Griffith UNIVERSITY
Queensland, Australia

SATA: Spatial Autocorrelation Token Analysis for Enhancing the Robustness of Vision Transformers

Nick Nikzad, Yi Liao, Yongsheng Gao, Jun Zhou

Griffith UNIVERSITY
Queensland, Australia

Current ViTs Robustness improvement limitations

- limited success
- training strategies, input augmentation
- network structural enhancements
- involve extensive training and fine-tuning

Spatial Autocorrelation: Moran's metric

- Assessing the spatial interdependence of entities based on their **locations** and **values**

Let $\mathbf{X}$ be a set of N observations (here, tokens) presented by embedded vectors $\mathbf{x}_i \in \mathbb{R}^d$, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]$, and an associated attribute, $\mathbf{a} = [a_1, a_2, ..., a_N]$, the local Moran's I metric can be defined as:

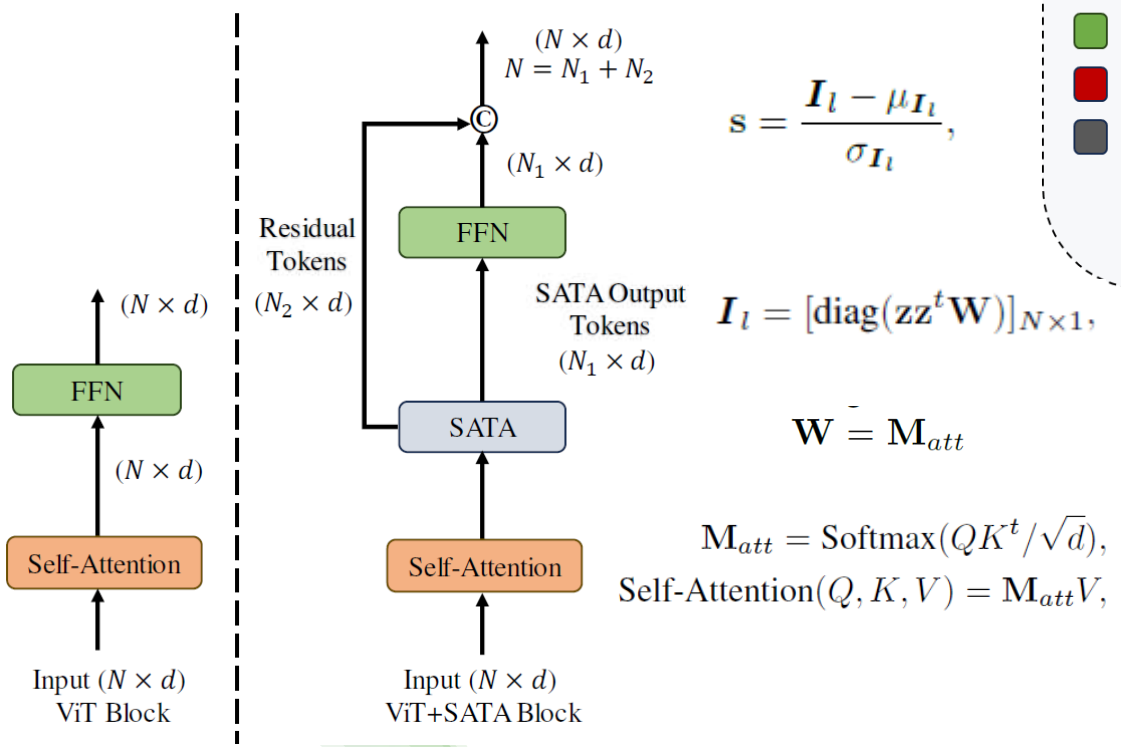
$$\boldsymbol{I}_l = [\text{diag}(\mathbf{z}\mathbf{z}^t\mathbf{W})]_{N\times 1}, \tag{3}$$

$$\mathbf{z} = \frac{\mathbf{a} - \mu}{\sigma}, \qquad \mathbf{a} = \left[a_i = \frac{1}{d}\sum_{t=1}^{d} \mathbf{x}_i(t) \right]_{N\times 1},$$

Nick Nikzad, Yongsheng Gao, and Jun Zhou. CSA-Net: Channel-wise Spatially Autocorrelated Attention Networks. *arXiv preprint arXiv:2405.05755*, 2024.

