Brain-Inspired Mechanisms for Encoding Temporal Information in the Transformer Architecture

Breck Emert, University of Kansas

Abstract

In this paper, we explore brain-inspired mechanisms for encoding temporal information within the transformer architecture, focusing on innate interpretability of the mechanism. The human brain is characterized by localized regions performing specialized functions, and produces miraculously complex behavior. Our reliance on combinatorial calculations in self-attention and feed-forward layers is intentionally naïve and could benefit in interpretability if split up. Current transformer architectures primarily rely on attention, which we show model only a weak correlation to actual temporal dynamics. We propose focusing on how the brain processes short sequences into events, and those events into larger episodes, showing that this mechanism is relatively simple to implement.

1 Introduction

Many field-leading experts in machine learning believe that mechanisms adjacent to the 2017-style transformer (a dominant version of a Large Language Model) will be unable to achieve AI or superhuman intelligence. Others such as Ilya Sutskever believe that next-token prediction models which achieve a great level of accuracy must understand the underlying properties of what generated its dataset rather than just the relations of it and can as such perform out-of-sample (superhuman) inference (Patel, 2023).

Either of these perspectives means that exploring novel attention mechanisms explicitly designed to model time is not productive if current implementations are capable of efficient representations of temporal relations. A perspective independent of this is that the expectation of alignment and interpretability is rising: mechanisms which have interpretable structure and clear loss metrics are paramount. This literature review aims to explore how the brain can predict, infill and order sequences, inference causally, and how it does or does not distinguish between spatial patterns and temporal ones. By locating these mechanisms, further experiments and implementations can be made which result in a shorter path towards interpretable and performant models. Large Language Models (LLMs) appear to show all these capabilities, but a consistent and underlying mechanism has not been identified.

2 Background

2.1 What Is Temporal Information?

The motivation to capture temporal information in our machine learning models hinges on temporal information being relevant for the task of information processing. Temporality is just order, and this attribute must be encoded in the input and maintained throughout calculation for a model to produce output driven causally by this information. Neuronal firing is not unique in that it can be described with space and time coordinates, leaving us to decide what pattern these coordinates encode.

2.2 What things are temporal?

Sara Walker insightfully notes, "We want to reduce biology to physics, but physics is an emergent property of biology"; it isn't by default that explicit temporal encodings are always necessary, and this idea should not be forgotten over the course of searching for one. Our brain consistently spatializes time, learning to use space to understand even basic and short temporal sequences (Karşılar & Van Rijn, 2024). However, interpretation of this information may rely on the underlying mechanism understood by the human brain, which because it structures language to maximize a goal, the sequence chosen may not be aligned with a linear or equivalent interpretation of time in the receiver. "The human species seems to have evolved the ability to understand the concept of time by coopting the circuits devoted to understanding space" (Buonomano, 2017). Further, concepts of time are not tied to the content – there is a "fundamental conditional independence between the temporal structure of succession and the events that succeed each other" (Friston & Buzsáki, 2016).

Every sequential computational process is inherently temporal, as it unfolds over time. Static computations can represent information timelessly, without regard to the sequential dependencies that may have created the data that they aim to model. Static models do not incorporate recurrent structures, which are efficient for capturing the complexity that temporal progression allows. Dependencies that rely on the timing of previous activities or events may be more akin to objects in the real world, allowing the relational structure to more easily encode how objects behave in the real world. An argument supported by this review will be that modelling temporal information without temporal mechanisms is drastically inefficient, rather than strictly necessary. RNNs show us that a single weight matrix, when receiving unique sets of both the previous hidden state and the next inputs, is capable of modelling the latent structure consistent within the input.

