BriLLM:
Brain-inspired Large Language Model

Hai Zhao[*]**, Hongqiu Wu, Dongjie Yang, Anni Zou, Jiale Hong**
Computer School, Shanghai Jiao Tong University
zhaohai@cs.sjtu.edu.cn, {wuhongqiu, djyang.tony, annie0103, hongjiale}@sjtu.edu.cn

Abstract

This paper reports the first brain-inspired large language model (BriLLM). This is a non-Transformer, non-GPT, non-traditional machine learning input-output controlled generative language model. The model is based on the Signal Fully-connected flowing (SiFu) definition on the directed graph in terms of the neural network, and has the interpretability of all nodes on the graph of the whole model, instead of the traditional machine learning model that only has limited interpretability at the input and output ends. In the language model scenario, the token is defined as a node in the graph. A randomly shaped or user-defined signal flow flows between nodes on the principle of ”least resistance” along paths. The next token or node to be predicted or generated is the target of the signal flow. As a language model, BriLLM theoretically supports infinitely long n-gram models when the model size is independent of the input and predicted length of the model. The model's working signal flow provides the possibility of recall activation and innate multi-modal support similar to the cognitive patterns of the human brain. At present, we released the first BriLLM version in Chinese, with 4000 tokens, 32-dimensional node width, 16-token long sequence prediction ability, and language model prediction performance comparable to GPT-1. More computing power will help us explore the infinite possibilities depicted above. [1] [2]

1 Introduction

Large language models (LLMs) are igniting the prospect of AGI (artificial general intelligence). However, even SOTA LLMs are still in terms of Transformer architecture and GPT training scheme unlikely to laugh at the final termination of AGI due to the huge difficulties in their scalability and interpretability, let alone the way Transformer or GPT-based LLM works is a far cry from the human brain, the alternative intelligence machine already existing in the nature for millions of years, showing how a true AGI must be.

The Transformer (Vaswani et al., 2017) has been a fundamental and indispensable framework for building SOTA LLM backbones. Although Transformers have demonstrated remarkable generalization capabilities across diverse tasks and scalability to achieve higher intelligence, the quadratic computational complexity of the attention mechanism over input sequences poses significant efficiency challenges, particularly for long sequences. This computational bottleneck has spurred research into more efficient attention variants, such as linear attention mechanisms, and RNN-like Transformers. While these studies focus on preserving model performance and lowering computational costs, they merely mitigate the issue without resolving the computational bottleneck at its core, since they remain dependent on attention-based mechanisms or attention variants.

Furthermore, the Transformer architecture exhibits limited parameter-level interpretability due to its complex self-attention mechanisms and opaque parameter interactions, a characteristic that renders

[*] Main Contribution. This work was not supported by any funding like NSFC. If any one of my six funding proposals on the idea in this paper more or less in the last seven years had been funded, then this work would have been accomplished much earlier with a much more solid basis. So all thanks only go to insistence, enthusiasm and dream in the hearts of our team.

[1] Code at: `https://github.com/brillm05/BriLLM0.5`

[2] Model at: `https://huggingface.co/BriLLM/BriLLM0.5`

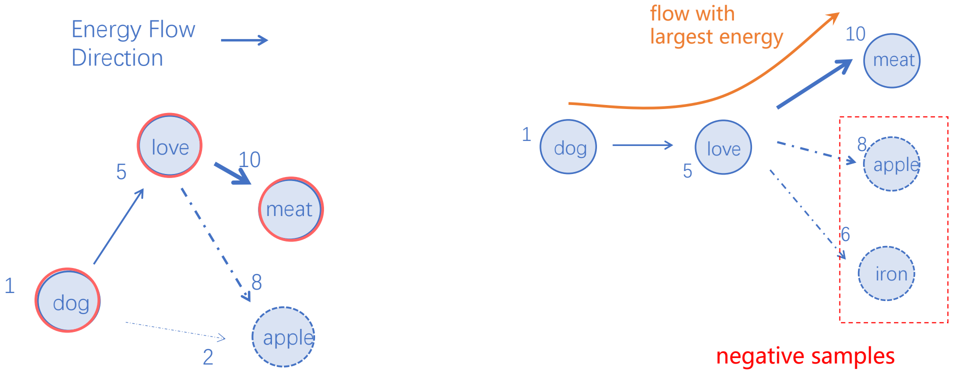
it functionally analogous to a black-box system. Many studies endeavor to reveal the black box by interpreting the intrinsic mechanism of self-attention or enhancing the interpretability of the model through visualization, attribution methods, and probing tasks. However, the complicated interaction of attention between hidden states still remains poorly understood.

