Subhankar Mishra Lab Weekly Talks

PIDformer: Transformer Meets Control Theory

PANKAJ KUMAR

SUBHANKAR MISHRA LAB

Oct 07, 2024

Outline / Table of Contents

Rank Collapse

- Effective rank of the representation matrices in a model diminishes
- Limits capacity to capture diverse patterns and relationships
- Cause
  - Over Parametrization
  - Redundancy in Representations

Adversarial Attacks

- Exploits non-robustness in model
- We use FGSM, PGD, SPSA, SLD to check the robustness of model

Approach

- Linking with controlled SSM (Novel Control Framework)
- PIDformer
  - Enhancing robustness and addressing rank collapse
  - Improves response time
- Energy Optimization
  - Connection with controlled SSMs further enhances the understanding of models
- Proved that softmax self-attention is noise-sensitive and leads to low-rank outputs

SSM and Self-attention

The attention mechanism computes the output of token i at ℓ-th layer as follows

$$\boxed{\mathbf{u}^{\ell}(i) = \sum_{j=1}^{N} \text{softmax}\left(\frac{\mathbf{q}^{\ell}(i)^{\top}\mathbf{k}^{\ell}(j)}{\sqrt{D_{qk}}}\right)\mathbf{v}^{\ell}(j)}$$

SSM approaches it by assuming the value matrix $\mathbf{V}^{\ell}$ discretizes the function v(x,t) as

$$\frac{\mathrm{d}\mathbf{v}(x,t)}{\mathrm{d}t} = \int_{\Omega} (\mathbf{v}(y,t) - \mathbf{v}(x,t))K(x,y,t)\,dy + \mathbf{z}(x,t)$$

$$\mathbf{v}(x,0) = \mathbf{v}_0(x), \quad \mathbf{z}(x,t) = 0, \quad \forall x \in \Omega, \forall t \geq 0$$

SSM contd.

By using a proximity kernel K,

$$K(x, y, t) := \frac{\exp\left(\frac{\mathbf{q}(x,t)^\top \mathbf{k}(y,t)}{\sqrt{D_{qk}}}\right)}{\int_\Omega \exp\left(\frac{\mathbf{q}(x,t)^\top \mathbf{k}(y',t)}{\sqrt{D_{qk}}}\right) dy'}$$

And, applying the Euler method to discretize it with the time step $\boldsymbol{\Delta} t(x) = 1$

$$\mathbf{v}(x, t+1) \approx \int_\Omega \frac{\exp\left(\frac{\mathbf{q}(x,t)^\top \mathbf{k}(y,t)}{\sqrt{D_{qk}}}\right)}{\int_\Omega \exp\left(\frac{\mathbf{q}(x,t)^\top \mathbf{k}(y',t)}{\sqrt{D_{qk}}}\right) dy'} \mathbf{v}(y', t)\, dy.$$

PIDformer

- Proportional (P)
  - scales the control input proportionally to the error
  - Larger the errors, stronger the control inputs
- Integral (I)
  - accumulates past errors, ensuring that even small, persistent errors are addressed over time
  - improves long term stability
- Derivative (D)
  - considers the rate of error change, allowing the controller to anticipate future deviations
  - improves model responsiveness to sudden changes