2.2 Existing Mechanisms in LLMs

There are mechanisms in the transformer architecture which may help maintain temporal information throughout the learning process. Self-attention allows tokens to take on meaning without regard to their order in the input sequence, and thereby with limited regard to the tokens' original temporal coordinates. This representation is static, processed in an instantaneous block of computation, but is performed iteratively. It also depends on the values which get embedded into the input tokens. We may hope that the iterative interaction of the attention mechanism's value vectors with input vectors modifies the information into a structured representation that uncovers and preserves temporal relationships even in the absence of more explicit temporal mechanisms. Several examples of this emergent behavior are mentioned in this paper.

The memory of machine learning models is like human memory in that their controversial ability to output exact sequences is independent of the concept of overfitting. A human is thought of as having learned a concept when they can translate symbolic logic to a novel scenario, and their memorization is fragmented throughout processing and supporting brain regions (Nadeau, 2020; Robertson, 2001). LLMs, similarly, do not necessarily have their declarative and procedural knowledge tied. A model can produce arbitrary bits of memorized content, such as a publicized and blatant case of ChatGPT producing a Stack Overflow author's name along with its code answer, or creatively explaining mathematical concepts through a hypothetical squirrel teacher. These are bidirectionally independent, and this is suggested in a more formal way by reminding ourselves of a model's learning process. LLMs can both be memorizing and generalizing. Ilya Sutskever covers this idea well in an interview by highlighting that all forms of the model, represented in a Bayesian posterior distribution, memorize the information. This is what we are asking it to do. But these Bayesian posteriors are being averaged, sampled randomly by stochastic gradient descent (SGD) throughout the iterative training procedure. As a result, we necessarily get a model which has memorized the output to some degree, yet generalizes (Patel, 2023). That is, while each individual function within the posterior may be tailored to the training data, the aggregate output, which the model uses to make predictions, is capable of generalization.

One way this limitation is inherently addressed is through Recurrent Neural Networks (RNNs), which allow for past information to be incorporated into the current processing steps through feedback loops. This goes beyond just maintaining the information for access, and instead dynamically alters its behavior based on the accumulation and evolution of input over time. RNNs have been consistently shown to contain interpretable neurons, such as one which performs line-break tracking (Karpathy et al., 2015). The temporal information highlighted by some neurons is paralleled in the brain with internal clocks.

Whether or not we find the current iteration of transformers to understand time does not mean a new data or model paradigm is necessary to encode it. There are then two facts which must be true to get emergent temporal behavior from a model: Language itself must be capable of representing causal relationships, and a model which is learning from this language must be capable of resembling, at an abstract level, how the data are actually generated (Lake et al., 2016).

3 Neuroscientific Knowledge of Time and Attention

The brain has been shown to encode 'what' and 'where' through the ventral and dorsal streams, but does it have an explicit 'when' pathway? Charles Sanders Peirce understood language through his unique categorization of meaning into icons, indices, and symbols. He formalized a system which can efficiently map the sign tokens that we communicate (such as a word or sub-word, further referred to as just 'token') to physical objects. Briefly, icons are token-object resemblances (such as visual similarity), indices are causally linked or associated in space or time, and symbols are marked by social convention, tacit agreement, or explicit code which establishes the relationship or link of the token-object (Deacon, 1997). Support for this system has been shown in several neuroimaging studies, showing an increase in semantic activity in the left prefrontal cortex when words were categorized on the basis of semantic rather than physical attributes (Wagner et al., 1998).

Under this framework, the primary utilization of temporal information is through processing symbols. Unlike icons and indices, which are directly correlated with their referents, symbols interact uniquely with time. They are part of a dynamic system that incorporates past experiences and a local context window to forecast future states and the actions by which we achieve these states.

3.1 Regions of Our 'When' Pathway

Must the 'when' pathway look like the 'what' and 'where' pathways? There are a variety of contributors of a 'when' pathway in the brain, and we will briefly cover regions associated with more dedicated processing and their more unique behavior. Following the finding that any rat neuron can be used as a timer (Johnson et al., 2010), several papers dug further to argue that population clocks emerge from internal dynamics of the brain and are inherent to the recurrent networks of the brain (Buonomano & Laje (2010); Ivry & Schlerf, 2008).

