Biologically-Inspired Deep Predictive Learning for Episodic Memory Event Segmentation

A THESIS PRESENTED
BY
ZERGHAM AHMED
TO
THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS
IN THE SUBJECT OF
COMPUTER SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2022

Thesis advisor: Professor Gabriel Kreiman Zergham Ahmed

Biologically-Inspired Deep Predictive Learning for Episodic Memory Event Segmentation

Abstract

Computational modeling for event segmentation in the literature has been directed towards detailing individual regions of the brain, mainly the hippocampus, prefrontal cortex and substantia nigra and the role they play in episodic memory formation. Modern predictive recurrent neural networks, which have been shown to perform well on naturalistic video frame prediction, have the potential to explain critical properties of neuronal responses and perception during event segmentation. However, they have not been examined in this light. This thesis explores the ability of state-of-the-art deep predictive learning approaches to explain properties of event segmentation. We identify similar artificial neural responses in our model and biological neural responses from Intracranial Electroencephalography data. These responses are compared in the context of detecting separation between different events in naturalistic video data input. Finding such parallels could provide insight for improving the biological plausibility of deep learning networks. Furthermore, such computational models can serve as a biological proxy and a testing ground for mechanistic hypotheses that bridge neural computation to observable behavior.

Contents

To finding meaning in life.

Acknowledgments

I am very thankful to Professor Gabriel Kreiman for giving me the opportunity to join his lab. He has been very welcoming and helpful in guiding me towards a direction of research that I enjoy greatly. Being a part of the Kreiman lab has been one of the experiences that I cherish the most of my undergraduate career.

I am also extremely thankful to Dr. Jie Zheng without whom this project would not have been possible. The initial idea for this work came directly from her and she worked closely with me to adapt it and narrow its scope. Jie dedicated a great deal of time, meeting with me a couple of times a week and always being there for me. Jie has taught me so much about how to conduct research from finding and developing an idea to analyzing and documenting results. She truly went above and beyond by giving me guidance, encouraging and inspiring me to pursue graduate studies. I am very lucky to have a mentor like her.

Finally, I am thankful for all the faculty and friends at Harvard that I have made including my blockmates that I lived with who have always supported me through the toughest of times: Aidan Carey, Christian Tabash, Gerry Ascensio, James Caven, Patrick Magahis, Kidus Negesse, Thomas Burr.

Listing of figures

Listing of tables

1

Introduction

Previous theoretical and experimental studies have shown that the Hippocampus (HPC), prefrontal cortex (PFC) and substantia nigra (SN) play critical roles in episodic memory formation and event segmentation [7, 16], a process of partitioning continuous experience into temporally organized sequences of events. This project explores a multi-network deep learning model called PredNet that has similar compartments [8] to find whether its modules, mainly the analagous HPC module, plays a similar role in event segmentation.

The main purpose of this project is to explore whether current predictive recurrent neural networks used for unsupervised learning can explain critical properties of biological neuronal responses and perception during event segmentation. We start from the comparison between the prediction module of PredNet and the HPC since the HPC region has been proposed to be critical for prediction error [3]. Comparably PredNet propagates prediction errors throughout its network. The architecture of PredNet will be described later. To answer our question of whether predictive recurrent neural networks can explain event segmentation, we first ensure that the model behaves similarly to the brain at the neuronal level. One key neuronal behavior we use for comparison comes from recent work that shows the existence of hippocampal neurons that can differentiate between three different types of event boundaries [20]. Therefore, we attempt to find whether our model can differentiate between these three types of event boundaries as well and whether specific artificial neurons are responsible for identifying this differentiation. We begin by using PredNet since it incorporates temporal and spatial predictive coding principles [8] into its deep learning framework and performs well on naturalistic data in the form of video clips. Thus, PredNet provides an opportunity to test prediction error and find the artificial and biological neuronal similarities in the context of event segmentation of naturalistic stimuli which is the primary data perceived by biological networks. The motivation for comparing neuron responses at event boundaries to artificial neuron responses follows up from work showing the existence of neurons, referred to as boundary cells and event cells, that demarcate different types of episodic transitions [20]. Patients were given a task of viewing Stimulus set 1, composed of 90 different video clips containing three types of event boundaries or in other words event transitions: No boundary, Soft Boundary, and Hard Boundary. Patients were then evaluated on scene recognition and time discrimination memory tests for the clips. During these tests, their neural activity was recorded. The results showed that certain neurons, referred to as boundary cells (B cells), transiently increased their firing rates after the occurrence of Soft and Hard Boundaries in video clips. Another set of neurons, referred to as event cells (E cells) transiently increased their firing rate only in response to Hard Boundaries [20]. These results are important as they highlight the neural activity during the event segmentation process, specifically at the boundaries of where events are partitioned. Finding corresponding artificial neuron activation in our computational model allows us to evaluate the similarity between the biological and artificial networks at the single neuron level. For example, we can do this by comparing the intracranial electroencephalogram (iEEG) recordings of neuron responses at the event boundary to the artificial neuron responses in the layers of the deep learning model. The list of similarities we examined for are detailed in the experiments section.

1.1 Our Contributions

The novel contributions of this work are that it addresses the 1) functional role of the HPC system in computational modeling of event segmentation and attempts to establish a connection between an analogous artificial deep learning network. Provided that our computational model is able to reproduce the neural results identified from iEEG recordings and analogous neuronal activation to B and E cells are found, our deep learning network can serve as a mechanistic hypothesis testing ground for linking neural computation to behavior. By matching the single neuron activity and eventually circuit-level activity, our model can serve as a plausible proxy for biological circuits. In turn, manipulation of the components of the model through computational "lesions" can be used to suggest electrical stimulation of the real HPC-PFC-SN neural circuit. An example of this lesion process would be manipulating one component of the deep learning network, for instance the deviance detection module, and in turn analyzing its performance and neural activity to see how it responds to event segmentation phenomena (such as detection of event transitions). The insights from this can be used to suggest electrical manipulation of the biological HPC-PFC-SN circuit or experiments on individual regions.