Workflow of PIDformer

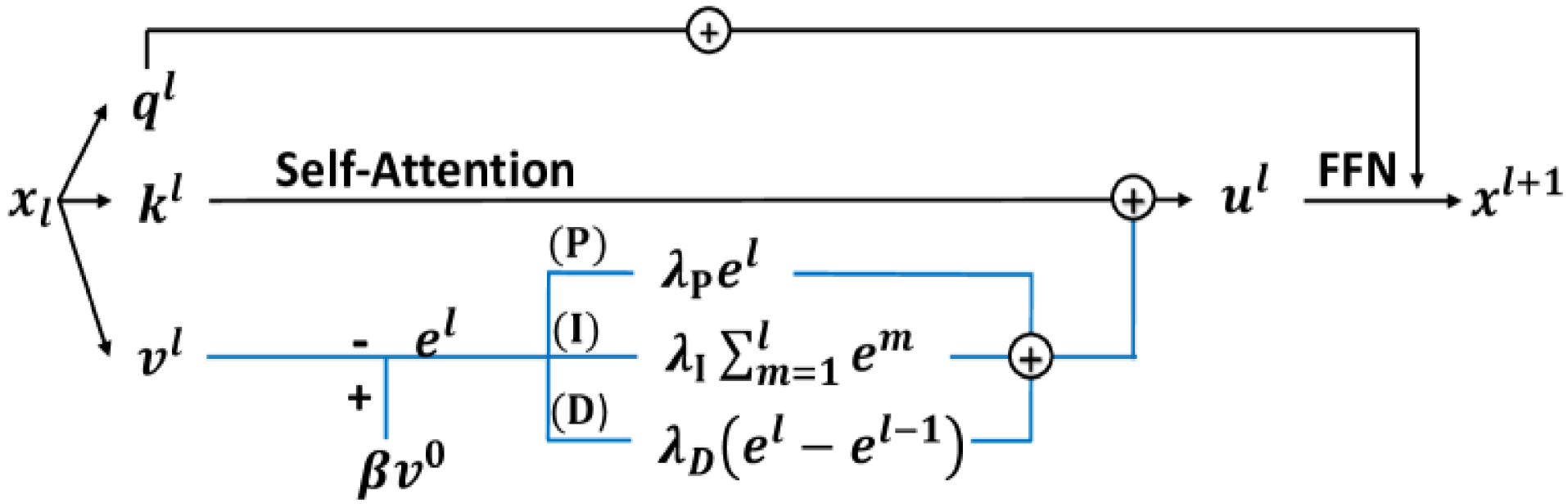
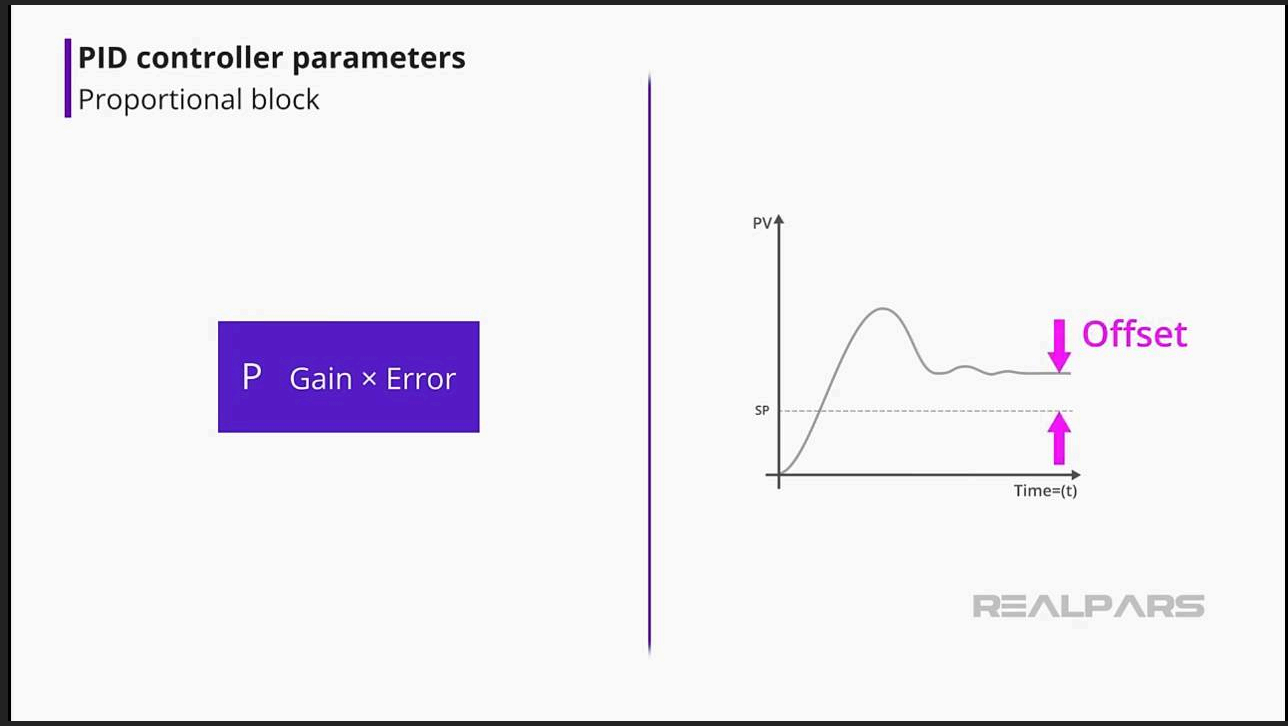


Figure 1: Our proposed PIDformer model at each layer.

