mood = sad”, “location = a house”, “weather = rainy”}. If the first sequence was stored as a single, complete memory, the model will successfully avoid recalling the first situation during the second sequence, as the very first observation {“mood = sad”} mismatches the content of the stored memory. Now, consider what happens if – in addition to storing the complete memory – the model also stored an incomplete memory of the first sequence at the subevent level: {“location = a

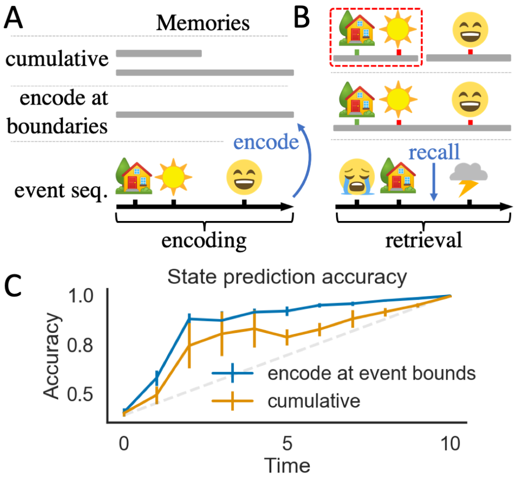

Figure 5: Encoding event sequences in a single episodic memory enhances mismatch detection. **A)** The resulting memory chunks under the two encoding regimes; **B)** Storing incomplete chunks at the sub-event level can cause subsequent false recall (boxed in red). When the model has partial knowledge about the current situation (e.g. "location" = "a house", "mood = sad"), lures are easier to reject if all information is connected. **C)** Models that encoded episodic memories at the end of an event (and not before) had better event prediction performance.

house", "weather = sunny"} (Fig. 5 B, red box). This incomplete memory can not be rejected based on mismatch, hence the model might recall it and wrongly predict that the weather will be sunny. Thus, encoding at event boundaries (but not beforehand) facilitates target-lure disambiguation during subsequent retrieval, which (in turn) benefits event prediction.

Conclusion

We proposed that during event processing, the cortex actively maintains a situation model of the ongoing events, and the hippocampus stores and retrieves these situation models. We instantiated this idea as a memory-augmented neural network and showed how the cortex can learn to interact with the hippocampus and leverage episodic memories to support event prediction. In particular, our simulations showed that i) during retrieval, cortex learns to adaptively trade off between waiting to accumulate information about the current situation versus retrieving episodic memories to support event prediction; ii) encoding at event boundaries produces event memories with a more complete specification of the features of the situation, which makes target-lure disambiguation easier during subsequent retrieval.

Our results may have useful implications for machine learning (ML). Memory-augmented neural networks are being used increasingly often in ML research (Pritzel et al., 2017; Ritter et al., 2018). Our simulations show that – for sequential prediction tasks like the one modeled here – optimizing the timing of encoding can lead to substantial performance benefits.

In the future, we would like to extend these principles of event memory to more realistic environments, where event sequences have hierarchical structure spanning multiple timescales, and are generated by multiple event schema that are potentially compositional and non-stationary. Also, a key limitation of this work is that – while we demonstrated better performance when encoding was limited to event boundaries – the model did not learn when to encode on its own. We are presently extending the model with a reinforcement learning objective to optimize encoding policy.

Acknowledgments

This work was supported by a Multi-University Research Initiative grant to KAN and UH (ONR/DoD N00014-17-1-2961).

References

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721.e5.

Ben-Yakov, A., & Henson, R. N. (2018). The hippocampal film editor: Sensitivity and specificity to event boundaries in continuous experience. *The Journal of Neuroscience*, *38*(47), 10057–10068.

Ben-Yakov, A., Rubinson, M., & Dudai, Y. (2014). Shifting gears in hippocampus: temporal dissociation between familiarity and novelty signatures in a single event. *The Journal of Neuroscience*, *34*(39), 12973–12981.

Chen, J., Honey, C. J., Simony, E., Arcaro, M. J., Norman, K. A., & Hasson, U. (2016). Accessing real-life episodic information from minutes versus hours earlier modulates hippocampal and high-order cortical dynamics. *Cerebral Cortex*, *26*(8), 3428–3441.

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, *116*(1), 129–156.

Pritzel, A., Uria, B., Srinivasan, S., Badia, A. P., Vinyals, O., Hassabis, D., ... Blundell, C. (2017). Neural episodic control. *Proceedings of Machine Learning Research*, *70*, 2827–2836.

Radvansky, G. A., & Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences*, *17*, 133–140.

Ritter, S., Wang, J. X., Kurth-Nelson, Z., Jayakumar, S. M., Blundell, C., Pascanu, R., & Botvinick, M. (2018, May). Been there, done that: Meta-Learning with episodic recall. In *Proceedings of the international conference on machine learning.*

Silva, M., Baldassano, C., & Fuentemilla, L. (2019). Rapid memory reactivation at movie event boundaries promotes episodic encoding. *bioRxiv*.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592.