PIDformer: Transformer Meets Control Theory

Tam Nguyen [1] **César A. Uribe** [1] **Tan M. Nguyen** [† 2] **Richard G. Baraniuk** [† 1]

Abstract

In this work, we address two main shortcomings of transformer architectures: input corruption and rank collapse in their output representation. We unveil self-attention as an autonomous state-space model that inherently promotes smoothness in its solutions, leading to lower-rank outputs and diminished representation capacity. Moreover, the steady-state solution of the model is sensitive to input perturbations. We incorporate a Proportional-Integral-Derivative (PID) closed-loop feedback control system with a reference point into the model to improve robustness and representation capacity. This integration aims to preserve high-frequency details while bolstering model stability, rendering it more noise-resilient. The resulting controlled state-space model is theoretically proven robust and adept at addressing the rank collapse. Motivated by this control framework, we derive a novel class of transformers, PID-controlled Transformer (PIDformer), aimed at improving robustness and mitigating the rank-collapse issue inherent in softmax transformers. We empirically evaluate the model for advantages and robustness against baseline transformers across various practical tasks, including object classification, image segmentation, and language modeling.

1. Introduction

Transformer models (Vaswani et al., 2017) have shown remarkable achievements across various domains such as reinforcement learning (Chen et al., 2021a; Janner et al., 2021), computer vision (Dosovitskiy et al., 2021b; Touvron et al., 2021; Zhao et al., 2021; Guo et al., 2021), natural language processing (Devlin et al., 2018; Al-Rfou et al., 2019; Child et al., 2019; Raffel et al., 2020) and other practical applications (Zhang et al., 2019; Gulati et al., 2020). At the core of transformers lies the self-attention mechanism, which computes weighted averages of token representations within a sequence based on the similarity scores between pairs of tokens, thus capturing diverse syntactic and semantic relationships effectively (Cho et al., 2014; Parikh et al., 2016). This flexibility in capturing relationships has been identified as a key factor contributing to the success of transformers.

1.1. Background: Self-Attention

Given a sequence of tokens $\mathbf{X}^\ell := [\boldsymbol{x}^\ell(1), \cdots, \boldsymbol{x}^\ell(N)]^\top$, $\mathbf{X}^\ell \in \mathbb{R}^{N\times D_x}$, the query, key and value matrices at layer ℓ-th are $\mathbf{Q}^\ell = \mathbf{X}\mathbf{W}_Q^{\ell\ \top}$; $\mathbf{K}^\ell = \mathbf{X}\mathbf{W}_K^{\ell\ \top}$; and $\mathbf{V}^\ell = \mathbf{X}\mathbf{W}_V^{\ell\ \top}$, respectively. The weight matrix $\mathbf{W}_Q^\ell, \mathbf{W}_K^\ell \in \mathbb{R}^{D_{qk}\times D_x}$ and $\mathbf{W}_V^\ell \in \mathbb{R}^{D\times D_x}$. The attention mechanism computes the output of token i at layer ℓ-th as follows

$$\boldsymbol{u}^\ell(i) = \sum_{j=1}^{N} \mathrm{softmax}\Big(\boldsymbol{q}^\ell(i)^\top \boldsymbol{k}^\ell(j)/\sqrt{D_{qk}}\Big)\boldsymbol{v}^\ell(j), \quad (1)$$

where $\boldsymbol{q}^\ell(i)$ is the row i-th of $\mathbf{Q}^\ell$ and $\boldsymbol{k}^\ell(j), \boldsymbol{v}^\ell(j)$ are the row j-th of $\mathbf{K}^\ell, \mathbf{V}^\ell$, respectively. The softmax function computes the attention score between token i and j, for all $i, j = 1, \ldots, N$. The self-attention (1) is referred to as softmax attention. Our work refers to a transformer that uses softmax attention as a softmax transformer.

Despite their remarkable success, transformers exhibit practical performance issues in their robustness and representation capacity. For example, recent studies (Mahmood et al., 2021; Madry et al., 2017; Zhou et al., 2022) have provided empirical evidence of Vision Transformer's susceptibility to adversarial attacks and common input perturbations, such as noise or blur. Additionally, deep transformer-based models have been observed to suffer from rank-collapse in their outputs, wherein token embeddings become increasingly similar as the model depth increases (Shi et al., 2022; Dong et al., 2021; Wang et al., 2022). This issue severely constrains the representation capacity of transformers, hindering their performance in various tasks. Addressing these issues is crucial for ensuring the reliability and effectiveness of transformer models across different applications.

[†]Co-last authors [1]Department of Electrical & Computer Engineering, Rice University, Houston, USA [2]Department of Mathematics, National University of Singapore, Singapore. Correspondence to: Tam Nguyen <mn72@rice.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

1.2. Contribution

We introduce self-attention as a self-evolving state-space model (SSM) and provide insights into the non-robustness and rank-collapse issues inherent in transformers. Specifically, we demonstrate that self-attention can be seen as a discretization of an SSM from a gradient flow, minimizing the nonlocal total variation (Gilboa & Osher, 2008) of an input signal and promoting smoothness. This characteristic leads to rank collapse and diminishes the output's representation capacity. Additionally, the steady-state solution of the SSM is sensitive to input perturbation. Motivated by this novel understanding, we propose the Proportional-Integral-Derivative (PID) control transformer, PIDformer, as a new transformer class that mitigates both issues. PIDformer is derived as a discretization of a PID-control integrated SSM proven to enhance the model's stability and representation capacity. Our contributions are four-fold.

1. We present a novel control framework for self-attention mechanisms, unveiling the connection between self-attention and the state-space model. Our analysis sheds light on the shortcomings of transformers, which exhibit non-robust behavior to input perturbations and are prone to rank collapse.

2. Motivated by these analyses, we propose PIDformer, a new class of transformers, that integrates a Proportional-Integral-Derivative (PID) controller into transformers. PIDformer enhances model robustness and effectively mitigates the rank-collapse issue.

3. We demonstrate how the connection between energy optimization and our controlled SSMs enhances the understanding of these models.

4. We theoretically prove that employing softmax self-attention is inherently sensitive to noise and tends to produce low-rank outputs. In contrast, our controlled SSM is guaranteed to exhibit superior robustness and avoid the rank-collapse issue.

We empirically demonstrate the advantages of PIDformers on various large-scale applications, including the ImageNet object classification (Deng et al., 2009) (under diverse input perturbations and robustness benchmarks), ADE20K image segmentation (Zhou et al., 2018), and WikiText-103 language modeling (Merity et al., 2017). tasks.

Organization. We structure our paper as follows: In Section 2, we introduce a control framework for self-attention, offering insights into the non-robustness and rank-collapse issues in transformer-based models. In Section 3, we incorporate a PID controller into the SSM, providing theoretical guarantees of its stability and ability to mitigate the rank-collapse issue. Subsequently, we developed PIDformer, a discretization of the PID-controlled SSM, and established the connection between these dynamics and energy optimization for further understanding. In Section 4, we empirically validate the benefits of PIDformer. We review related work in Section 5. Finally, we summarize our main contributions and provide additional results, details, and proofs in the Appendix.

2. A Control Framework for Self-Attention

Consider the value matrix of layer ℓ-th $\mathbf{V}^\ell := [\boldsymbol{v}^\ell(1), \cdots, \boldsymbol{v}^\ell(N)]^\top \in \mathbb{R}^{N\times D}$ in Section 1.1. Let $\Omega \subset \mathbb{R}$, $x \in \Omega$, and $\boldsymbol{v}(x,t) := [v_1(x,t), \ldots, v_D(x,t)]^T$ be a real vector-valued function, $\boldsymbol{v} : \Omega \times [0,\infty) \to \mathbb{R}^D$, $\boldsymbol{v} \in L^2(\Omega \times [0,\infty))$. Assume the value matrix $\mathbf{V}^\ell$ discretizes the function $\boldsymbol{v}(x,t)$ on the spatial and time dimension. In the context of a control system, $\boldsymbol{v}(x)$ can be considered as the state signal of the following state-space model:

$$\frac{d\boldsymbol{v}(x,t)}{dt} = \int_\Omega (\boldsymbol{v}(y,t) - \boldsymbol{v}(x,t))K(x,y,t)dy + \boldsymbol{z}(x,t)$$
$$\boldsymbol{v}(x,0) = \boldsymbol{v}^0(x), \boldsymbol{z}(x,t) = \mathbf{0}, \forall x \in \Omega, \forall t \geq 0 \qquad (2)$$

where $\boldsymbol{z} \in L^2(\Omega \times [0,\infty))$ is a control input and $\boldsymbol{v}^0$ is the initial state. The function $K(x,y,t)$ is the kernel function that captures the proximity of the signal $\boldsymbol{v}$ at positions x, y at time t. Here, the SSM is autonomous, as no control inputs or feedback are fed into the system. In this section, we illustrate that system in (2) induces smoothness to the signal by minimizing the nonlocal total variation (Gilboa & Osher, 2008) of the signal, hence losing detailed information as it evolves. Subsequently, we show that self-attention serves as a discretization of this dynamic. Lastly, we theoretically demonstrate that the SSM in 2 is vulnerable to input perturbation and representation collapse.

2.1. Connection between State Space Model and Nonlocal Variational Minimization

We show that the gradient flow aimed at minimizing the following nonlocal functional is a case of our SSM described in (2)

$$J(\boldsymbol{v}) = \frac{1}{2}\int_{\Omega\times\Omega} \|\boldsymbol{v}(x) - \boldsymbol{v}(y)\|_2^2 k(x,y)dxdy. \qquad (3)$$

Here, $J(\boldsymbol{v})$, the sum of the square of the nonlocal derivative on the spatial dimension $\partial_y \boldsymbol{v}(x) = \big(\boldsymbol{v}(x) - \boldsymbol{v}(y)\big)\sqrt{k(x,y)}$ (Gilboa & Osher, 2008) , represents the non-local variation of the signal $\boldsymbol{v}$. $k(x,y)$ captures the proximity between position x and y in the signal. Minimizing $J(\boldsymbol{v})$ promotes the smoothness of $\boldsymbol{v}$ and penalizes high-frequency in the signal.

The gradient of J with respect to $\boldsymbol{v}$ is then given by

$$\nabla_{\boldsymbol{v}} J(\boldsymbol{v}) = \left[\frac{\partial J}{\partial v_1}, \frac{\partial J}{\partial v_2}, \ldots, \frac{\partial J}{\partial v_D}\right]^T. \qquad (4)$$

As shown in the Appendix B.10, the Frechet derivative of J with respect to v_j is

$$\frac{\partial J}{\partial v_j} = \int_\Omega (v_j(x) - v_j(y)(k(x,y) + k(y,x))dy. \quad (5)$$

Substituting the formula for $\partial J/\partial v_j$ in (5) into (4) for $\nabla_{\boldsymbol{v}} J(\boldsymbol{v})(x)$, we obtain the following gradient flow

$$\begin{aligned}\frac{d\boldsymbol{v}(x,t)}{dt} &= -\nabla_{\boldsymbol{v}} J(\boldsymbol{v}) \\ &= \int_\Omega \big(\boldsymbol{v}(y,t) - \boldsymbol{v}(x,t)\big)\big(k(x,y) + k(y,x)\big)dy,\end{aligned} \quad (6)$$

The autonomous state-space representation in (2) simplifies to this dynamic when $K(x,y,t) := k(x,y)+k(y,x)$, which is symmetric and time-invariant. In this scenario, the model reduces the total nonlocal variance of the signal, resulting in a smoother solution. This renders the model susceptible to rank collapse in the output representation. In Section 2.2, we prove that the model suffers from rank collapse regardless of whether $K(x,y,t)$ is symmetric.

Connection between SSM and self-attention. We show that a discretization of our SSM recovers the self-attention mechanism. Let $\boldsymbol{q}, \boldsymbol{k} : \Omega \times [0,\infty) \to \mathbb{R}^{D_{qk}}$, $\boldsymbol{q}, \boldsymbol{k} \in L^2(\Omega \times [0,\infty))$ be real vector-valued functions. Similar to $\boldsymbol{v}(x,t)$, we can discretize $\boldsymbol{q}(x,t), \boldsymbol{k}(x,t)$ on spatial dimension to attain the query vectors $\boldsymbol{q}^\ell(1), \ldots, \boldsymbol{q}^\ell(N) \in \mathbb{R}^{D_{qk}}$, and the key vectors $\boldsymbol{k}^\ell(1), \ldots, \boldsymbol{k}^\ell(N) \in \mathbb{R}^{D_{qk}}$ of layer ℓ-th. We define the proximity kernel as

$$K(x,y,t) := \frac{\exp\big(\boldsymbol{q}(x,t)^T \boldsymbol{k}(y,t)/\sqrt{D_{qk}}\big)}{\int_\Omega \exp\big(\boldsymbol{q}(x,t)^T \boldsymbol{k}(y',t)/\sqrt{D_{qk}}\big)dy'}.$$

Applying the Euler method to discretize (2) with the time step $\Delta t(x) := 1$, the update step of the system becomes

$$\begin{aligned}&\boldsymbol{v}(x,t+1) \\ &\approx \int_\Omega \frac{\exp\big(\boldsymbol{q}(x,t)^T \boldsymbol{k}(y,t)/\sqrt{D_{qk}}\big)}{\int_\Omega \exp\big(\boldsymbol{q}(x,t)^T \boldsymbol{k}(y',t)/\sqrt{D_{qk}}\big)dy'} \boldsymbol{v}(y,t)dy.\end{aligned} \quad (7)$$

Using the Monte-Carlo method (Metropolis & Ulam, 1949) to approximate the integrals in the spatial dimension in (7), we attain

$$\boldsymbol{v}^{\ell+1}(i) \approx \sum_{j=1}^{N} \mathrm{softmax}\Big(\boldsymbol{q}^\ell(i)^\top \boldsymbol{k}^\ell(j)/\sqrt{D_{qk}}\Big)\boldsymbol{v}^\ell(j).$$

which recovers $\boldsymbol{u}^\ell(i)$, the output token i of self-attention at layer ℓ-th as in (1). As self-attention discretizes the SSM outlined in (2), it inherits the characteristics of the model, making it susceptible to input corruption and output rank collapse. These properties are theoretically demonstrated in Section 2.2.