Griffith UNIVERSITY
Queensland, Australia

SATA

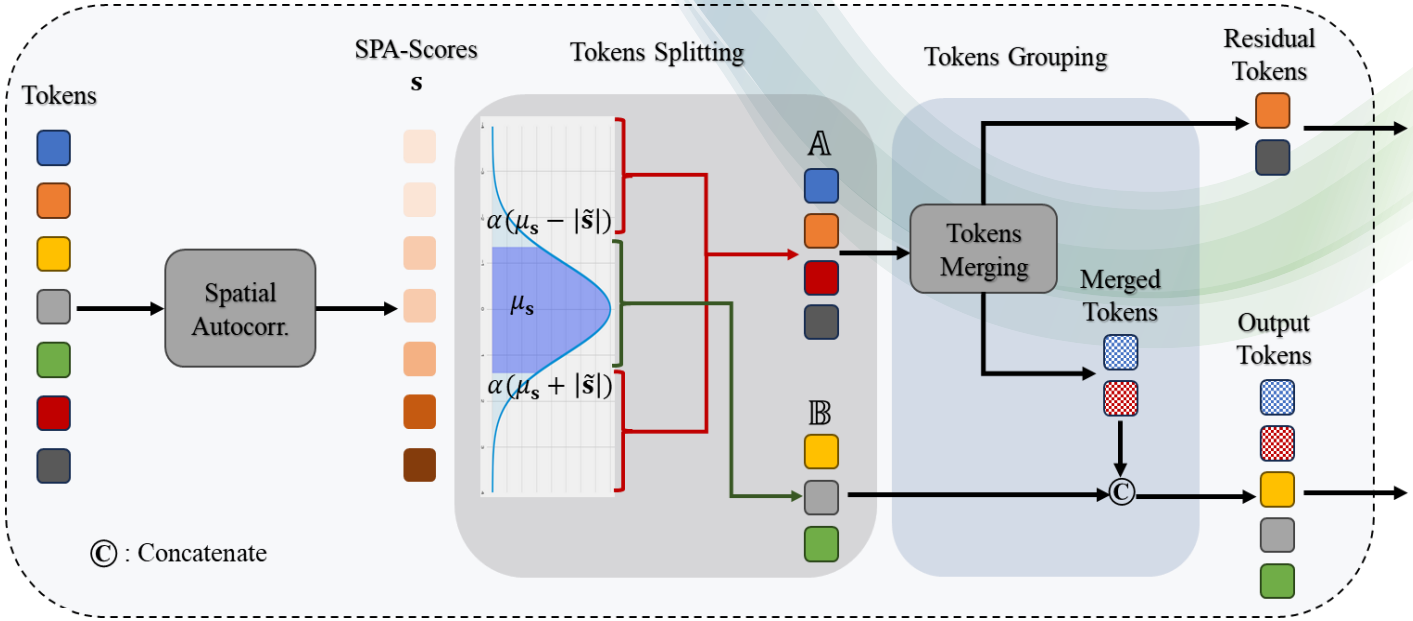


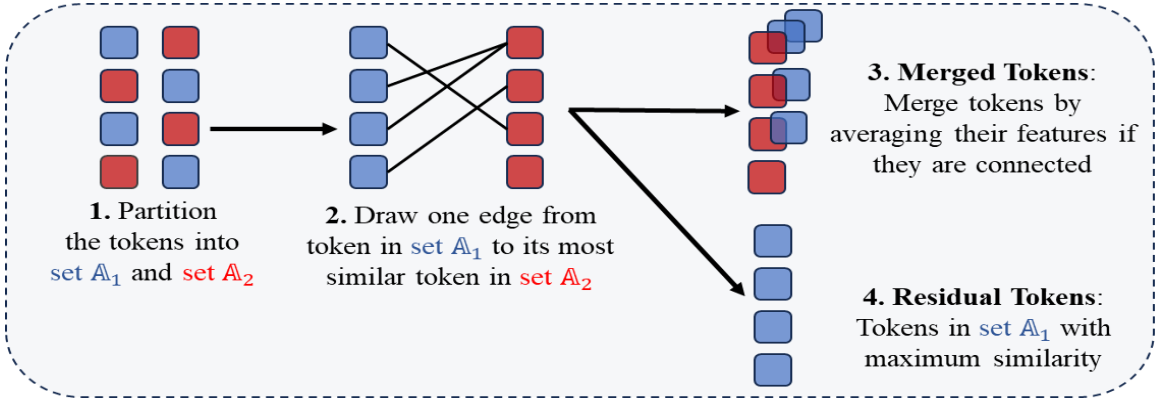
$$\mathbf{s} = \frac{\boldsymbol{I}_l - \mu_{\boldsymbol{I}_l}}{\sigma_{\boldsymbol{I}_l}},$$

$$\boldsymbol{I}_l = [\mathrm{diag}(\mathbf{z}\mathbf{z}^t\mathbf{W})]_{N\times 1},$$

$$\mathbf{W} = \mathbf{M}_{att}$$

