The orbits of neural networks

William Chuang

September 30, 2023

1 Introduction

Inspired by [5, 2, 3], consider the following question: let a training set and a trained neural network are given where the trained neural network already converges to a local minimum, meaning given a positive real number $\epsilon > 0$, then each sample X_i in the given training set, there is $N \in \mathbb{N}$ such that for the number of epoch greater than N, the model is in the epsilon ball around the local minimum of the loss surface (i.e. its parameter space or moduli space). Then, is there a way to find equivalent models of this trained model?

Definition 1. *Two models are equivalent if they both converge on a fixed training set.*

It is necessary to ask this question so that the trained model can be understood better to a point that the distribution of its weights can be studied using random matrix theory, and if all the other equivalent models can be derived instantaneously, then this is also equivalent to apply quantum computing to find all those equivalent solutions in parallel.

If a model can be replaced by another equivalent model, then the other model might be at the different point on the loss surface, then the barrier between the local minimum to the nearest lower local minimum might have a lower height or in some better case with a negative height, meaning the model can be trained to a better model with a smaller error by using a new set of training data to descent to that lower local minimum. Hence, it is also sufficient to ask the question: how to find equivalent models of a given trained model? If the answer to this question is unknown, suppose a trained model cannot distinguish the fake data generated by an adversary model from the true

data, or if a model needs to go through a debugging process for some false classified input, then the model not only has to be retrained, but it cannot be justified it is the best solution due to the process to reach the solution is random. Thus, on the contrary, if all of its equivalent models can be found immediately, then chances are, within the set of equivalent models, there might exist a model that will not take the fake input or falsely classified a corner case, meaning a model can be replaced instantly without going through a new training process.

2 Algorithm

In Euclidean space, equivalent models of a trained model can be obtained by redefining matrix multiplication using Euclidean distance $d_E(x, y) = |x - y|$ for all $x, y \in \mathbb{R}$. For instance, let A be in $\mathbb{R}^{n\times m}$, B be in $\mathbb{R}^{d\times m}$, , and their entries are denoted by a_{ki} and b_{li} where $k \in \{1, ..., n\}$, $l \in \{1, ..., d\}$, and $i \in \{1, ..., m\}$. Then, $\left(AB^T\right)_{kl} = \sum_{i=1}^{m} a_{ki}b_{il}$.

Instead of using matrix multiplication, a new operation using Euclidean distance is defined in the following:

$$\left(A \odot_E B^T\right)_{kl} := \sum_{i=1}^{m} d_E\left(a_{ki}, b_{il}\right).$$

Now, let the above A and B be any two matrices that need to be multiplied in a given model that already converges under the above new operation $\odot_E$ between matrices A and B. Then by adding a real number α to A and B, infinite many equivalent models can be derived, since

$$\left((\alpha + A) \odot_E (\alpha + B)\right)_{kl} := \sum_{i=1}^{m} d_E\left(\alpha + a_{ki}, \alpha + b_{il}\right) = \sum_{i=1}^{m} d_E\left(a_{ki}, b_{il}\right) = \left(A \odot_E B\right)_{kl}.$$

In other words, the original given model is unchanged. This operation results in a translation of the loss surface.

Alternatively, every entry of the matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{d\times m}$ can be embedded into the upper half-plane $\mathbb{H}^2$. That is, the entries of A and B become complex numbers with positive imaginary part: $a_{ki} \in \mathbb{H}^2$ and $b_{il} \in \mathbb{H}^2$.

Define a new binary operation between A and B using hyperbolic length on the upper-half

plane $\mathbb{H}^2$[1]:

$$(A \odot_{\mathbb{H}^2} B)_{kl} := \sum_{i=1}^{m} d_{\mathbb{H}^2}(a_{ki}, b_{il}).$$

Then on the upper-half plane, the projective special linear group $\mathrm{PSL}(2,\mathbb{R})$ is its automorphism group. Furthermore, elements of the group are conformal mapping that preserve not only angles but hyperbolic distance. Hence, $\mathrm{PSL}(2,\mathbb{R})$ is also the isometry group of the upper-half plane.