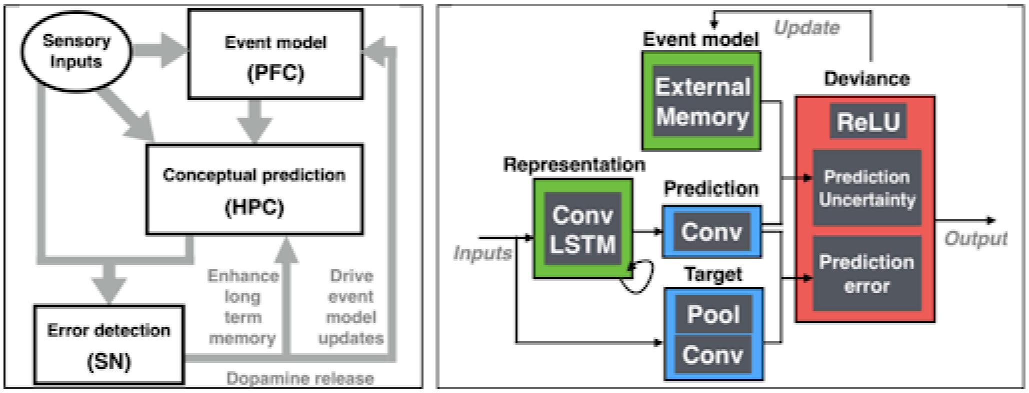

Figure 1.1: A diagram of the extended PredNet network compared to the analogous biological Network. The left diagram shows the biological HPC-PFC-SN network involved in event segmentation, while the right diagram shows the similar modules in PredNet which have the same functional roles.

In sum, we focus on two goals which will tell us whether predictive neural networks can explain critical properties of neuronal responses during event segmentation:

1. Neuronal behavior, specifically that it can differentiate between the NB, SB, and HB.
2. Neuronal functionality, such that there is a similar activation profile as the neuronal responses of E and B cells as the brain.

The first goal will tell us that the deep learning model is performing or behaving similarly as the biological neurons. The second goal will tell us whether the compartments of the deep learning model has similar functionality. By keeping track of the first goal we know that the model is explaining the same properties of event segmentation while the second goal highlights the internal workings of our model so that we may change it to better mimic the biological network. By better mimicking the biological network, the artificial network will have more potential to explain the same properties and match the performance of the biological one.

1.2 Thesis Outline

Chapter 2 will review literature that provides background on the process of event segmentation and work that deals with existing computational models for it. **Chapter 3** will detail the experiments we conducted to search for similarities in the biological and artificial networks as well as baseline performance tests. **Chapter 4** will discuss the future implications of the work and conclude with further directions for research.

2

Related Work

2.1 Theoretical Background

We encode a continuous stream of perceived experience into a sequence of segmented events [19]. Learning the underlying themes and associations between these events, we use them to guide future behavior. Episodic memory is defined by this characteristic [19]. Due to this functionality, event segmentation serves as a powerful predictive tool [1, 17]. Event segmentation has been shown to lead to better recall memory [15], even up to one month later [4]. The associations and themes learned from memories in order to segment them are referred to as event models [19]. Event Segmentation Theory is a key mechanistic hypothesis underlying event segmentation [19]. It says that the working event models are updated in response to transient increases in prediction error [15]. Many computational models incorporate this hypothesis in some way, including PredNet [8]. Event segmentation is also important due its effects on Alzheimer's disease and maintaining memory [18]. Considering all of this information, we assume that Event Segmentation is an important biological characteristic of memory.

This work will deal with video clips which can represent the continous stream of perceived experiences. The segmentation of this clip into separate events will be denoted by the event transition as given by the boundary type of the clip (either HB, SB, or NB as we will see later on).

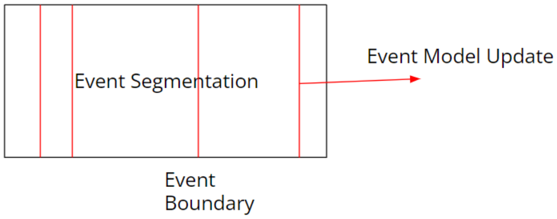

Figure 2.1: The Event Segmentation Process. This is a basic representation of the event segmentation process. We can take the box to represent a video clip with the red lines representing the event transition which demarcates two different scenes. The creation of each event transition can be thought of as the result of the updated event model which can be represented by error propagation or more sophisticated methods in computational models.

2.2 Computational Models in Literature

PredNet is a deep learning approach using multiple neural networks and utilizing predictive coding [13] from neuroscience by propagating error throughout the layers. PredNet includes repeated and stacked models with memory, prediction and deviance detection components. It operates by forming a recurrent representation of each frame using a convolutional long short-term memory network (Conv LSTM) and then sending those representations to a deep convolutional network to make predictions. The network also incorporates a deviance detection module which calculates the prediction error and prediction uncertainty. An error signal is propagated forward through a convolutional layer to become an input to the next layer [8]. This follows the predictive coding principle from neuroscience [13].

PredNet has also been shown to capture single-unit response dynamics in the visual cortex [9], which further supports its potential for providing similar neural responses for event segmentation.

The structured event memory model (SEM) has a similar multi-network model with analogous components to the biological circuits involved in event segmentation [5]. This model uses probabilistic generative approaches to represent the dynamics of events. SEM has also been shown to perform well on naturalistic video data [5]. SEM is focused on scene representation and event dynamics. However, many of its features incorporate features such as vector representation of scenes that are suited for representing event dynamics.

Other computational models such as [14, 10, 11] explore gating signals to work with transient increases in prediction error [14]. Some models [10], also specifically incorporate the LSTM into their network [6]. This LSTM is a recurrent neural network (RNN) with gating mechanisms [10].

3

Neural Behavior Experiments

3.1 Methods

Our computational model, initially PredNet, is given the task of next-frame video prediction for naturalistic datasets. During this task, we identify artificial neuron activation at event transitions in the video clips to find similarities between neural responses of B and E cells [20]. The datasets we use are stimulus set 1 and 2 which will be detailed below. These stimulus sets were given to patients during frame recognition and temporal order tasks while recording B cell and E cell activity. Therefore, we use this dataset in order to make a direct comparison with the performance and behavior of PredNet. We use the error metric Mean Squared Error (MSE) to assess the difference between the actual next-frame and the next-frame predicted by PredNet. We proceed with the hypothesis that PredNet will associate event boundaries with high prediction error, in other words MSE near the event transition will be high relative to other frames.