To address these challenges, we propose BriLLM, a novel architecture for language modeling that is inspired by signal propagation among neurons in the brain. The BriLLM architecture is structured as a bi-directional graph with multiple nodes and edges. Each node (currently set as a hidden layer of neurons) represents a token, and BriLLM leverages fully-connected neural networks as edges to construct the relationship between these nodes. Like neural signal propagation through biological pathways, BriLLM predicts subsequent tokens by identifying the optimal pathway for energy tensor propagation across nodes. Central to this process is the energy tensor — a dynamic signal representation within BriLLM — which guides the selection of the next node (token). At each step, the model evaluates candidate edges (transitions) and selects the one that maximizes the energy tensor's value, ensuring coherent and contextually relevant token generation.

The proposed mechanism termed Signal Fully-connected Flowing (SiFu) systematically models the entire signal propagation process. This SiFu architecture comprises three core components: (1) a fully-connected directed graph topology where each node maintains bidirectional connections with all other nodes, (2) a dynamic weighting system that modulates signal transmission intensity between nodes based on their functional correlations, and (3) a nonlinear activation module that enables hierarchical relationship extraction during signal propagation.

2 SiFu Mechanism

Inspired by the working mode of the brain, we propose *Signal Fully-connected Flowing (SiFu)* on the Directed Graph, a novel input-output stream control mechanism for machine learning, serving as the core design of BriLLM. As shown in Figure 1a, *SiFu* model is a graph composed of multiple nodes, which are sparsely activated and utilize tensors to transmit a nominal signal. Each node (ideally, a layer of neurons) represents a certain concept or word, e.g., a noun, a verb, etc. Each edge models the relationship between every node pair. The signal is transmitted by the magnitude of the energy. The energy will be strengthened, i.e., maximized, if it is in the right route. Or, at least, the right path always keeps the maximal energy for the transmitted signal. Each node is sequentially activated in terms of the maximized energy. Route or path is determined in a competitive way, i.e., the next node will be activated only if the energy can be maximally delivered in this node.



(a) Forward

(b) Optimization

Figure 1: An illustration of SiFu Directed Graph.

SiFu model works in a straightforward way, after choosing a series of tokens as input, let a signal continuously transmit from the the beginning node in order, all the tokens represented by each node along the right path that the signal energy keeps the maximal compared to other alternative paths will be collected as the output.

For example, as shown in Figure 1a, the path "dog → love → meat" has the largest energy. As shown in Figure 1b, the correct sequence should yield the largest energy. For example, to calculate the loss for the sequence "love → meat": multiple negative samples in the vocabulary, such as "apple" and "iron," are selected. Energy tensors are computed for both the ground-truth node ("meat") and negative nodes ("apple", "iron"). The cross-entropy loss maximizes the energy associated with the node "meat" while minimizing energies from the negative nodes.

3 BriLLM Formulation

BriLLM implements *SiFu* neural network for language modeling, as shown in Figure 3. Each token in the vocabulary is modeled as a node, which is defined by a hidden layer of neurons in the neural network. Similarly to other sequence transduction models, we convert each token into a learned embedding vector $x_i \in \mathbb{R}^{d_{model}}$. An edge connecting token u and v is modeled as a fully-connected matrix $W_{u,v} \in \mathbb{R}^{d_{model} \times d_{model}}$ and a bias $b_{u,v} \in \mathbb{R}^{d_{model}}$. Two fully-connected matrices $W_{u,v}$ and $W_{v,u}$ play the roles of the bidirectional edges between tokens. The signal tensors are fitted into matrices. The forward process begins with an initial signal shape:

$$e_1 = \mathbf{1}_n = [1, 1, \dots, 1]^\top \in \mathbb{R}^{d_{model}} \tag{1}$$