PID-Controller

P-Component

- Adding the P-term alone makes solution independent of the initial input (V_0) for a fixed reference function
- This robustness can be further controlled by adjusting the scaling factor (β) of the reference function
- need for careful tuning of λ_P



PID-Controller

I-Component

- tackles long term information loss
- useful when loss accumulates slowly over layers

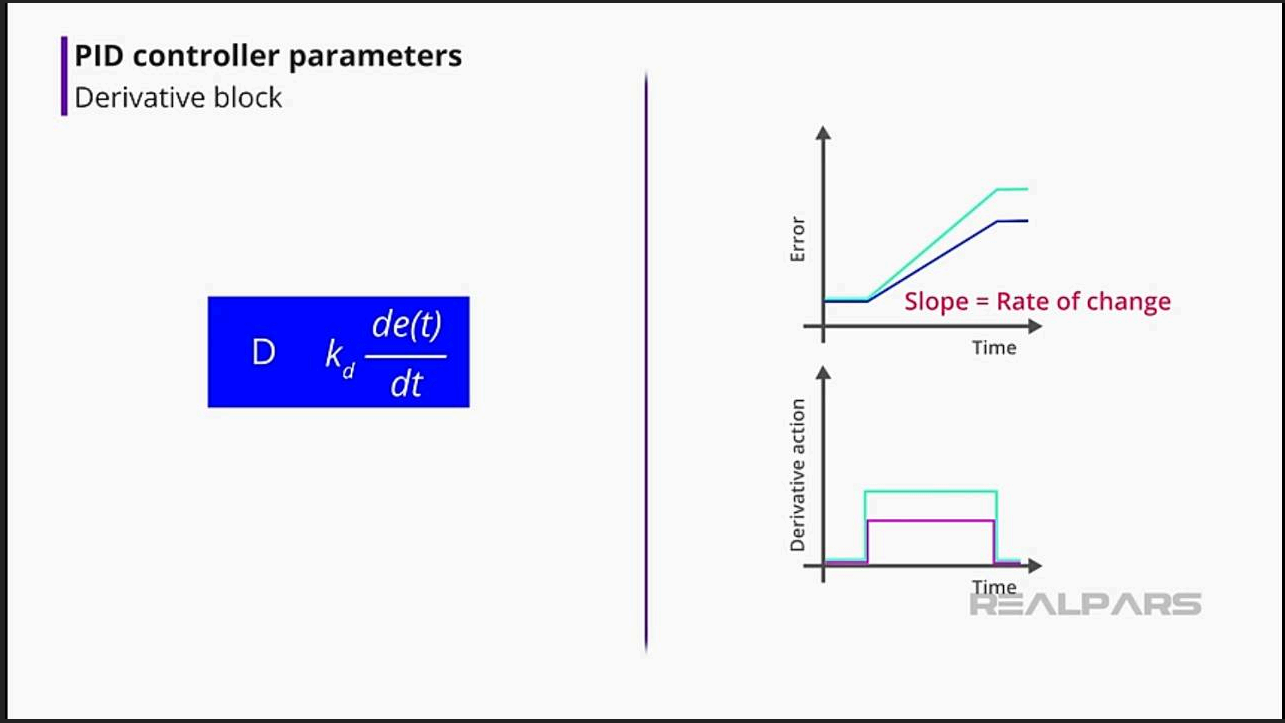
D-component

- rapid increase in rate $\Longrightarrow$ potential instability in future
- which triggers stronger control signal

PID-Controller

PD Control

- Improvements
  - Adding the D-term doesn't change solution achieved with P-control
  - It stabilizes the system by damping rapid fluctuation
  - maintains bounded error
- sensitive to noise
  - Higher λ_D leads to instability



