The Representations of Deep Neural Networks Trained on Dihedral Group Multiplication

Do deep neural networks (DNNs) reuse the same algorithmic primitives to learn group multiplications? For modular addition, [1] describe network representations as (*approximate) cosets*, proving *O(log n)* such representations are learned. They conjecture that these same primitives will be used in DNNs learning dihedral group multiplication (DGM). Thus, we reverse engineer DNNs trained on DGM, finding representations are *exact* or *approximate cosets* by studying the activation geometry of clusters of neurons, finding manifolds aligned with (*approximate) coset structures*.

Background. The *dihedral group* D_n is the symmetries of a regular *n*-gon, containing *2n* elements: *n* rotations r^k for $k \in \{0, \ldots, n-1\}$ that rotate the *n*-gon by $2\pi/n$ radians, and *n* reflections sr^k reflecting about *n* distinct axes. The rotation r^0 is the *identity* element, denoted *e*, for which $ex = xe = x$ for any $x \in D_n$. These operations form a non-commutative group multiplication when $n \geq 3$, meaning the order in which operations are multiplied matters—for instance, $sr \neq rs$. **DGM**: $a \cdot b = C$, $a,b \in D_n$ involves composing two symmetries in sequence (notation: applied right to left): $r^a \cdot r^b = r^{(a+b) \bmod n}$ *(rotation)*, $sr^a \cdot r^b = sr^{(a+b) \bmod n}$ *(reflection)*, $r^a \cdot sr^b = sr^{(b-a) \bmod n}$ *(reflection)*, $sr^a \cdot sr^b = r^{(b-a) \bmod n}$ *(rotation)*. Cayley graphs encode geometric structure. A Cayley graph of D_n is expressed via a *generating set {r, s}*, where nodes are group elements and (directed) edges are labeled by *{r, s}*. Particularly, an edge labeled $x \in \{r,s\}$ between nodes *a, b* exists if $b = xa$. Fig. 1 shows a Cayley graph for D_3. We will train on D_{18}*: note that* D_{18} *can be decomposed into* six D_3 graphs, with each one corresponding to a coset. To contrast, approximate cosets arise when neurons fail to decompose D_{18} into subgraphs, *i.e.* approximate cosets are when all of D_{18}, is learned.

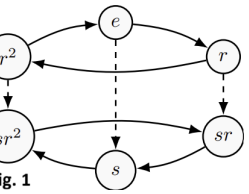

Fig. 1

Results. We cluster neurons by their group-Fourier-basis similarities. For each cluster, we then apply PCA to the activation matrix (neuron × datum). This analysis shows that the representations learned by DNNs are either cosets (Fig. 2) or approximate cosets (Fig. 3). In Fig. 2, six disjoint hexagrams are learned, each corresponding to a different D_3, which together compose D_{18}. In contrast, Fig. 3 shows that a single non-disjoint approximate coset structure is learned, covering all of D_{18} without decomposition. Fig. 3 shows that if the answer to the DGM is a rotation (panel 1) it's embedded perpendicularly compared to if the answer C is a reflection (panel 2). We also study 1000 random seeds, quantitatively finding models prefer precise coset representations, *i.e.* trained DNNs learn precise cosets much more frequently (Fig. 4).

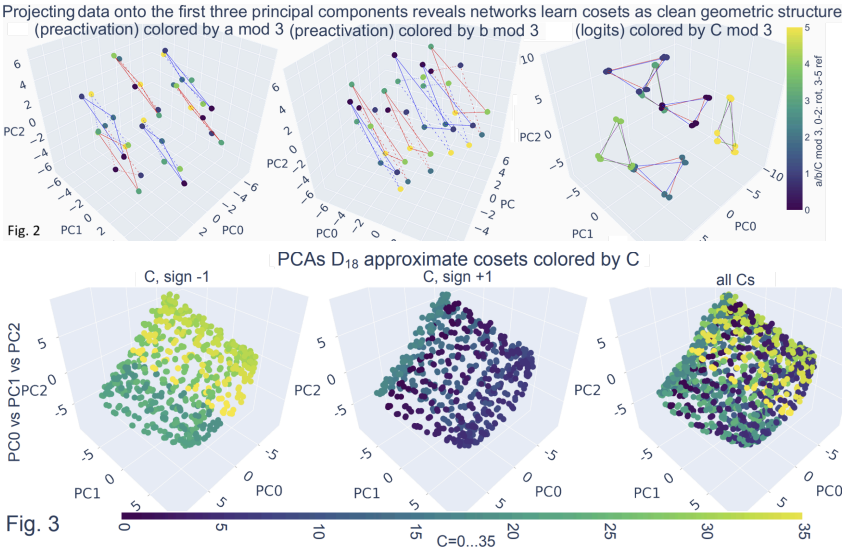

Fig. 2

Fig. 3

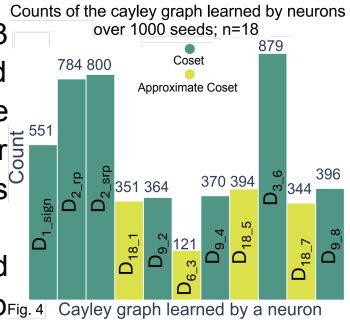

Fig. 4 Cayley graph learned by a neuron

Discussion. Our results show that DNNs *can utilize coset* and *approximate coset* structures to learn non-commutative group multiplications. Thus, our work provides empirical evidence toward proving the conjecture that DNNs will always use (*approximate) coset* structure to learn group multiplications. Also, we perform our analysis in a setting that's significantly different to [1], which only investigated commutative multiplications, extending our understanding of DNNs trained on group multiplications.

[1] Gavin McCracken, Gabriela Moisescu-Pareja, Vincent Letourneau, Doina Precup, and Jonathan Love. Uncovering a universal abstract algorithm for modular addition in neural networks, 2025. URL https://arxiv.org/abs/2505.18266.