2.2. Stability and Representation Collapse of the State Space Model

Model robustness is its ability to maintain high performance despite encountering uncertain or challenging scenarios such as noisy data, distribution shifts, or adversarial attacks (Wang & Bansal, 2018; Dong et al., 2020). Robustness also entails stability, wherein the model's output remains relatively unchanged even when the input is perturbed.

For the theoretical analysis of our SSMs, we assume that the kernel K is time-invariant, i.e., $K(x,y,t) = K(x,y)$. This assumption is practical in the context of transformers, particularly in deep transformer models, where the attention matrix tends to remain similar after the initial layers (Shi et al., 2022). The discretization of model in (2) on the spatial dimension gives

$$\frac{d\boldsymbol{v}(i,t)}{dt} = \sum_{j=1}^{N} (\boldsymbol{v}(j,t) - \boldsymbol{v}(i,t))K(i,j),$$

for $i,j = 1,2,\ldots,N$ By choosing $K(i,j) := \mathrm{softmax}\big(\boldsymbol{q}(i)^T \boldsymbol{k}(j)/\sqrt{D_{qk}}\big)$, its corresponding matrix representation is obtained as

$$\mathbf{V}'(t)dt = \mathbf{K}\mathbf{V}(t) - \mathbf{V}(t), \mathbf{V}(0) = \mathbf{V}^0, \quad (8)$$

where $\boldsymbol{K}$ is a right-stochastic matrix with all positive entries. In the context of transformer, $\boldsymbol{K}$ is the attention matrix and $\mathbf{V} = [\boldsymbol{v}^0(1), \ldots, \boldsymbol{v}^0(N)]^T$ is the value matrix at the first layer. Lemma 1 sheds light on the stability and representation collapse of the solution for the SSM in (2).

Lemma 1. *Given $\{\alpha_1, \alpha_2, \ldots, \alpha_M\}, M \leq N$, is the complex spectrum of $\mathbf{K} - \mathbf{I} \in \mathbb{R}^{N\times N}$. The solution of the ordinary differential equation (ODE) (8) is given by*

$$\mathbf{V}(t) = \boldsymbol{P}\exp(\boldsymbol{J}t)\boldsymbol{P}^{-1}\mathbf{V}^0, \quad (9)$$

where $\boldsymbol{P}\boldsymbol{J}\boldsymbol{P}^{-1}$ is the Jordan decomposition of $\boldsymbol{K} - \boldsymbol{I}$, $\boldsymbol{P}$ is invertible and contains the generalized eigenvectors of $\boldsymbol{K} - \boldsymbol{I}$, and $\boldsymbol{J} = \mathbf{diag}(\boldsymbol{J}_{\alpha_1,m_1}, \boldsymbol{J}_{\alpha_2,m_2}, \ldots, \boldsymbol{J}_{\alpha_M,m_M})$ is the Jordan form of matrix $\boldsymbol{K} - \boldsymbol{I}$ with,

$$\boldsymbol{J}_{\alpha_i,m_i} = \begin{bmatrix} \alpha_i & 1 & \ldots & 0 \\ \vdots & \ddots & & \vdots \\ & & \alpha_i & 1 \\ 0 & \ldots & & \alpha_i \end{bmatrix} \in \mathbb{R}^{m_i \times m_i}, \text{ for } i = 1, \ldots, M$$

are Jordan blocks. Here, $\sum_{i=1}^{M} m_i = N$.

The proof of Lemma 1 is shown in the Appendix B.2. Since $\mathbf{K}$ is a positive right-stochastic matrix, its largest and unique eigenvalue α_1 is 1 and $|\alpha_i| < 1$ (see Theorem 4.1 in (Bandeira et al., 2020)), meaning $\mathrm{Re}(\alpha_i) \in [-1, 1)$, for $i = 2, \ldots, M$. Hence, the matrix $\mathbf{K} - \mathbf{I}$, whose eigenvalues are $\alpha_1 - 1, \ldots, \alpha_M - 1$, has a unique largest eigenvalue of 0 and the real part of other eigenvalues in $[-2, 0)$. This leads to the rank collapse of the steady-state solution, as stated in the following Lemma 2.

Lemma 2. $\lim_{t\to\infty} \mathbf{V}(t) = \big[c_{1,1}\boldsymbol{p_1}, \ \ldots, \ c_{1,D_x}\boldsymbol{p_1}\big]$, *where* $\boldsymbol{p}_1$ *is the eigenvector corresponds with the eigenvalue* $(\alpha_1 - 1) = 0$ *of* $\boldsymbol{K} - \boldsymbol{I}$*, and* $c_{1,1}, \ldots, c_{1,D_x}$ *are the coefficients w.r.t* $\boldsymbol{p}_1$ *of the decomposition of* $\boldsymbol{V}^0$*'s columns in the Jordan basis (column vectors of* $\boldsymbol{P}$*).*

The proof of Lemma 2 is shown in the Appendix B.3. This yields two insights. Firstly, the steady-state solution of the system depends on the initial $\boldsymbol{V}^0$, implying that any perturbation in the input results in changes in the output. Secondly, the solution experiences rank collapse, with the rank of its steady state solution being 1 as $t \to \infty$. This indicates that our SSM in (2) not only yields a non-robust solution but also experiences information loss (low-rank output representation). As the self-attention mechanism discretizes the model in (2), it inherently exhibits both issues.

3. Transformer with PID-Controller for State-Space Representation

To counteract the loss of detailed information caused by smoothness and to bolster model stability, a PID controller is integrated into the state-space representation as follows:

$$\begin{aligned}
\frac{d\boldsymbol{v}(x,t)}{dt} &= \int_\Omega (\boldsymbol{v}(y,t) - \boldsymbol{v}(x,t))K(x,y,t)dy + \boldsymbol{z}(x,t) \\
\boldsymbol{z}(x,t) &= \lambda_P \boldsymbol{e}(x,t) + \lambda_I \int_0^t \boldsymbol{e}(x,t) + \lambda_D \frac{d\boldsymbol{e}(x,t)}{dt} \\
\boldsymbol{v}(x,0) &= \boldsymbol{v}^0(x), \boldsymbol{z}(x,0) = \boldsymbol{0}.
\end{aligned} \tag{10}$$

The regularizer term, denoted as $\boldsymbol{e}(x,t) = \boldsymbol{f}(x) - \boldsymbol{v}(x,t)$, encapsulates the loss of information as $\boldsymbol{v}(x,t)$ becomes smoother over time. Here, the reference function $\boldsymbol{f}(x)$ represents a high-frequency signal containing detailed information about the original inputs. We select $\boldsymbol{f}(x)$ as the scaled initial value function, denoted as $\beta\boldsymbol{v}(x,0)$. In the context of a transformer, we set $\boldsymbol{f}(i) = \beta\boldsymbol{v}^0(i)$, representing the value vector embedding at token index i of the first layer. This choice is motivated by our desire to have flexibility in determining the detailed information from the input signal we wish to preserve. This flexibility is governed by the parameter $\beta \in (0, 1]$. The regularizer $\boldsymbol{e}(x,t)$ is fed back into the system, guiding the model to reintegrate the lost information while maintaining stability through three components: (P), (I), and (D).

- The (P) term is directly proportional to the regularizer, $e(x,t)$. In cases of substantial information loss, the control input $\boldsymbol{z}(x,t)$ should be proportionately large, determined by the gain factor λ_P, to reintroduce the lost information into the system. A small choice of λ_P results in slow convergence, while a large choice may lead to overshooting issues, causing instability in reaching the reference point.
- The (I) term accumulates all past errors, given by $\lambda_I \int_0^t \boldsymbol{e}(x,t)$. This component aids in reintroducing any persistent, long-term loss of information that might persist despite proportional control.
- Finally, the (D) term, $\lambda_D \frac{d\boldsymbol{e}(x,t)}{dt}$, anticipates future losses of information by considering the rate at which the error is changing. A more rapid change in error prompts a greater control effect, and the derivative term proves beneficial in enhancing the stability and responsiveness of the control system.

In this section, we unveil a connection between the two components, (P) and (I), of the SSM in (10) and different optimization methods applied to minimize a regularized functional. This functional is tailored to preserve the detailed information of the solution. Moreover, we show that the P-control (where $\lambda_I = \lambda_D = 0$), PD-control ($\lambda_I = 0$), and PID-controlled SSM in (10) are theoretically guaranteed to be more robust and mitigate the issue of rank collapse. Subsequently, we introduce the PID-controlled transformer (PIDformer), a novel architecture that enhances performance and robustness.

3.1. Connection between (P) and (I) Components with Different Optimization Methods

In Section 2.1, we have shown that the SSM in (2) implicitly performs a gradient descent to minimize the nonlocal variation $J(\boldsymbol{v})$, which leads to the loss of signal information. Now, we illustrate that the feedback-controlled state-space in (10), without the derivative (D) ($\lambda_D = 0$), implicitly minimizes the following functional:

$$\begin{aligned}
E(\boldsymbol{v}, \boldsymbol{f}) &= J(\boldsymbol{v}) + G(\boldsymbol{v}, \boldsymbol{f}) \\
&= \frac{1}{2}\int_{\Omega\times\Omega} \|\boldsymbol{v}(x) - \boldsymbol{v}(y)\|_2^2 k(x,y)dxdy \\
&+ \frac{\lambda}{2}\int_\Omega \|\boldsymbol{v}(x) - \boldsymbol{f}(x)\|_2^2 dx.
\end{aligned} \tag{11}$$

where the data fidelity term $G(\boldsymbol{v}, \boldsymbol{f}) = \frac{\lambda}{2}\int_\Omega \|\boldsymbol{v}(x) - \boldsymbol{f}(x)\|_2^2 dx$ (Gilboa & Osher, 2008; 2007) is introduced to penalize significant information loss. This observation further validates that systems in (10) retains relevant information from the reference signal $\boldsymbol{f}$.

P-controlled SSM as gradient descent to minimize $E(\boldsymbol{v}, \boldsymbol{f})$. The gradient of E w.r.t $\boldsymbol{v}$ is $\nabla_{\boldsymbol{v}} E(\boldsymbol{v}) = \nabla_{\boldsymbol{v}} J(\boldsymbol{v}) + \lambda\big(\boldsymbol{v}(x) - \boldsymbol{f}(x)\big)$. The derivation of the derivative is given in Appendix B.10. Using the gradient descent method, we obtain the gradient flow:

$$\begin{aligned}
\frac{d\boldsymbol{v}(x,t)}{dt} &= -\nabla_{\boldsymbol{u}} E(\boldsymbol{v}) \\
&= \int_\Omega \big(\boldsymbol{v}(y,t) - \boldsymbol{v}(x,t)\big)\big(k(x,y) + k(y,x)\big)dy \\
&+ \lambda\big(\boldsymbol{f}(x) - \boldsymbol{v}(x,t)\big).
\end{aligned} \tag{12}$$

If we set $K(x,y,t) := k(x,y) + k(y,x)$ to be symmetric and time-invariant, and $\lambda_P = \lambda, \lambda_I = \lambda_D = 0$, the controlled system in (10) simplifies to the gradient flow of E in (12). It suggests that integrating the (P) component into the system in (2) minimizes the functional E and reintroduces the lost information to the system.

PI-controlled SSM as Bregman iteration to minimize $E(\boldsymbol{v}, \boldsymbol{f})$. An alternative to gradient descent, Bregman iteration (Yin et al., 2008; Zhang et al., 2010) iteratively refines the solution by minimizing a Bregman divergence, which measures the discrepancy between the current solution and a reference point. Given the convex functional $J(\boldsymbol{v})$, the Bregman divergence of J between $\boldsymbol{v}$ and $\boldsymbol{s} \in L^2(\Omega)$ is $D_J^p(\boldsymbol{v}, \boldsymbol{s}) := J(\boldsymbol{v}) - J(\boldsymbol{s}) - \langle \boldsymbol{p}, \boldsymbol{v} - \boldsymbol{s} \rangle$, where $\boldsymbol{p} \in \partial J(\boldsymbol{s})$, the subgradient of J at $\boldsymbol{s}$. $D_J^p(\boldsymbol{v}, \boldsymbol{s})$ captures the difference between $J(\boldsymbol{v})$ and the tangent plane $J(\boldsymbol{s}) - \langle \boldsymbol{p}, \boldsymbol{v} - \boldsymbol{s} \rangle$. The $\ell+1$-th Bregman iteration to minimize $\min_{\boldsymbol{v}} J(\boldsymbol{v})$ with the contraint $G(\boldsymbol{v}, \boldsymbol{f})$ is given by:

$$\boldsymbol{v}^{\ell+1} = \arg\min_{\boldsymbol{v}} D_J^{p^\ell}(\boldsymbol{v}, \boldsymbol{v}^\ell) + G(\boldsymbol{v}, \boldsymbol{f}), p^\ell \in \partial J(\boldsymbol{v}^\ell) \quad (13)$$

The following Lemma 3 shows that the optimization problem in (13) can be turned into solving iterative subproblems.