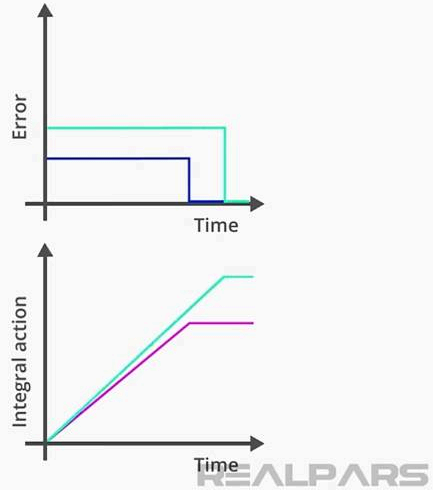
PID-Controller

PID Control

- ensures robustness against input corruptions
- resolves sensitivity issue

PID controller parameters
Integral block

$$I \quad k_i \int_0^t e(t)dt$$



PIDformer as discretization of PID-control SSM

$$\frac{\mathrm{d}\mathbf{v}(x,t)}{\mathrm{d}t} = \int_{\Omega} (\mathbf{v}(y,t) - \mathbf{v}(x,t))K(x,y,t)\,dy + \mathbf{z}(x,t)$$

$$\mathbf{z}(x,t) = \lambda_P \mathbf{e}(x,t) + \lambda_I \int_0^t \mathbf{e}(x,t)\,dt + \lambda_D \frac{\mathrm{d}\mathbf{e}(x,t)}{\mathrm{d}t}$$

$$\mathbf{v}(x,0) = \mathbf{v}^0(x), \quad \mathbf{z}(x,0) = 0.$$

Note: SSM implicitly performs gradient descent to minimize nonlocal variation J(v), resulting in signal information loss.

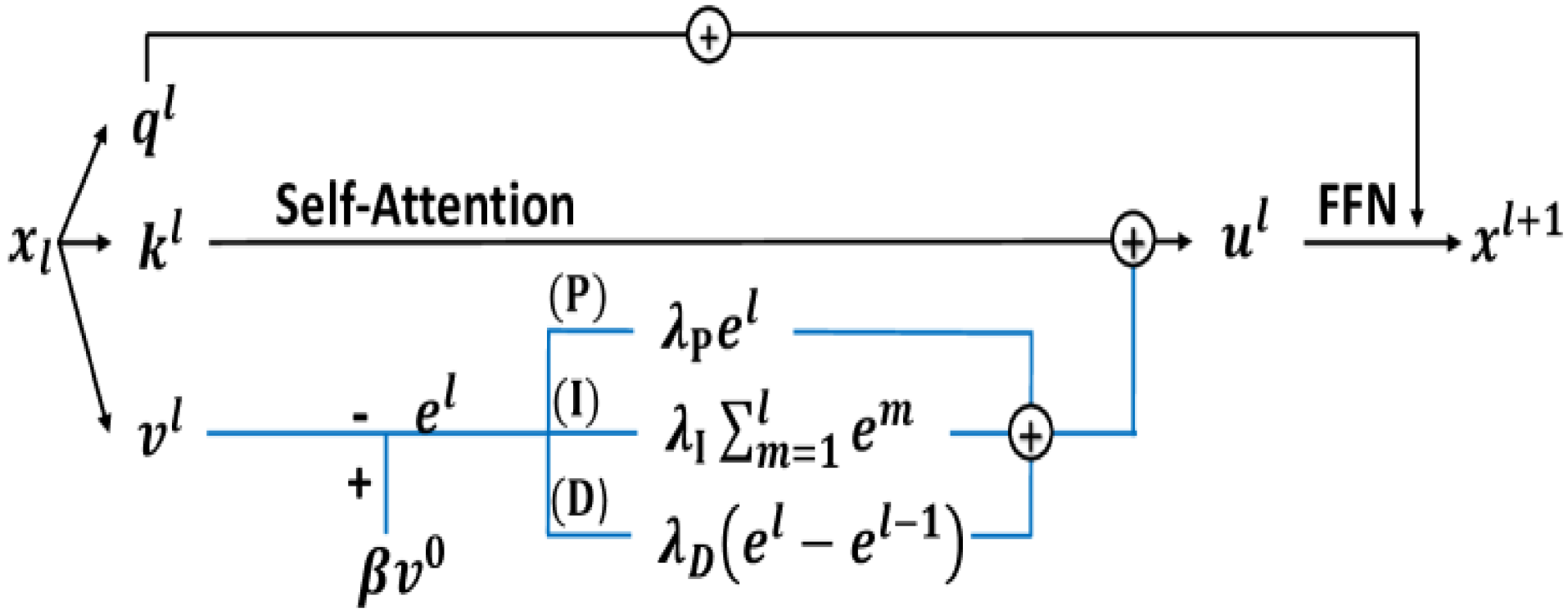


Figure 1: Our proposed PIDformer model at each layer.