$$\mathbf{M}_{att} = \mathrm{Softmax}(QK^t/\sqrt{d}),$$
$$\text{Self-Attention}(Q,K,V) = \mathbf{M}_{att}V,$$





Griffith UNIVERSITY
Queensland, Australia

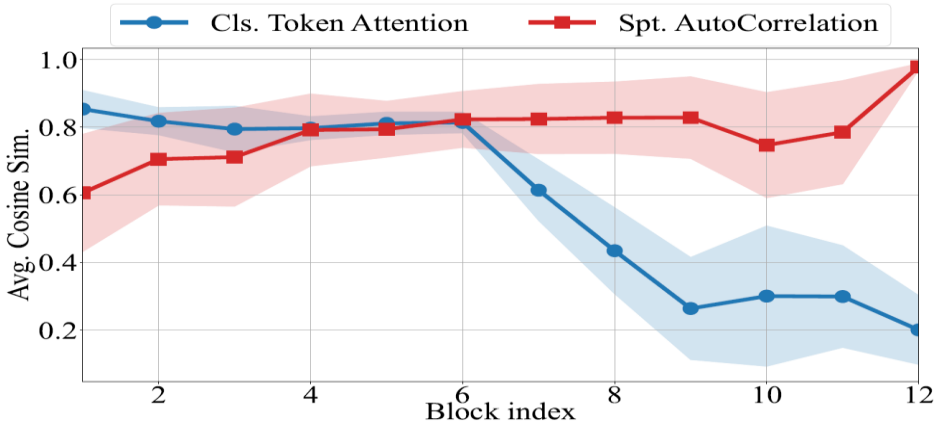
New STOA on ViTs Robustness Performance

*all without any additional **training or fine-tuning** of baseline models.