Basal ganglia: links perception to motor control. The timing mechanism that this region performs is relevant to how the brain decodes the sequential nature of language. Involvement of the basal ganglia with our understanding time come from studies of several different animals. One involving rats showed that lesions to their substantia nigra (SN), as well as to their caudate-putamen (CPu), significantly affect temporal processing (Meck, 2006). These regions show the same effect when lesioned: the rats are capable of modifying their behavior through differentiating response rates to different length signals, but were incapable of understanding the relative reward value of these signals. They differed in that the effects of lesions to the SN can be restored through administering L-DOPA.

It should be noted that the CPu and SN are a part of the nigrostriatal system, which is one of the four major dopamine pathways of the brain. The SNc is the origin of this pathway and projects dopaminergic axons into the striatum, which includes the CPu. A model of interval timing that covers this behavior is the striatal beat-frequency (SBF) model. Models of how interval timing occurs within the brain are useful in that linguistic-symbolic correlates of the mechanisms that the brain relies on can help accurately identify temporal patterns in language, but are not covered further here.

We must be cautious in thinking that the brain is (close to) the root of the reason why we understand time. That is, we understand time on small, parochial scales rather than as one underlying circuit. When presented stimuli with something bright on the screen, we perceive the duration to be longer. The same goes for when the stimulus is a large number.

Cerebellum: Usually characterized by its strong structural connectivity and involvement with motor control, studies are now also finding consistent activity during non-motor activities (Ackermann, 2008; Tomlinson et al., 2013). Its activity during speech was examined further by Schwartze & Kotz (2015), who reviewed evidence for a link between temporal and speech-processing systems. They propose that speech perception is (micro) event-based, supported by findings that the cerebellum receives input from temporal areas and the cochlear nucleus during early stages of processing. Its output preserves the signal's onsets/offsets and steeply rising spectral edges (Schwartze & Kotz, 2013). Rather than cerebellar activity during language processing being predominantly elicited by words per se, the auditory system responds to signal properties that change as a function of time (Kluender et al., 2003). Further research is needed to clarify the causal location of these interactions.

Hippocampus: The hippocampus is central to encoding, storing, and retrieving episodic memory, which our use of is inherently temporal when we make comparisons with these memories. It is a crucial hub in the medial temporal lobe that integrates the events encoded by this form of memory with semantic knowledge into a narrative. This integration proves useful in forming a common ground of social discourse more resilient to errors (Duff & Brown-Schmidt, 2012). This is in alignment with isomorphic approaches that believe the brain needs to model language redundantly (Smith, 2005), which is a general conclusion of the field of linguistics. A key function adjacent to this process is its involvement in pattern separation and pattern completion, which is the mechanism by which differentiation of events can occur (Moscovitch et al., 2016).

Autonoetic consciousness, or mental time travel, is a difficult concept to cover in relation to LLMs. Curriculum learning is not popular for their training, meaning that models do not learn information in a particular order, and are thereby not incentivized to learn when a piece of information was learned by the model. However, as this paper relies on a semantic understanding of language, and because humans rely on episodic memory for actual utility, it should be covered. Humans have been shown to overcome their immediacy bias (temporal discounting) by imagining spending the larger sum of money in the future (Benoit et al, 2011). The activation of the PFC found during this study was closely coupled with the hippocampus, suggesting that the episodic simulation provided by the hippocampus was used. This perspective makes autonoetic thinking by a LLM more reasonable, but does not imply that it does or is likely to happen in current models. Gonzalez and DiPaola (2024) argue in a recent paper frameworks which can support this cognitive strategy for LLMs.

Hippocampal circuits are largely blind to inputs (Lisman et al., 2017), returning a generally learned encoding pattern to several regions of the brain. Coupled with its required involvement in processing sequences (O'Keefe & Reece, 1993), the hippocampus functions not merely as a memory unit but as a complex integrative system for temporal and spatial contexts. Constructing coherent narratives out of disjointed events seems certainly useful for brain regions and may implicate its involvement in so many brain regions.