3.2 Datasets

Stimulus set 1 includes clips with No Boundary (NB), Soft Boundary (SB), and Hard Boundary (HB). NB clips do not have any labeled event transitions, while SB clips have an event transition in the form of changing camera angles while in the same general setting of the clip. HB clips transition to a completely different clip at the event transition. There are 30 of each of these types of clips with varying frame rates. Stimulus set 2 includes 25 raw clips with no boundaries as well. These clips are longer in time length which allow us to see how the MSE fluctuates over its frames. Each video clip is around 30 seconds long, with a frame rate of 29.38 frames/second and with a frame width and height of 1920 by 1080. Code was written for frame extraction to extract all the frames for each clip in the stimulus sets. We utilized the downscale function from the PredNet repository, to transform the frames into valid inputs of size 128 x 160 into PredNet. The stimulus sets were used as input in the PredNet model and the MSE was returned. The MSE was then plotted over the number of frames which corresponded to the number of frames in the clip. This gives us an understanding of

PredNet's prediction error over the course of the clip. Different scripts were used to generate graphs that summarized the neural behavior of PredNet. These summary graphs will be described in the next sections.

3.3 Event Transition Detection

We proceed with our hypothesis that PredNet will associate event boundaries with high prediction error, in other words MSE near the event transition will be high relative to other frames. Therefore, the SB clips should have higher MSE near the soft boundary as should HB clips. The MSE should be relatively smaller when not crossing these boundaries. After crossing them, MSE should become a local maximum and then gradually decrease to recover. However, the MSE for NB should fluctuate and be less predictable. Our initial results show that the PredNet model follows our expected results, namely that the MSE is highest at the event transition and the frame following it for SB and HB clips. The MSE reaches maximum at this point and the frames following the event transition remain with high MSE values which gradually decrease and stabilize for both SB and HB clips. Most of the MSE plots start out with relatively low MSE and then increase significantly at the event transition, then gradually decrease for both of these clip types. For NB clips, the graph fluctuates and there is not as clear a pattern. However, we saw that MSE was higher at beginning and end points for NB clips. Comparing the graphs to the clips in video format, we saw that points where unlabeled event transitions occurred had greater MSE. The clips with no boundaries or cuts in stimulus set 2 show the same behavior to this. Figure 2 shows example graphs that include labels pointing to

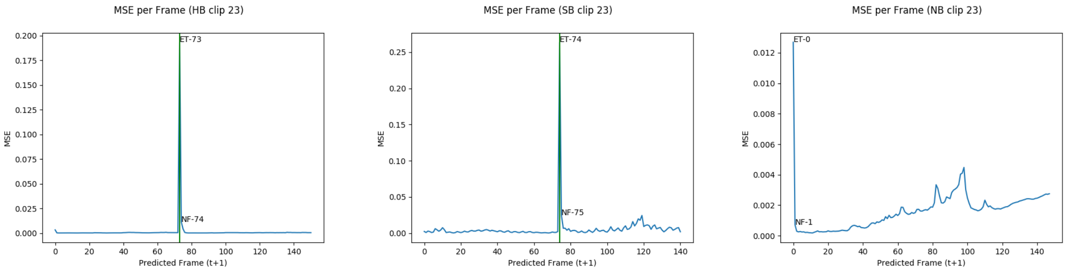

Figure 3.1: Human Labeled vs PredNet Labeled Event Transition This figure shows each clip type with the human labeled event transition marked with a green vertical line and the PredNet predicted event transition frame written in black near it. The NB clip does not have event transitions and thus does not have a human label.

the event transition and the next-frame after it. This was the general shape of the graphs for each clip. These results imply that our hypothesis that the MSE of the prediction will be higher at the event boundary holds. Thus, this suggests that PredNet is able to detect and respond to event transitions.

Another feature to note of these graphs are the green vertical line for clips in stimulus set 1 and the multiple colored vertical lines for clips in set 2. These vertical lines show the actual labeled event transition marked by a human. Note that for most of the graphs these labeled eent transitions are close to the maximum MSE of the prediction which is also labeled in the graph. Therefore, we can think of this as comparing the actual event transition with the event transition calculated by PredNet. For stimulus set 2, since it is composed of continuous 30 second clips, have actual event transitions marked relative to how the scene is changing as determined by the volunteer labeling it. Note that these transitions vary with respect to the MSE. However, for some of the graphs one of them is located relatively close to a local maximum. This comparison of actual and predicted event transition is summarized in the graph below.

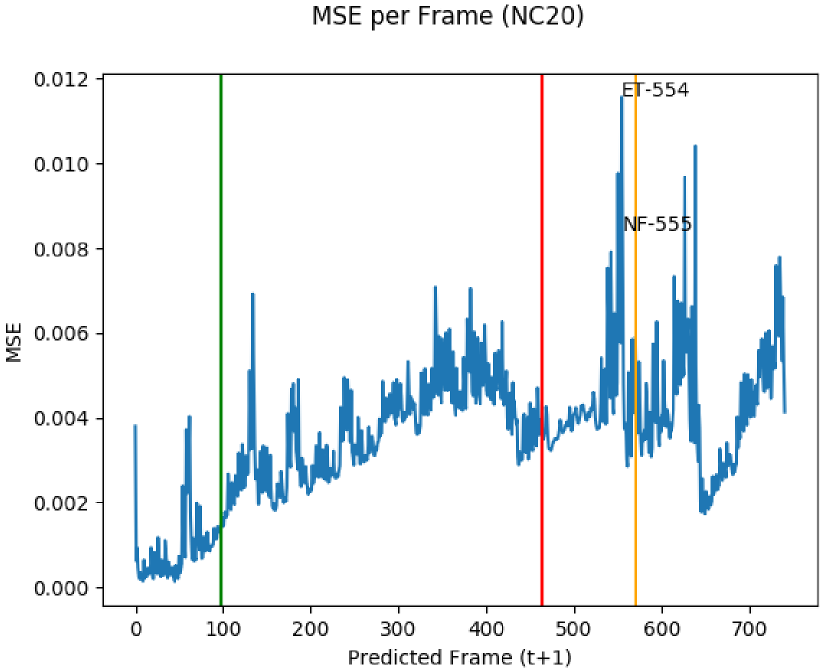

Figure 3.2: Graph of MSE per frame for a Stimulus set 2 clip. This graph shows the MSE over time for a clip with No Cuts which is the condition as the NB clips in Stimulus set 1.