Lemma 3. *Applying Bregman iteration to minimize* $E(\boldsymbol{v}, \boldsymbol{f})$ *involves solving iterative subproblems:*

$$\begin{aligned} \boldsymbol{v}^{\ell+1} &= \arg\min_{\boldsymbol{v}} J(\boldsymbol{v}) + \frac{\lambda}{2} \int_\Omega \|\boldsymbol{v}(x) - \boldsymbol{f}(x) - \boldsymbol{e}_{\mathrm{a}}^\ell(x)\|_2^2 dx \\ \boldsymbol{e}_{\mathrm{a}}^\ell(x) &= \sum_{m=1}^{\ell} \boldsymbol{e}^m(x) = \sum_{m=1}^{\ell} \big(\boldsymbol{f}(x) - \boldsymbol{v}^m(x)\big), \end{aligned} \quad (14)$$

The proof of Lemma 3 is in Appendix B.4. Here, the term $\boldsymbol{e}_{\mathrm{a}}^\ell(x)$ captures the accumulated information loss between the original and the smoothed signals $\boldsymbol{v}^m(x)$ of each iteration $m = 1, \ldots, \ell$. Taking a one-step update in the direction of gradient descent (see Appendix B.11), we obtain

$$\begin{aligned} \boldsymbol{v}^{\ell+1}(x) = &\int_\Omega \big(\boldsymbol{v}^\ell(y) - \boldsymbol{v}^\ell(x)\big)\big(k(x,y) + k(y,x)\big) dy \\ &+ \boldsymbol{v}^\ell(x) + \lambda \boldsymbol{e}^\ell(x) + \lambda \boldsymbol{e}_{\mathrm{a}}^\ell(x). \end{aligned} \quad (15)$$

On the other hand, the Euler discretization with $\Delta t = 1$ of the PI-controlled state-space in (10) (as $\lambda_D = 0$) is:

$$\begin{aligned} \boldsymbol{v}^{\ell+1}(x) = \boldsymbol{v}^\ell(x) &+ \int_\Omega \big(\boldsymbol{v}^\ell(y) - \boldsymbol{v}^\ell(x)\big) K(x,y) dy \\ &+ \lambda_P \boldsymbol{e}^\ell(x) + \lambda_I \sum_{m=1}^{\ell} \boldsymbol{e}^m(x). \end{aligned} \quad (16)$$

By selecting a time-independent $K(x,y,t) := k(x,y) + k(y,x)$ and setting $\lambda_P = \lambda_I = \lambda$, the update step of the PI-controlled model in (16) simplifies to the update step of Bregman iteration in (15). This connection suggests that the PI-controlled SSM minimizes $E(\boldsymbol{v}, \boldsymbol{f})$.

3.2. Stability and Representation Collapse of PID-Controlled State Space Model

In this section, we aim to show that: (i) Integrating the (P) term enhances robustness against input perturbations and mitigates rank collapse of the output representation; (ii) Adding the (D) term in PD-control further stabilizes the system by mitigating rapid and unstable changes of $\mathbf{V}(t)$, (iii) finally, integrating the (I) term in the PID-controlled SSM described in (10) guarantees system stability, making it robust to input corruption. Following the same assumption in Section 2.2, we assume that $K(x,y,t)$ is time-invariant for our theoretical analysis in this section.

3.2.1. Analysis of P-control SSM

Robustness of P-controlled SSM. From the SSM in (10), by choosing $\lambda_I = \lambda_D = 0$, and applying Euler discretization on the spatial domain, the P-control model is given as:

$$\begin{aligned} \frac{d\boldsymbol{v}(i,t)}{dt} = &\sum_{j=1}^{N} (\boldsymbol{v}(j,t) - \boldsymbol{v}(i,t)) K(i,j) \\ &+ \lambda_P \big(\boldsymbol{f}(i) - \boldsymbol{v}(i,t)\big), \end{aligned} \quad (17)$$

for $i, j = 1, 2, \ldots, N$, and $K(i,j) := \mathrm{softmax}\big(\boldsymbol{q}(i)^T \boldsymbol{k}(j) / \sqrt{D_{qk}}\big)$. The corresponding matrix representation is given as

$$\frac{d\mathbf{V}(t)}{dt} = \mathbf{K}\mathbf{V}(t) - (\lambda_P + 1)\mathbf{V}(t) + \lambda_P \boldsymbol{F}, \mathbf{V}(0) = \mathbf{V}^0. \quad (18)$$

where $\boldsymbol{F} = [\boldsymbol{f}(1), \ldots, \boldsymbol{f}(N)]^T$. The following Lemma 4 help us analyze the stability and representation collapse of the solution for the SSM in (18). Here, since the eigenvalues of $\mathbf{K}$ the has the real part in $[0, 1]$, $\lambda_P + 1$ ($\lambda_P > 0$) can not be one of them. This implies that $\det(K - (\lambda_P + 1)I) \neq 0$ hence the matrix is non-singular.

Lemma 4. *Let* $\boldsymbol{B} := \mathbf{K} - (\lambda_P + 1)\mathbf{I} \in \mathbb{R}^{N \times N}$, *the solution of the ordinary differential equation (18) is*

$$\mathbf{V}(t) = \exp(\boldsymbol{B}t)(\mathbf{V}^0 + \boldsymbol{B}^{-1}\boldsymbol{F}) - \lambda_P \boldsymbol{B}^{-1}\boldsymbol{F}. \quad (19)$$

If $\boldsymbol{B}$ *has only eigenvalues with negative real parts, then* $\lim_{t\to\infty} \boldsymbol{V}(t) = -\lambda_P \boldsymbol{B}^{-1}\boldsymbol{F}$.

The proof of Lemma 4 is shown in the Appendix B.5. As shown in Section 2.2, since the eigenvalues of $\boldsymbol{K}$ has $\mathrm{Re}(\alpha_i) \in [-1, 1], i = 1, \ldots, M$, therefore the real parts of eigenvalues of $\boldsymbol{B}$ must be in the range $[-2 - \lambda_P, -\lambda_p]$, which are all negative. As the result of 4, the steady state solution in (19) $\lim_{t\to\infty} \boldsymbol{V}(t) = -\lambda_P \boldsymbol{B}^{-1}\boldsymbol{F}$. Therefore, adding any perturbation to the initial state $\mathbf{V}^0$ does not change the steady state solution. However, in our context of a transformer, the perturbation also affects the reference point $\boldsymbol{F}$, which is chosen to be a scaled $\beta\mathbf{V}^0$, leading to the steady state solution becomes $-\lambda_P \beta \boldsymbol{B}^{-1}\mathbf{V}^0$. Fortunately,

the P-control system allows the error caused by perturbation to be as neglectable as desired. The following Proposition 1 confirms the robustness of the P-control SSM.

Proposition 1. *Given the coefficient $\lambda_P > 0$ in (10), and any arbitrary $\bar{\epsilon}, \delta > 0$, by adding the perturbation $\boldsymbol{\epsilon} \in \mathbb{R}^{N\times D}$, $\|\boldsymbol{\epsilon}\|_\infty \leq \bar{\epsilon}$ to $\mathbf{V}^0$, the corresponding change in the steady state solution of the system in (18) is independent of λ_P and becomes negligible with an amount of at most δ if*

$$\beta \leq \delta/\bar{\epsilon}. \tag{20}$$

The proof of Proposition 1 is shown in the Appendix B.6. Proposition 1 suggests that we can select the hyper-parameter β to make the impact of input perturbation on the output as small as desired.

P-controlled SSM on representation collapse. Since $\boldsymbol{B}^{-1}$ is full rank ($\boldsymbol{B}$ is non-singular), hence $\mathrm{rank}(-\lambda_P \boldsymbol{B}^{-1}\boldsymbol{F}) = \mathrm{rank}(\boldsymbol{F})$ (Strang, 2006). In the case of a transformer, when choosing $\boldsymbol{F} = \beta\mathbf{V}^0$, the rank of the steady state solution equals the rank of the input $\mathbf{V}^0$. This implies that the P-control dynamic in (18) prevents rank collapse.

3.2.2. Analysis of PD-controlled SSM

Since $\lambda_D \frac{d\boldsymbol{e}(x,t)}{dt} = \lambda_D \frac{d}{dt}(\boldsymbol{f}(x) - \boldsymbol{v}(x,t)) = -\lambda_D \frac{d\boldsymbol{v}(x,t)}{dt}$, from the SSM in (10), by choosing $\lambda_I = 0$, $K(i,j) := \mathrm{softmax}\big(\boldsymbol{q}(i)^T\boldsymbol{k}(j)/\sqrt{D_{qk}}\big)$ for $i,j = 1,2,\ldots,N$, and applying Euler discretization on the spatial domain, the PD-control model can be represented in the matrix form:

$$\begin{aligned}\mathbf{V}'(t) &= \mathbf{K}\mathbf{V}(t) - (\lambda_P+1)\mathbf{V}(t) + \lambda_P\boldsymbol{F} - \lambda_D\mathbf{V}'(t)\\ &= \frac{1}{1+\lambda_D}\big(\boldsymbol{K} - (\lambda_P+1)\boldsymbol{I}\big)\mathbf{V}(t) + \frac{\lambda_P}{1+\lambda_D}\boldsymbol{F},\end{aligned} \tag{21}$$

with $\mathbf{V}(0) = \mathbf{V}^0$. The solution of (21) is provided in the following Lemma 5.

Lemma 5. *Let $\boldsymbol{B} := \mathbf{K} - (\lambda_P+1)\mathbf{I} \in \mathbb{R}^{N\times N}$, the solution of the ordinary differential equation (21) is*

$$\mathbf{V}(t) = \exp\Big(\frac{1}{1+\lambda_D}\boldsymbol{B}t\Big)(\mathbf{V}^0 + \boldsymbol{B}^{-1}\boldsymbol{F}) - \lambda_P\boldsymbol{B}^{-1}\boldsymbol{F}.$$

and $\lim_{t\to\infty}\boldsymbol{V}(t) = -\lambda_P\boldsymbol{B}^{-1}\boldsymbol{F}$.

The proof of Lemma 5 is provided in Appendix B.7. This intriguing result suggests two key insights. Firstly, incorporating the (D) component into the P-control system does not alter the steady state of the solution. Consequently, the solution of the PD-controlled SSM retains the advantages of a P-control model, including avoiding rank collapse and ensuring bounded error. Secondly, the derivative term offers an additional benefit of further stabilizing the system by decreasing the eigenvalue by a factor of $1/(1+\lambda_D)$, thereby mitigating rapid changes in $\mathbf{V}(t)$.

3.2.3. Analysis of PID-controlled SSM

Following the same analysis in Section 3.2.1, by choosing $K(i,j) := \mathrm{softmax}\big(\boldsymbol{q}(i)^T\boldsymbol{k}(j)/\sqrt{D_{qk}}\big)$ and discretizing on the spatial domain, the matrix representation of the PID-controlled SSM reduced from (10) becomes

$$\begin{aligned}\mathbf{V}'(t) =& \frac{1}{\lambda_D+1}\Bigg(\big(\mathbf{K} - (\lambda_P+1)\boldsymbol{I}\big)\mathbf{V}(t)\\ &+ \lambda_I\int_0^t(\boldsymbol{F} - \mathbf{V}(t))dt + \lambda_P\boldsymbol{F}\Bigg),\end{aligned} \tag{22}$$

where $\mathbf{V}(0) = \mathbf{V}^0$. To deal with the integral in (22), we take the derivative of both sides, the equation becomes $\mathbf{V}''(t) = \frac{1}{\lambda_D+1}\big(\big(\mathbf{K} - (\lambda_P+1)I\big)\mathbf{V}'(t) - \lambda_I\mathbf{V}(t)\big)$, which is turned into a system of 1-st order differential equation:

$$\begin{bmatrix}\mathbf{V}'(t)\\ \mathbf{V}''(t)\end{bmatrix} = \begin{bmatrix}\boldsymbol{0} & \boldsymbol{I}\\ -\frac{\lambda_I\boldsymbol{I}}{\lambda_D+1} & \frac{\boldsymbol{K} - (\lambda_P+1)\boldsymbol{I}}{\lambda_D+1}\end{bmatrix}\begin{bmatrix}\mathbf{V}(t)\\ \mathbf{V}'(t)\end{bmatrix}, \tag{23}$$

where $\mathbf{V}(0) = \mathbf{V}^0$, and $\mathbf{V}'(0) = \frac{1}{\lambda_D+1}\big((\boldsymbol{K} - (\lambda_P+1))\mathbf{V}^0 + \lambda_P\boldsymbol{F}\big)$. To gain robustness, the steady state solution of the model should be independent of any perturbation of the input $\mathbf{V}_0$ The following Proposition 2 illustrates the stability of the system.

Proposition 2. *For any $\lambda_P, \lambda_I, \lambda_D > 0$, the system in (23) has a stable solution.*

The proof of Proposition 2 is in the Appendix B.8. The Proposition implies that the PID-controlled SSM in (10) remains robust and stable for any selection of positive values for $\lambda_P, \lambda_I, \lambda_D$.

