

Building Transformers from Neurons and Astrocytes

Leo Kozachkov^{a,b}, Ksenia V. Kastanenka^c, and Dmitry Krotov^a

^aMIT-IBM Watson AI Lab, IBM Research

^bDepartment of Brain and Cognitive Sciences, MIT

^cDepartment of Neurology, MassGeneral Institute for Neurodegenerative Diseases, Massachusetts General Hospital and Harvard Medical School

Glial cells account for roughly 90% of all human brain cells, and serve a variety of important developmental, structural, and metabolic functions. Recent experimental efforts suggest that astrocytes, a type of glial cell, are also directly involved in core cognitive processes such as learning and memory. While it is well-established that astrocytes and neurons are connected to one another in feedback loops across many time scales and spatial scales, there is a gap in understanding the computational role of neuron-astrocyte interactions. To help bridge this gap, we draw on recent advances in artificial intelligence (AI) and astrocyte imaging technology. In particular, we show that neuron-astrocyte networks can naturally perform the core computation of a Transformer, a particularly successful type of AI architecture. In doing so, we provide a concrete and experimentally testable account of neuron-astrocyte communication. Because Transformers are so successful across a wide variety of task domains, such as language, vision, and audition, our analysis may help explain the ubiquity, flexibility, and power of the brain's neuron-astrocyte networks.

Artificial Intelligence | Neuroscience | Transformers | Glia | Astrocytes

Correspondence: leokoz8@mit.edu, krotov@ibm.com

Significance Statement

Transformers have become the default choice of neural architecture for many machine learning applications. Their success across multiple domains such as language, vision, and speech raises the question: how can one build Transformers using biological computational units? At the same time, in the glial community there is a gradually accumulating evidence that astrocytes, formerly believed to be passive house-keeping cells in the brain, in fact play important role in brain's information processing and computation. In this work we hypothesize that neuron-astrocyte networks can naturally implement the core computation performed by the Transformer block in AI. The omnipresence of astrocytes in almost any brain area may explain the success of Transformers across a diverse set of information domains and computational tasks.

Introduction

Astrocytes, one kind of glia, are a ubiquitous cell type in the central nervous system. There is increasing evidence that astrocytes play an active and flexible role in cognition (1–3). Specifically, astrocytes actively contribute to circuit activity underlying cognitive function. It is empirically well-established that astrocytes and neurons communicate with one another via feedback loops that span many spatial and temporal scales (4, 5). However, a firm computational interpretation of neuron-astrocyte communication is missing.

Transformers, a particular type of AI architecture, have become influential in machine learning (6) and, increasingly, in computational neuroscience (7–13). They are currently the choice model for tasks across many disparate domains, including natural-language processing, vision, and speech (14). A typical Transformer block uses four basic operations: attention, feed-forward neural network, layer normalization, and skip connections. These operations are arranged in a certain way so that the entire block can learn relationships between the tokens, which represent the data (e.g. in the language domain tokens correspond to words in a sentence, in the image domain tokens correspond to portions of the image). Interestingly, several recent reports suggested architectural similarities between Transformers and the hippocampus (8, 12), and representational similarities with human brain recordings (7, 9, 13). However, unlike more traditional neural networks, such as convolutional networks (15) or Hopfield networks (16), which have a long tradition of biological implementations, Transformers are only at the beginning of their interpretation in terms of known biological processes.

In this work we show that neuron-astrocyte networks perform the core computations of a Transformer. We arrive at this result by explicitly constructing a neuron-astrocyte network whose output approximates that of a Transformer with high probability. The main novel computational element of our network is the tripartite synapse, the ubiquitous three-factor connection between an astrocyte, a pre-synaptic neuron, and a post-synaptic neuron (17). We show that tripartite synapses perform the role of normalization in the Transformer's self-attention operation. As such, neuron-astrocyte networks are natural candidates for the biological “hardware” that can be used for conducting computation with Transformers.

A primer on the biology of astrocytes

Glial cells are the other major cell type in the brain besides neurons. The exact ratio of glia to neurons is disputed, but it is somewhere between 1:1 and 10:1 (18). The most well-studied type of glial cell is the astrocyte. A defining feature of astrocytes is that a single astrocyte cell forms connections with thousands to millions of nearby synapses (19). For example, a single human astrocyte can cover between 270,000 to 2 million synapses within a single domain (20). Astrocytes are for the most part electrically silent, encoding information in the dynamics of intracellular calcium ions (Ca^{2+}). In most parts of the brain, neurons and astrocytes are closely intertwined. For example, in the hippocampus as many as 60% of all axon-dendrite synapses are wrapped by astrocyte cell membranes (21). In the cerebellum, the number is even

higher. This three-way arrangement (pre-synaptic axon, post-synaptic dendrite, astrocytes membrane) is so common that it has been given a name: the tripartite synapse (17).

Astrocyte membranes contain receptors corresponding to the neurotransmitters released at the synaptic sites they ensheath. For example, astrocytes in the basal ganglia are sensitive to dopamine, whereas in the cortex astrocytes are sensitive to glutamate (22). Despite being affected by the same pre-synaptic neurotransmitters, post-synaptic neurons and astrocytes respond very differently: neurons primarily encode information using action potentials, but astrocytes encode information via elevations in free intercellular calcium. Importantly, neuron-to-astrocyte signaling can trigger a response in the opposite astrocyte-to-neuron direction thus establishing a feedback loop between neural cells and astrocytes. Astrocytes can either depress or facilitate synapses, depending on the situation (23). For example, astrocytes in the hypothalamus have been observed to multiplicatively scale the excitatory synapses they ensheath by the same common factor (24).

