Understanding Self-attention Mechanism via Dynamical System Perspective

Zhongzhan Huang\textsuperscript{1*} Mingfu Liang\textsuperscript{2*} Jinhui Qin\textsuperscript{3} Shanshan Zhong\textsuperscript{1} Liang Lin\textsuperscript{1†}
\textsuperscript{1}Sun Yat-sen University \textsuperscript{2}Northwestern University, USA \textsuperscript{3}Guangdong University of Technology
huangzhzh23,zhongshsh5@mail2.sysu.edu.cn, mingfuliang2020@u.northwestern.edu, scape1989@gmail.com, linliang@ieee.org

Abstract

The self-attention mechanism (SAM) is widely used in various fields of artificial intelligence and has successfully boosted the performance of different models. However, current explanations of this mechanism are mainly based on intuitions and experiences, while there still lacks direct modeling for how the SAM helps performance. To mitigate this issue, in this paper, based on the dynamical system perspective of the residual neural network, we first show that the intrinsic stiffness phenomenon (SP) in the high-precision solution of ordinary differential equations (ODEs) also widely exists in high-performance neural networks (NN). Thus the ability of NN to measure SP at the feature level is necessary to obtain high performance and is an important factor in the difficulty of training NN. Similar to the adaptive step-size method which is effective in solving stiff ODEs, we show that the SAM is also a stiffness-aware step size adaptor that can enhance the model’s representational ability to measure intrinsic SP by refining the estimation of stiffness information and generating adaptive attention values, which provides a new understanding about why and how the SAM can benefit the model performance. This novel perspective can also explain the lottery ticket hypothesis in SAM, design new quantitative metrics of representational ability, and inspire a new theoretic-inspired approach, StepNet. Extensive experiments on several popular benchmarks demonstrate that StepNet can extract fine-grained stiffness information and measure SP accurately, leading to significant improvements in various visual tasks.

1. Introduction

The self-attention mechanism (SAM) [41, 15, 16, 9, 43, 4] is widely used in various artificial intelligence fields and has successfully improved the models’ performance in a number of vision tasks, including image classification [22, 60, 47], object detection [37, 56, 25], instance segmentation [8, 52], image super-resolution [64, 46, 49], etc. However, most previous works lay emphasis on designing a new self-attention method, and intuitively or heuristically exploring how the self-attention mechanism helps the performance. For example, many popular channel attention methods [22, 47, 56, 35] consider the attention values as the soft weight of the channels, leading to the importance reassignment of feature maps. These soft weights can also...
be seen as a gate mechanism [60, 28] to control the forward transmission of information flow, which are usually applied to neural network pruning and neural architecture search [40, 65]. Another viewpoint [38] argues that the self-attention mechanism can help to regulate the noise by enhancing instance-specific information to obtain a better regularization effect. Moreover, the receptive field [62, 63, 67] and long-range dependency [69, 17, 57] are also used to understand the role of self-attention. Although these explanations describe the behavior of self-attention mechanisms to some extent, the relationship between the SAM and model performance is still ambiguous.

To establish a more specific modeling between the SAM and model performance, in this paper, we rethink the role of the SAM with the dynamical systems perspective of neural networks (NN) with residual blocks. Specifically, we first define the stiffness phenomenon (SP) and ground truth (GT) trajectory of NNs at the feature level based on stiffness metrics and the ground truth solution of ODEs.

Next, we find that the intrinsic SP observed in the high-precision solutions of ODEs is also prevalent in high-performance NNs. This observation implies that the representational ability of NN to measure SP at the feature level is necessary to obtain high performance, as shown in Fig 1, which is hindering the learning of neural networks, and advanced training strategies are needed to achieve this requirement.

Similar to the adaptive step-size method which is effective in solving stiff ODEs, we theoretically and empirically demonstrate that the SAM is a stiffness-aware step size adaptor that can refine the estimation of stiffness information and generate the adaptive attention values to measure intrinsic SP for approaching a GT trajectory which has the upper bound performance, leading to high accuracy. This novel perspective can also explain the lottery ticket hypothesis in SAM (LTH4SA) [27], design new quantitative metrics of representational ability, and inspire a new theoretic-inspired approach, StepNet. Extensive experiments on several popular benchmarks show the effectiveness of StepNet in various vision tasks, including image classification and object detection. Our contributions are summarized as follows:

1. We propose a novel understanding of the SAM and reveal a close connection between the SAMs and the numerical solution of stiff ODEs, which is an effective explanation for understanding why and how the SAM enhances the performance of NNs.

