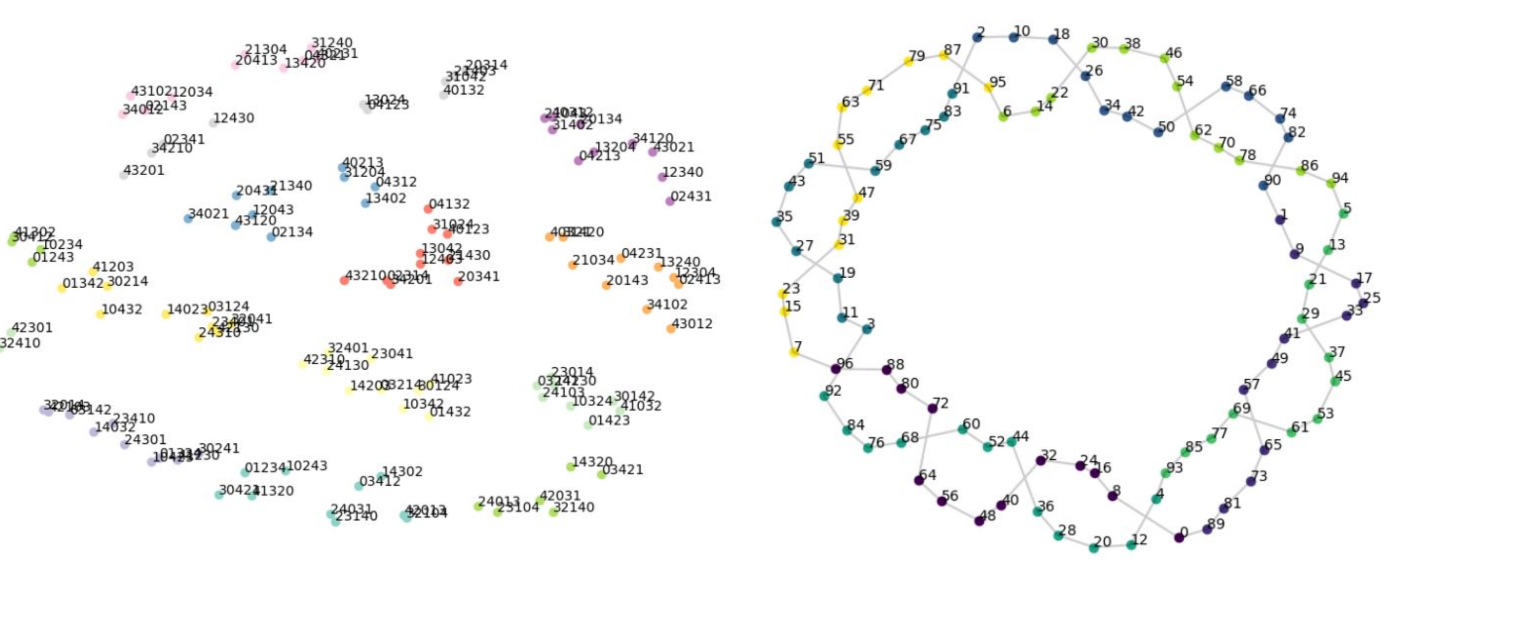
Fourier Circuits in Neural Networks: Unlocking the Potential of Large Language Models in Mathematical Reasoning and Modular Arithmetic