Interestingly, there is also extensive astrocyte-to-astrocyte communication in the brain. Astrocytes form large-scale networks with one another (19). These networks are spatially tiled, with regular intercellular spacing of a few tens of micrometers (25). Unlike neurons, which communicate primarily with spikes, astrocytes communicate via calcium waves that propagate between their cell bodies, processes, and endfeet (26). These waves have speeds of a few tens of micrometers per second. It is thought that these waves could be used to synchronize neural populations and coordinate important neural processes (27).

Biological Implementation of a Transformer Block

This section begins by introducing Transformers using AI notations. Then, the neuron-astrocyte network is described, and it is shown that the attention operation can be implemented using the proposed model. At the end of the section it is explained how the remaining operations of the Transformer block (layer normalization and feed-forward network) can be included into the neuron-astrocyte circuit.

A. Transformer Notation. Consider a sequence of N token embeddings. Each token can correspond to a word (or a part of the word) if the Transformer is used in the language domain, or a patch of an image in the vision domain. Each embedding is of dimension d . The tokens are streamed into the network one-by-one (online setting), and the time of the token’s presentation is denoted by t . The t^{th} embedding is given by a vector $\mathbf{x}_t \in \mathbb{R}^d$. In the Transformer block each of these tokens is converted to a key, value, or query vector via a corresponding linear transformation, $W_K, W_Q \in \mathbb{R}^{D \times d}$ and $W_V \in \mathbb{R}^{d \times d}$:

$$\begin{aligned} \mathbf{k}_t &= W_K \mathbf{x}_t \\ \mathbf{v}_t &= W_V \mathbf{x}_t \\ \mathbf{q}_t &= W_Q \mathbf{x}_t \end{aligned} \quad (1)$$

where D is the internal size of the attention operation. These vectors are collected into matrices X, K, V, Q by stacking them columnwise. The nonlinearities are applied to matrices columnwise:

$$\phi(K) \equiv \begin{bmatrix} | & | & \dots & | \\ \phi(\mathbf{k}_1) & \phi(\mathbf{k}_2) & \dots & \phi(\mathbf{k}_N) \\ | & | & \dots & | \end{bmatrix}$$

Written in this way, the Transformer self-attention operator becomes:

$$\text{SelfAttn}(X) = V \text{softmax}(K^T Q) \quad (2)$$

where the softmax normalization is computed along the *columns* of $K^T Q$. The output of this self-attention operation is then passed along to a feed-forward neural network (FFN) that acts separately on each token (each column of its input). Without loss of generality, a single-headed attention Transformer is studied. In this case the output of the full Transformer block may be written as a two-step process:

$$\begin{aligned} Y &= \text{LayerNorm}(\text{SelfAttn}(X) + X) \\ \text{Transformer}(X) &= \text{LayerNorm}(\text{FFN}(Y) + Y) \end{aligned} \quad (3)$$

where FFN refers to a feedforward network, applied to each token separately and identically. It is not obvious how or if it is possible to write Eq. (3) in terms of simple mathematically abstracted biological operations (artificial neurons and perhaps astrocytes). Note that the main difficulty is that the SelfAttn operation contains interaction terms *between* token embeddings (28), and not just between elements of a single embedded token. In the next few sections, we show that by defining Hebbian-based “read” and “write” operations, we can capture these interactions in a biologically plausible way. In order to gain theoretical insight into Transformers, it is common to tie the weights (29, 30). This tying can be within a single Transformer block, between blocks, or both. In this section we will tie the weights within a single block, but not between blocks. We will relax this weight sharing constraint in the later sections. In particular, we tie W_Q, W_K, W_V as follows:

$$W_Q = W_K = W, \quad W_V = I$$

for some arbitrary matrix W and the identity matrix, I . In general, we will not require that $d = D$. We include this constraint now to fully analyze the simplest version of our model that captures the essential elements of our argument. Without loss of generality, we will ignore layer normalization steps for now, returning to it in section F.

B. Neuron-Astrocyte Network. A high-level overview of our circuit is shown in Figure 1. The network consists of a perceptron with an input layer, a hidden layer, and an output layer. The d -dimensional inputs are passed to the hidden layer with m units, as well as to the last layer via a skip connection (not shown). The hidden layer applies a fixed non-linearity to incoming inputs. The outputs of the hidden layer are passed to the last layer via a linear mapping $H \in \mathbb{R}^{d \times m}$. The synapses in the matrix H are tripartite synapses, meaning