3.3. Transformer with PID Control

By applying the Euler discretization with time step $\Delta t = 1$, initializing $\boldsymbol{v}$ at $t = 0$ as $\boldsymbol{v}(x,0) = v^0(x)$, and choosing

$$K(x,y,t) := \frac{\exp\big(\boldsymbol{q}(x,t)^T\boldsymbol{k}(y,t)/\sqrt{D_{qk}}\big)}{\int_\Omega \exp\big(\boldsymbol{q}(x,t)^T\boldsymbol{k}(y',t)/\sqrt{D_{qk}}\big)dy'},$$

the update step of PID-controlled SSM in (10) becomes:

$$\begin{aligned}&\boldsymbol{v}^{\ell+1}(x)\\ &\approx \int_\Omega\big(\boldsymbol{v}^\ell(y) - \boldsymbol{v}^\ell(x)\big)\frac{\exp\big(\boldsymbol{q}^\ell(x)^T\boldsymbol{k}^\ell(y)/\sqrt{D_{qk}}\big)}{\int_\Omega\exp\big(\boldsymbol{q}^\ell(x)^T\boldsymbol{k}^\ell(y')/\sqrt{D_{qk}}\big)dy'}dy\\ &+ \boldsymbol{v}^\ell(x) + \lambda_P\boldsymbol{e}^\ell(x) + \lambda_I\sum_{m=1}^{\ell}\boldsymbol{e}^m(x) + \lambda_D(\boldsymbol{e}^\ell(x) - \boldsymbol{e}^{\ell-1}(x)),\end{aligned} \tag{24}$$

where $\boldsymbol{e}^m(x) = \boldsymbol{f}(x) - \boldsymbol{v}^m(x)$ for $m = 1,\ldots,\ell$. Applying the Monte-Carlo method to approximate the integrals in (24) and discretizing $\boldsymbol{v}^{l+1}(x)$, $\boldsymbol{v}^m(x)$, and $\boldsymbol{v}_0(x)$ on a spatial

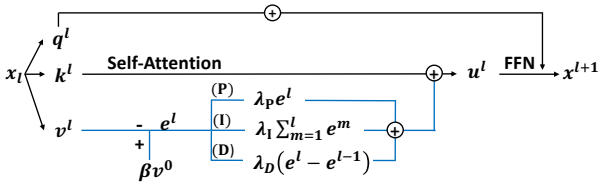

Figure 1. Our proposed PIDformer model at each layer.

dimension, and by choosing $\boldsymbol{f}(x) = \boldsymbol{v}(x)$, we attain the output of the following novel PID-attention at layer ℓ-th is defined as

Definition 1 (PID-control Transformer (PIDformer))**.** *Given a set of key and value vectors $\{\boldsymbol{k}^\ell(j), \boldsymbol{v}^\ell(j)\}_{j=1}^N$ in each layer ℓ, $\ell = 1, \ldots, L$, for each query vector $\boldsymbol{q}^\ell(i)$, $i = 1, \ldots, N$, in the same layer, the self-attention unit at layer ℓ in a PID-control Transformer (PIDformer) computes the corresponding output vector $\boldsymbol{u}^\ell(i)$ of the query $\boldsymbol{q}^\ell(i)$ by the following attention formula:*

$$\boldsymbol{u}^\ell(i) = \sum_{j=1}^{N} \text{softmax}\Big(\boldsymbol{q}^\ell(i)^\top \boldsymbol{k}^\ell(j)/\sqrt{D_{qk}}\Big)\boldsymbol{v}^\ell(y) + \lambda_P \boldsymbol{e}^\ell(i) + \lambda_I \sum_{m=1}^{\ell} \boldsymbol{e}^m(i) + \lambda_D(\boldsymbol{e}^\ell(i) - \boldsymbol{e}^{\ell-1}(i)), \tag{25}$$

where $\boldsymbol{e}^\ell = \boldsymbol{v}^0 - \boldsymbol{v}^\ell$, $\boldsymbol{v}^0(1), \ldots, \boldsymbol{v}^0(N) \in \mathbb{R}^D$ are the value vectors in the first layer of PIDformer.

Since PID-attention is a discretization of the controlled SSM in (10), it is inherently a more robust attention mechanism. Fig. 1 illustrates the architecture of PIDformer.

4. Experimental Results

In this section, we empirically demonstrate the advantages of our proposed PIDformer approach across multiple tasks, including ImageNet classification (Deng et al., 2009), ADE20K image segmentation (Zhou et al., 2018), and language modeling on WikiText-103 (Merity et al., 2017). Our objectives are to: (i) illustrate that PIDformer significantly outperforms the transformer baseline with softmax-attention across diverse tasks, (ii) highlight that the PID DeiT model exhibits significantly higher robustness than softmax attention under various adversarial attacks, and for out-of-distribution generalization, (iii) demonstrate that PID DeiT does not suffer from rank collapse in output representation. Throughout our experiments, we compare the performance of our proposed models with baselines of the same configuration. For additional details regarding datasets, models, and training procedures, please refer to Appendix A.

Object Classification on ImageNet. To demonstrate the advantage of our PIDformer, we compare it with the DeiT baseline (Touvron et al., 2021) on the ImageNet image classification task. Our PID DeiT surpasses the DeiT baseline, as shown in Table 1. Notably, our model performs significantly better than the baseline under white-box attacks, including fast gradient sign method (FGSM) (Dong et al., 2020), projected gradient descent method (PGD) (Tramer & Boneh, 2019b); score-based black-box attack method SPSA (Uesato et al., 2018); and sparse L_1 descent (SLD) (Tramer & Boneh, 2019a) method as well as noise-adding attack. Moreover, the last four rows of Table 1 demonstrate that PID DeiT is consistently more robust than the DeiT baseline under other adversarial examples and out-of-distribution dataset, including the Imagenet-C (common data corruption and perturbations, such as adding noise and blurring the images) (Hendrycks & Dietterich, 2019), Imagenet-A (adversarial examples) (Hendrycks et al., 2021b), Imagenet-R (out of distribution generalization) (Hendrycks et al., 2021a), and Imagenet-O(out-of-distribution detection) (Hendrycks et al., 2021b) datasets. Furthermore, in Appendix C.1, we visualize the performance gap between PID DeiT and the baseline DeiT model under attacks with escalating perturbation levels. This result demonstrates the significant advantages PIDformer has over the baseline model in terms of robustness, further confirming the benefits of our model.

Table 1. Evaluation of PID DeiT versus Softmax DeiT on the clean ImageNet validation set, as well as under various adversarial attacks and out-of-distribution datasets.

Attack	Metric/Model	*Softmax DeiT*	PID DeiT (%)
Clean	Top-1 Acc (%)	72.17	**73.13**
	Top-5 Acc (%)	91.02	**91.76**
FGSM	Top-1 Acc (%)	33.64	**38.52**
	Top-5 Acc (%)	68.18	**72.53**
PGD	Top-1 Acc (%)	12.02	**15.08**
	Top-5 Acc (%)	34.99	**39.69**
SPSA	Top-1 Acc (%)	65.75	**67.98**
	Top-5 Acc (%)	90.07	**90.58**
SLD	Top-1 Acc (%)	69.32	**70.84**
	Top-5 Acc (%)	90.8	**91.43**
Noise	Top-1 Acc (%)	69.2	**70.87**
	Top-5 Acc (%)	89.67	**90.77**
Imagenet-A	Top-1 Acc (%)	6.90	**8.82**
Imagenet-R	Top-1 Acc (%)	32.83	**34.89**
Imagenet-C	mCE (↓)	71.20	**68.41**
Imagenet-O	AUPR	17.47	**19.22**

Image Segmentation on ADE20K dataset. We evaluate the performance of Segmenter models (Strudel et al., 2021) using both PID DeiT and DeiT backbones on the ADE20K image segmentation task (Zhou et al., 2017), as outlined in Table 2. The outcomes illustrate significant performance enhancements obtained by employing the PID DeiT backbone instead of the DeiT backbone across both single-scale (SS) and multi-scale (MS) Mean Intersection over Union (MIoU) metrics.

Language Model on WikiText-103. In addition to computer vision tasks, we evaluate our model's performance

Table 2. Single-scale (SS) MIoU and multi-scale MIoU (MS) of the PID DeiT vs. the DeiT on the ADE20K image segmentation.

Model/Metric	SS MIoU	MS MIoU (%)
Softmax DeiT	35.72	36.68
PID DeiT	**37.42**	**38.28**

Table 3. Test and valid perplexity (Test PPL and Valid PPL) on WikiText-103 of PIDformer compared to the softmax transformer.

Method/Metric	Valid PPL	Test PPL
Softmax Transformer	33.15	34.29
PIDformer	**32.44**	**33.45**

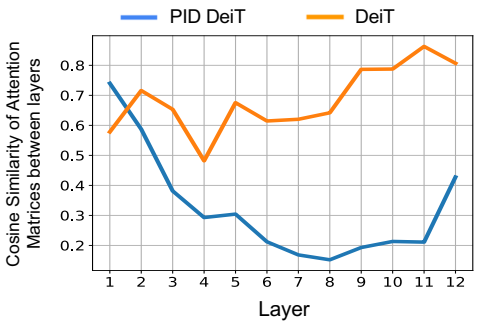

Figure 2. The cosine *similarity* of token representations in PID DeiT compared to baseline DeiT models across layers for ImageNet classification. The DeiT baseline demonstrates representation rank collapse as tokens become increasingly similar as depth increases. In contrast, PID DeiT models exhibit significantly greater diversity in tokens, indicating a mitigation in rank-collapse.

in the language modeling task on the WikiText-103 dataset (Table 3). Our PIDformer language model surpasses the softmax transformer model (Xiong et al., 2021) in test and valid perplexity. These results, combined with findings across various tasks, empirically demonstrate the significant advantages of PIDformer models.

Representation Collapse Analysis. We empirically show PIDformer's effectiveness in addressing rank collapse in transformers. In Fig. 2, we compare token representation cosine similarity across layers in PID DeiT and softmax baseline models pretrained on Imagenet. PID DeiT exhibits significantly lower similarity, especially in later layers, alleviating rank collapse and enhancing token embedding diversity. Further details are in Appendix A.6.

5. Related Work

Robust transformer. Ensuring the generalization and robustness of both vision transformer and language model remains an ongoing research focus. Large language models are vulnerable to input corruption (Wang et al., 2021; Peyrard et al., 2022; Jin et al., 2020; Zang et al., 2019), posing a challenge in developing robust real-world applications that can withstand unforeseen adversarial threats. For ViTs, investigations into model robustness against adversarial attacks, domain shifts, and out-of-distribution data are crucial for real-world deployment. Techniques such as data augmentation, regularization, and adversarial training are actively explored to enhance the robustness and generalization capabilities of ViTs. Many investigations (e.g., (Yuan et al., 2023; Paul & Chen, 2022; Mahmood et al., 2021; Bhojanapalli et al., 2021; Madry et al., 2017; Mao et al., 2022; Zhou et al., 2022)) have attempted to explain and improve the resilience of ViT models against typical adversarial attacks. For example, (Mahmood et al., 2021) empirically mitigates ViT's vulnerability to white-box adversarial attacks by introducing a simple ensemble defense strategy that notably enhanced robustness without sacrificing accuracy on clean data.

Rank-collapse in transformer. Rank collapse in deep transformers, observed across domains from natural language processing (Shi et al., 2022) to computer vision (Wang et al., 2022; Dong et al., 2021), is evident. In computer vision, Zhou et al. (2021) find that adding more layers to the Vision Transformer (ViT) (Dosovitskiy et al., 2021a) quickly saturates its performance. Moreover, their experiments show that a 32-layer ViT performs worse than a 24-layer ViT, attributed to token representations becoming identical with increasing model depth. To address this matter, (Wang et al., 2022) discovers that self-attention functions as a low-pass filter, causing token representations in ViTs to be smoothed. Furthermore, (Shi et al., 2022) identifies a similar phenomenon in BERT (Devlin et al., 2018), and investigates rank-collapse from a graph perspective. Their work shows that the self-attention matrix is like a normalized adjacency matrix of a graph and layer normalization is crucial in addressing the over-smoothing issue in Transformer models. If the standard deviation of layer normalization is too large, Transformer outputs converge to a low-rank subspace, causing over-smoothing. To mitigate this, the authors use hierarchical fusion strategies to adaptively combine representations from different layers, ensuring more diverse outputs. Our work is orthogonal to the existing method as we develop a control framework to tackle the non-robustness and rank-collapse issues in transformers. Our work is orthogonal to these works since we explain the rank-collapse of transformers from the point of view of control theory.

Control theory in deep learning There are exisiting works using control theory to design other network structures in our revision. Among these works, (Chen et al., 2021b) introduces the Close-Loop Control Neural Network (CLC-NN), which utilizes an additional control signal to implicitly minimize a running loss that measures discrepancies between true and observed features under input perturbation at each layer, hence promoting robustness

of the model.

Additionally, (Luo et al., 2023) explores the use of optimal control to design multi-round prompt tuning for large language models. This approach provides a different perspective on leveraging control theory for neural networks, focusing on enhancing dynamics multi-round interactions in prompt engineering rather than directly on network robustness.

Orthogonal to these existing approaches, our work introduces a unique control framework for self-attention mechanisms, establishing a novel connection between self-attention and the state-space model (SSM). Through a detailed analysis of the SSM, we elucidate the vulnerabilities inherent in transformer models and propose a solution by integrating PID control into the transformer. This integration not only addresses identified vulnerabilities but also contributes to the broader discourse on applying control theory principles to deepen our understanding and enhance the robustness of neural network architectures.

Another work that integrate control to improve deep learning models is ControlVAE framework, proposed by (Shao et al., 2020). ControlVAE employs Proportional-Integral (PI) control to tune a hyperparameter in the Variational Autoencoder (VAE) objective automatically. Our approach, however, integrates PID control directly into the architecture of transformers, marking a conceptual and practical departure from adjusting hyperparameters.

6. Concluding Remarks

In this paper, we present a novel control framework for self-attention mechanisms, revealing their inherent non-robustness and susceptibility to rank collapse in token representation. Leveraging this control perspective, we introduce the PIDformer, a novel PID-control Transformer designed to enhance robustness and mitigate the rank-collapse issue. Empirical validation across a range of large-scale applications, including ImageNet object classification (under various input perturbations and robustness benchmarks), ADE20K object segmentation, and WikiText-103 language modeling, confirms PIDformer's benefits. A limitation of our paper is the oversight regarding the privacy-preserving aspects of PIDformer. Exploring the potential of controlled transformers in enhancing privacy-preserving techniques is an intriguing avenue for future research.