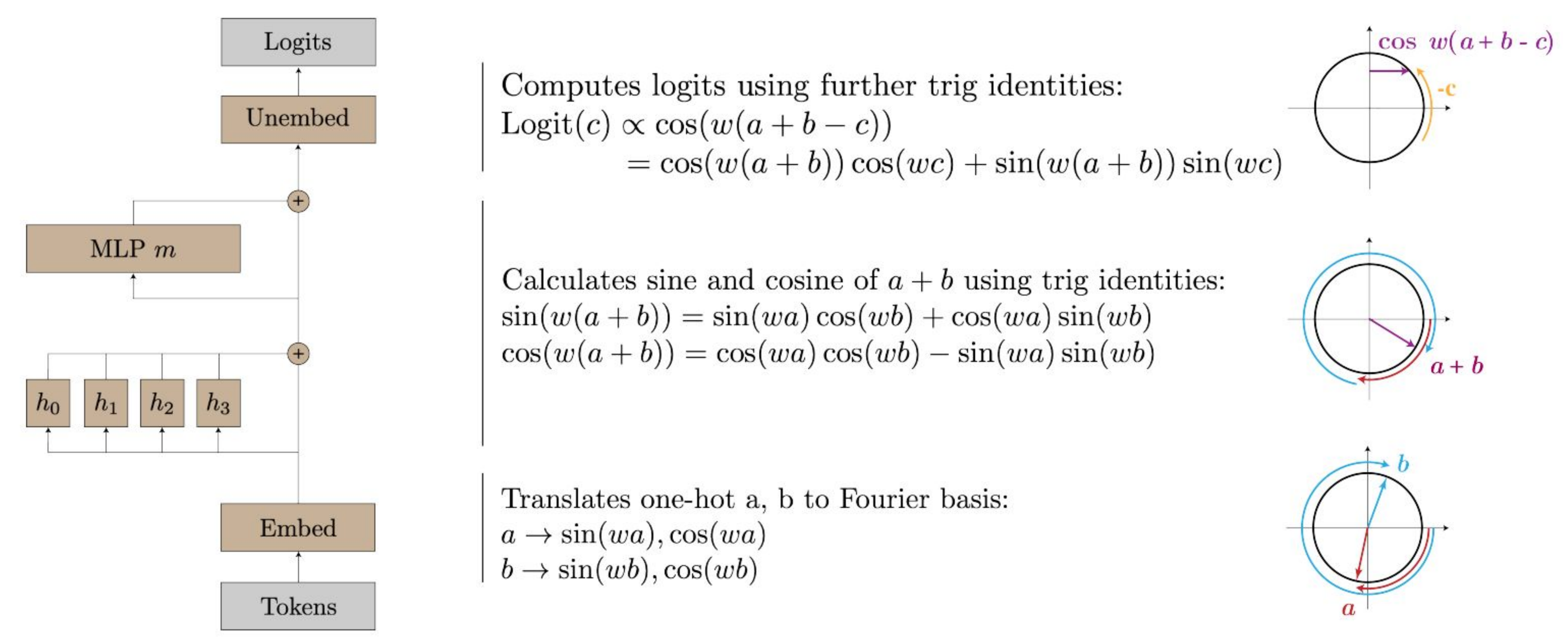
Jiuxiang Gu, Chenyang Li, Yingyu Liang, Zhenmei Shi, Zhao Song, Tianyi Zhou

Background



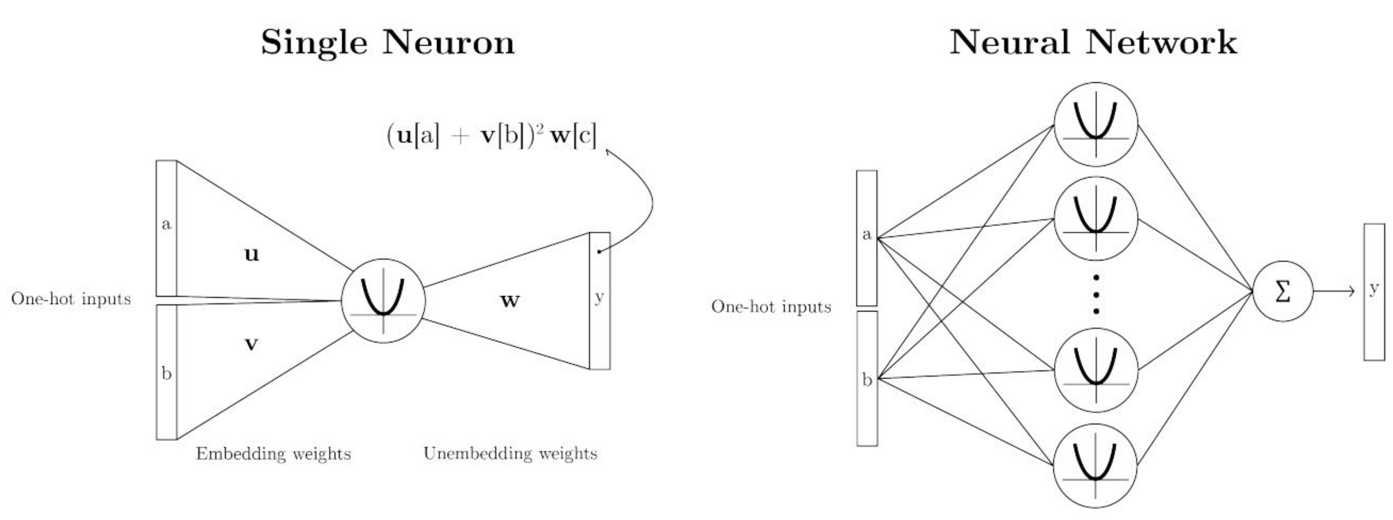
Visualizing the mathematical operations learning