The results of **figure 3.3** show that the predicted event transition varies a lot for the no boundary clips of set 2. It shows that the SB and HB clips have correctly predicted event transitions. Some of the values appear to be negative which indicate that the actual event transition was before the maximum MSE output by PredNet. These results imply that PredNet is able to demarcate the transition between events comparable to those labeled by humans as shown by the low frame difference for SB and HB. Another metric we note is the maximum MSE for every clip. This is shown in the MSE magnitude graph below. The results show that MSE is highest in order of HB clips, then SB, and lastly NB. This also supports our hypothesis since we expect error to be greater for HB clips which introduce a completely different clip and would be more difficult for PredNet to predict. Whereas,

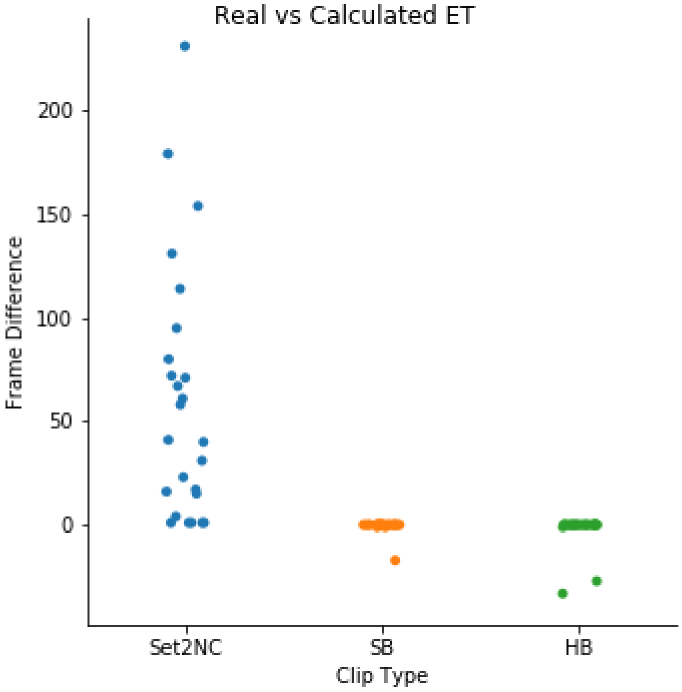

Figure 3.3: Stimulus set 1 Actual and Predicted Event Transitions. The frame difference between actual human labeled event transition and the event transitions predicted by PredNet (based on Maximum MSE) for all clips in stimulus set 1.

SB would be a bit easier since the scene transition is similar in setting. Finally, due to the continuity of general scene structure in NB clips, we would expect its MSE magnitude to be lowest. This is comparable to the theoretical performance of human predictions of the SB, HB, and NB clips for the same difficulties of having increasingly variable frames for clips with more intense event transitions.

Finally, we note the recovery threshold for PredNet near the event transition for each clip. We do this by calculating the average MSE of the 10 frames before the frame that was predicted to be the event transition. Then, we found the number of frames it took for MSE to drop below this value. We show in figure 6 that SB seem to have the most difficulty recovering from the event tran-

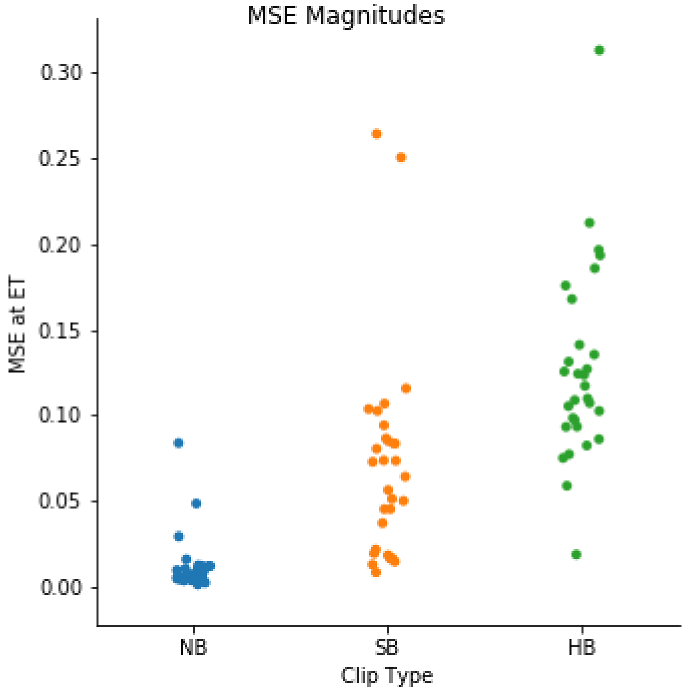

Figure 3.4: PredNet maximum MSE values for each clip across each boundary type in Stimulus set 1. This plot shows the maximum MSE of the next-frame prediction made by PredNet for each clip in stimulus set 1.

sition and HB and NB have similar recovery. These results could mean that PredNet is able to predict more effectively in the long run from a more intense form of event boundary such as HB. This explanation does not exactly match up with human performance in the short-term as SB have smoother and more predictable scene transitions. It does however match up with the results from [20] which implied that recognition accuracy was greater after crossing event boundaries which in turn suggest that the ability to recognize such boundaries are beneficial to memory performance.

The frame around the event transition appears distorted due to the temporary increase in prediction error. Moreover, after the event transitions for HB and SB clips, the prediction for the frames following the transition are also blurry. This supports our hypothesis of prediction error increasing