Then an expanded signal tensor $\mathcal{E}_i \in \mathbb{R}^{d_{model}}$ is computed as a linear weighted sum of previous signals using learnable weights $w \in \mathbb{R}^{maxlen}$:

$$\mathcal{W} = \mathrm{softmax}(w_{1:i}) \tag{2}$$

$$\mathcal{E}_i = \sum_{k=1}^{i} \mathcal{W}_k e_k \tag{3}$$

The signal flowing from token x_i to all other tokens $e_{i+1,v} \in \mathbb{R}^{d_{model}}$ will be computed:

$$e_{i+1,v} = \mathrm{GeLU}(W_{x_i,v}\mathcal{E}_i + b_{x_i,v})$$

During decoding, the next token x_{i+1} is determined by the edge with the largest energy:

$$x_{i+1} = \arg\max_v \|e_{i+1,v}\|_2 \tag{4}$$

$$e_{i+1} = \max \|e_{i+1,v}\|_2$$

The L2 norm of the signal tensor represents its energy magnitude.

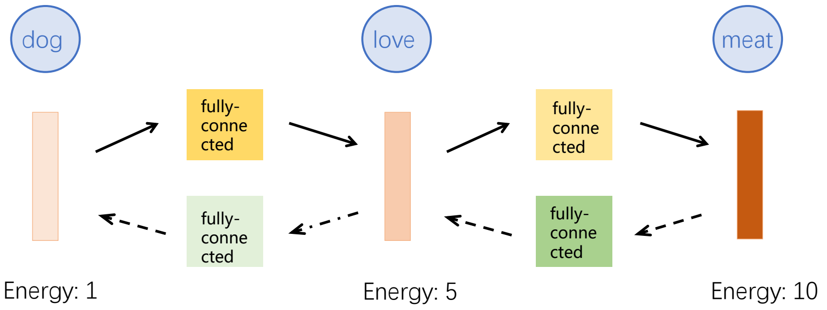


Figure 2: The architecture of BriLLM.

To train a sample in BriLLM, every time we build an individual common neural network to perform the regular BP training. This network consists of two parts, in which the front part connects all input nodes (i.e., tokens), then it follows the rear parts which connect all possible paths in order. At last, a softmax layer collects all paths' energy tensors to indicate the right path with a 0-1 ground truth vector. We adopt a cross-entropy loss for training.

4 Experiments

We released BriLLM-Chinese-V0.5 model.

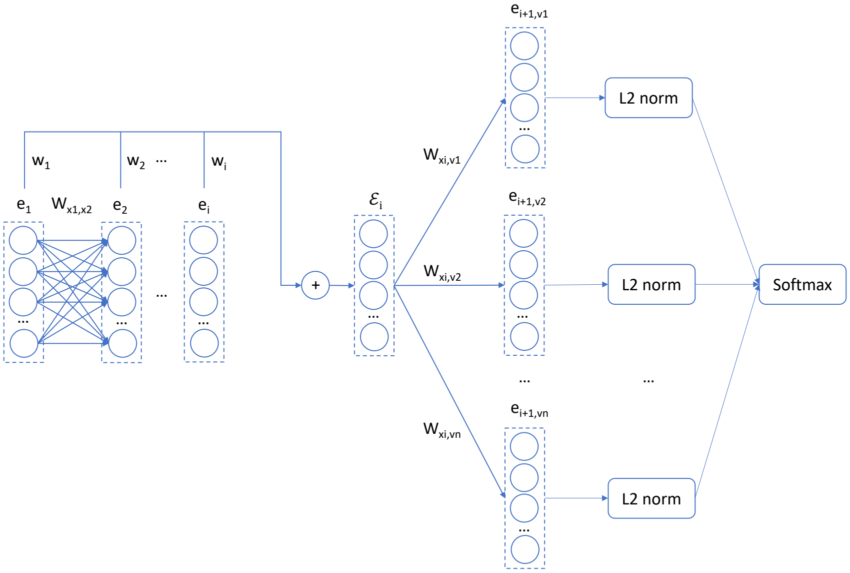


Figure 3: The training network of BriLLM for one training sample .