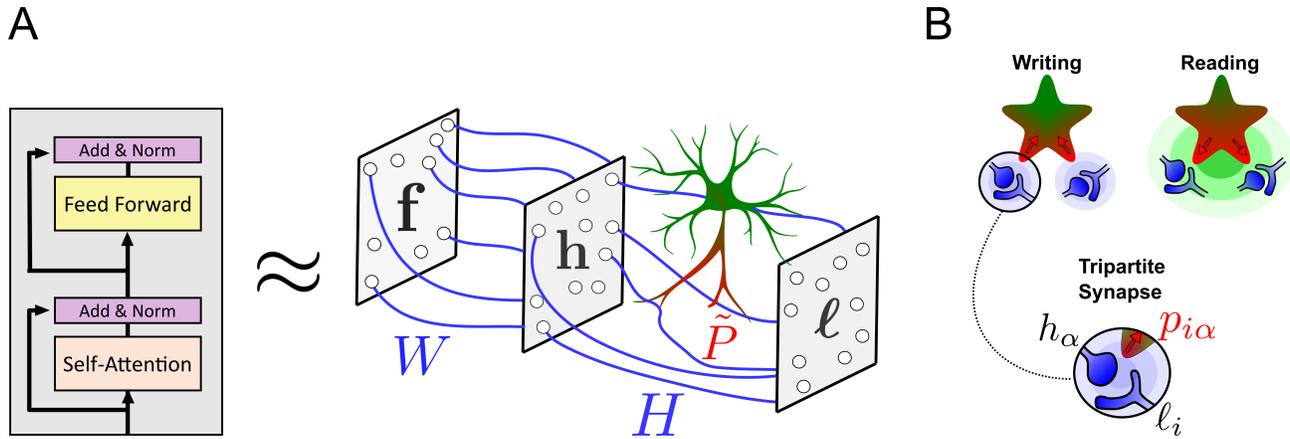


Fig. 1. (Panel A) A high-level overview of the proposed neuron-astrocyte network. The Transformer block is approximated by a feed-forward network with an astrocyte unit that ensheaths the synapses between the hidden and last layers (matrix H). Data is constantly streamed into the network. (Panel B) During the writing phase the neuron-to-neuron weights are updated using Hebbian learning rule and the neuron-to-astrocyte weights are updated using a pre-synaptic plasticity rule. During the reading phase the data is forwarded through the network, and the astrocyte modulates the synaptic weights H .

that each of the md synapses is associated with an astrocyte process $p_{i\alpha}$. The Latin indices i, j are used to enumerate neurons in the first and last layers, while the Greek indices α, β are reserved for the hidden neurons. With these notations the strength of the synapse between a hidden neuron α and the output neuron i is denoted by $H_{i\alpha}$ and the activity of the astrocyte process that ensheathes this synapse is described by $p_{i\alpha}$. The layers are denoted from left to right as $\mathbf{f}, \mathbf{h}, \mathbf{l}$ (first, hidden, last), respectively. Our network is described by the following equations:

$$\begin{aligned} \mathbf{f} &= \mathbf{x} && \in \mathbb{R}^d \\ \mathbf{h} &= \phi(W\mathbf{f}) && \in \mathbb{R}^m \\ \mathbf{l} &= r(H \odot \tilde{P})\mathbf{h} + \mathbf{f} && \in \mathbb{R}^d \end{aligned} \quad (4)$$

The scalar $r = \{0, 1\}$ stands for ‘read’, and is zero during the writing phase and unity during the reading phase. The symbol \odot denotes the Hadamard product (element-wise multiplication) between two matrices. The matrix $\tilde{P} \in \mathbb{R}^{d \times m}$ captures the effect of the astrocyte processes, and is defined as follows:

$$\tilde{P}_{i\alpha} = \frac{1}{p_{i\alpha}}$$

This inverse modulation of synaptic weights by astrocytes has been observed, for example, in studies involving tumour necrosis factor-alpha (TNF- α), wherein astrocytes will up-scale synaptic weights in response to low neural activity and downscale weights in response to high neural activity. More generally, many studies have observed that astrocytes can both depress and facilitate synapses, depending on the situation (4, 31–34)

C. Astrocyte Process Dynamics. As discussed in the introduction, astrocyte processes are sensitive to pre-synaptic neural activity. To capture this mathematically, we assume that the astrocyte process Ca^{2+} response is linearly proportional to the pre-synaptic neuron activation h_α of neuron α in layer \mathbf{h} . The constant of proportionality between the astrocyte process activation and the pre-synaptic neural activity is

denoted as $g_{i\alpha}$. This constant is in general different for every astrocyte process. Upon presentation of an embedded token to the network, astrocyte process $p_{i\alpha}$ initially responds with a local calcium elevation $g_{i\alpha}h_\alpha$. This Ca^{2+} response is then spatially averaged with the responses of other nearby astrocyte processes so that, after transients, the processes have the same value once a token is presented:

$$p_{i\alpha} = \frac{1}{md} \sum_{j=1}^d \sum_{\beta=1}^m g_{j\beta} h_\beta \quad (5)$$

The neuron-to-astrocyte signalling pathway in our circuit is completely described by Eq. (5).

D. Writing Phase. During the writing phase, r is set to zero. Biologically, this condition could correspond to some global neuromodulator being released into the local environment, for example acetylcholine, as suggested in (10, 35). Plugging $r = 0$, Eq. (4) becomes:

$$\begin{aligned} \mathbf{f}_t &= \mathbf{x}_t \\ \mathbf{h}_t &= \phi(W\mathbf{f}_t) = \phi(\mathbf{k}_t) \\ \mathbf{l}_t &= \mathbf{f}_t = \mathbf{v}_t \end{aligned} \quad (6)$$

As the embedded tokens are passed into Eq. (6) sequentially, the weight matrix H is updated via Hebbian plasticity with a learning rate of $\frac{1}{m}$. Upon presentation of token t , the matrix H is:

$$H_t = H_{t-1} + \frac{1}{m} \mathbf{l}_t \mathbf{h}_t^T = H_{t-1} + \frac{1}{m} \mathbf{v}_t \phi^T(\mathbf{k}_t)$$

After all the embeddings are passed through the network, assuming H is initially the zero matrix, the final value of H is:

$$H = \frac{1}{m} V \phi^T(K)$$

At the same time, the neuron-to-astrocyte weights are updated via pre-synaptic plasticity. Upon presentation of token t , these weights are:

$$g_{t,i\alpha} = g_{t-1,i\alpha} + \phi(W\mathbf{x}_t)_\alpha \rightarrow g_{i\alpha} = \sum_{t=1}^N \phi(\mathbf{k}_t)_\alpha$$