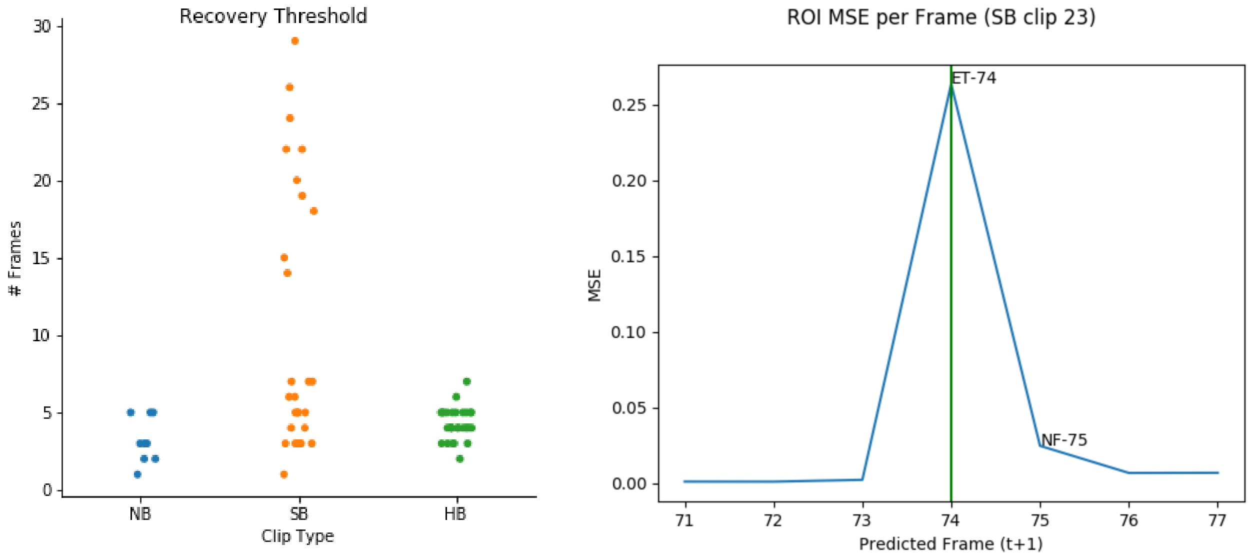

Figure 3.5: Frames needed for MSE to recover The plot on the left shows the amount of frames needed to recover to the average MSE of the previous 10 clips prior to the event transition. The plot to the right shows a view zooming into the region of interest (ROI) near the event transition to show the gradual MSE recovery.

near the transitions.

3.4 Control Task 1: PredNet SVM Decoding

For our first control task we decode the three boundary types HB, SB, and NB using only the MSE values from PredNet. Our hypothesis is that the HB and NB clips will be easiest to distinguish by MSE values alone. This is because the HB clips will have much higher MSE magnitudes than NB clips. Then, SB and NB will be the second easiest for the same reason of MSE magnitude difference. Finally, the HB and SB clips should have the lowest decoding accuracy. To test this hypothesis, we run an experiment using a support vector machine (SVM) [12] which is a useful machine learning method for binary classification tasks. For all of our SVM experiments we used Linear Support

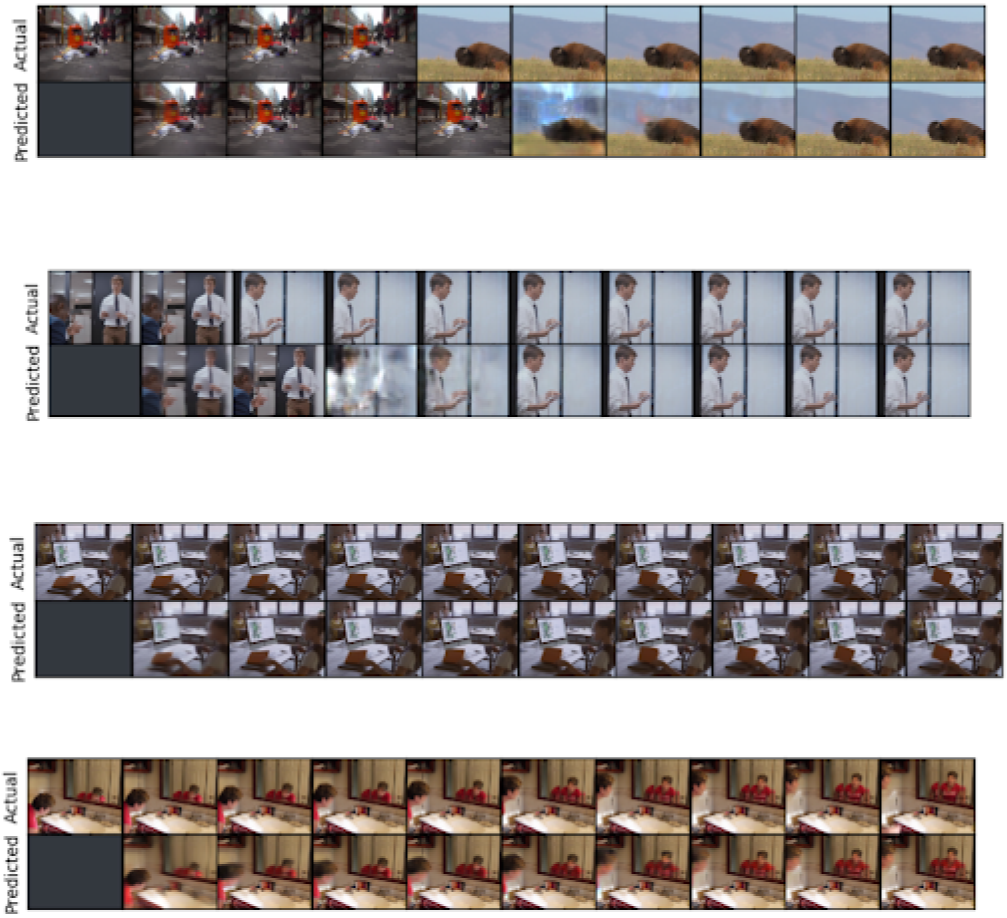

Figure 3.6: PredNet prediction error increases near event transitions From top to bottom: Stimulus set 1 SB, HB, NB, Stimulus set 2. Here we show the next-frame predicted by PredNet compared with the actual next-frame in the video clip.

Vector Classification (LinearSVC) from [12], which uses a linear kernel function. All LinearSVC settings were kept default. We proceeded by setting up a binary classification task to decode 3 pairings: HB and NB, SB and NB, HB and SB. For each clip type (HB, SB, and NB), The X input array of the LinearSVC included the maximum MSE value of that particular clip, thereby indicating the MSE at the event transition. The corresponding Y input array contained the labels for the MSE, whether it came from HB, SB, or NB. First, we shuffled the data and then split it into a 80:20 training and test data split ratio. For example, on one trial of the HB and NB pairing we would have an X array of just HB and NB MSEs and the Y array would contain the label for each element. If, the first element in the X array was the MSE indicating the event transition of a HB clip, then the first element of the Y array would be the string ”HB”. We used 10-fold cross-validation for a total of 10 trials for each pairing. **Table**3.1 shows that the LinearSVC supports our hypothesis for our control experiment. We see that the binary classification task decoding between HB and NB gives the best accuracy. This is related to the neural results from [20] which showed that boundary cells respond to both SB and HB while event cells only respond to HB. Our control task illustrates that the SVM can decode based on MSE values output by PredNet. This shows that PredNet's behavior is similar to the HPC neurons since it outputs responses (the MSE values) which indicate that it can distinguish the HB, SB and NB boundary types.