2. Based on our novel views of SAMs, we explain the lottery ticket hypothesis in SAM, design new quantitative metrics of representational ability, and propose a powerful theoretic-inspired approach, StepNet.

2. The Stiffness and Self-attention Mechanism

In this section, we first introduce the concepts of stiffness in ODEs, SAM, and the dynamical system perspective for the NN with residual blocks. Then we further explore the SP in NNs and connect it with the SAM, which finally motivates us to propose a theoretic inspire approach.

2.1. Preliminaries and Related Works

2.1.1 The Dynamical System Perspective of NN

There are many well-known network architectures that have the residual blocks, like ResNet [18], UNet [23], Transformer [41], ResNeXt [61], etc. The residual blocks in one stage can be written as

\[ x_{t+1} = x_t + f(x_t; \theta_t), \]  

where \( x_t \in \mathbb{R}^d \) is the input of NN \( f(\cdot; \theta_t) \) with the learnable parameters \( \theta_t \) in \( t^{th} \) block. Many recent works [3, 42, 5, 59, 51, 68, 44] have established an insightful connection between residual blocks and dynamical systems, which reveal that the residual blocks can be interpreted as one step of a forward numerical method, i.e.,

\[ u_{t+1} = u_t + S(u_t; f, \Delta t) \Delta t, \]

for the numerical solution of an ODE as Eq.(2):

\[ \frac{du(t)}{dt} = f[u(t)], \quad u(0) = c_0, \]

where \( c_0 \) represents an initial condition, which corresponds to the input of the residual network. \( u(t) \equiv u_t \) is a time-dependent \( d \)-dimensional state, which is used to describe the input feature \( x_t \) in \( t^{th} \) block. The output of neural network \( f(\cdot; \theta_t) \) in \( t^{th} \) block can be regarded as an integration \( S(u_t; f, \Delta t) \) with step size \( \Delta t \) using a numerical method \( S \), e.g., the Forward Euler method [53].

2.1.2 The Stiffness in ODEs

In mathematics, a stiff equation is a differential equation [33], like Eq.(2), for which the numerical methods for solving that equation are numerically unstable, leading to poor prediction. For most ODEs, the stiffness is universal and intrinsic [34]. When the solution is unstable, we can use a fine step size \( \Delta t \) instead of a coarse step size to obtain finer differentiation, resulting in high-precision integration. Therefore, utilizing an adaptive step size based on a specific numerical method is the most straightforward way to solve stiffness ODEs, like the improvement from Forth order Runge–Kutta method [2] to Runge–Kutta–Fehlberg method [58, 14]. However, there is no universally accepted mathematical definition of stiffness [34], but the main idea is that the equation includes some terms that can lead to rapid variation in the solution.

Therefore, to quantify the stiffness to a certain extent, some simplified indexes are proposed, like the versatile
stiffness index (SI) \( \zeta_{SI} \) and the stiffness-aware index (SAI) \( \zeta_{SAI} \). Specifically, for the dynamics of the state \( u(t) \) in Eq.(2), the SI at the state \( u(t) \equiv u^i \) is defined by \( \zeta_{SI}(u^i) = \max(\|\text{Re}(\lambda_i)\|) \), where \( \lambda_i \) is the eigenvalue of the Jacobian matrix at \( u^i \) for the right-hand side of Eq.(2) \[1\]. The maximum eigenvalue (real part) of the Jacobian matrix represents the speed of the change for the solution. In some data-driven settings, the analytic expression of the right-hand side of Eq.(2), i.e., \( f(u^i; \theta) \cdot \Delta_t, t = 0, 1, \ldots, L - 1 \). The\( \zeta_{SAI}(u^i) \) \[1\] captures a more apparent stiffness phenomenon on the feature trajectory presented in Section 2.1, we define the stiffness phenomenon \( \zeta \) as usual. Thus, \( \zeta_{SAI}(u^i) \) can be used to measure the relative stiffness information, such as the rank of the stiffness, we can use a simplified SAI, i.e., \( \zeta_{SAI}(u^i) = \frac{\|u^{i+1} - u^i\|^2}{\|t_{i+1} - t_i\|^2} \), as a new kind of stiffness information measurement for analysis \[39\].