Optimization step

$$E(\mathbf{v}, \mathbf{f}) = J(\mathbf{v}) + G(\mathbf{v}, \mathbf{f}) : (\text{with } \lambda_D = 0)$$

$$= \frac{1}{2} \int_\Omega \|\mathbf{v}(x) - \mathbf{v}(y)\|_2^2 k(x, y)\, dx\, dy + \frac{\lambda}{2} \int_\Omega \|\mathbf{v}(x) - \mathbf{f}(x)\|_2^2\, dx$$

Using gradient descent and Frechet derivative with respect to v_j we have:

$$\frac{\mathrm{d}\mathbf{v}(x,t)}{\mathrm{d}t} = -\nabla_{\mathbf{v}} E(\mathbf{v}, \mathbf{f}) = -\nabla_{\mathbf{v}} J(\mathbf{v})(x) + \lambda(\mathbf{f}(x) - \mathbf{v}(x))$$

$$= \int_\Omega (\mathbf{v}(y,t) - \mathbf{v}(x,t))\,(k(x,y) + k(y,x))dy + \lambda\,(\mathbf{f}(x) - \mathbf{v}(x,t))$$

Using Bregman Iteration, this PI controlled SSM minimises $E(v, f)$

Transformer with PID control

The update step of the PID-controlled SSM becomes:

$$\mathbf{v}^{\ell+1}(x) \approx \int_{\Omega} \left(\mathbf{v}^{\ell}(y) - \mathbf{v}^{\ell}(x)\right) \frac{\exp\left(\frac{q^{\ell}(x)^{\top} k^{\ell}(y)}{\sqrt{D_{qk}}}\right)}{\int_{\Omega} \exp\left(\frac{q^{\ell}(x)^{\top} k^{\ell}(y')}{\sqrt{D_{qk}}}\right) dy'} dy$$

$$+\mathbf{v}^{\ell}(x) + \lambda_P \cdot \mathbf{e}^{\ell}(x) + \lambda_I \sum_{m=1}^{\ell} \mathbf{e}^{m}(x) + \lambda_D \left(\mathbf{e}^{\ell}(x) - \mathbf{e}^{\ell-1}(x)\right)$$

where

$$\mathbf{e}^{m}(x) = \mathbf{f}(x) - \mathbf{v}^{m}(x) \quad \text{for } m = 1, \ldots, \ell$$

PIDformer

PIDformer computes the corresponding output vector $u^{\ell}(i)$ of the query $q^{\ell}(i)$ by the following attention formula:

$$\mathbf{u}_{\ell}(i) = \sum_{j=1}^{N} \text{softmax}\left(\frac{\mathbf{q}^{\ell}(i)^{\top}\mathbf{k}^{\ell}(j)}{\sqrt{D_{qk}}}\right) \cdot \mathbf{v}^{\ell}(y) + \lambda_P \mathbf{e}^{\ell}(i) + \lambda_I \sum_{m=1}^{\ell} \mathbf{e}^{m}(i) + \lambda_D \left(\mathbf{e}^{\ell}(i) - \mathbf{e}^{\ell-1}(i)\right),$$

where

$$\mathbf{e}^{\ell} = \mathbf{v}^{0} - \mathbf{v}^{\ell}, \quad \mathbf{v}^{0}(1), \ldots, \mathbf{v}^{0}(N) \in \mathbb{R}^{D}$$

are the value vectors in the first layer of PIDformer.