Trial	HB,NB	SB,NB	HB,SB
1	0.917	0.75	0.667
2	1.000	1.000	0.667
3	0.917	0.833	0.583
4	1.000	0.75	0.583
5	0.917	0.75	0.333
6	0.917	0.833	0.750
7	0.917	0.833	0.750
8	0.917	0.750	0.917
9	0.917	0.917	0.667
10	1.000	0.833	0.583
Average	**0.942**	**0.825**	**0.650**

Table 3.1: PredNet Binary LinearSVC results for decoding boundary types using MSE values. This table shows the decoding accuracy results of LinearSVC ran on the test data for each binary classification task. Notably, HB and NB have the highest average accuracy and HB and SB have lowest.

3.5 Control Task 2: Pixel-Level Subtraction SVM Decoding

In this experiment, we compare the decoding accuracy of the three pairings of boundary types using pixel-level subtractions as inputs. The pixel-level subtraction is given by the following equation which calculates the error frame Δ_f stored in the array as input:

$$\Delta_f = \sum_{i=1}^{n} (y_E - y_{E-1})^2. \tag{3.1}$$

In **equation** 3.1, y_E represents the actual video frame at the event transition. This event transition is the maximum MSE of that particular clip as outputed by PredNet during next-frame pre-

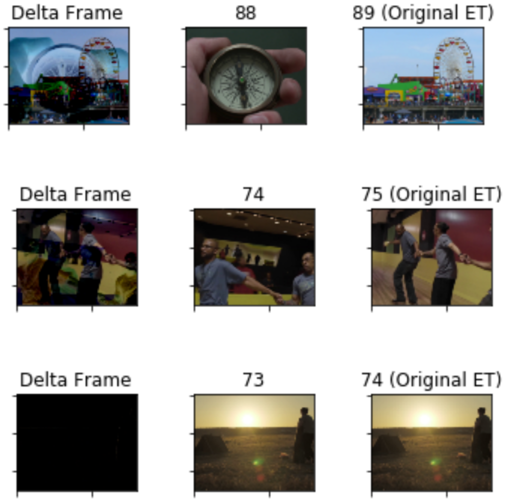

Figure 3.7: Δ_f **as outputed by Pixel-level subtraction of actual frames** From top to bottom: Stimulus set 1 HB, SB, NB clips are shown with their PredNet calculated event transition for HB and SB. For NB the middle of clip is taken as the event transition and the frame before it is subtracted from it. The pixel-level subtractions provide a baseline to compare the PredNet performance with.

diction. y_{E-1} represents the frame before occurring before it. For NB clips, since there is no event transition we do not use the maximum MSE value from PredNet. Instead, we divide the length of the frames by 2 and use that as the event transition. So, a clip with 148 frames will have an event transition at frame 74. Using this pixel-level subtraction method, we receive a baseline prediction of what the next frame before the event transition *should* be.

We use these delta frames as the X input into the SVM along with their corresponding HB, SB, or NB label as the Y input. We first reshape the X input based on the number of features of the delta frame. The following code is used for reshaping the X input based on the length of the training images:

```
X = numpy.reshape(.reshape(LenofXtrain,-1))
```

Again, linearSVC from [12] was used for binary classification using the same default settings.

Trial	HB,NB	SB,NB	HB,SB
1	0.917	0.917	0.667
2	1.000	0.667	0.500
3	0.917	0.833	0.583
4	1.000	0.750	0.583
5	1.000	0.917	0.250
6	1.000	0.833	0.250
7	0.917	0.917	0.500
8	1.000	0.917	0.583
9	1.000	0.833	0.333
10	1.000	0.750	0.500
Average	**0.975**	**0.833**	**0.475**

Table 3.2: Pixel-level subtraction LinearSVC results for decoding boundary types. This table shows the decoding accuracy results of LinearSVC ran on the test data for each binary classification task. Notably, HB and NB have the highest average accuracy and HB and SB have lowest. The pixel-level subtraction frames, which were obtained by subtracting the frame at the event transition by the frame occuring before it, were used as input.

Here is a combined overview of both averages:

PredNet	HB,NB	SB,NB	HB,SB
Average	**0.942**	**0.825**	**0.650**

PxSub	HB,NB	SB,NB	HB,SB
Average	**0.975**	**0.833**	**0.475**

Table 3.3: PredNet SVM vs Pixel-Subtraction SVM Average results for linearSVC using PredNet MSEs on the left and the pixel-subtraction method (PxSub) on the right.

From **Table** 3.3, we see that PredNet performs better than the baseline pixel-level subtraction in distinguishing between HB and SB. These results align with our hypothesis since the HPC neurons are able to distinguish between HB and SB.

Pixel-level subtraction performs well on the NB since a subtract will result in a very similar frame to the actual. For the same reason, it is reasonable that this method will perform worse on HB since there is a greater difference between the subtracted frames near the event transitions.

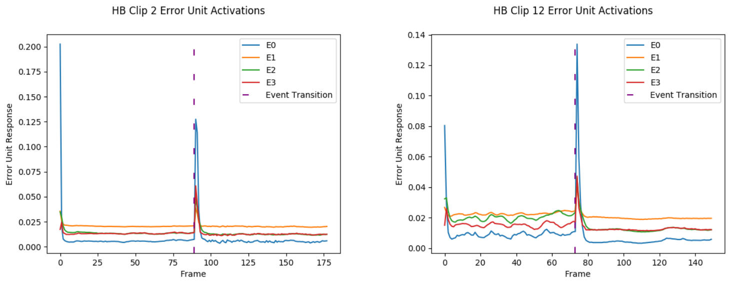

Figure 3.8: Error Unit Activation during HB clip next-frame prediction This figure shows two plots for randomly selected HB clips. We accumulated these results for all 30 HB clips, which produced similar plots as these ones. Notably, the error unit response was highest at the event transition predicted by PredNet

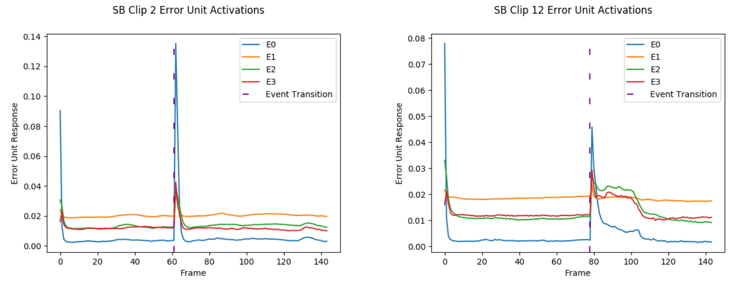

Figure 3.9: Error Unit Response during SB clip next-frame prediction This is a similar error unit response plot which shows that the response was highest at the event transition predicted by PredNet.