2.2. Measuring Stiffness Phenomenon (SP) in NN

In this section, based on the definition of stiffness in ODEs and the dynamical systems view of residual blocks presented in Section 2.1, we define the stiffness phenomenon (SP) in the feature trajectory \( [x_1, x_2, \ldots] \) generated by Eq.(1) from two perspectives. Firstly, (1) from a local perspective, we propose the Neural Stiffness Index (NSI) using the idea of Eq.(3) and Definition 1 to measure the SP based on the change of adjacent features. NSI can be used to visualize the SP qualitatively and intuitively. However, according to the related works about stiff ODEs \[24, 39\], NSI needs to be greater than a certain threshold, i.e., the local change needs to be large enough, to be considered that the SP happens. However, the threshold varies in different scenarios, and it is hard to pick a unified threshold \[39\].

Therefore, secondly, (2) from a global perspective, we propose a versatile metric called Total Neural Stiffness (TNS) using NSI to quantitatively measure the stiffness phenomenon of any neural networks on a given dataset. The TNS defined in Eq.(5) considers all threshold settings, including the relative threshold \( \mu(1+M_1) \) and the absolute threshold \( M_2, M \in \mathbb{R}^+ \). A larger TNS value indicates a more apparent stiffness phenomenon on the feature trajectory. The convergence of TNS is guaranteed by Theorem 1. Furthermore, from Eq.(4), the calculation of NSI requires the 2-norm of the feature. However, previous works \[36, 29, 13, 30\] show that the norm in the first block in each stage are usually extremely large and sensitive, which may affect the calculation of Eq.(5). Therefore, we will exclude these features while measuring TNS.

Definition 1. \( \text{(Neural Stiffness Index)} \) For the feature trajectories \( x_1, x_2, x_3, \ldots, x_L \) generated by a neural network with \( L \) residual blocks, i.e., \( x_{t+1} = x_t + f(x_t; \theta_t) \cdot \Delta_t, t = 0, 1, \ldots, L - 1 \). The Neural Stiffness Index (NSI) at \( x_t \) is

\[
\zeta_{NSI}(x_t) = \frac{1}{\|x_t\|^2} \left\| \frac{x_{t+1} - x_t}{\Delta_t} \right\|^2, \tag{4}
\]

e.g., for NN in Eq.(1), \( \zeta_{NSI}(x_t) = \|x_{t+1} - x_t\|^2/\|x_t\|^2 \).

Definition 2. \( \text{(Total Neural Stiffness)} \) In Definition 1, the feature trajectory has the stiffness \( \zeta_{NSI}(x_t; M) \) with degree \( M = (M_1, M_2) \) when \( \exists \) \( t \) such that \( \zeta_{NSI}(x_t) \geq \max(\mu(1 + M_1), M_2) \), where \( \mu \) is the mean of all features’ NSI from stage \( S \) of \( x_t \). For input \( x_0 \) sampled from test distribution \( P(x_0) \), the Total Neural Stiffness is \( \int_{\mathbb{R}^M} \delta(M) dM \), where

\[
\delta(M) = \mathbb{E}_{x_0 \sim P(x_0)} |_{_{\exists t, s.t. \zeta_{NSI}(x_t) \geq \max(\mu(1 + M_1), M_2)}}, \tag{5}
\]

and \( I \) is the characteristic function, i.e., when \( \zeta_{NSI}(x_t) \) not less than \( \max(\mu(1 + M_1), M_2) \), \( I = 1 \), otherwise, \( I = 0 \), \( M \in \mathbb{R}^+ \times \mathbb{R}^+ \).