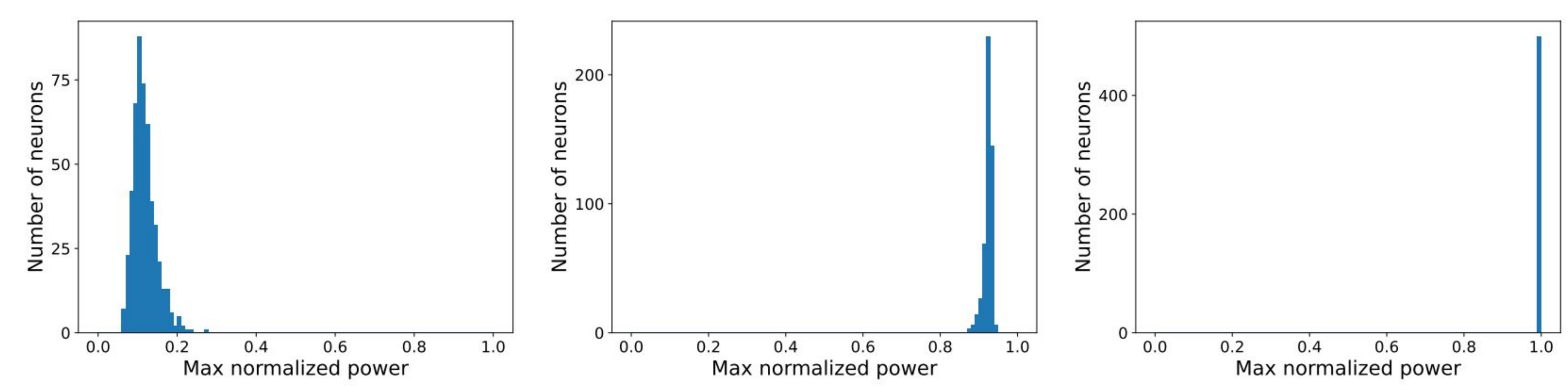
Source "Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets." (arXiv 2022)



The attention and MLP module in the Transformer imbues the neurons with Fourier circuit-like properties

Source "Progress measures for grokking via mechanistic interpretability." (arXiv 2023)