Acknowledgement

This work is supported by NSF grants CCF-1911094, IIS-1838177, and IIS-1730574;ONR grants N00014-18-12571, N00014-20-1-2534, and MURI N00014-20-1-2787; AFOSR grant FA9550-22-1-0060; and a Vannevar Bush Faculty Fellowship, ONR grant N00014-18-1-2047, the National Science Foundation Grant #2211815 and the Google Research Scholar Award. This research/project is supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-TC-2023-012-SGIL).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Al-Rfou, R., Choe, D., Constant, N., Guo, M., and Jones, L. Character-level language modeling with deeper self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3159–3166, 2019.

Bandeira, A. S., Singer, A., and Strohmer, T. *Mathematics of Data Science*. 2020. URL https://people.math.ethz.ch/~abandeira/BandeiraSingerStrohmer-MDS-draft.pdf.

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., and Veit, A. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10231–10241, 2021.

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021a.

Chen, Z., Li, Q., and Zhang, Z. Towards robust neural networks via close-loop control. *arXiv preprint arXiv:2102.01862*, 2021b.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*, 2019.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://www.aclweb.org/anthology/D14-1179.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dong, Y., Fu, Q.-A., Yang, X., Pang, T., Su, H., Xiao, Z., and Zhu, J. Benchmarking adversarial robustness on image classification. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 321–331, 2020.

Dong, Y., Cordonnier, J.-B., and Loukas, A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pp. 2793–2803. PMLR, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=YicbFdNTTy.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=YicbFdNTTy.

Gilboa, G. and Osher, S. Nonlocal linear image regularization and supervised segmentation. *Multiscale Model. Simul.*, 6:595–630, 2007.

Gilboa, G. and Osher, S. Nonlocal operators with applications to image processing. *Multiscale Model. Simul.*, 7: 1005–1028, 2008.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Guo, M.-H., Cai, J.-X., Liu, Z.-N., Mu, T.-J., Martin, R. R., and Hu, S.-M. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

Janner, M., Li, Q., and Levine, S. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34: 1273–1286, 2021.

Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 8018–8025, 2020.

Luo, Y., Tang, Y., Shen, C., Zhou, Z., and Dong, B. Prompt engineering through the lens of optimal control. *arXiv preprint arXiv:2310.14201*, 2023.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mahmood, K., Mahmood, R., and Van Dijk, M. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7838–7847, 2021.

Mao, X., Qi, G., Chen, Y., Li, X., Duan, R., Ye, S., He, Y., and Xue, H. Towards robust vision transformer. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 12042–12051, 2022.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Metropolis, N. and Ulam, S. The monte carlo method. *Journal of the American Statistical Association*, 44(247): 335–341, 1949. ISSN 01621459. URL http://www.jstor.org/stable/2280232.

Morača, N. Upper bounds for the infinity norm of the inverse of sdd and s-sdd matrices. *Journal of Computational and Applied Mathematics*, 206(2):666–678, 2007. ISSN 0377-0427. doi: https://doi.org/10.1016/j.cam.2006.08.013. URL https://www.sciencedirect.com/science/article/pii/S0377042706005139.

Parikh, A., Täckström, O., Das, D., and Uszkoreit, J. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2249–2255, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1244. URL https://www.aclweb.org/anthology/D16-1244.

Paul, S. and Chen, P.-Y. Vision transformers are robust learners. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 36, pp. 2071–2081, 2022.

Peyrard, M., Ghotra, S., Josifoski, M., Agarwal, V., Patra, B., Carignan, D., Kiciman, E., Tiwary, S., and West, R. Invariant language modeling. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5728–5743, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.387. URL https://aclanthology.org/2022.emnlp-main.387.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015.

Schlag, I., Irie, K., and Schmidhuber, J. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, pp. 9355–9366. PMLR, 2021.

Shao, H., Yao, S., Sun, D., Zhang, A., Liu, S., Liu, D., Wang, J., and Abdelzaher, T. Controlvae: Controllable variational autoencoder. In *International conference on machine learning*, pp. 8655–8664. PMLR, 2020.

Shi, H., GAO, J., Xu, H., Liang, X., Li, Z., Kong, L., Lee, S. M. S., and Kwok, J. Revisiting over-smoothing in BERT from the perspective of graph. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=dUV91uaXm3.

Silvester, J. R. Determinants of block matrices. *The Mathematical Gazette*, 84(501):460–467, 2000. ISSN 00255572. URL http://www.jstor.org/stable/3620776.

Strang, G. *Linear algebra and its applications*. Thomson, Brooks/Cole, Belmont, CA, 2006. ISBN 0030105676 9780030105678 0534422004 9780534422004. URL http://www.amazon.com/Linear-Algebra-Its-Applications-Edition/dp/0030105676.

Strudel, R., Garcia, R., Laptev, I., and Schmid, C. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7262–7272, 2021.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. Training data-efficient image

transformers distillation through attention. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

Tramer, F. and Boneh, D. Adversarial training and robustness for multiple perturbations. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/5d4ae76f053f8f2516ad12961ef7fe97-Paper.pdf.

Tramer, F. and Boneh, D. Adversarial training and robustness for multiple perturbations. *Advances in neural information processing systems*, 32, 2019b.

Uesato, J., O'Donoghue, B., Kohli, P., and van den Oord, A. Adversarial risk and the dangers of evaluating against weak attacks. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5025–5034. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/uesato18a.html.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Wang, B., Wang, S., Cheng, Y., Gan, Z., Jia, R., Li, B., and Liu, J. Infobert: Improving robustness of language models from an information theoretic perspective. 2021. Publisher Copyright: © 2021 ICLR 2021 - 9th International Conference on Learning Representations. All rights reserved.; 9th International Conference on Learning Representations, ICLR 2021 ; Conference date: 03-05-2021 Through 07-05-2021.

Wang, P., Zheng, W., Chen, T., and Wang, Z. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=O476oWmiNNp.

Wang, Y. and Bansal, M. Robust machine comprehension models via adversarial training. In Walker, M., Ji, H., and Stent, A. (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 575–581, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2091. URL https://aclanthology.org/N18-2091.

Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., and Singh, V. Nyströmformer: A Nyström-based Algorithm for Approximating Self-Attention. 2021.

Yin, W., Osher, S., Goldfarb, D., and Darbon, J. Bregman iterative algorithms for l(1)-minimization with applications to compressed sensing. *Siam Journal on Imaging Sciences - SIAM J IMAGING SCI*, 1, 01 2008. doi: 10.1137/070703983.

Yuan, Z., Zhou, P., Zou, K., and Cheng, Y. You are catching my attention: Are vision transformers bad learners under backdoor attacks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24605–24615, 2023.

Zang, Y., Qi, F., Yang, C., Liu, Z., Zhang, M., Liu, Q., and Sun, M. Word-level textual adversarial attacking as combinatorial optimization. *arXiv preprint arXiv:1910.12196*, 2019.

Zhang, S., Yao, L., Sun, A., and Tay, Y. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1):1–38, 2019.

Zhang, X., Burger, M., Bresson, X., and Osher, S. Bregmanized nonlocal regularization for deconvolution and sparse reconstruction. *SIAM Journal on Imaging Sciences*, 3(3):253–276, 2010. doi: 10.1137/090746379. URL https://doi.org/10.1137/090746379.

Zhao, H., Jiang, L., Jia, J., Torr, P. H., and Koltun, V. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16259–16268, 2021.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., and Torralba, A. Semantic understanding of scenes through the ade20k dataset, 2018.

Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., and Feng, J. Deepvit: Towards deeper vision transformer, 2021.

Zhou, D., Yu, Z., Xie, E., Xiao, C., Anandkumar, A., Feng, J., and Alvarez, J. M. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pp. 27378–27394. PMLR, 2022.

A. Additional Details on the Experiments in Section 4

This section provides datasets, models, and training details for experiments in Section 4.

A.1. Image Classification on Imagenet

Datasets and Metrics. The ImageNet dataset, as described in (Deng et al., 2009; Russakovsky et al., 2015), consists of 1.28 million images for training and $50,000$ images for validation, covering the classification of 1000 different categories. Performance evaluation is based on top-1 and top-5 accuracies.

Models and Baselines. Our baseline model is the DeiT-tiny model (Touvron et al., 2021), which consists of 12 transformer layers, 3 attention heads per layer, and a model dimension of 192. For model setting and setting and configuration, we follow (Touvron et al., 2021). Their implementation is available at https://github.com/facebookresearch/deit. The $\lambda_P, \lambda_I, \lambda_D$, and β used for our PID DeiT method is $0.8, 0.5, 0.05$, and 0.1, respectively.

A.2. Image Segmentation on ADK20 dataset

Datasets and Metrics. The ADE20K dataset is renowned for its incorporation of complex scenes featuring detailed labels, establishing it as one of the most rigorous semantic segmentation datasets. It comprises a training set of $20,210$ images covering 150 semantic classes. Furthermore, the dataset includes $2,000$ images in the validation set and $3,352$ images in the test set. Performance in this task is evaluated using the Single-scale mean Intersection over Union (SS mIoU) and Multi-scale (MS mIoU) metrics.

Models and baselines. The training configuration and setting for our models are followed by (Strudel et al., 2021). The baseline model is finetuned with the pretrained DeiT-tiny backbone while our segmenter model used the pretrained PID DeiT-tiny, with $\lambda_P, \lambda_I, \lambda_D$, and β are $0.5, 0.3, 0.05$, and 1, respectively.

A.3. Language Modeling on WikiText-103

Datasets and Metrics. The WikiText-103 dataset is composed of Wikipedia articles tailored to capture extensive contextual dependencies. Its training set includes roughly $28,000$ articles, totaling around 103 million words. Each article consists of text blocks averaging about $3,600$ words. The validation and test sets contain $218,000$ and $246,000$ words, respectively, divided into 60 articles per set and approximately $268,000$ words each. Our experiment adheres to the standard setup outlined in (Merity et al., 2017; Schlag et al., 2021), which entails segmenting the training data into independent long segments of length L words. For evaluation, we utilize a batch size of 1 and process the text sequence using a sliding window of size L. When calculating perplexity (PPL), we only consider the last position, except for the first segment where all positions are evaluated, consistent with the methodology in (Al-Rfou et al., 2019; Schlag et al., 2021).

Models and baselines. For our language modeling implementation, we rely on the publicly available code https://github.com/IDSIA/lmtool-fwp developed by (Schlag et al., 2021). In our experiments, we set the dimensions of keys, values, and queries to 128, while the training and evaluation context length is set to 256. In this experiment, $\lambda_P, \lambda_I, \lambda_D$, and β being set to $0.4, 0.5, 0.1$ and 0.3, respectively, yields the best performance of PIDformer language model.

A.4. Adversarial Examples and Out-of-distribution datasets

Imagenet-C To assess robustness against typical image corruptions, we employ the ImageNet-C dataset (Hendrycks & Dietterich, 2019), which comprises 15 categories of artificially generated corruptions spanning five levels of severity. ImageNet-C evaluates models using the mean corruption error (mCE) metric, where a lower mCE value indicates greater resilience to corruption.

Imagenet-A ImageNet-A (Hendrycks et al., 2021b) is a dataset consisting of authentic images that have been filtered to deceive ImageNet classifiers. Specifically, it focuses on a subset of 200 classes chosen from the original 1000 classes in ImageNet-1K. Errors made within these 200 classes are regarded as significant, encompassing a wide range of categories represented in ImageNet-1K.

Imagenet-O This dataset comprises examples that have been adversarially filtered to challenge out-of-distribution detectors for ImageNet (Hendrycks et al., 2021b). It includes samples from the larger ImageNet-22K dataset but excludes those from ImageNet1K. Specifically, samples are chosen if they are confidently misclassified as an ImageNet-1K class by a ResNet-50

model. The evaluation metric utilized is the area under the precision-recall curve (AUPR).

Imagenet-R This dataset comprises diverse artistic interpretations of object classes found in the original ImageNet dataset, a practice discouraged by the creators of the original ImageNet (Hendrycks et al., 2021a). ImageNet-R encompasses 30,000 renditions of images representing 200 classes from the ImageNet dataset, with a selection made from a subset of the ImageNet-1K classes.

A.5. Adversarial Attacks

We employ fast gradient sign method (FGSM) (Dong et al., 2020), projected gradient descent method (PGD) (Tramer & Boneh, 2019b); and Sparse L_1 descent method as well as noise-adding attack These attacks were applied to the entire validation set of ImageNet. FGSM and PGD attacks distort the input image with a perturbation budget $\epsilon = 3/255$, and $\epsilon = 0.1$ for SPSA, under l_∞ norm, while the PGD attack uses 20 steps with a step size of $\alpha = 0.15$. For the SLD and noise attack, we follow the same setting in https://github.com/cleverhans-lab

A.6. Rank-collapse Analysis

The average cosine similarity between all pairs of token's representations $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in a sequence is computed as

$$\frac{1}{N(N-1)} \sum_{i \neq j} \frac{\boldsymbol{x}_i^T \boldsymbol{x}_j}{\|\boldsymbol{x}_i\|_2 \|\boldsymbol{x}_j\|_2}.$$

The result is then averaged over 1000 randomly chosen test data in ImageNet. The result is then reported for each layer, as in Fig. 2.