Note that as a consequence of the pre-synaptic plasticity, the weight $g_{i\alpha}$ does not depend on the index i . Therefore, we will only refer to the vector $\mathbf{g} \in \mathbb{R}^m$, which—through the pre-synaptic plasticity—is simply the sum over all token presentations of the hidden layer neural activations:

$$\mathbf{g} = \sum_{t=1}^N \phi(\mathbf{k}_t)$$

E. Reading Phase. During the reading phase, the read gate is set to $r = 1$ in Eq. (4), and the inputs are forwarded through the network. The astrocyte process activation value p , which according to Eq. (5) does not depend on indices i and α , is given by:

$$p = \frac{1}{m} \mathbf{h}^T \mathbf{g} = \frac{1}{m} \phi(\mathbf{q}_t)^T \sum_{t=1}^N \phi(\mathbf{k}_t) \quad (7)$$

To obtain the last equality we have used $\mathbf{h}_t = \phi(W\mathbf{x}_t) = \phi(\mathbf{q}_t)$. Plugging in all the steps of Eq. (4), we see that the last layer has the following value:

$$\ell_t = \frac{1}{p} H \phi(\mathbf{q}_t) + \mathbf{x}_t = \frac{V \phi^T(K) \phi(\mathbf{q}_t)}{\phi(\mathbf{q}_t)^T \sum_{t=1}^N \phi(\mathbf{k}_t)} + \mathbf{x}_t \quad (8)$$

If we compute ℓ for every input \mathbf{x}_t and stack the results column-wise into a matrix L , we have:

$$L = V \phi^T(K) \phi(Q) G^{-1} + X \quad (9)$$

where G is a diagonal matrix such that $G_{zz} = \phi(\mathbf{q}_z)^T \sum_{t=1}^N \phi(\mathbf{k}_t)$.

At this point, Eq. (9) is equivalent to a non-autoregressive version of the Linear Transformer (36). However, to make our circuit approximate a traditional Transformer with softmax attention, one more piece is required. We need to choose ϕ carefully so that we can reproduce the exponential dot-product in the softmax operation. In other words, we need ϕ to satisfy:

$$\phi(\mathbf{x})^T \phi(\mathbf{y}) \approx \exp(\mathbf{x}^T \mathbf{y})$$

This can be achieved using random feature attention (RFA) (37). RFA relies on a well-known result in kernel approximation theory, which is that the radial basis function (RBF) kernel can, with high probability, be approximated very well using random projections and cosines (38). To map the RBF kernel to the exponential dot-product kernel used in Transformers, one can simply scale the RBF approximation as follows:

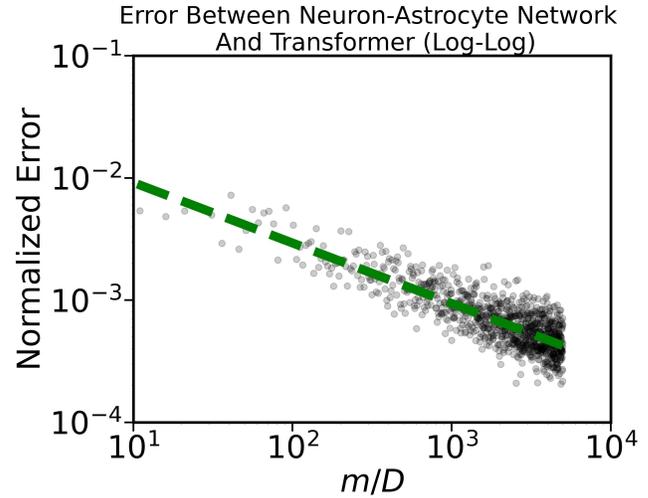


Fig. 2. Error vs number of hidden units (m) in our network. As m increases, the difference between the output of the neuron-astrocyte circuit and the AI Transformer block decreases.

$$\phi(\mathbf{x}) = \exp\left(\frac{\|\mathbf{x}\|^2}{2}\right) \phi_{\text{RBF}}(\mathbf{x}) = \exp\left(\frac{\|\mathbf{x}\|^2}{2}\right) \cos(\Pi \mathbf{x} + \mathbf{b}) \quad (10)$$

where the elements of $\Pi \in \mathbb{R}^{m \times D}$ are drawn from a standard normal distribution, and the elements of $\mathbf{b} \in \mathbb{R}^m$ are drawn from the uniform distribution on $[0, 2\pi]$. The cosine is applied element-wise. Note that we have dropped the usual factor of $\sqrt{\frac{2}{m}}$ from the RBF approximation of (38)—this is because all constant prefactors of ϕ get canceled out in the fraction appearing in Eq. (8). Plugging Eq. (10) into Eq. (9) yields the following:

$$L \approx \text{SelfAttn}(X) + X \quad (11)$$

F. The Role of Normalization and Feed-forward Layers. Thus far we have neglected the role of layer normalization and feedforward layers. That is because these are operations that are performed on a *per-token* basis, and can therefore be dropped into our model without changing anything in our arguments. In the Transformer model, layer normalization occurs directly preceding and succeeding the FFN layer. With layer normalization and FFN layer, our circuit becomes:

$$Y = \text{LayerNorm}(L) \approx \text{LayerNorm}(\text{SelfAttn}(X) + X) \\ \text{Transformer}(X) = \text{LayerNorm}(\text{FFN}(Y) + Y)$$

Note that both the FFN and layer norm operations are biologically plausible. Layer normalization is a form of divisive normalization, which is well-studied in the neuroscience literature (39, 40). In the context of Transformers, it can be achieved by inserting a hidden layer (with lateral connections to implement the normalization) directly before and after the feed-forward layer.