Experimental Results

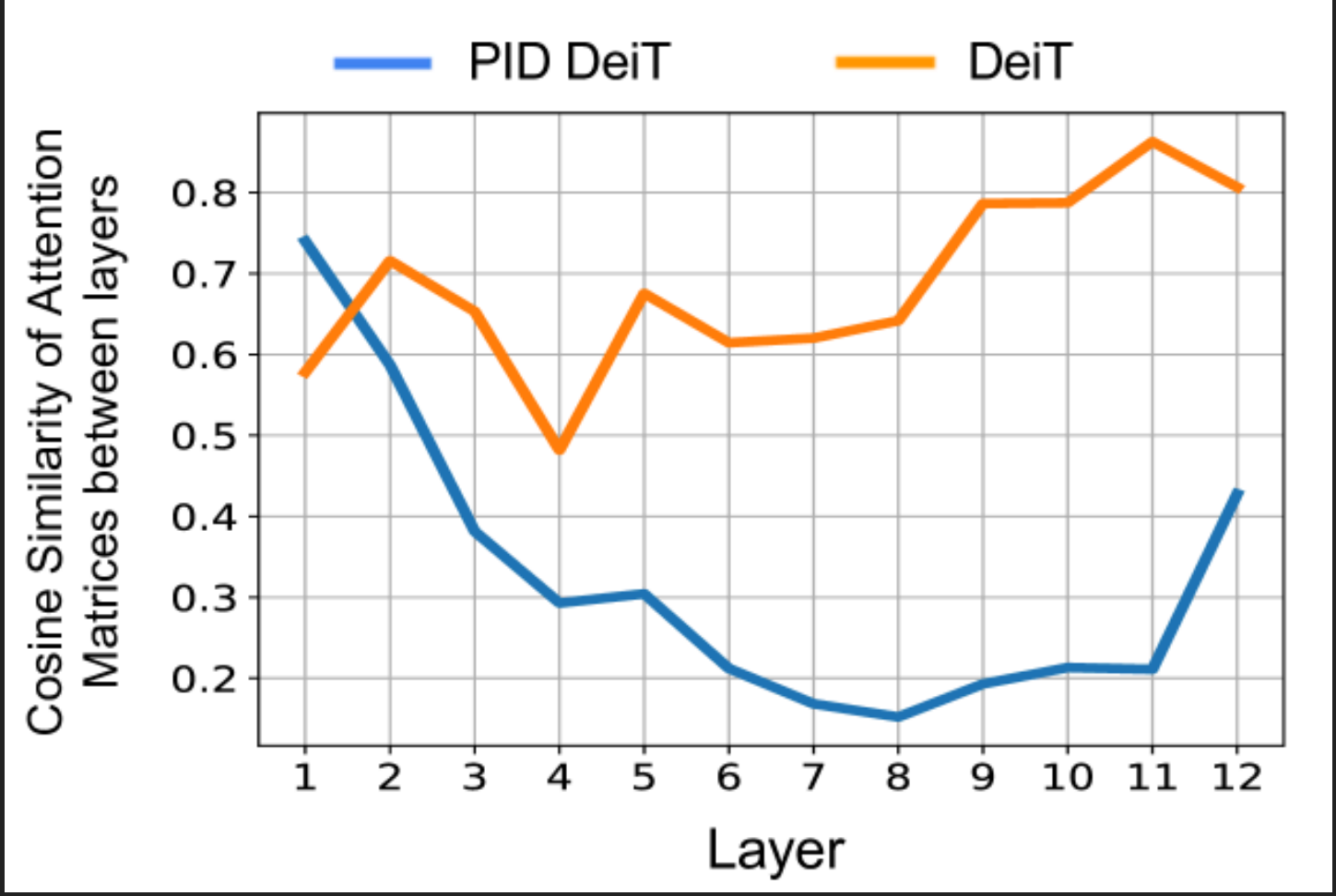
Model/Metric	SS MIoU	MS MIoU(%)
Softmax DeiT	35.72	36.68
PID DeiT	37.42	38.28

Model/Metric	SS MIoU	MS MIoU(%)
Softmax Transformer	33.15	34.29
PIDformer	32.44	33.45

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**

Experimental Results



Limitations

- System Complexity
  - need to manage interactions between different components
- Hyperparameter Tuning
  - tuning $\lambda_P, \lambda_D, \lambda_I, \beta$ can take a longer time than expected
- Privacy Preservation
  - potential of controlled transformers is not discussed
- Limited Experiments on ViT
  - Exploring fields other than Vision Transformers can uncover different challenges or advantages of model

References

- Main papers:
  1. Nguyen et al; PIDformer: Transformer Meets Control Theory
  2. https://openreview.net/pdf?id=3fd776zKmo

Image sources

1. https://www.realpars.com/blog/pid-controller
2. Nguyen et al; PIDformer: Transformer Meets Control Theory

Subhankar Mishra Lab Weekly Talks



SML