B. Technical Proofs

B.1. Solution of the first order ODE

Given $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$, $\boldsymbol{Y}(t) \in \mathbb{R}^{N \times P}, t > 0$, we are interested in the solution of the first order ODE stated as:

$$\boldsymbol{Y}'(t) = \boldsymbol{Q}\boldsymbol{Y}(t), \boldsymbol{Y}(0) = \boldsymbol{Y}^0. \tag{26}$$

The general solution of (26) is $\boldsymbol{Y}(t) = \exp(\boldsymbol{Q}t)\boldsymbol{C}$, where $\boldsymbol{C} \in \mathbb{R}^{n \times p}$ is an any constant matrix. Indeed,

$$\begin{aligned} \boldsymbol{Y}'(t) &= (\boldsymbol{I} + \boldsymbol{Q}t + \frac{\boldsymbol{Q}^2 t^2}{2!} + \frac{\boldsymbol{Q}^3 t^3}{3!} + \dots)'\boldsymbol{C} \\ &= (\boldsymbol{Q} + \boldsymbol{Q}^2 t + \frac{\boldsymbol{Q}^3 t}{2!} + \dots)\boldsymbol{C} \\ &= \boldsymbol{Q}\exp(\boldsymbol{Q}t)\boldsymbol{C} = \boldsymbol{Q}\boldsymbol{Y}(t). \end{aligned} \tag{27}$$

To satisfy the intitial condition, $\boldsymbol{Y}(0) = \boldsymbol{Q}\exp(\boldsymbol{Q}0)\boldsymbol{C} = \boldsymbol{Y}^0$. Hence, $\boldsymbol{C} = \boldsymbol{Y}^0$ and the solution for the intial value problem in (26) is $\exp(\boldsymbol{Q}t)\boldsymbol{Y}^0$.

Every square matrix can be Jordan decomposed as the form of $\boldsymbol{Q} = \boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1}$, where $\boldsymbol{S}$ is invertible and contains the generalized eigenvectors of $\boldsymbol{Q}$, and $\boldsymbol{J} = \mathbf{diag}(\boldsymbol{J}_{\eta_1,m_1}, \boldsymbol{J}_{\eta_2,m_2}, \dots, \boldsymbol{J}_{\eta_M,m_M})$ is the Jordan form of matrix $\boldsymbol{Q}$ with,

$\boldsymbol{J}_{\eta_i,m_i} = \begin{bmatrix} \eta_i & 1 & \dots & 0 \\ \vdots & \ddots & & \vdots \\ & & \eta_i & 1 \\ 0 & \dots & & \eta_i \end{bmatrix} \in \mathbb{R}^{m_i \times m_i}$, for $i = 1, \dots, M$ are Jordan blocks and $\eta_1, \dots \eta_M$ are eigenvalues of $\boldsymbol{Q}$.

We rewrite the solution of (26) using the Jordan decomposition as

$$\begin{aligned} \boldsymbol{Y}(t) &= \exp(\boldsymbol{Q}t)\boldsymbol{Y}^0 = \exp(\boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1}t)\boldsymbol{Y}^0 \\ &= (\boldsymbol{S}\boldsymbol{S}^{-1} + \boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1}t + \frac{(\boldsymbol{S}\boldsymbol{J}\boldsymbol{S}^{-1})^2 t^2}{2!} + \dots)\boldsymbol{Y}^0 \\ &= \boldsymbol{S}\exp(\boldsymbol{J}t)\boldsymbol{S}^{-1}\boldsymbol{Y}^0. \end{aligned} \tag{28}$$

We are now interested in the asymptotic behavior of the solution in (28) as $t \to \infty$.
When Q only has eigenvalues negative real parts. As $\eta_1, \dots \eta_M < 0$, we consider

$$\begin{aligned}
\exp(\boldsymbol{J}_{\eta_i,m_i} t) &= \sum_{k=0}^{\infty} \frac{(\boldsymbol{J}_{\eta_i,m_i} t)^k}{k!} \\
&= \begin{bmatrix} \sum_{k=0}^{\infty} \frac{t^k \eta_i^k}{k!} & \sum_{k=1}^{\infty} \frac{t^k \eta_i^{k-1}}{(k-1)!} & \dots & \sum_{k=m_i}^{\infty} \frac{t^k \eta_i^{k-m_i+1}}{(k-m_i+1)!} \\ \vdots & \ddots & & \\ 0 & \dots & \sum_{k=0}^{\infty} \frac{t^k \eta_i^k}{k!} & \sum_{k=1}^{\infty} \frac{t^k \eta_i^{k-1}}{(k-1)!} \\ 0 & \dots & 0 & \sum_{k=0}^{\infty} \frac{t^k \eta_i^k}{k!} \end{bmatrix} \\
&= \begin{bmatrix} e^{\eta_i t} & te^{\eta_i t} & \dots & t^{m_i-1} e^{\eta_i t} \\ \vdots & \ddots & & \\ 0 & \dots & e^{\eta_i t} & te^{\eta_i t} \\ 0 & \dots & 0 & e^{\eta_i t} \end{bmatrix}
\end{aligned} \tag{29}$$

which is derived from the result $\boldsymbol{J}^k_{\eta_i,m_i} = \begin{bmatrix} \eta_i^k & \binom{j}{1} \eta_i^{k-1} & \dots & \binom{j}{m_i-1} \eta_i^{k-m_i+1} \\ \vdots & \ddots & & \\ 0 & \dots & \eta_i^k & \binom{j}{1} \eta_i^{k-1} \\ 0 & \dots & 0 & \eta_i^k \end{bmatrix}$

Therefore, when $t \to 0, \exp(\boldsymbol{J}_{\eta_i,m_i} t) \to \mathbf{0}$, making $\exp(\boldsymbol{J}t) \to \mathbf{0}$ and hence the solution in (28) will goes to $\mathbf{0}$ or being stable.
When Q only has at least one eigenvalue with positive real part. Without the loss of generalization, let $\mathrm{Re}(\eta_1) > 0$. Hence $\|\exp(\boldsymbol{J}_{\eta_1,m_i} t)\| \to \infty$ when $t \to \infty$. In other words, the solution of (26) will explode or unstable.

B.2. Proof of Lemma 1

The proof of Lemma 1 is the direct result in Appendix B.1. The solution of the ordinary differential equation (ODE) in (8) is $\mathbf{V}(t) = \boldsymbol{P}\exp(\boldsymbol{J}t)\boldsymbol{P}^{-1}\mathbf{V}^0$, where $\boldsymbol{P}\boldsymbol{J}\boldsymbol{P}^{-1}$ if the Jordan decomposition of $\boldsymbol{K} - \boldsymbol{I}$, $\boldsymbol{J} = \mathbf{diag}(\boldsymbol{J}_{\alpha_1,m_1}, \boldsymbol{J}_{\alpha_2,m_2}, \dots, \boldsymbol{J}_{\alpha_M,m_M})$ and $\alpha_1 \geq \alpha_2 \geq \dots, \geq \alpha_M, M \leq N$ are eigenvalues $\mathbf{K} - \mathbf{I}$. Consequently, we have proved the Lemma 1

B.3. Proof of Lemma 2

In Section 2.2, we have shown that $\boldsymbol{K} - \boldsymbol{I}$ has a unique largest eigenvalue $\lambda_1 = 0$. This means that the Jordan blocks corresponding with other eigenvalues which has negative real parts will approach $\exp(\boldsymbol{J}_{\eta_i,m_i} t) \to \mathbf{0}$, for $i = 1, \dots, M;, i \neq 1$, as $t \to \infty$. As the consequence, $\exp(\boldsymbol{J}t)$ are fill with 0 for all entries except for the first entry $\exp(\boldsymbol{J}t)(0,0) = 1$. Hence, the solution in (9) becomes

$$\begin{bmatrix} c_{1,1}\boldsymbol{p_1}, & \dots, & c_{1,D_x}\boldsymbol{p_1} \end{bmatrix}.$$

This concludes the proof.

B.4. Proof of Lemma 3

For $\boldsymbol{v}^{\ell+1}$ to be the solution of the optimization problem in (13), since $\mathbf{0} \in \partial J(\boldsymbol{v}^{\ell+1}) - \boldsymbol{p}^{\ell} + \partial G(\boldsymbol{v}^{\ell+1}, \boldsymbol{f})$, hence the iteration becomes:

$$\begin{cases} \boldsymbol{v}^{\ell+1} = \arg\min_{\boldsymbol{v}} J(\boldsymbol{v}) - <\boldsymbol{p}^{\ell}, \boldsymbol{v}> + G(\boldsymbol{v}, \boldsymbol{f}) \\ \boldsymbol{p}^{\ell+1} \in \boldsymbol{p}^{\ell} - \partial G(\boldsymbol{v}^{\ell+1}, \boldsymbol{f}). \end{cases}$$

When $G(\boldsymbol{v}, \boldsymbol{f}) = \frac{\lambda}{2}\int_{\Omega} \|\boldsymbol{v}(x) - \boldsymbol{f}(x)\|_2^2 dx$,

$$\begin{aligned} G(\boldsymbol{v}, \boldsymbol{f}) - \langle \boldsymbol{p}^{\ell}, \boldsymbol{v}\rangle &= \frac{\lambda}{2}\int_{\Omega}\left(\left(\|\boldsymbol{v}(x)\|_2^2 - 2\langle \boldsymbol{v}(x), \boldsymbol{f}(x)\rangle + \|\,\boldsymbol{f}(x)\|_2^2\right) + \lambda\langle \sum_{m=1}^{\ell}\left(\boldsymbol{v}^m(x) - \boldsymbol{f}(x)\right), \boldsymbol{v}(x)\rangle \right) dx \\ &= \frac{\lambda}{2}\int_{\Omega}\left(\|\boldsymbol{v}(x)\|_2^2 - \lambda\langle \boldsymbol{f}(x) - \sum_{m=1}^{\ell}\Big(\boldsymbol{v}^m(x) - \boldsymbol{f}(x)\Big), \boldsymbol{v}(x)\rangle \right) dx + \text{constant} \\ &= \frac{\lambda}{2}\int_{\Omega} \|\boldsymbol{v}(x) - \boldsymbol{f}^{\ell}(x)\|_2^2 dx + \text{constant}, \end{aligned}$$

where $\boldsymbol{f}^{\ell}(x) = \boldsymbol{f}^{\ell-1}(x) + \boldsymbol{f}(x) - \boldsymbol{v}^{\ell}(x)$.
Substituting $G(\boldsymbol{v}, \boldsymbol{f}) - \langle \boldsymbol{p}^{\ell}, \boldsymbol{v}\rangle$ into the iteration, it becomes

$$\begin{cases} \boldsymbol{v}^{\ell+1} = \arg\min_{\boldsymbol{v}} J(\boldsymbol{v}) + \frac{\lambda}{2}\int_{\Omega} \|\boldsymbol{v}(x) - \boldsymbol{f}^{\ell}(x)\|_2^2 dx \\ \boldsymbol{f}^{\ell}(x) = \boldsymbol{f}^{\ell-1}(x) + \boldsymbol{f}(x) - \boldsymbol{v}^{\ell}(x). \end{cases} \tag{30}$$

The iteration in Lemma 3 can be reformulated as:

$$\boldsymbol{v}^{\ell+1} = \arg\min_{\boldsymbol{v}} J(\boldsymbol{v}) + \frac{\lambda}{2}\int_{\Omega} \|\boldsymbol{v}(x) - \boldsymbol{f}(x) - \boldsymbol{e}_{\text{a}}^{\ell}(x)\|_2^2 dx$$

where $\boldsymbol{e}_{\text{a}}^{\ell}(x) = \sum_{m=1}^{\ell} \boldsymbol{e}^m(x) = \sum_{m=1}^{\ell}\left(\boldsymbol{f}(x) - \boldsymbol{v}^m(x)\right)$ we conclude the proof for Lemma 3.

B.5. Proof of Lemma 4

To find the solution of Eqn 18, firstly, we find the solution for the homogenous ODE:

$$\mathbf{V}^{(h)\prime}(t) = \big(\mathbf{K} - (\lambda_P + 1)\boldsymbol{I}\big)\mathbf{V}^{(h)}(t) \tag{31}$$

From the result in Appendix B.1, the solution for this ODE is $\exp(\boldsymbol{B}t)\boldsymbol{C}$ where $\boldsymbol{B} = \mathbf{K} - (\lambda_P + 1)\boldsymbol{I}$ and $\boldsymbol{C} \in \mathbb{R}^{N\times D_x}$ is any constant matrix. Secondly, we find a particular solution for (18) by solving $\mathbf{V}^{(p)\prime}(t) = \boldsymbol{B}\mathbf{V}(t)^{(p)} + \lambda_P \boldsymbol{F} = \mathbf{0}$. Since $\boldsymbol{B}$ is invertible, the solution for this equation is $\mathbf{V}^{(p)}(t) = -\lambda_P \boldsymbol{B}^{-1}\boldsymbol{F}$. It is easy to check that $\mathbf{V}(t) = \mathbf{V}^{(h)}(t) + \mathbf{V}^{(p)}(t)$ is the solution of the $\mathbf{V}'(t) = \boldsymbol{B}\mathbf{V}(t) + \lambda_P \boldsymbol{F}$. Applying the initial condition, $\mathbf{V}(0) = \boldsymbol{C} - \lambda_P \boldsymbol{B}^{-1}\boldsymbol{F} = \mathbf{V}^0$, we find $\boldsymbol{C} = \mathbf{V}^0 + \lambda_P \boldsymbol{B}^{-1}\boldsymbol{F}$. Therefore, we have proved that the solution for the IVP problem in (18) is indeed $\mathbf{V}(t) = \exp(\boldsymbol{B}t)(\mathbf{V}^0 + \boldsymbol{B}^{-1}\boldsymbol{F}) - \lambda_P \boldsymbol{B}^{-1}\boldsymbol{F}$.

In Section 3.2.1, we show that $\boldsymbol{B}$ has only eigenvalues with negative real parts. As the result in Appendix B.1, when $t \to 0$, the $\exp(\boldsymbol{B}t) \to \mathbf{0}$, leading to the vanishing of the $\mathbf{V}^{(h)}(t)$. Hence the steady state solution for the ODE in (18) becomes $-\lambda_P \boldsymbol{B}^{-1}\boldsymbol{F}$.

This concludes the proof.