General Case of Untied Weights

In this section we relax the weight tying condition, and generalize our construction to the case when $D \neq d$. Whereas in the previous sections r acted as a gatekeeper for weight matrix H , we will now *also* have r act a gatekeeper for a few other weight matrices. Using the same variable names, consider the following neuron-astrocyte forward equations:

$$\begin{aligned} \mathbf{f} = \mathbf{x} & \in \mathbb{R}^d \\ \mathbf{h} = (1-r)\phi(W_K\mathbf{f}) + r\phi(W_Q\mathbf{f}) & \in \mathbb{R}^m \\ \ell = r(H \odot \tilde{P})\mathbf{h} + (1-r)W_V\mathbf{f} + r\mathbf{f} & \in \mathbb{R}^d \end{aligned} \quad (12)$$

When $r = 0$, we recover the writing phase of Eq. (6); when $r = 1$, we recover the reading phase equations of Eq. (8). When we impose the weight tying constraint of $W_K = W_Q = W$ and $W_V = I$, we recover the original equations of Eq. (4). Eq. (12) describe the neuron-astrocyte implementation of the general Transformer block without the weight sharing constraint imposed.

Numerical Validation

The results derived above have also been checked numerically. In Figure 2 one can see the error between the proposed neuron-astrocyte network and the actual AI Transformer block as a function of the ratio of the width of the hidden layer to the size of the token embedding. As expected from the theoretical analysis, the error between the two networks rapidly decreases as the hidden layer becomes wider. In practice, as the width of the hidden layer becomes 5-10 times the embedding dimension the two networks produce very similar outputs.

Do We Need Astrocytes?

Although we are interested in relating Transformers to glial cells, a natural question when positing any new brain mechanism is: "Can the same behavior be achieved with the old, well-known mechanisms?". This section demonstrates that an equivalent Transformer circuit can be constructed using just neurons, together with divisive normalization achieved via shunting inhibition. The circuit is similar to Eq. (4):

$$\begin{aligned} \mathbf{f} = \mathbf{x} & \in \mathbb{R}^d \\ \mathbf{h} = \phi(W\mathbf{f}) & \in \mathbb{R}^m \\ R = \mathbf{g}^T \mathbf{h} & \in \mathbb{R} \\ \ell = \frac{r}{R} H \mathbf{h} + \mathbf{f} & \in \mathbb{R}^d \end{aligned} \quad (13)$$

The only difference between Eq. (13) and Eq. (4) is the addition of a new element, R , and the removal of the astrocyte processes. Here R is an inhibitory neuron that divisively normalizes feed-forward inputs into layer ℓ . However, it does not inhibit all feedforward inputs equally. Despite both \mathbf{h} and \mathbf{f} being feed-forward inputs to layer ℓ , the divisive inhibition is only implemented on the inputs coming from layer

\mathbf{h} . This can happen, for example, if the feed-forward synaptic inputs coming from layer \mathbf{h} arrive at the dendritic tree close to where inhibitory inputs from neuron R shunt current flow, while the feed-forward inputs coming from layer \mathbf{f} synapse far away from the shunting (41). Leaving the reading and writing phases untouched, circuit Eq. (13) implement the same forward pass as Eq. (4).

Time Scale of Transformers in the Brain

A notable component explicitly missing from Transformers, as well as from our biological circuit, is time. Unlike the brain, Transformers are not dynamical systems. Except in the special instance of autoregressive linear Transformers (which may be seen as a kind of recurrent neural network), Transformers do not maintain a hidden state via recurrent weights (36). Nevertheless, timing is a critical aspect of information processing in the brain, so in this section we describe a plausible timescale for our circuit in terms of known biophysics parameters. The following paragraph argues that the total compute time for our circuit is of the order of $(N + 1) \times 100$ milliseconds, where N is the number of tokens. Our circuit operates in two distinct phases: a reading phase and a writing phase. The reading phase does not involve any plasticity, so the only relevant timescale to compute is how long it takes to traverse the neuron \rightarrow astrocyte \rightarrow synapse pathway. Recent data indicate that this entire signaling pathway can be traversed on the order of a few hundreds of milliseconds (3). The speed of the writing phase is limited by the speed of plasticity. There are two types of plasticity used in our model: 1) Hebbian plasticity between neurons and 2) pre-synaptic plasticity between neuronal and astrocytic processes. The former can happen on relatively fast timescale, for example through spike-timing-dependent plasticity (STDP). In that case of STDP, the timescale can be as fast as a few tens of milliseconds (42). The latter timescale is harder to establish, due to limitations in calcium recording technology. While fast calcium transients in astrocyte processes have been recently recorded (3), and neuron-astrocyte plasticity has been experimentally observed (43), fast (e.g. $< 1s$) neuron-astrocyte plasticity has not been observed yet, possible due to limitations of the calcium imaging. Assuming the timescale of this plasticity is also a few hundred milliseconds, then an estimate for the overall timescale of our model would be a few hundred milliseconds for the read phase, plus a few hundred milliseconds per token for the write phase. This leads to a total compute time of $\tau_{compute} = (100 + N \times 100)$ milliseconds. Finally, note that most "real-world" Transformers are composed of many blocks. Our circuit is a model of a single block. Thus, the total compute time for a k block transformer modelled using our framework is $k\tau_{compute}$.

Discussion

Here we have built a computational neuron-astrocyte model which is functionally equivalent to an important AI architecture: the Transformer. This model serves a dual purpose. The first purpose is to provide a concrete, computational account

of how the communication between astrocytes and neurons subserves brain function. The second purpose is to provide a biologically-plausible account of how Transformers might be implemented in the brain. While the feedback loop between neurons and astrocytes is well-studied from an experimental perspective, there is comparatively little work studying it from a computational perspective (1). Most of the computational studies, with notable exceptions (34, 44, 45), are focused on understanding the biophysics of neuron-astrocyte or astrocyte-astrocyte signaling. An important feature of our model is that it is flexible enough to approximate any Transformer. In other words, we do not only show how to model a particular Transformer (i.e. one with weights that have already been trained for some specific task) – rather, we show how to approximate all *possible* Transformers using neurons and astrocytes. Given the demonstrated power and flexibility of Transformers, this generality can help to explain why astrocytes are so prevalent across disparate brain areas and species. Our model has several immediate applications. First, as calcium imaging technologies improve, it will become increasingly feasible to explicitly compare artificial representations in AI networks to representations in biological astrocyte networks – as is already done when comparing AI networks to biological neural networks (9, 15, 46). Our work suggests that a reasonable place to start may be to compare Transformer normalization terms to astrocytic calcium responses. Second, our work suggests that experimenters may perform causal manipulations on Transformers to generate putative hypotheses for how astrocyte function goes astray in brain disorders and diseases (47). These hypotheses can in turn be potentially used to suggest novel treatments and routes for clinical success.