Theorem 1. For \( \delta(M) \) defined as Eq.(5), the TNS is \( \int_{\mathbb{R}^M} \delta(M) dM \) is convergent. See proof in the appendix.

2.3. The Ground Truth (GT) Trajectory

To explore the properties of the SP in the NN, it is desirable to have a “ground truth” such that we can analyze the relationship between the properties of the SP and the model’s performance to propose specific improvements for the model. Given a dataset \( \mathcal{D}(x, y) \) and a network \( A(\theta_0, s) \) with residual blocks, where \( \theta_0 \) is initialization parameters and \( s \) denotes training setting (e.g., learning rate, weight decay, etc.). We consider some feature trajectories whose corresponding model \( A(\theta_0, s) \) have supremum performance as GT trajectories in Definition 3.

Definition 3. \( \text{(GT trajectory)} \) For the dataset \( \mathcal{D}(x, y) \), a task-oriented performance metrics \( \kappa \), and a residual neural network \( A(\theta_0, s) \) with initialization parameters \( \theta_0 \) and training setting \( s \). After training, there are an infinite number of feature trajectories \( [x_1, x_2, \ldots, i] = 1, 2, \ldots, \infty \), from a given input \( x_0 \) to the corresponding output \( y_0 \). The GT trajectories are the feature trajectories whose performance of \( A(\theta_0, s) \) on \( \mathcal{D}(x, y) \) can reach

\[
\sup_{\theta_0, s} \kappa(A(\theta_0, s), \mathcal{D}(x, y)). \tag{6}
\]

First, the GT trajectories introduced in Definition 3 exist. Let’s consider the non-empty set \( K \) whose elements are the metrics \( \kappa \) under all initialization parameters \( \theta_0 \) and training setting \( s \). Note that the task-oriented performance metrics \( \kappa \) are usually bounded in various deep learning tasks, especially in supervised learning, e.g., the upper
bound for the metrics of the classification task is 100%, so \( \kappa(A(\theta_0, s), D(x, y)) < +\infty \). According to the well-known theorem [12] that every non-empty subset \( \subset \mathbb{R} \) which has an upper bound has the supremum, thus the set \( K \) has the supremum, which means the GT trajectories exist. However, this kind of trajectory given an input \( x_0 \) may not be unique, which will be shown in Appendix. Moreover, as mentioned in Section 2.2, the feature trajectory is data-driven, and the corresponding analytic expression of the right-hand side of Eq.(2) in a NN is unknown. Thus it is infeasible to obtain the analytical form of the GT trajectories, and instead, trajectories from the model with high enough performance, like some advanced self-attention networks, can be seen as the proxies to empirically approach the properties of GT trajectories. Therefore, we take ResNet164 as the backbone and select four high-performance self-attention networks, i.e., SENet [22], FCANet [50], ECANet [56], and SRMNet [35], to approximately analyze the properties of GT trajectories. In Fig.2, we show the TNS for each network with different \( M_1, M_2 \).

For the CIFAR100 and STL10 datasets, from the results of all high-performance networks in Fig.2, we observe that the GT trajectories have a significantly large TNS. In other words, for most inputs, the GT trajectories have the SP. Although the TNS of the original residual neural network is relatively small, it can still measure the SP to some extent.
Moreover, as we can empirically observe that SP widely exists in these high-performance networks, we conjecture that the existence of SP is an intrinsic property of GT trajectories (More discussions are shown in Section 3.1). Thus if a NN structure can better estimate the stiffness information to measure and capture the SP, e.g., with high TNS, then such a structure can generate feature trajectories that are closer to the GT trajectories and thus achieve high performance. Otherwise, if a NN structure cannot inherently measure the stiffness information, it may produce large deviations between its feature trajectory and the GT trajectories. Such a deviation will gradually accumulate with the forward process of NN, leading to poor prediction. Lastly, in Fig.3, we provide the intuitive visualization of the SP by NSI in Definition 1. Obviously, on both CIFAR100 and STL10 datasets, we can observe that the GT trajectory measured by the SENet has significant and rapid oscillations in each stage, leading to the SP. The trajectories measured by “Org” are relatively smooth but also have some oscillations that are consistent with the observation in Fig.2.