B.6. Proof of Proposition 1

We first show that $\boldsymbol{B}$ is a strictly diagonal dominant (SDD) matrix, i.e., $|\boldsymbol{B}(i,i)| > |\sum_{j\neq i}^{N} \boldsymbol{B}(i,j)|$, for $i, j = 1, \ldots, N$. In fact, $|\boldsymbol{B}(i,i)| = |\boldsymbol{K}(i,i) - \lambda_p - 1| > |1 - \boldsymbol{K}(i,i)| = |\sum_{j\neq i}^{N} \boldsymbol{K}(i,j)| = |\sum_{j\neq i}^{N} \boldsymbol{B}(i,j)|$ because $\boldsymbol{K}$ is a right-stochastic

matrix with all entries in $(0, 1]$ and sum of each row is 1.
Hence, following (Morača, 2007), the upper bound of $\|\boldsymbol{B}^{-1}\|_\infty$, when $\boldsymbol{B}$ is an SDD matrix, is given as

$$\|\boldsymbol{B}^{-1}\|_\infty \leq \frac{1}{\min\limits_{i\in N}(|\boldsymbol{B}(i,i)| - |\sum\limits_{j\neq i}^{N}\boldsymbol{B}(i,j)|)} \tag{32}$$

$$= \frac{1}{|\boldsymbol{K}(i,i) - \lambda_p - 1| - |1 - \boldsymbol{K}(i,i)|} = \frac{1}{\lambda_P}, \tag{33}$$

where $\|\boldsymbol{B}^{-1}\|_\infty = \max\limits_{i=1}^{N}\sum\limits_{j=1}^{N}|\boldsymbol{B}^{-1}(i,j)|$.

On the other hand,

$$\begin{aligned}\| - \lambda_P\beta\boldsymbol{B}^{-1}\boldsymbol{\epsilon}\|_\infty &\leq \lambda_P\beta\|\boldsymbol{B}^{-1}\|_\infty\|\boldsymbol{\epsilon}\|_\infty \\ &= \lambda_P\beta\frac{1}{\lambda_P}\bar{\epsilon} = \beta\bar{\epsilon}\end{aligned} \tag{34}$$

For the bounded error get arbitrarily small, we constraint $\beta\bar{\epsilon} \leq \delta$, making $\beta \leq \frac{\delta}{\bar{\epsilon}}$.
Here in the proof, we used the submultiplicity property of $\|.\|_\infty$ norm of matrices, which is proved as follow:

$$\begin{aligned}\|\boldsymbol{B}^{-1}\boldsymbol{\epsilon}\|_\infty = \sup_{\boldsymbol{x}}\frac{\|\boldsymbol{B}^{-1}\boldsymbol{\epsilon}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} &= \sup_{\boldsymbol{x}}\frac{\|\boldsymbol{B}^{-1}\boldsymbol{\epsilon}\boldsymbol{x}\|_\infty\|\boldsymbol{\epsilon}\boldsymbol{x}\|_\infty}{\|\boldsymbol{\epsilon}\boldsymbol{x}\|_\infty\|\boldsymbol{x}\|_\infty} \\ &\leq \sup_{\boldsymbol{x}}\frac{\|\boldsymbol{B}^{-1}\boldsymbol{\epsilon}\boldsymbol{x}\|_\infty}{\|\boldsymbol{\epsilon}\boldsymbol{x}\|_\infty}\sup_{\boldsymbol{x}}\frac{\|\boldsymbol{\epsilon}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} \\ &\leq \sup_{\boldsymbol{x}}\frac{\|\boldsymbol{B}^{-1}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty}\sup_{\boldsymbol{x}}\frac{\|\boldsymbol{\epsilon}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} \\ &= \|\boldsymbol{B}^{-1}\|_\infty\|\boldsymbol{\epsilon}\|_\infty\end{aligned}$$

With this, we conclude the proof of Proposition 1

B.7. Proof of Lemma 5

To find the solution of (21), firstly, we find the solution for the homogenous ODE:

$$\mathbf{V}^{(h)\prime}(t) = \frac{1}{1+\lambda_D}\big(\mathbf{K} - (\lambda_P+1)\boldsymbol{I}\big)\mathbf{V}^{(h)}(t)$$

From the result in Appendix B.1, the solution for this ODE is $\exp(\frac{1}{\lambda_D+1}\boldsymbol{B}t)\boldsymbol{C}$ where $\boldsymbol{B} = \mathbf{K} - (\lambda_P+1)\boldsymbol{I}$ and $\boldsymbol{C} \in \mathbb{R}^{N\times D_x}$ is any constant matrix. Secondly, we find a particular solution for (21) by solving $\mathbf{V}^{(p)\prime}(t) = \frac{1}{\lambda_D+1}(\boldsymbol{B}\mathbf{V}(t)^{(p)} + \lambda_P\boldsymbol{F}) = \mathbf{0}$. Since $\boldsymbol{B}$ is invertible, the solution for this equation is $\mathbf{V}^{(p)}(t) = -\lambda_P\boldsymbol{B}^{-1}\boldsymbol{F}$. The solution is $\mathbf{V}(t) = \mathbf{V}^{(h)}(t) + \mathbf{V}^{(p)}(t)$. Applying the initial condition, $\mathbf{V}(0) = \boldsymbol{C} - \lambda_P\boldsymbol{B}^{-1}\boldsymbol{F} = \mathbf{V}^0$, we find $\boldsymbol{C} = \mathbf{V}^0 + \lambda_P\boldsymbol{B}^{-1}\boldsymbol{F}$. Therefore, we have proved that the solution for the IVP problem in (21) is indeed $\mathbf{V}(t) = \exp(\frac{1}{\lambda_D+1}\boldsymbol{B}t)(\mathbf{V}^0 + \boldsymbol{B}^{-1}\boldsymbol{F}) - \lambda_P\boldsymbol{B}^{-1}\boldsymbol{F}$.

In Section 3.2.1, we show that $\boldsymbol{B}$ has only eigenvalues with negative real parts. As the result in Appendix B.1, when $t \to 0$, the $\exp(\frac{1}{\lambda_D+1}\boldsymbol{B}t) \to \mathbf{0}$, leading to the vanishing of the $\mathbf{V}^{(h)}(t)$. Hence the steady state solution for the ODE in (21) becomes $-\lambda_P\boldsymbol{B}^{-1}\boldsymbol{F}$. We have proved Lemma 5.

B.8. Proof of Proposition 2

Let

$$\boldsymbol{M} = \begin{bmatrix}\mathbf{0} & \boldsymbol{I} \\ -\frac{\lambda_I\boldsymbol{I}}{\lambda_D+1} & \frac{\boldsymbol{K} - (\lambda_P+1)\boldsymbol{I}}{\lambda_D+1}\end{bmatrix} \tag{35}$$

For the solution of (23) to be stable, the real part of eigenvalues of $\boldsymbol{M}$ must be all negative. Let $\boldsymbol{B} := \boldsymbol{K} - (\lambda_P + 1)\boldsymbol{I}$, for any eigenvalue γ of $\boldsymbol{M}$

$$\begin{aligned}\det(\boldsymbol{M} - \gamma\boldsymbol{I}) &= \det\left(\begin{bmatrix} -\gamma\boldsymbol{I} & \boldsymbol{I} \\ -\frac{\lambda_I}{\lambda_D+1}\boldsymbol{I} & \frac{1}{\lambda_D+1}(\boldsymbol{B} - \gamma\boldsymbol{I}) \end{bmatrix}\right) \\ &= \det\Big(\frac{1}{\lambda_D+1}(-\gamma\boldsymbol{B} + \gamma^2\boldsymbol{I} + \lambda_I\boldsymbol{I})\Big), \qquad \text{(since } \boldsymbol{B} - \gamma\boldsymbol{I} \text{ and } -\lambda_I\boldsymbol{I} \text{ commute, see (Silvester, 2000))} \\ &= 0\end{aligned} \tag{36}$$

Notice that $\gamma = 0$ is not a solution of (36). This fact is proved by contradiction. If $\gamma = 0$ is a solution, $\det(-\gamma\boldsymbol{B} + \gamma^2\boldsymbol{I} + \lambda_I\boldsymbol{I}) = \det(\lambda_I\boldsymbol{I}) = (\lambda_I)^N\det(\boldsymbol{I}) = (\lambda_I)^N > 0$ because $\lambda_I > 0$. This is contradict to (36). Since $\gamma \neq 0$, we can rewrite (36) as:

$$(-\frac{\gamma}{\lambda_D+1})^N\det(\boldsymbol{B} - (\gamma + \frac{\lambda_I}{\gamma})\boldsymbol{I}) = 0 \tag{37}$$

$$\iff \det(\boldsymbol{B} - (\gamma + \frac{\lambda_I}{\gamma})\boldsymbol{I}) = 0. \tag{38}$$

Therefore, $\gamma + \dfrac{\lambda_I}{\gamma}$ are eigenvalues of $\boldsymbol{B}$. Given κ_i, for $i = 1, \ldots, m; m \leq N$ are eigenvalues of $\boldsymbol{B}$. For each i, we find the solution of

$$\gamma_i + \frac{\lambda_I}{\gamma_i} = \kappa_i \tag{39}$$

$$\iff \gamma_i^2 - \kappa\gamma_i + \lambda_I = 0 \tag{40}$$

Let $\gamma_{i,1}, \gamma_{i,1}$ are the solution of (39), and then

$$\begin{cases} \gamma_{i,1} + \gamma_{i,2} = \kappa_i \\ \gamma_{i,1}\gamma_{i,2} = \lambda_I \end{cases} \iff \begin{cases} \mathrm{Re}(\gamma_{i,1}) + \mathrm{Re}(\gamma_{i,2}) = \mathrm{Re}(\kappa_i) \\ \mathrm{Im}(\gamma_{i,1}) + \mathrm{Im}(\gamma_{i,2}) = \mathrm{Im}(\kappa_i) \\ \mathrm{Re}(\gamma_{i,1})\mathrm{Re}(\gamma_{i,2}) - \mathrm{Im}(\gamma_{i,1})\mathrm{Im}(\gamma_{i,2}) = \lambda_I \\ \mathrm{Re}(\gamma_{i,1})\mathrm{Im}(\gamma_{i,2}) + \mathrm{Im}(\gamma_{i,1})\mathrm{Re}(\gamma_{i,2}) = 0 \end{cases} \tag{41}$$

In Section 3.2.1, we show that $\boldsymbol{B}$ has only eigenvalues with negative real parts. Hence, $\mathrm{Re}(\kappa_i) < 0$. Firstly, without any loss of generalization, suppose that $\mathrm{Re}(\gamma_{i,1}) = 0$. This means

$$\begin{cases} \mathrm{Re}(\gamma_{i,2}) = \mathrm{Re}(\kappa_i) < 0 \\ -\mathrm{Im}(\gamma_{i,1})\mathrm{Im}(\gamma_{i,2}) = \lambda_I \\ \mathrm{Im}(\gamma_{i,1})\mathrm{Re}(\gamma_{i,2}) = 0 \end{cases} \Rightarrow \begin{cases} \mathrm{Im}(\gamma_{i,1}) = 0 \\ -\mathrm{Im}(\gamma_{i,1})\mathrm{Im}(\gamma_{i,2}) = 0 \neq \lambda_I > 0 \end{cases} \tag{42}$$

which causes contradiction. Therefore, $\mathrm{Re}(\gamma_{i,1}) \neq 0$. As the result, $\mathrm{Im}(\gamma_{i,2}) = -\dfrac{\mathrm{Im}(\gamma_{i,1})\mathrm{Re}(\gamma_{i,2})}{\mathrm{Re}(\gamma_{i,1})}$, substituting to (41), we obtain

$$\mathrm{Re}(\gamma_{i,1})\mathrm{Re}(\gamma_{i,2}) = \lambda_I - \mathrm{Im}(\gamma_{i,1})^2\frac{\mathrm{Re}(\gamma_{i,2})}{\mathrm{Re}(\gamma_{i,1})}. \tag{43}$$

Suppose that $\mathrm{Re}(\gamma_{i,1})\mathrm{Re}(\gamma_{i,2}) < 0$, hence $\dfrac{\mathrm{Re}(\gamma_{i,2})}{\mathrm{Re}(\gamma_{i,1})} < 0$ leading to $-\mathrm{Im}(\gamma_{i,1})^2\dfrac{\mathrm{Re}(\gamma_{i,2})}{\mathrm{Re}(\gamma_{i,1})} > 0$, (because $\mathrm{Im}(\gamma_{i,1})^2 > 0$). Therefore the RHS of (43) is greater than 0 (since λ_I also greater than 0), which contradicts our assumption that $\mathrm{Re}(\gamma_{i,1})\mathrm{Re}(\gamma_{i,2}) < 0$. As a consequence, we obattain the following result:

$$\begin{cases} \mathrm{Re}(\gamma_{i,1}) + \mathrm{Re}(\gamma_{i,2}) = \mathrm{Re}(\kappa_i) < 0 \\ \mathrm{Re}(\gamma_{i,1})\mathrm{Re}(\gamma_{i,2}) > 0 \end{cases} \iff \begin{cases} \mathrm{Re}(\gamma_{i,1}) < 0 \\ \mathrm{Re}(\gamma_{i,2}) < 0, \end{cases} \tag{44}$$

for $i = 1, \ldots, m$. Therefore, all eigenvalues of $\boldsymbol{M}$ as negative real parts. Combined with result in Appendix B.1, we have the system described by (23) has stable solution when $t \to 0$, for all $\lambda_P, \lambda_I, \lambda_D > 0$. This concludes our proof.

B.9. The Fretchet derivation of the derivative of J w.r.t v_j.