ACKNOWLEDGEMENTS

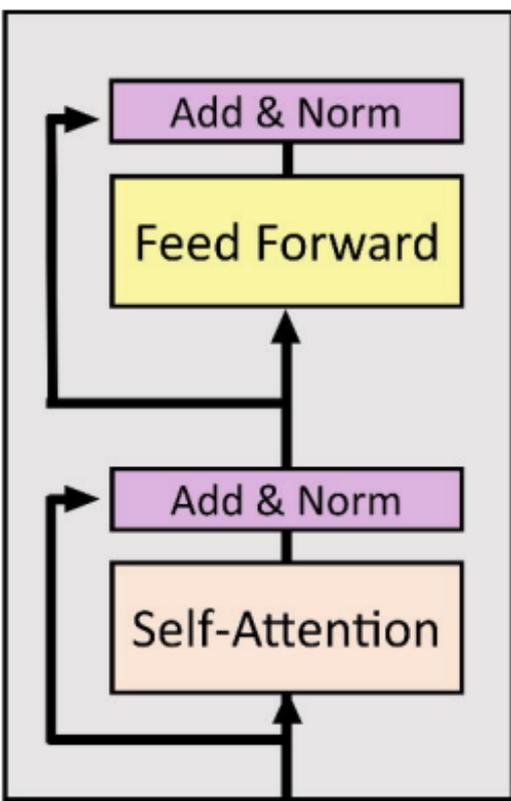
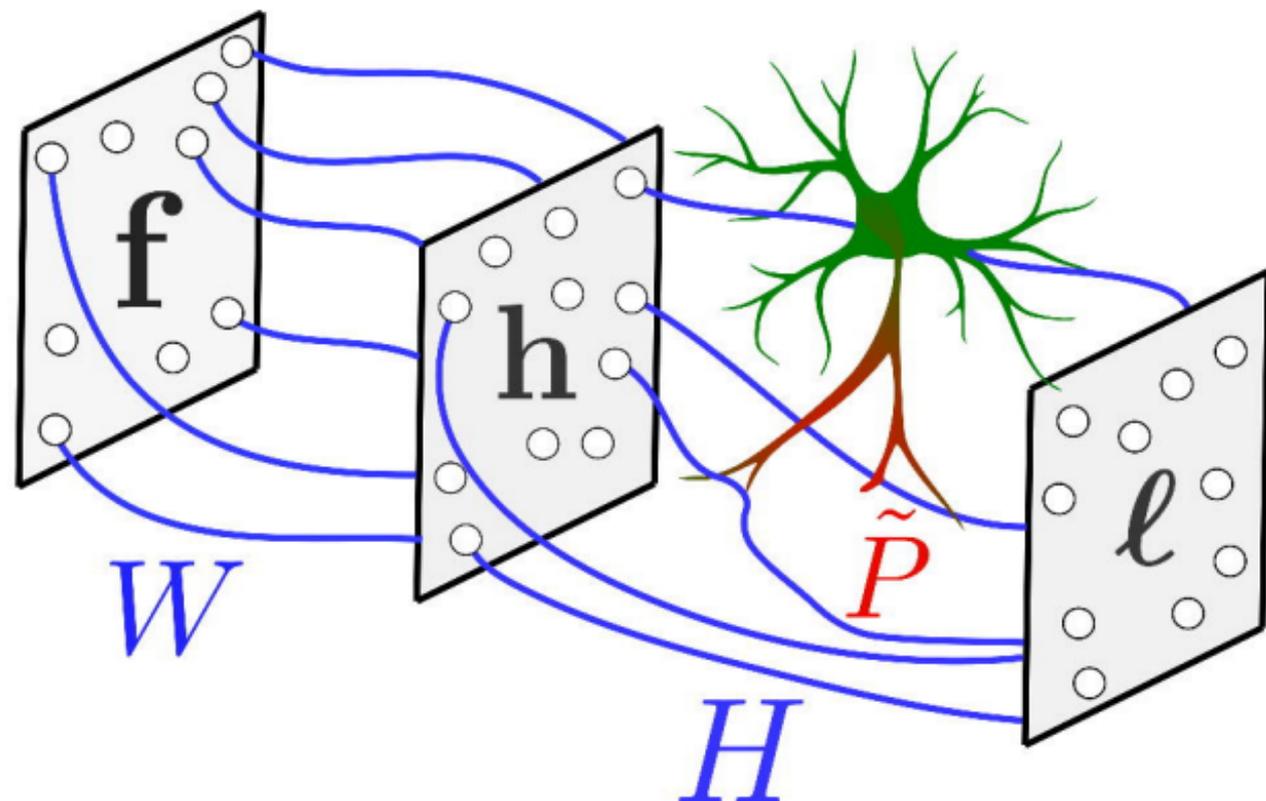
We thank Dan Gutfreund, John Hopfield, and Mriganka Sur for helpful comments and feedback. This work was completed while LK was an MIT-IBM Watson AI Lab Summer 2022 Intern. KVK acknowledges funding from the following sources: BrightFocus Foundation Grant A2020833S, and National Institutes of Health Grant R01AG066171.