3.2 Memory

Gazzaniga reminds us that "The outcome of learning is memory" which suggests much of what we see as time-dependent learning is developing an automatic ability to perform the best action given the inputs. Dissecting this output leads to the forms of sensory memory, short-term memory, and working memory, which play different roles in our ability to perform a task. Sensory memory lasts only very briefly once encoded, continuously cycling out with new information. Short-term and working memory provide a workspace for manipulation, also crucial for understanding and organizing events in time. Long-term memory, which is historically broken down into declarative and procedural memories, stores facts and memories acquired through learning. An event-based model of the world supports adaptive behavior better tuned to the current environment (Clewett et al., 2019). We have highlighted event-based decoding of memory, but not how memories are encoded in a way which can later be manipulated in this way. In Clewett et al. (2019), they further highlight the need to identify conditions that lead to separation as opposed to integration of the context from memory. The brain shows bias towards integration because as previously stated, iconic and indexical relationships are how Hebbian learning operates. They further show a study by Ezzyat & Davachi (2014) which measured participants' estimation of temporal distance of events, finding that scene changes during encoding lead to a further assignment of temporal distance.

4 How these biological mechanisms can inform ML

We've shown how the brain reacts to a changing environment not to recreate time, but to illicit internal semantics in an order which, as merged, align a modeled goal with the state of the environment and vice-versa. Thomas G. Deitterich says "The fundamental problem is that our large language models, although we want to interpret them and use them as if they are knowledge bases, they are actually not knowledge bases, they are statistical models of knowledge bases." (valgrAI, 2023). Modelling knowledge of the world requires an exhaustive database of events, which the brain performs not by memorization, but by reliance on association of its biased networks with the current stimuli surrounding hidden state. We would like further research to propose a mechanism which introduces only linear computational cost, is innately interpretable, and requires no additional infrastructure to any code environment.

5 Conclusion

The exploration of brain-inspired mechanisms for encoding temporal information within the transformer architecture brings us closer to understanding how models could potentially mimic the human brain's ability to process and remember sequential information as coherent events. This research underlines the inherent limitations of current transformer models which predominantly process sequences in a flattened and static manner. This is not indicative of an approach which will continue to approach 100% accuracy in next-token prediction.

Our proposition to integrate event-based architecture into a transformer is nuanced to the temporal world and avoids inefficient mappings to and from this space. This is a jump towards models which are not subject to the nines of reliability problem, and is fully interpretable.

References

Benoit, R. G., Gilbert, S. J., & Burgess, P. W. (2011). A neural mechanism mediating the impact of episodic prospection on farsighted decisions. the Journal of Neuroscience/the Journal of Neuroscience, 31(18), 6771–6779. https://doi.org/10.1523/jneurosci.6559-10.2011

Buonomano, D. (2017). *Your brain is a time machine: the neuroscience and physics of time*. W. W. Norton & Company.

Buonomano, D. V., & Laje, R. (2010). Population clocks: motor timing with neural dynamics. Trends in Cognitive Sciences, 14(12), 520–527. https://doi.org/10.1016/j.tics.2010.09.002

Clearer Thinking with Spencer Greenberg. (2023, January 18). What, if anything, do AIs understand? with ChatGPT Co-Founder Ilya Sutskever [Video]. YouTube. https://www.youtube.com/watch?v=NLjS1UOr8Nc

Clewett, D., DuBrow, S., & Davachi, L. (2019). Transcending time in the brain: How event memories are constructed from experience. Hippocampus, 29(3), 162–183. https://doi.org/10.1002/hipo.23074

D‘Angelo, E., & De Zeeuw, C. I. (2009). Timing and plasticity in the cerebellum: focus on the granular layer. Trends in Neurosciences, 32(1), 30–40. https://doi.org/10.1016/j.tins.2008.09.007

De Zeeuw, C. I., Hoebeek, F. E., & Schonewille, M. (2008). Causes and consequences of oscillations in the cerebellar cortex. *Neuron*, *58*(5), 655–658. https://doi.org/10.1016/j.neuron.2008.05.019

Deacon, T. W. (1997). *The symbolic species: The Co-evolution of Language and the Brain*.