The partial derivative $\partial J/\partial v_j$, $j = 1, 2, \ldots, D$, is defined through its dot product with an arbitrary function $h_j \in L^2(\Omega \times [0, \infty))$ as follows

$$\begin{aligned}
\frac{\partial J}{\partial v_j} \cdot h_j(x,t) &= \frac{d}{d\tau} J(v_j + \tau h_j)\big|_{\tau=0} \\
&= \frac{1}{2}\left(\frac{d}{d\tau}\int_{\Omega\times\Omega}(v_j(x) - v_j(y) + \tau h_j(x) - \tau h_j(y))^2 k(x,y)dxdy\right)\Big|_{\tau=0} \\
&= \left(\int_{\Omega\times\Omega}(v_j(x,t) - v_j(y) + \tau h_j(x) - \tau h_j(y,t))(h_j(x) - h_j(y))k(x,y)dxdy\right)\Big|_{\tau=0} \\
&= \int_{\Omega\times\Omega}(v_j(x) - v_j(y))(h_j(x) - h_j(y))k(x,y)dxdy \\
&= \int_{\Omega\times\Omega}(v_j(x) - v_j(y))h_j(x)k(x,y)dxdy - \int_{\Omega\times\Omega}(v_j(x) - v_j(y))h_j(y)k(x,y)dxdy
\end{aligned}$$

Applying a change of variables $(x, y) \to (y, x)$ to the second term of the above integral, we have

$$\begin{aligned}
\frac{\partial J}{\partial v_j} \cdot h_j(x) &= \int_{\Omega\times\Omega}(v_j(x) - v_j(y))h_j(x)k(x,y)dxdy - \int_{\Omega\times\Omega}(v_j(y) - v_j(x))h_j(x,t)k(y,x)dxdy \\
&= \int_{\Omega\times\Omega}(v_j(x) - v_j(y)(k(x,y) + k(y,x))dy h_j(x)dx
\end{aligned}$$

Thus, the Frechet derivative of J with respect to v_j is given by

$$\frac{\partial J}{\partial v_j} = \int_\Omega (v_j(x) - v_j(y)(k(x,y) + k(y,x))dy.$$

B.10. The derivation of the gradient flow of $E(v, f)$

Taking the gradient of $E(\boldsymbol{v}, \boldsymbol{f})$ with respect to $\boldsymbol{v}$, we obtain

$$\nabla_{\boldsymbol{v}} E = \nabla_{\boldsymbol{v}} J + \left[\frac{\partial G}{\partial u_1}, \frac{\partial G}{\partial u_2}, \ldots, \frac{\partial G}{\partial u_D}\right]^T. \tag{45}$$

The partial derivative $\partial G/\partial v_j$, $j = 1, 2, \ldots, D$, is defined through its dot product with an arbitrary function $h_j \in L^2(\Omega)$ as follows

$$\begin{aligned}
\frac{\partial G}{\partial v_j} \cdot h_j(x) &= \frac{d}{d\tau} G(v_j + \tau h_j)\big|_{\tau=0} \\
&= \frac{\lambda}{2}\left(\frac{d}{d\tau}\int_\Omega (v_j(x) - f_j(x) + \tau h_j(x))^2 dx\right)\Big|_{\tau=0} \\
&= \lambda \int_\Omega (v_j(x) - f_j(x))h_j(x)dx.
\end{aligned}$$

Thus, the Frechet derivative of F with respect to v_j is given by

$$\frac{\partial G}{\partial v_j} = \lambda(v_j(x) - f_j(x)) \tag{46}$$

Substituting the formula for $\partial G/\partial v_j$ in (46) into (45) for $\nabla_{\boldsymbol{v}} E(\boldsymbol{v}, \boldsymbol{f})$, we obtain the following gradient flow

$$\frac{d\boldsymbol{v}(x,t)}{dt} = -\nabla_{\boldsymbol{v}} E(\boldsymbol{v}, \boldsymbol{f}) = -\nabla_{\boldsymbol{v}} J(\boldsymbol{v})(x) + \lambda\big(\boldsymbol{f}(x) - \boldsymbol{v}(x)\big). \tag{47}$$

This concludes the derivation.

B.11. The derivation of (15)

Denote $H(\boldsymbol{v}, \boldsymbol{f}) := \frac{\lambda}{2}\int_\Omega \|\boldsymbol{v}(x) - \boldsymbol{f}(x) - \boldsymbol{e}^\ell(x)\|_2^2 dx$. Taking the gradient of $J(\boldsymbol{v}) + H(\boldsymbol{v}, \boldsymbol{f})$ with respect to $\boldsymbol{v}$, we obtain

$$\nabla_{\boldsymbol{v}} E = \nabla_{\boldsymbol{v}} J + \left[\frac{\partial H}{\partial v_1}, \frac{\partial H}{\partial v_2}, \ldots, \frac{\partial H}{\partial v_D}\right]^T. \tag{48}$$

The partial derivative $\partial H/\partial v_j$, $j = 1, 2, \ldots, D$, is defined through its dot product with an arbitrary function $h_j \in L^2(\Omega)$ as follows

$$\begin{aligned}
\frac{\partial H}{\partial v_j} \cdot h_j(x) &= \frac{d}{d\tau} H(v_j + \tau h_j)\Big|_{\tau=0} \\
&= \frac{\lambda}{2}\left(\frac{d}{d\tau}\int_\Omega (v_j(x) - f_j(x) - e_j^\ell(x) + \tau h_j(x))^2 dx\right)\Bigg|_{\tau=0} \\
&= \lambda \int_\Omega (v_j(x) - f_j(x) - e_j^\ell) h_j(x) dx.
\end{aligned}$$

Thus, the Frechet derivative of F with respect to v_j is given by

$$\frac{\partial H}{\partial v_j} = \lambda(v_j(x) - f_j(x) - e_j^\ell) \tag{49}$$

Substituting the formula for $\partial H/\partial v_j$ in (49) into (48) for $\nabla_{\boldsymbol{v}} E(\boldsymbol{v}, \boldsymbol{f})$, we obtain the following gradient flow at iteration $\ell + 1$

$$\begin{aligned}
\frac{d\boldsymbol{v}(x,t)}{dt} = &\int_\Omega \big(\boldsymbol{v}(y,t) - \boldsymbol{v}(x,t)\big)\big(k(x,y) + k(y,x)\big) dy \\
&+ \lambda\big(\boldsymbol{f}(x) - \boldsymbol{v}(x,t) + \boldsymbol{e}^\ell(x)\big).
\end{aligned} \tag{50}$$

Applying Euler method to discretize (50) with $\Delta t = 1$ and $\boldsymbol{v}(x, 0) = \boldsymbol{v}^\ell(x)$, we approximate the $\boldsymbol{v}^{\ell+1}$ with one-step gradient descent:

$$\begin{aligned}
\boldsymbol{v}^{\ell+1}(x) = &\int_\Omega \big(\boldsymbol{v}^\ell(y) - \boldsymbol{v}^\ell(x)\big)\big(k(x,y) + k(y,x)\big) dy \\
&+ \boldsymbol{v}^\ell(x) + \lambda \boldsymbol{e}^\ell(x) + \lambda \boldsymbol{e}_{\mathrm{a}}^\ell(x).
\end{aligned}$$

This concludes the derivation.

C. Additional Experiment results

C.1. PID DeiT and softmax DeiT under escalating perturbation attacks.

We evaluate PID DeiT and softmax DeiT under FGSM and PGD attack methods with increasing perturbation budgets (see Fig. 3) (scaled by 255). The proposed PID DeiT exhibits stronger defense in both attack methods and various perturbation budgets.

C.2. Combine PIDformer with other defense model

To further demonstrate the advantages of PID control in enhancing model robustness, we employ the Fully Attention Network (FAN) (Zhou et al., 2022), a state-of-the-art robust vision transformer, as a baseline, detailed in Table 4 below. Our experiments illustrate the significant increase in model robustness against various adversarial attacks and out-of-distribution datasets when our PID control is integrated with the FAN baseline. We use the publicly available code https://github.com/NVlabs/FAN for the implementation and model configuration.

C.3. PIDformer with different hyperparameters

We have performed an extensive study of hyperparameters on the ADE20K image segmentation task (see Table 3), where we investigate the impact of different settings for the PIDformer's hyperparameters, i.e., λ_P, λ_I, λ_D, and β. In this study,

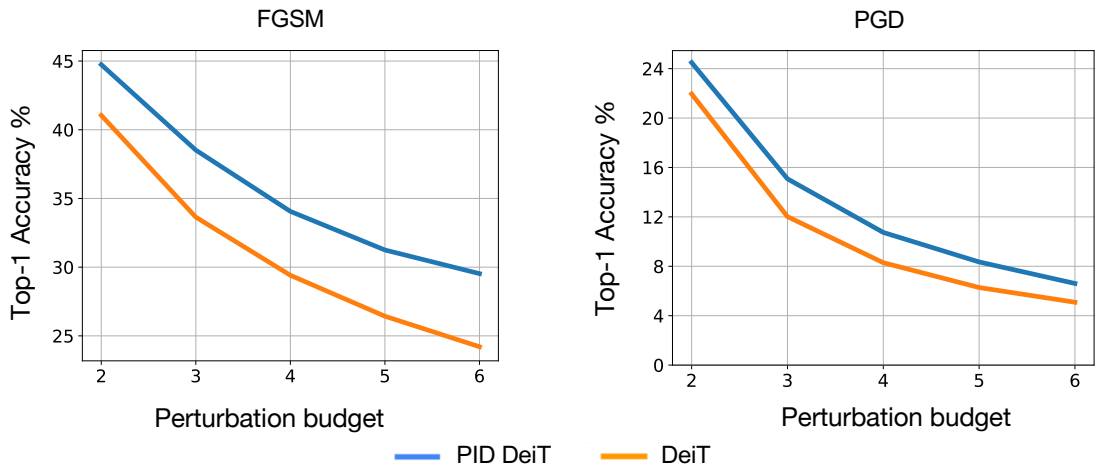

Figure 3. The top-1 classification accuracy curves on ImageNet against FGSM and PGD attack methods, plotted against perturbation budgets (scaled by 255).

Table 4. Evaluation of PID FAN versus FAN on the clean ImageNet validation set, as well as under various adversarial attacks and out-of-distribution datasets.

Attack	Metric/Model	*FAN*	PID FAN
Clean	Top-1 Acc (%)	77.11	**77.40**
	Top-5 Acc (%)	93.71	**93.85**
FGSM	Top-1 Acc (%)	38.27	**39.61**
	Top-5 Acc (%)	71.62	**73.74**
PGD	Top-1 Acc (%)	12.87	**15.5**
	Top-5 Acc (%)	29.16	**34.64**
SLD	Top-1 Acc (%)	75.6	**76.1**
	Top-5 Acc (%)	93.56	**93.68**
Noise	Top-1 Acc (%)	75.2	**75.9**
	Top-5 Acc (%)	92.52	**92.71**
Imagenet-A	Top-1 Acc (%)	13.96	**15.65**
Imagenet-R	Top-1 Acc (%)	41.45	**42.95**
Imagenet-C	mCE (↓)	60.06	**58.66**
Imagenet-O	AUPR	18.46	**19.67**

λ_P is in $[0.2, 0.5, 0.8]$, λ_I is in $[0.3, 0.6, 0.9]$, λ_D from $[0.05, 0.3, 0.6]$, and β from $[0.3, 0.6, 1.0]$. The evaluation reports model performance on clean/attacked datasets to assess robustness under various conditions. Our findings indicate that the model's performance is generally stable across a wide range of hyperparameter settings, suggesting that PIDformer is not overly sensitive to specific parameter values. However, we note an exception with higher values of λ_D, where performance as increasing λ_D from 0.05 to 0.3 then to 0.6 decrease the performance of the model. This insight suggests that there is some sensitivity to the derivative term (D), which could guide practitioners in tuning PIDformer for specific applications.

C.4. Compare with other baselines

In order to further illustrate the benefits of our model, we compare our model with FeatScale, a state-of-the-art model designed to address oversmoothing in vision transformers. The new results in Table 6 reveal that PIDformer significantly outperforms FeatScale (Wang et al., 2022). Furthermore, when combining PIDformer with FeatScale, we observe substantial improvements compared to DeiT plus FeatScale, underscoring our approach's compatibility and additive benefits when integrated with existing methods targeting oversmoothing.

Table 5. Single-scale (SS) and multi-scale (MS) DeiT and PID DeiT (different hyperparameters) under clean data and FGMS-attacked data on the ADE20K image segmentation

Model	λ_P	λ_I	λ_D	β	**Clean data**		**FGMS**	
					SS	**MS**	**SS**	**MS**
Softmax	0	0	0	0	35.72	36.68	27.26	32.27
PID DeiT	0.5	0.3	0.05	1.0	**37.42**	**38.28**	28.7	33.87
	0.8	0.3	0.05	1.0	36.56	37.37	28.0	33.68
	0.2	0.3	0.05	1.0	35.77	36.63	28.01	33.23
	0.5	0.6	0.05	1.0	36.72	37.77	29.01	33.85
	0.5	0.9	0.05	1.0	36.09	36.82	27.99	32.69
	0.5	0.3	0.3	1.0	36.01	36.98	28.11	32.83
	0.5	0.3	0.6	1.0	34.27	35.17	27.61	32.54
	0.5	0.3	0.05	0.3	37.19	38.17	**29.11**	**34.81**
	0.5	0.3	0.05	0.6	37.06	37.92	28.13	32.63

Table 6. We compare PID DeiT with DeiT combined with FeatScale (Wang et al., 2022) and incorporate our method with FeatScale model.

Model/Metric	Top-1 Acc (%)	Top-5 Acc (%)
PID DeiT	**73.13**	**91.76**
DeiT + FeatScale	72.346	91.22
PID DeiT + FeatScale	72.93	91.55