2.4. Self-attention Mechanism (SAM) and Stiffness-aware Step Size Adaptor

In this section, we introduce the self-attention mechanism (SAM) and reveal the role it plays in improving model performance based on the dynamical perspective of the neural networks introduced in the above sections. For the analysis in the main paper, we use the channel attention neural networks [22, 56, 28] as an example, and the transformer-based self-attention models [41, 54] will be discussed in Appendix. For the channel self-attention networks, they can be written as follows by comparing with Eq.(1):

$$\hat{x}_{t+1} = x_t + F(x_t; \theta_2) \otimes \mathbf{F}(x_t; \theta_1; \phi_1), \tag{7}$$

where $\otimes$ is the Hadamard product and $\mathbf{F}(\cdot; \phi_1)$ is the self-attention module with learnable parameters $\phi_1$ in $l^{th}$ blocks based on different attention methods. For example, in SENet, $\mathbf{F}(\cdot; \phi_1) = \mathbf{W}_{t1}(\text{ReLU}(\mathbf{W}_{t2}(\cdot)))$, where $\mathbf{W}_{t1} \in \mathbb{R}^{d \times r}$ and $\mathbf{W}_{t2} \in \mathbb{R}^{r \times d}$ are learnable matrices, $r < d$.

Compared with the forward numerical method mentioned in Section 2.1.2, the attention value $\mathbf{F}(f(x_t; \theta_2); \phi_1)$ in Eq.(7) can be referred to as the step size in solving ODEs, and the original residual neural networks can be rewritten as

$$x_{t+1} = x_t + f(x_t; \theta_2) \otimes \Delta t, \tag{8}$$

where the step size $\Delta t = 1$. By comparing Eq.(7) and Eq.(8), we can readily find that the attention value generated by the SAM exactly serves as an adaptive step size, i.e., $\Delta t = \mathbf{F}(f(x_t; \theta_2); \phi_1)$. Moreover, since the last layers of various self-attention modules are usually Sigmoid function or Softmax function, thus the attention values in Eq.(7) are less than 1, i.e., $\Delta t = \mathbf{F}(f(x_t; \theta_2); \phi_1) < 1$. This implies that with SAM, we can provide a smaller and more flexible step size than that from the original residual neural network. Now we show that the step size generated by the SAM is also stiffness-aware. As mentioned in Section 2.1.2, we can use $(x_{t+1} - x_t)/\Delta t$ to measure the stiffness information in feature trajectory at $x_t$. Then for the original residual neural networks in Eq.(1), we have

$$f(x_t; \theta_1) = \frac{x_{t+1} - x_t}{\Delta t} |_{\Delta t=1} \equiv \mathbf{\hat{c}}_{\text{NSI}}(x_t), \tag{9}$$

and hence the $f(x_t; \theta_1)$ can be regarded as a kind of coarse stiffness information with step size $\Delta t = 1$, which is consistent with the discussions of Fig.2 and Fig.3 in Section 2.3. For the SAM, we have

$$\mathbf{F}(f(x_t; \theta_2); \phi_1) = \mathbf{F}(\frac{1}{\Delta t} (x_{t+1} - x_t) |_{\Delta t=1}; \phi_1) \tag{10}$$

Coarse stiffness information

$$= \mathbf{F}(\mathbf{\hat{c}}_{\text{NSI}}(x_t); \phi_1).$$

From Eq.(10), we can summarize how the self-attention module helps to improve the performance of the original residual neural networks: (1) Capture the stiffness information, the self-attention module $\mathbf{F}(\cdot; \phi_1)$ take the accessible and coarse stiffness information $f(x_t; \theta_1)$ from Eq.(9) as input. Then as shown in Fig.2 and Fig.3, the self-attention module can refine this coarse information to obtain a finer estimation of stiffness information; (2) Generate the adaptive step size. Based on this finer estimation, the module $\mathbf{F}(\cdot; \phi_1)$ outputs suitable attention values $\mathbf{F}(f(x_t; \theta_2); \phi_1)$ to adaptively measure the SP in the neural network, which means the SAM can enhance the representational ability of NN. For example, if the feature trajectory at $x_t$ needs to measure a large NSI from Eq.(4), the attention value $\mathbf{F}(f(x_t; \theta_2); \phi_1)$ can be small to get the large $\mathbf{\hat{c}}_{\text{NSI}}(x_t) = \frac{1}{\|x_t\|} \|\frac{x_{t+1} - x_t}{\mathbf{F}(f(x_t; \theta_2); \phi_1)}\|_2^2$.