Ezzyat, Y., & Davachi, L. (2014). Similarity Breeds Proximity: Pattern Similarity within and across Contexts Is Related to Later Mnemonic Judgments of Temporal Proximity. Neuron, 81(5), 1179–1189. https://doi.org/10.1016/j.neuron.2014.01.042

Friston, K. J., & Buzsáki, G. (2016). The Functional Anatomy of time: what and when in the brain. Trends in Cognitive Sciences, 20(7), 500–511. https://doi.org/10.1016/j.tics.2016.05.001

Gazzaniga, M. S., & Mangun, G. R. (Eds.). *The Cognitive Neurosciences (5th ed.).* (2014). In The MIT Press eBooks. https://doi.org/10.7551/mitpress/9504.001.0001

Ivry, R. B., & Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. Trends in Cognitive Sciences, 12(7), 273–280. https://doi.org/10.1016/j.tics.2008.04.002

Johnson, H. A., Goel, A., & Buonomano, D. V. (2010). Neural dynamics of in vitro cortical networks reflects experienced temporal patterns. Nature Neuroscience, 13(8), 917–919. https://doi.org/10.1038/nn.2579

Karpathy, A., Johnson, J., & Li, F. (2015). Visualizing and understanding recurrent networks. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1506.02078

Kluender, K. R., Coady, J. A., & Kiefte, M. (2003). Sensitivity to change in perception of speech. Speech Communication, 41(1), 59–69. https://doi.org/10.1016/s0167-6393(02)00093-6

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. Behavioral and Brain Sciences, 40. https://doi.org/10.1017/s0140525x16001837

Lisman, J., Buzsáki, G., Eichenbaum, H., Nadel, L., Ranganath, C., & Redish, A. D. (2017). Viewpoints: how the hippocampus contributes to memory, navigation and cognition. Nature Neuroscience, 20(11), 1434–1447. https://doi.org/10.1038/nn.4661

Moscovitch, M., Cabeza, R., Winocur, G., & Nadel, L. (2016). Episodic Memory and Beyond: The hippocampus and neocortex in transformation. Annual Review of Psychology, 67(1), 105–134. https://doi.org/10.1146/annurev-psych-113011-143733

O'Keefe, J., & Recce, M. (1993). Phase relationship between hippocampal place units and the EEG theta rhythm. Hippocampus, 3(3), 317–330. https://doi.org/10.1002/hipo.450030307

Patel, D. (2023, March 27). Ilya Sutskever (OpenAI Chief Scientist) - Building AGI, alignment, & Future Models. Dwarkesh Podcast. https://www.dwarkeshpatel.com/p/ilya-sutskever

Sean Carroll. (2020, January 13). Mindscape 79 | Sara Imari Walker on Information and the Origin of life [Video]. YouTube. https://www.youtube.com/watch?v=hcG5IBHrvYo

Smith, N. (2005). Chomsky's science of language. In J. McGilvray (Ed.), The Cambridge Companion to Chomsky (pp. 21–41). chapter, Cambridge: Cambridge University Press.

valgrAI. (2023, July 10). "What's wrong with LLMs and what we should be building instead" - Tom Dietterich - #VSCF2023 [Video]. YouTube. https://www.youtube.com/watch?v=cEyHsMzbZBs

Wagner AD, Schacter DL, Rotte M, et al. Building memories: remembering and forgetting verbal experiences as predicted by brain activity. Science 1998;281:1188-91.