Bibliography

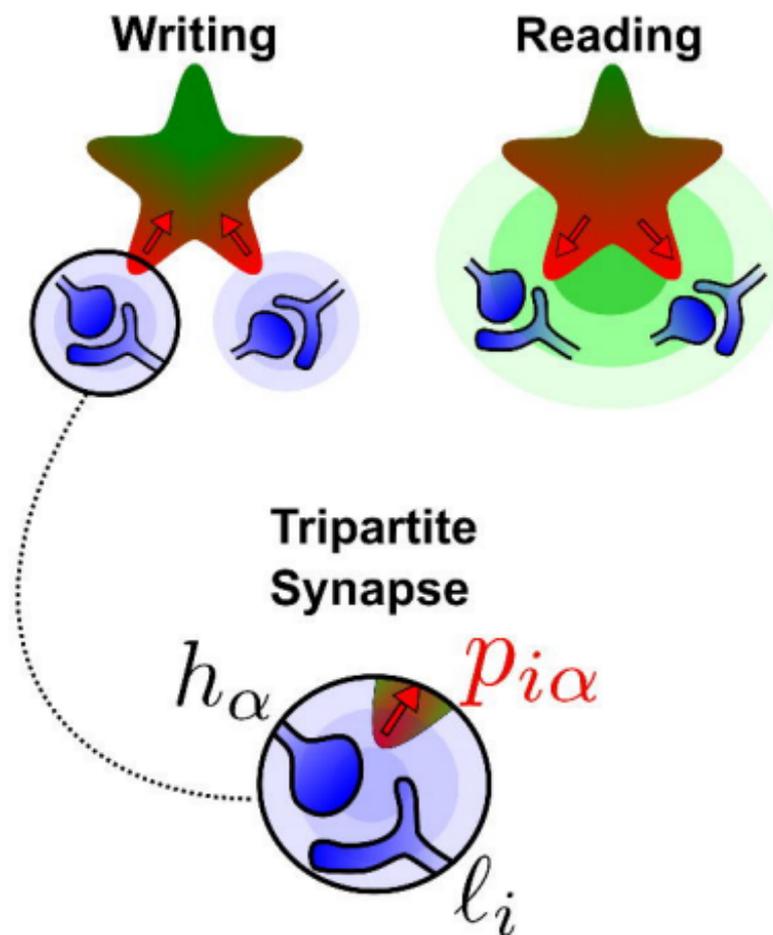
1. Ksenia V Kastanenko, Rubén Moreno-Bote, Maurizio De Pittà, Gertrudis Perea, Abel Erasos-Pichot, Roser Masgrau, Kira E Poskanzer, and Elena Galea. A roadmap to integrate astrocytes into systems neuroscience. *Glia*, 68(1):5–26, 2020.
2. Mónica López-Hidalgo, Vered Kellner, and James Schummers. Astrocyte subdomains respond independently in vivo. *bioRxiv*, page 675769, 2019.
3. Jillian L Stobart, Kim David Ferrari, Matthew JP Barrett, Chaim Glück, Michael J Stobart, Marc Zuend, and Bruno Weber. Cortical circuit activity evokes rapid astrocyte calcium signals on a similar timescale to neurons. *Neuron*, 98(4):726–735, 2018.
4. Paulo Kofuji and Alfonso Araque. Astrocytes and behavior. *Annual Review of Neuroscience*, 44(1):49–67, Jul 2021. ISSN 0147-006X, 1545-4126. doi: 10.1146/annurev-neuro-101920-112225.
5. Barbara Lykke Lind, Alexey R. Brazhe, Sanne Barsballe Jessen, Florence C. C. Tan, and Martin J. Lauritzen. Rapid stimulus-evoked astrocyte ca^{2+} elevations and hemodynamic responses in mouse somatosensory cortex in vivo. *Proceedings of the National Academy of Sciences*, 110(48):E4678–E4687, Nov 2013. doi: 10.1073/pnas.1310065110.
6. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
7. Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *Advances in Neural Information Processing Systems*, 32, 2019.
8. Dmitry Krotov and John J Hopfield. Large associative memory problem in neurobiology and machine learning. In *International Conference on Learning Representations*, 2020.
9. Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.
10. Danil Tyulmankov, Ching Fang, Annapurna Vadaparty, and Guangyu Robert Yang. Biological learning in key-value memory networks. In *Advances in Neural Information Processing Systems*, volume 34, page 22247–22258. Curran Associates, Inc., 2021.
11. Trenton Bricken and Cengiz Pehlevan. Attention approximates sparse distributed memory. *Advances in Neural Information Processing Systems*, 34:15301–15315, 2021.
12. James C. R. Whittington, Joseph Warren, and Timothy E. J. Behrens. Relating transformers to models and neural representations of the hippocampal formation. *arXiv:2112.04035 [cs, q-bio]*, Mar 2022. arXiv: 2112.04035.
13. Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):1–10, 2022.
14. Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *arXiv preprint arXiv:2106.04554*, 2021.
15. Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, Jun 2014. doi: 10.1073/pnas.1403112111.
16. John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.
17. Gertrudis Perea, Marta Navarrete, and Alfonso Araque. Tripartite synapses: astrocytes process and control synaptic information. *Trends in Neurosciences*, 32(8):421–431, Aug 2009. ISSN 0166-2236. doi: 10.1016/j.tins.2009.05.001.
18. Christopher S Von Bartheld, Jami Bahney, and Suzanaerculano-Houzel. The search for true numbers of neurons and glial cells in the human brain: A review of 150 years of cell counting. *Journal of Comparative Neurology*, 524(18):3865–3895, 2016.
19. Michael M Halassa, Tommaso Fellin, Hajime Takano, Jing-Hui Dong, and Philip G Haydon. Synaptic islands defined by the territory of a single astrocyte. *Journal of Neuroscience*, 27(24):6473–6477, 2007.
20. N. A. Oberheim, T. Takano, X. Han, W. He, J. H. C. Lin, F. Wang, Q. Xu, J. D. Wyatt, W. Pilcher, J. G. Ojemann, B. R. Ransom, S. A. Goldman, and M. Nedergaard. Uniquely hominid features of adult human astrocytes. *Journal of Neuroscience*, 29(10):3276–3287, Mar 2009. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.4707-08.2009.
21. Alexey Semyanov and Alexei Verkhratsky. Astrocytic processes: from tripartite synapses to the active milieu. *Trends in neurosciences*, 44(10):781–792, 2021.
22. Alexei Verkhratsky and Arthur Butt. *Glial neurobiology: a textbook*. John Wiley & Sons, 2007.
23. Eric A Newman. New roles for astrocytes: Regulation of synaptic transmission. *Trends in Neurosciences*, 26(10):536–542, Oct 2003. ISSN 0166-2236. doi: 10.1016/S0166-2236(03)00237-6.