2.5. Theoretic-inspired Approach: StepNet

From Section 2.4 and Eq.(10), we know that the ability to properly estimate the stiffness information is essential for the performance of the self-attention module. Thus if we want to obtain better model performance, we can consider estimating other accessible and better stiffness information in the self-attention module. Now we introduce a better self-attention formulation to capture better stiffness information, which is motivated by the asymptotic analysis between SI and SAI as follows. In Section 2.1.2, SAI is used as a proxy for the versatile index of stiffness (SI) to measure stiffness information in ODEs for tackling the computational difficulties of SI in data-driven problems. In Theorem 2, we first show how SAI can approximate the SI.

**Theorem 2.** For an ODE $du(t)/dt = f(u(t))$ defined at Eq.(2), if the Jacobian matrix $J_u$ at $u_t$ is a $n \times n$ symmetric real matrix and $\{\lambda_i\}_{i=1}^n$ are its $n$ distinct eigenvalues, and
Table 1. The classification accuracy on CIFAR10, CIFAR100, and STL10. “#P(M)” means the number of parameters (million). Bold and underline indicate the best results and the second best results, respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>CIFAR10</th>
<th>CIFAR100</th>
<th>STL10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#P(M)</td>
<td>top-1 acc. (%)</td>
<td>acc. ↑</td>
</tr>
<tr>
<td>ResNet [18]</td>
<td>1.70</td>
<td>93.35 ± 0.18</td>
<td>-</td>
</tr>
<tr>
<td>+SE [22]</td>
<td>1.91</td>
<td>94.26 ± 0.14</td>
<td>0.91</td>
</tr>
<tr>
<td>+CBAM [60]</td>
<td>1.92</td>
<td>93.91 ± 0.14</td>
<td>0.56</td>
</tr>
<tr>
<td>+ECA [56]</td>
<td>1.70</td>
<td>94.15 ± 0.23</td>
<td>0.80</td>
</tr>
<tr>
<td>+SEM [66]</td>
<td>1.95</td>
<td>94.52 ± 0.11</td>
<td>1.17</td>
</tr>
<tr>
<td>+SRM [35]</td>
<td>1.74</td>
<td>94.39 ± 0.12</td>
<td>1.24</td>
</tr>
<tr>
<td>+FCA [50]</td>
<td>1.91</td>
<td>94.66 ± 0.12</td>
<td>1.31</td>
</tr>
<tr>
<td>+IEBN [38]</td>
<td>1.73</td>
<td>94.70 ± 0.13</td>
<td>1.35</td>
</tr>
<tr>
<td>+StepNet (Ours)</td>
<td>1.75</td>
<td>95.14 ± 0.24</td>
<td>1.79</td>
</tr>
</tbody>
</table>

From Eq.(12), the calculation of our StepNet has two phases: (1) In Fig.4 (a), we first estimate a coarse \((t+1)\)th feature map \(x_{t+1} = x_t + f(x_t; \theta_t)\) by Eq.(1); (2) After that, the Adaptor \(\tilde{F}(\cdot; \cdot; \phi_t)\) take both the \(x_t\) and \(x_{t+1}\) as input to generate the adaptive attention values, which can better measure the stiffness information to generate finer step size for better capturing the SP to enhance the representational ability of the model and boost the performance. The network architecture of the Adaptor is shown in Fig.4 (b) and more training details are provided in Appendix.

3. Experiment

In this section, we use several popular vision benchmarks to verify the effectiveness of the proposed StepNet, including image classification and object detection. All experiments are verified 5 times with random seeds, and the average performances with standard deviations are reported. The experimental settings can be found in Appendix.