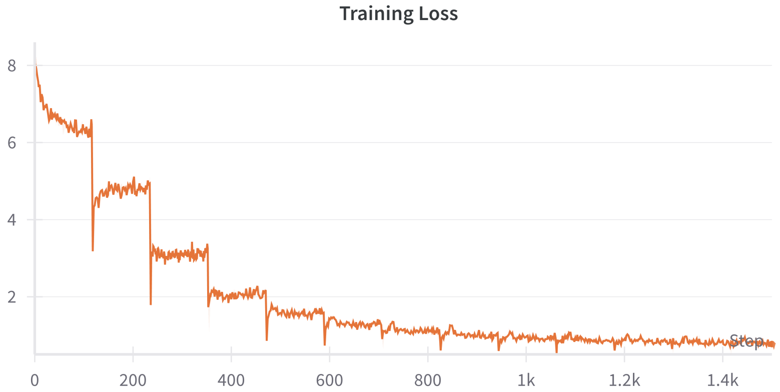


Figure 4: The training loss.

Dataset We use the subset from the Chinese version of Wikipedia, which contains over 100M Chinese characters. We truncate the long sentences into small sentences with a maximum length of 16. We select a vocabulary of 4,000 tokens consisting of the most frequently used Chinese characters.

Implementation Details. BriLLM is implemented using PyTorch. It uses sinusoidal positional encoding, GeLU as the activation function, cross-entropy loss for next-token prediction, and an embedding size of $d_{model} = 32$. We used the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The model size is about $512 + 4000 * 4000 * (32 * 32 + 32) \approx 16B$. We trained our models on one machine with 8 NVIDIA A800 GPUs for 1.5k steps. The training loss is shown in Figure 4.

Complexity n is the sequence length, v is the vocabulary size, and d is the representation dimension. The computational complexity is $O(n \cdot v \cdot d^2)$.

Case Study Table 1 presents some of the decoding results, including both training samples and test samples.

Input	**Completion**
Training samples	
《幽明录》，亦作	《幽明录》，亦作《幽冥录》、《我
《罗马》描述了	《罗马》描述了古罗马从共和国走下
众多哥萨克领导人开始	众多哥萨克领导人开始改变阵营运动
阿根廷探戈是起源于	阿根廷探戈是起源于阿根廷或乌拉圭
惠安第一中学是	惠安第一中学是位于福建省泉州市天
提琴本泛指	提琴本泛指有琴杆，需一手持杆另一
塞人定义为	塞人定义为阿塞拜疆人或南高加索地
袁乃宽旧居是	袁乃宽旧居是清朝及中华民国军事情
飞翔公园站是	飞翔公园站是广州地铁二号线的一。
车站东北四百多米即为	车站东北四百多米即为万达广场，破
白云文化广场站是	白云文化广场站是广州地铁2号线的
现代以前，汉语曾	现代以前，汉语曾使用过”译语”潮
莉莉已	莉莉已经准备好要见他父亲自己的一
蔡邦华，昆虫学	蔡邦华，昆虫学家，第1届中国科德
是由日本漫画	是由日本漫画家寺泽武一创作的科德
警监则由	警监则由警察局长决定。另外两岸气
，他的小提琴	，他的小提琴演奏技巧备受关注。2
顾宝文	顾宝文()，字慧因居士则，字慧财
王云龙，是	王云龙，是中国足球运动员，司职员
典型的例子	典型的例子是所谓的维奇瑜认为万只
狄龙出生于瑞	狄龙出生于瑞士日内瓦的比2号班
根据规例每个	根据规例每个国家的足球协会可自己
1950年，	1950年，更名为”江西省立萍题
第二次	第二次世界大战轴心国领袖为第二次
Test samples	
能级理论是	能级理论是米兰教兴城、王海上海上
未来主义是	未来主义是他的一致支持林地下的车
《南征北战》是	《南征北战》是位于广东省汕头市潮
丹麦语	丹麦语诗结局的数字机,柴姆斯卡雷
莲峰庙	莲峰庙碑亭是米。”,设立为那亚州
他也不认为	他也不认为一个地区()是一个地区
卓越工程师	卓越工程师评量大陆的固的选择权—
群众只能够	群众只能够喷嘴能随即在宗,每年去
晚些时候	晚些时候阮惠安岭林斯·罗力发的第
他是	他是日返自行车特的一部,但没有的

Table 1: Case study of decoding results.