To derive an equivalent model, take any $M \in \mathrm{PSL}(2,\mathbb{R})\backslash\{\mathrm{id}\}$, then there exists an isomorphism to map M to a linear fractional (Möbius) map $T_M(z) = \frac{az+b}{cz+d}, a, b, c, d, z \in \mathbb{C}$.

Definition 2 (Hyperbolic-Orbit)**.** *Define the hyperbolic-orbit algorithm to be the following new operation:*

$$(T_M(A) \odot_{\mathbb{H}^2} T_M(B))_{kl} := \sum_{i=1}^{m} d_{\mathbb{H}^2}(T_M(a_{ki}), T_M(b_{il})).$$

Proposition 1. *With the above notations, then the newly defined product is an invariant:*

$$(T_M(A) \odot_{\mathbb{H}^2} T_M(B))_{kl} = \sum_{i=1}^{m} d_{\mathbb{H}^2}(T_M(a_{ki}), T_M(b_{il})) = \sum_{i=1}^{m} d_{\mathbb{H}^2}(a_{ki}, b_{il}) = (A \odot_{\mathbb{H}^2} B)_{kl}.$$

Proof. This result follows immediately from the property of the Möbius mapping T_M when it is corresponding to a matrix M in the projective special linear group $\mathrm{PSL}(2,\mathbb{R})$, then T_M is an isometry under the hyperbolic metric. □

Let $\mathbb{B}^2 := \{z \in \mathbb{C} : |z| < 1\}$ be the Poincare's disc and Ψ be the Cayley's transformation $\Psi : \mathbb{H}^2 \to \mathbb{B}^2$.

Proposition 2. *With the above definition and proposition, then the value of the product has the same value when the points a_{ki} and b_{il} are sent to Poincare's disc:*

$$(T_M(A) \odot_{\mathbb{H}^2} T_M(B))_{kl} = \sum_{i=1}^{m} d_{\mathbb{H}^2}(T_M(a_{ki}), T_M(b_{il})) = \sum_{i=1}^{m} \rho_{\mathbb{B}^2}(T_M(a_{ki}), T_M(b_{il})).$$

Proof. Since for each $z_1, z_2 \in \mathbb{H}^2$, $d_{\mathbb{H}^2}(z_1, z_2) = \rho_{\mathbb{B}^2}(\Psi(z_1), \Psi(z_2))$. □

3 Example

All layers of a given neural network model can be implemented with the above hyperbolic-orbit algorithm. The following is an example when the hyperbolic-orbit algorithm is applied to the self-attention mechanism, i.e. transformer-based models. Let $A = Q \in \mathbb{C}^{n \times d_q}$ and $B = K^T \in \mathbb{C}^{d_k \times n}$ in self-attention mechanism[4] using dynamic scaling factor β (defined in our another work in progress), and by convention let $d_q = d_k$.

4 Discussion

With the hyperbolic-orbit algorithm, it is possible to find an appropriate Möbius mapping $M \in \mathrm{PSL}(2, \mathbb{R})$ such that training a 3-node neural network become not necessarily NP-hard in [2]. Further, if every layer of a given neural network is implemented with the hyperbolic-orbit algorithm introduced in this paper, by combining with perturbations and parallel computing, the steepest descent path toward the global minimum could be designed. The algorithm introduced by the paper can help to save training time for parallel computing in the ensemble method, that is, instead of starting over from the beginning to train N models in parallel, $N \in \mathbb{N}$, the training can start with N models that are equivalent to a trained model that already has a small error. Then, in each epoch, the best model could be selected and instantaneously its N equivalent models could be obtained using the hyperbolic-orbit algorithm.

References

[1] Alan F. Beardon, *The geometry of discrete groups*, Graduate Texts in Mathematics, vol. 91, Springer-Verlag, New York, 1995, Corrected reprint of the 1983 original. MR 1393195

[2] Avrim Blum and Ronald Rivest, *Training a 3-node neural network is np-complete*, Advances in neural information processing systems **1** (1988).

[3] J Stephen Judd, *Neural network design and the complexity of learning*, MIT press, 1990.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, *Attention is all you need*, Advances in neural information processing systems **30** (2017).

[5] Chulhee Yun, Suvrit Sra, and Ali Jadbabaie, *Global optimality conditions for deep neural networks*, arXiv preprint arXiv:1707.02444 (2017).