**Image classification.** We compare the proposed StepNet with several existing self-attention modules on four datasets for image classification. These four datasets are CIFAR10 [32], CIFAR100 [32], STL-10 [11] and ImageNet [55], and the results of these datasets are listed in Table 1 and 2, which show that the StepNet improves the accuracy significantly over the original networks and consistently compared with other existing self-attention modules under different datasets and backbones.

**Object detection.** We further conduct experiments for
is removed, respectively. The results show that input with \(x_t\) and \(x_{t+1}\) simultaneously is necessary to estimate fine stiffness information as discussed in Section 2.5. Moreover, \(F(x_{t+1} - x_t)\) means that we only use the adaptor in StepNet as the self-attention module following the normal paradigm without considering the better estimation of stiffness information in Eq.(12). Comparing the result of \(F(x_{t+1} - x_t)\) and our \(F(x_{t+1}, x_t)\), we can see that a finer estimation of stiffness information like ours is necessary to achieve better performance. For (2), as shown in Fig.5, we constructed four alternative adaptor structures. The experimental results in Table 4 illustrate that all four alternatives are inferior to ours. However, the best structure to generate the adaptive step sizes is still unknown, and in the future, we can still improve the design of the adaptor, e.g., to better utilize \(x_t\) and \(x_{t+1}\) through a neural network effectively.

### The property of stiffness phenomenon (SP) and the LTH4SA
In fact, for any input \(x_0\) and its corresponding output \(y_0\), the feature trajectory built by a well-trained residual neural network has two properties: (1) For most inputs \(x_0\), their feature trajectories have SP; (2) For each trajectory, only a few features can cause SP. Specifically, for property (1), in Section 2.3, we approximate the GT trajectories with several advanced self-attention models. From Fig.2 and Fig.3, we can empirically observe that for most of the inputs, their feature trajectories have SP. In the appendix, we provide more visualizations of feature trajectories using SENet as an example, and these visualizations also provide empirical evidence for property (1). For property (2), we define the stiffness proportion \(\hat{p} = \frac{1}{|f|} \sum_{x_0 \sim F(x_0)} \max\{\rho(1 + M_1), M_2\}\) to measure the expected number that how many features from a feature trajectory have \(\xi_{NS}(x_t; M)\) with degree M. If the \(\hat{p}\) of a feature trajectory is close to 100%, it means that

#### 3.1. Ablation Study and Discussions

Now we further explore the property of the StepNet and the stiffness phenomenon in neural networks.

**The structure of StepNet.** In Section 2.4 we summarize that the self-attention mechanism can help the model performance in two ways, i.e., the extraction of stiffness information and the generation of adaptive step sizes. In fact, these two ways correspond to (1) the input of the self-attention module and (2) the design of the module. For (1), in Table 4, we explore the performance when \(x_t\) or \(x_{t+1}\)

<table>
<thead>
<tr>
<th>Model</th>
<th>CIFAR10</th>
<th>CIFAR100</th>
<th>STL10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Org</td>
<td>93.35 ±0.18</td>
<td>74.30 ±0.30</td>
<td>82.66 ±1.05</td>
</tr>
<tr>
<td>(F(x_t))</td>
<td>94.59 ±0.14</td>
<td>76.34 ±0.14</td>
<td>84.74 ±0.64</td>
</tr>
<tr>
<td>(F(x_{t+1}))</td>
<td>94.32 ±0.23</td>
<td>76.53 ±0.04</td>
<td>85.12 ±0.54</td>
</tr>
<tr>
<td>(F(x_{t+1} - x_t))</td>
<td>94.69 ±0.33</td>
<td>76.42 ±0.10</td>
<td>85.26 ±0.72</td>
</tr>
<tr>
<td>(F(x_{t+1}, x_t)) (ours)</td>
<td>95.14 ±0.24</td>
<td>77.04 ±0.22</td>
<td>86.08 ±0.11</td>
</tr>
<tr>
<td>Adaptor a</td>
<td>94.39 ±0.12</td>
<td>76.19 ±0.14</td>
<td>84.21 ±0.12</td>
</tr>
<tr>
<td>Adaptor b</td>
<td>94.05 ±0.11</td>
<td>75.45 ±0.12</td>
<td>84.10 ±0.24</td>
</tr>
<tr>
<td>Adaptor c</td>
<td>94.19 ±0.24</td>
<td>75.12 ±0.08</td>
<td>84.12 ±0.44</td>
</tr>
<tr>
<td>Adaptor d</td>
<td>94.92 ±0.24</td>
<td>76.83 ±0.14</td>
<td>85.62 ±0.44</td>
</tr>
</tbody>
</table>