3.6 PredNet Error Unit Response

In this section, we will show the error unit responses outputted by PredNet during next-frame prediction. We expect that the error unit responses will be highest at the event transitions since that is where the greatest prediction error occurs.

Figures 3.8-3.10 show that for the HB and SB clip types, the error unit responds to the event transition of the clip. This matches with our neural results where boundary cells respond to both SB and HB [20]. Therefore, this behavior matches that of the boundary cells of the HPC neurons. In

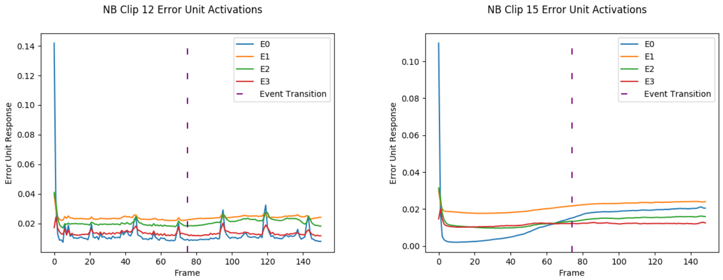

Figure 3.10: Error Unit Response during NB clip next-frame prediction This is a similar error unit response plot which shows that the response was highest at the event transition predicted by PredNet.

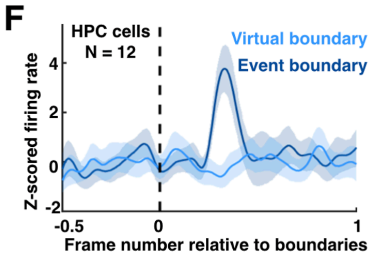

Figure 3.11: HPC cell Behavior. This figure comes from [20] and shows the similarity in response between the HPC neurons and the PredNet error unit responses.

addition, we see that for NB clips there is no transient increase in error unit response as in HB and NB clips. This confirms that the model is responding the changes in events during the time course of the clips. Figures 3 and 5 in [20] show similar results where E and B cells demarcate the HB, SB, and NB epsiodic event transitions.

4

Discussion

4.1 Analysis of Results

To sumarize, we ran the following experiments:

1. Event Transition Detection
2. Control Task 1: PredNet SVM Decoding
3. Pixel-level Subtraction SVM Decoding

4. PredNet Error Unit Responses

Our goal was to examine whether PredNet, a modern predictive recurrent neural network had the ability to explain the critical property of detecting event boundaries which was characteristic of the E and B cells in [20]. Specifically, we wanted to know whether the neurons of this deep learning network could produce the same neural results and the E and B cells by being able to distinguish between HB, SB, and NB clip types and demarcate the event transition that took place in the clips. **Experiment 1** was able to show this since PredNet clearly demarcated the boundary and for most clips it was exactly the same as the human marked event transition. **Experiments 2 and 3** further supported this hypothesis since we saw that PredNet was able to outperform the baseline pixel-level subtraction method when distinguishing between the HB and SB clips. Additionally, we wanted to see whether the neurons behaved similarly as shown by their inner workings, mainly their Error Unit Responses as shown by **Experiment 4**.

One of the shortcomings of this work is that it did not properly address whether there are neurons that have similar behavior to just boundary cells and event cells. We just showed whether the model responded to the differing boundary types. It would be beneficial to examine layers or groups of neurons carefully to see whether there differences among how they distinguish between the boundary types. For example, we could plot the time course of average response at each error output layers within a certain range of frames relative to the boundary occurrence and then look for layers that could differentiate the three boundary types. It would be interesting to see the degree to which certain layers contribute, if at all. This would more carefully match the neural results from [20] which showed that B cells responded to both HB and SB, whereas E cells only responded to

HB.

This work could further be improved by creating graphs that better summarized the response of the E_0-E_3 units of the model. There graphs we have showed the error unit response on a per clip bases but we did not create any graphs which averaged the results over all clips in each of the boundary types. For instance, an aggregate error unit graph for all HB clips would be summarize the results for HB conditions well.

Control task 1 could be improved by also adding the recovery frame numbers to the SVM decoding tasks. We can expect that adding the recovery frame numbers would give better decoding accuracy especially for distinguishing between SB and HB clips since the model will have more information of the MSE thresholds for the conditions.

Another control task that would have been interesting to test would be to use the boundary probabilities of the SEM model as input to the SEM and see whether PredNet or SEM provide better decoding accuracy.

Finally, the arguments presented in this work could be improved by incorporating a deeper analysis of the literature surrounding computational modeling. This work focuses on PredNet and very similarly structured computational models and does not thoroughly explore vastly different architectures such as SEM or networks that focus on the gating mechanism mentioned in **Chapter 2**.

4.2 Future Directions

This work can be taken further by implementing the extended PredNet architecture in **Figure 1.1**. This could allow for a greater circuit-level modeling of the biological circuit which would be useful for understanding circuit-level interactions. One implication of this would be that once a sufficiently plausible model is developed, it can be used to suggest stimulation or lesion experiments by first being done on the computational model. The external memory component in **Figure 1.1** was been conceptualized during the course of this work but we did not get a chance to implement it. We were considering incorporating the SEM model or taking inspiration from it to include in this external memory component. Another possible addition to the PredNet architecture would be to add word labels that describe the seen. Such labels would be added along with the input and a state-of-the-art RNN such as GPT-3 [2] could be trained to label the events before and after the event boundaries.

Two other neural results that were highlighted in [20] were that 1) B cells and E cells did not respond to clip onsets and offsets and 2) When responding to HB clips, B cells increased their firing rate response before E cells did. For 1) clip onsets and offsets refer to a fixation cross that is added during certain parts of the clip. It was noted that from human subject results, B and E cells did not respond to these clip onsets and offsets. When testing this we can expect that MSE and recovery frame at clip onsets and offsets will drop as PredNet learns the task structure.