5 Conclusion, Limitation and the Future

BriLLM introduces a novel framework for language modeling by replacing attention-based architectures with a brain-inspired dynamic signal propagation mechanism over a fully connected graph. By representing tokens as nodes and leveraging energy tensor dynamics to identify optimal pathways, the model is capable of doing non-autoregressive generation, full node-level interpretability, and theoretically infinite n-gram modeling. Its biologically plausible design decouples model size from sequence length, enabling efficient resource utilization while simulating neurocognitive processes like memory formation. This work challenges the dominance of attention mechanisms, offering a scalable, transparent alternative aligned with neural signaling principles.

Currently, due to our quite limited computational power for this work, we just reach an early model checkpoint with a moderate hyperparameter setting on the Chinese language. However, the current released model has demonstrated promising performance compared to GPT-1 (Radford et al., 2018). Meanwhile, so limited computational resource puts huge obstacles to let us explore much more potential over BriLLM.

Our existing BriLLM implementation has a size of $d_{model} \times d_{model} \times n \times n$, where n is the number of tokens (nodes) and d_{model} is embedding size. This quadratically increasing model size is indeed an inconvenience. However, as most model parameters come from the fully-connected matrices, it is possible to adopt a sort of sparse representation or shared parameters for those less active tokens, i.e., set a default non-updated matrix for all these inactive tokens. Our preliminary estimation shows such a strategy may save up to 90% parameters for BriLLM.

Both our BriLLM training practice and the SiFu mechanism show BriLLM is hard to efficiently train in parallel as every time the training has to be conducted in a different individual neural network. In addition, a theoretically accurate training objective needs the right predicted token to compare its energy to all the other tokens. When the token set is large, such ranking may result in a very wide softmax output layer, which further slows the training down and requires much larger training memory. It is lucky that such inconvenience may be alleviated by some sort of approximated ranking strategy. Namely, BriLLM training may be done locally only within those ‘necessary’ compared to counterpart tokens. When all these locally trained networks do not overlap, then all these local networks can be trained parallelly, so that the entire BriLLM model training can be done in a good parallel way.

Full model interpretability of BriLLM theoretically facilitates BriLLM to serve as a multi-modal model by nature. Each node in BriLLM does not have to be defined as tokens from languages, they are surely capable of being defined as alternative modal units or jointly defined among different modalities. It is different from LLM, in the case of node-redefinition, no matter one or many nodes, the BriLLM does not need to be re-trained from the very beginning. In one word, the full model interpretability enables BriLLM a natural multi-modal model design, helping the machine learning model closer to the cognition mode as the human brain.

Note that even though BriLLM theoretically supports an infinite-gram language model without increasing model size, in practice, the model during training has to cover long enough input sequences, otherwise BriLLM decoding cannot give good enough sequence prediction beyond the training sample length. However, facilitating longer sequence prediction in terms of BriLLM just depends on longer training without resizing the model itself.

So far, we adopt a uniform signal vector like Eq. (1). However, this shape of the signal is not necessary. We tried a randomly initialized signal, the BriLLM can be stably trained. According to the definition of BriLLM, the signal is indeed exploited nominally, however, it may differ the way for activating the input of BriLLM. In the future, we may explore the function of the signal as that of the pre-filled prompt in LLM. If the shape of the signal can be properly used as the primary scenario setting to specify the working of BriLLM, then this should be a much more natural way against in-context learning in the current LLM.

The last but not the least issue we need to explore about BriLLM is the possibility of supervised finetuning (SFT) like LLM. Note that as BriLLM does not need to resize the model for any sized input or output sequences and the size BriLLM has to be quadratically correlated to the embedding size and token numbers, it is not in an advantageous position when the model sizes are the same ’small’ or moderate as LLM. As we reported in this paper, a 16B BriLLM (our current released

Table 2: Comparison of LLM and BriLLM.

	LLM	BriLLM
model size	correlated to input context length	independent
interpretability	only in input & output	all nodes throughout the model
multi-modal implementation	limited to be joined from input/output	all nodes throughout the model

checkpoint) only gives comparable performance as 0.1B GPT-1. Thus, we have reasons to speculate that BriLLM has a very high emergent ability threshold. What's more, now we even do not know how to do SFT over BriLLM, which leaves a big future work.

References

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI blog*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL `https://arxiv.org/abs/1706.03762`.