Table 4. The ablation study on the structure of StepNet.
this trajectory has many features with large enough NSI. For Fig.4, we show the corresponding stiffness proportion in Fig.6. We can observe that for various $M_1$ and $M_2$, most of the $\hat{p} \leq 10\%$, which indicates only a few features in GT trajectory can cause SP.

Moreover, these two properties are also consistent with those intrinsic properties in physics dynamical systems. For instance, the close encounter is an important factor to cause the SP in the three-body motion. [38] considers 1,000 independent simulations of three-body trajectories following [10], where 91.4\% of the trajectories contain the close encounter, but on average only 4.2\% of the time intervals within the trajectories contain close encounter.

In addition, we identify that these two properties are closely related to the lottery ticket hypothesis [27] in the self-attention mechanism (LTH4SA). LTH4SA reveals that we only need to insert the self-attention module on a small number of blocks to achieve remarkable improvement for a NN. According to property (1), for most inputs, their GT trajectories have SP, and as mentioned in Section 2.4, the adaptive step size generated by the SAM can improve the representational ability of NN. Thus the SAM is valid for most of the inputs. Moreover, property (2) tells us that only a small part of the features in a feature trajectory can cause SP, and thus we only need to set the module on a small number of blocks to measure the SP of the whole trajectory. So if these two properties generally hold, we argue that LTH4SA may also be an intrinsic property of the SAM.

**Why do the GT trajectories have SP?** Now we attempt to understand why most GT trajectories have SP, which can help us design novel methods to boost the performance of representation learning. From Eq.(4) and Eq.(9), we know that for a well-trained residual neural network, $f(x_t; \theta_1)$ provides stiffness information and $\xi_{\text{NSI}}(x_t) = O(\|f(x_t; \theta_t)\|_2)$. When NSI is large, $\|f(x_t; \theta_t)\|_2$ is also large, which means that the elements (absolute values) of the output feature of the neural network $f(\cdot; \theta_t)$ at $t$-th block are relatively large. In some previous works [36, 21, 48], such kinds of features are considered important features and have major contributions to the model performance. In other words, a residual network can achieve high performance, i.e., it can approximate the GT trajectory, probably because the network has the ability to learn such important (stiff) features by the adaptive step sizes in a few blocks. So we further calculate the rank correlation (kendall correlation [31] and spearman correlation [45]) between the TNS and the model performance. The results are presented in Fig.7, which shows that the performance of the models and their representational ability to measure the SP are positively correlated. Moreover, as the TNS can reflect the ability of the model to measure the SP, thus the TNS can also be a novel representational ability metric to evaluate the neural network in practice and has the potential to be used in network formulation, such as neural architecture search [26, 40], network pruning [36, 20, 19] or other application [6, 7].
4. Conclusion

In this paper, we bridge the relationship between the self-attention mechanism (SAM) and the numerical solution of stiff ordinary differential equations, which reveals that the SAM is a stiffness-aware step size adaptor that can refine the estimation of stiffness information and generate suitable attention values for adaptively measuring the stiffness phenomenon in the neural network (NN) to enhance the representational ability of the NN and achieve high performance.

5. Acknowledgment

This work was supported in part by National Key R&D Program of China under Grant No. 2020AAA0109700, National Natural Science Foundation of China (NSFC) under Grant No.61836012, Guangdong Basic and Applied Basic Research Foundation No.2023A1515011374, National Natural Science Foundation of China (NSFC) under Grant No. 62206314, GuangDong Basic and Applied Basic Research Foundation under Grant No. 2022A1515011835.

References

[23] Huimin Huang, Lanfen Lin, Ruofeng Tong, Hongjie Hu, Qiaowei Zhang, Yutaro Iwamoto, Xianhua Han, Yen-Wei