24. Grant R.J. Gordon, Karl J. Iremonger, Srinivas Kantevari, Graham C.R. Ellis-Davies, Brian A. MacVicar, and Jaideep S. Bains. Astrocyte-mediated distributed plasticity at hypothalamic glutamate synapses. *Neuron*, 64(3):391–403, Nov 2009. ISSN 08966273. doi: 10.1016/j.neuron.2009.10.021.
25. Jai-Yoon Sul, George Orosz, Richard S. Givens, and Philip G. Haydon. Astrocytic connectivity in the hippocampus. *Neuron glia biology*, 1(1):3–11, Feb 2004. ISSN 1740-925X. doi: 10.1017/s1740925x04000031.
26. Nahoko Kuga, Takuya Sasaki, Yuji Takahara, Norio Matsuki, and Yuji Ikegaya. Large-scale calcium waves traveling through astrocytic networks in vivo. *The Journal of Neuroscience*, 31(7):2607–2614, Feb 2011. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.5319-10.2011.
27. Eliana Scemes and Christian Giaume. Astrocyte calcium waves: what they are and what they do. *Glia*, 54(7):716–725, 2006.
28. Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv:1906.02762 [cs, stat]*, Jun 2019. arXiv: 1906.02762.
29. Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR, 2022.
30. Yongyi Yang, Zengfeng Huang, and David Wipf. Transformers from an optimization perspective. *arXiv preprint arXiv:2205.13891*, 2022.
31. Gertrudis Perea and Alfonso Araque. Astrocytes potentiate transmitter release at single hippocampal synapses. *Science*, 317(5841):1083–1086, 2007.
32. Gertrudis Perea, Marta Navarrete, and Alfonso Araque. Tripartite synapses: astrocytes process and control synaptic information. *Trends in neurosciences*, 32(8):421–431, 2009.
33. Maurizio De Pittà, Nicolas Brunel, and Andrea Volterra. Astrocytes: Orchestrating synaptic plasticity? *Neuroscience*, 323:43–61, 2016.
34. Vladimir Ivanov and Konstantinos Michmizos. Increasing liquid state machine performance with edge-of-chaos dynamics organized by astrocyte-modulated plasticity. *Advances in Neural Information Processing Systems*, 34:25703–25719, 2021.
35. DD Rasmuson. The role of acetylcholine in cortical synaptic plasticity. *Behavioural brain research*, 115(2):205–218, 2000.
36. Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, page 5156–5165. PMLR, Nov 2020.
37. Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah A Smith, and Lingpeng Kong. Random feature attention. *arXiv preprint arXiv:2103.02143*, 2021.
38. Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.
39. Matteo Carandini and David J Heeger. Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62, 2012.
40. Yang Shen, Julia Wang, and Saket Naviakha. A correspondence between normalization strategies in artificial and biological neural networks. *Neural Computation*, 33(12):3179–3203, Nov 2021. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco_a_01439.
41. Frances S Chance and LF Abbott. Divisive inhibition in recurrent networks. *Network: Computation in Neural Systems*, 11(2):119, 2000.
42. Stefan Klampfl and Wolfgang Maass. Emergence of dynamic memory traces in cortical microcircuit models through stdp. *Journal of Neuroscience*, 33(28):11515–11529, 2013.
43. Wayne Croft, Katharine L Dobson, and Tomas C Bellamy. Plasticity of neuron-glia trans-

- mission: equipping glia for long-term integration of network activity. *Neural plasticity*, 2015, 2015.
44. Tao Song, Pan Zheng, ML Dennis Wong, and Xun Wang. Design of logic gates using spiking neural p systems with homogeneous neurons and astrocytes-like control. *Information Sciences*, 372:380–391, 2016.
 45. Guangzhi Tang, Ioannis E Polykretis, Vladimir A Ivanov, Arpit Shah, and Konstantinos P Michmizos. Introducing astrocytes on a neuromorphic processor: Synchronization, local plasticity and edge of chaos. In *Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop*, pages 1–9, 2019.
 46. Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2020.
 47. Carole Escartin, Elena Galea, András Lakatos, James P O’Callaghan, Gabor C Petzold, Alberto Serrano-Pozo, Christian Steinhäuser, Andrea Volterra, Giorgio Carmignoto, Amit Agarwal, et al. Reactive astrocyte nomenclature, definitions, and future directions. *Nature neuroscience*, 24(3):312–325, 2021.

A


 \approx


B



Error Between Neuron-Astrocyte Network
And Transformer (Log-Log)