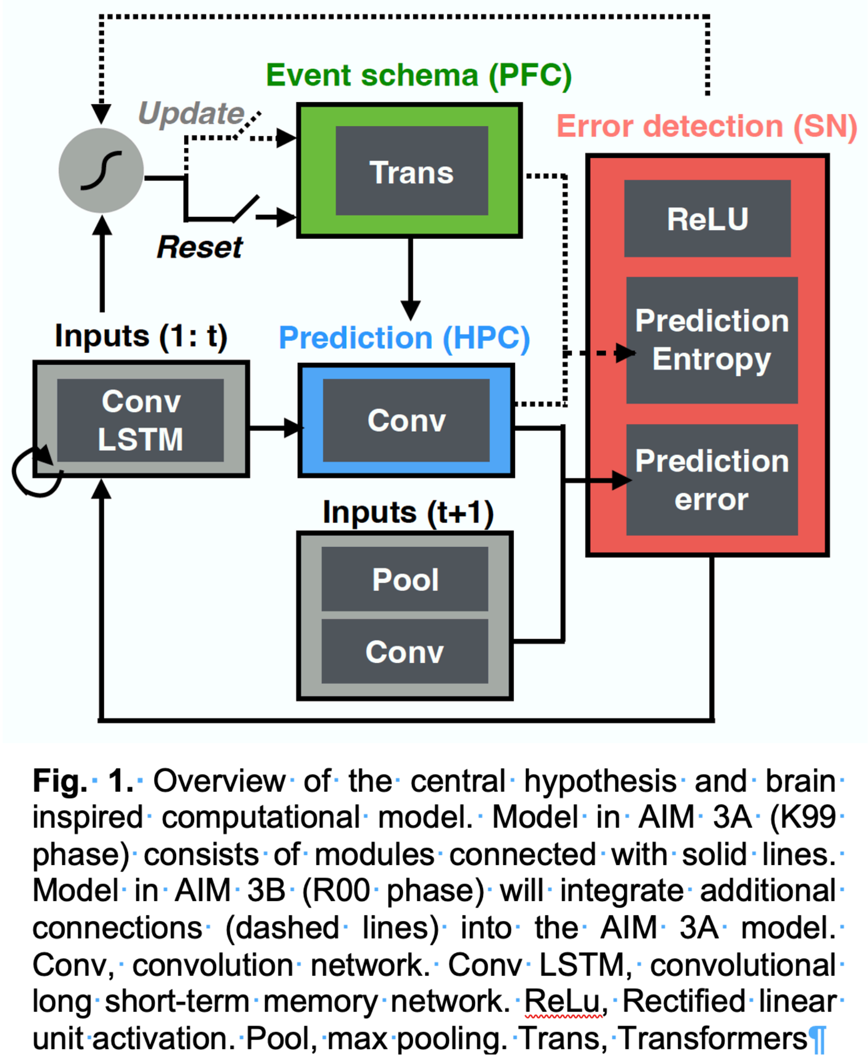

Fig. 1. Overview of the central hypothesis and brain inspired computational model. Model in AIM 3A (K99 phase) consists of modules connected with solid lines. Model in AIM 3B (R00 phase) will integrate additional connections (dashed lines) into the AIM 3A model. Conv, convolution network. Conv LSTM, convolutional long short-term memory network. ReLu, Rectified linear unit activation. Pool, max pooling. Trans, Transformers

Figure 4.1: A possible extension of the PredNet Model for future work. Conceived by Dr. Jie Zheng

4.3 Conclusion

This work has examined the ability of PredNet, a state-of-the-art predictive recurrent neural network, to reproduce similar neural results as hippocampal neurons as shown in recent literature. It has shown that PredNet has the ability to distinguish between different boundary conditions which indicate varying degrees of event transition. It has shown that these modern neural networks have matching behavior that is useful to investigate further. This work encourages examination of other characteristics of these deep learning networks in future studies. Such examination would benefit computer science, where architectures that perform well on large-scale problems similar to the brain. It would benefit Neuroscience as well due to the creation of biologically plausible models that can be used as a mechanistic testing ground for experiments.

References

[1] Barron, H. C., Auksztulewicz, R., & Friston, K. (2020). Prediction and memory: A predictive coding account. *Progress in neurobiology*, 192, 101821–101821.

[2] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners.

[3] Den Ouden, H., Kok, P., & De Lange, F. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in Psychology*, 3.

[4] Flores, S., Bailey, H. R., Eisenberg, M. L., & Zacks, J. M. (2017). Event segmentation improves event memory up to one month later. *Journal of experimental psychology. Learning, memory, and cognition*, 43(8), 1183–1202.

[5] Franklin, N., Norman, K., Ranganath, C., Zacks, J., & Gershman, S. (2020). Structured event memory: a neuro-symbolic model of event cognition.

[6] Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735–1780.

[7] Jin, J. & Maren, S. (2015). Prefrontal-hippocampal interactions in memory and emotion. *Frontiers in Systems Neuroscience*, 9.

[8] Lotter, W., Kreiman, G., & Cox, D. (2016). Deep predictive coding networks for video prediction and unsupervised learning.

[9] Lotter, W., Kreiman, G., & Cox, D. (2020). A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature machine intelligence*, 2(4), 210–219.

[10] Lu, Q., Hasson, U., & Norman, K. A. (2022). A neural network model of when to retrieve and encode episodic memories. *eLife*, 11.

[11] O'Reilly, R. C. & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, 18(2), 283–328.

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[13] Rao, R. P. N. & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1), 79–87.

[14] Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive science*, 31(4), 613–643.

[15] Richmond, L. L. & Zacks, J. M. (2017). Constructing experience: Event models from perception to action. *Trends in cognitive sciences*, 21(12), 962–980.

[16] Schott, B. H., Seidenbecher, C. I., Fenker, D. B., Lauer, C. J., Bunzeck, N., Bernstein, H.-G., Tischmeyer, W., Gundelfinger, E. D., Heinze, H.-J., & Duzel, E. (2006). The dopaminergic midbrain participates in human episodic memory formation: Evidence from genetic imaging. *The Journal of neuroscience*, 26(5), 1407–1417.

[17] Zacks, J. M., Kurby, C. A., Eisenberg, M. L., & Haroutunian, N. (2011). Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of cognitive neuroscience*, 23(12), 4057–4066.

[18] Zacks, J. M., Speer, N. K., Vettel, J. M., & Jacoby, L. L. (2006). Event understanding and memory in healthy aging and dementia of the alzheimer type. *Psychology and aging*, 21(3), 466–482.

[19] Zacks, J. M. & Swallow, K. M. (2007). Event segmentation. *Current directions in psychological science : a journal of the American Psychological Society*, 16(2), 80–84.

[20] Zheng, J., Schjetnan, A. G. P., Yebra, M., Gomes, B. A., Mosher, C. P., Kalia, S. K., Valiante, T. A., Mamelak, A. N., Kreiman, G., & Rutishauser, U. (2022). Neurons detect cognitive boundaries to structure episodic memories in humans. *Nature neuroscience*, 25(3), 358–368.