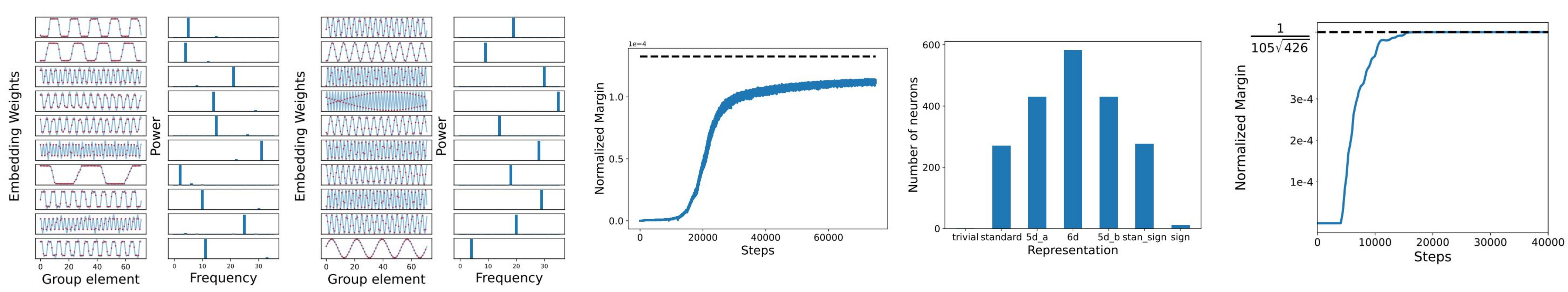




An individual neuron and one-hidden layer neural network learning

Source "Feature emergence via margin maximization: case studies in algebraic tasks." (arXiv 2023)

Motivation



Fourier power spectrum for a 1-hidden layer ReLU network and quadratic activation

Source "Feature emergence via margin maximization: case studies in algebraic tasks." (arXiv 2023)

Problem Setup

- The modular dataset

$$D_p := \{((a_1, \ldots, a_k), \sum_{i\in[k]} a_i) : a_1, \ldots, a_k \in \mathbb{Z}_p\}$$

- One-hidden layer networks $f(\theta, x) := \sum_{i=1}^{m} \phi(\theta_i, x)$
- A single neuron

$$\phi(\{u_1, \ldots, u_k, w\}, x_1, \ldots, x_k) := (u_1^\top x_1 + \cdots + u_k^\top x_k)^k w$$

- For input elements $(a_1, \ldots, a_k)$, a neuron simplifies to

$$\phi(\{u_1, \ldots, u_k, w\}, a_1, \ldots, a_k) = (u_1(a_1) + \cdots + u_k(a_k))^k w$$

- With $\theta = \{u_{i,1}, \ldots, u_{i,k}, w_i\}_{i=1}^m$, the network is denoted as:

$$f(\theta, a_1, \ldots, a_k) := \sum_{i=1}^{m} \phi(\{u_{i,1}, \ldots, u_{i,k}, w_i\}, a_1, \ldots a_k)$$

Theoretical Results

Theorem 1

If $m \geq 2^{2k-1} \cdot \frac{p-1}{2}$, then the max $L_{2,k+1}$-margin network satisfies:

- The maximum $L_{2,k+1}$-margin for a given dataset D_p is:

$$\gamma^* = \frac{2(k!)}{(2k+2)^{(k+1)/2}(p-1)p^{(k-1)/2}}$$

- For each neuron $\phi(\{u_1, \ldots, u_k, w\}; a_1, \ldots, a_k)$ there is a constant scalar $\beta \in \mathbb{R}$ and a frequency $\zeta \in \{1, \ldots, \frac{p-1}{2}\}$ satisfying

$$\begin{aligned} u_1(a_1) &= \beta \cdot \cos(\theta^*_{u_1} + 2\pi\zeta a_1/p) \\ u_2(a_2) &= \beta \cdot \cos(\theta^*_{u_2} + 2\pi\zeta a_2/p) \\ &\cdots \\ u_k(a_k) &= \beta \cdot \cos(\theta^*_{u_k} + 2\pi\zeta a_k/p) \\ w(c) &= \beta \cdot \cos(\theta^*_w + 2\pi\zeta c/p) \end{aligned}$$

Take-Home Message

Our research delves into the complexities of neural networks and Transformers, focusing on their strategies for solving modular addition with multiple inputs. We uncover that one-hidden layer networks, with a neuron count of $m \geq 2^{2k-2} \cdot (p-1)$, optimize an $L_{2,k+1}$-margin on modular arithmetic datasets, aligning each neuron with a unique Fourier spectrum for problem-solving. Corroborating empirical evidence further illuminates the computational mechanisms, notably in Transformers' attention matrices, marking a substantial advance in deciphering their algebraic operation sophistication.