Group	Model	FLOPs (G)	Params (M)	ImageNet-1K Top-1	ImageNet-1K Top-5	Robustness Benchmarks FGSM	PGD	IN-C (mCE↓)	IN-A	IN-R	IN-SK
ViT-Tiny	DeiT-Ti[46]	1.3	5.7	72.2	91.1	22.3	6.2	71.1	7.3	32.6	20.2
	ConvViT-Ti[10]	1.4	5.7	73.3	91.8	24.7	7.5	68.4	8.9	35.2	22.4
	PiT-Ti[23]	0.7	4.9	72.9	91.3	20.4	5.1	69.1	6.2	34.6	21.6
	PVT-Tiny[50]	1.9	13.2	75.0	92.5	10.0	0.5	79.6	7.9	33.9	21.5
	RVT-Ti [32]	1.3	8.6	78.4	94.2	34.8	11.7	58.2	13.3	43.7	**30.0**
	FAN-T-ViT [57]	1.3	7.0	79.2	-	-	-	57.5	-	42.5	-
	RVT-Ti+RSPC [16]	1.3	10.9	79.2	-	-	-	55.7	**16.5**	-	-
	SATA-T (ours)	1.0	5.7	**86.5**	**98.2**	**40.0**	10.9	**51.1**	14.6	**47.3**	25.2
ViT-Small	DeiT-S[46]	4.6	22.1	79.9	95.0	40.7	16.7	54.6	18.9	42.2	29.4
	ConvViT-S[10]	5.4	27.8	81.5	95.8	41.7	17.2	49.8	24.5	45.4	33.1
	PiT-S[23]	2.9	23.5	80.9	95.3	41.0	16.5	52.5	21.7	43.6	30.8
	PVT-Small[50]	3.8	24.5	79.9	95.0	26.6	3.1	66.9	18.0	40.1	27.2
	Swin-T[29]	4.5	28.3	81.2	95.5	33.7	7.3	62.0	21.6	41.3	29.1
	TNT-S[17]	5.2	23.8	81.5	95.7	33.2	4.2	53.1	24.7	43.8	31.6
	T2T-ViT_t-14[55]	6.1	21.5	81.7	95.9	40.9	11.7	53.2	23.9	45.0	32.5
	RVT-S [32]	4.7	22.1	81.7	95.7	51.3	26.0	50.1	24.1	46.9	35.0
	FAN-S-ViT [57]	5.3	28.0	82.9	-	-	-	47.7	29.1	50.4	-
	RVT-S+RSPC [16]	4.7	23.3	82.2	-	-	-	48.4	27.9	-	-
	SATA-S (ours)	3.9	22.1	**89.3**	**99.1**	**57.4**	18.0	**33.8**	**30.5**	**59.5**	**39.2**
ViT-Base	DeiT-B[46]	17.6	86.6	82.0	95.7	46.4	21.3	48.5	27.4	44.9	32.4
	ConvViT-B[10]	17.7	86.5	82.0	95.7	46.4	21.3	48.5	27.4	44.9	32.4
	PiT-B[23]	12.5	73.8	82.4	95.7	49.3	23.7	48.2	33.9	43.7	32.3
	PVT-Large[50]	9.8	61.4	81.7	95.9	33.1	7.3	59.8	26.6	42.7	30.2
	Swin-B[29]	15.4	87.8	83.4	96.4	49.2	21.3	54.4	35.8	46.6	32.4
	T2T-ViT_t-24[55]	15.0	64.1	82.6	96.1	46.7	17.5	48.4	28.9	47.9	35.4
	RVT-B [32]	17.7	86.2	82.5	96.0	52.3	27.4	47.3	27.7	48.2	35.8
	FAN-B-ViT [57]	10.4	54.0	83.6	-	-	-	44.4	35.4	51.8	-
	RVT-B+RSPC [16]	17.7	91.8	82.8	-	-	-	45.7	32.1	-	-
	TORA-ViT-B/16($\lambda = 0.1$) [26]	26.0	111.2	84.1	-	48.4	23.3	31.7	46.5	57.6	-
	TransNeXt [43]	18.4	89.7	84.8	-	-	-	43.5	50.6	53.9	41.4
	SATA-B (ours)	15.9	86.6	**93.9**	**99.7**	**63.9**	20.2	**28.7**	**63.5**	**70.0**	**49.8**
	SATA-B* (ours)	15.9	86.6	**94.9**	**99.8**	**65.6**	**28.3**	**13.6**	**63.6**	**79.2**	**57.9**

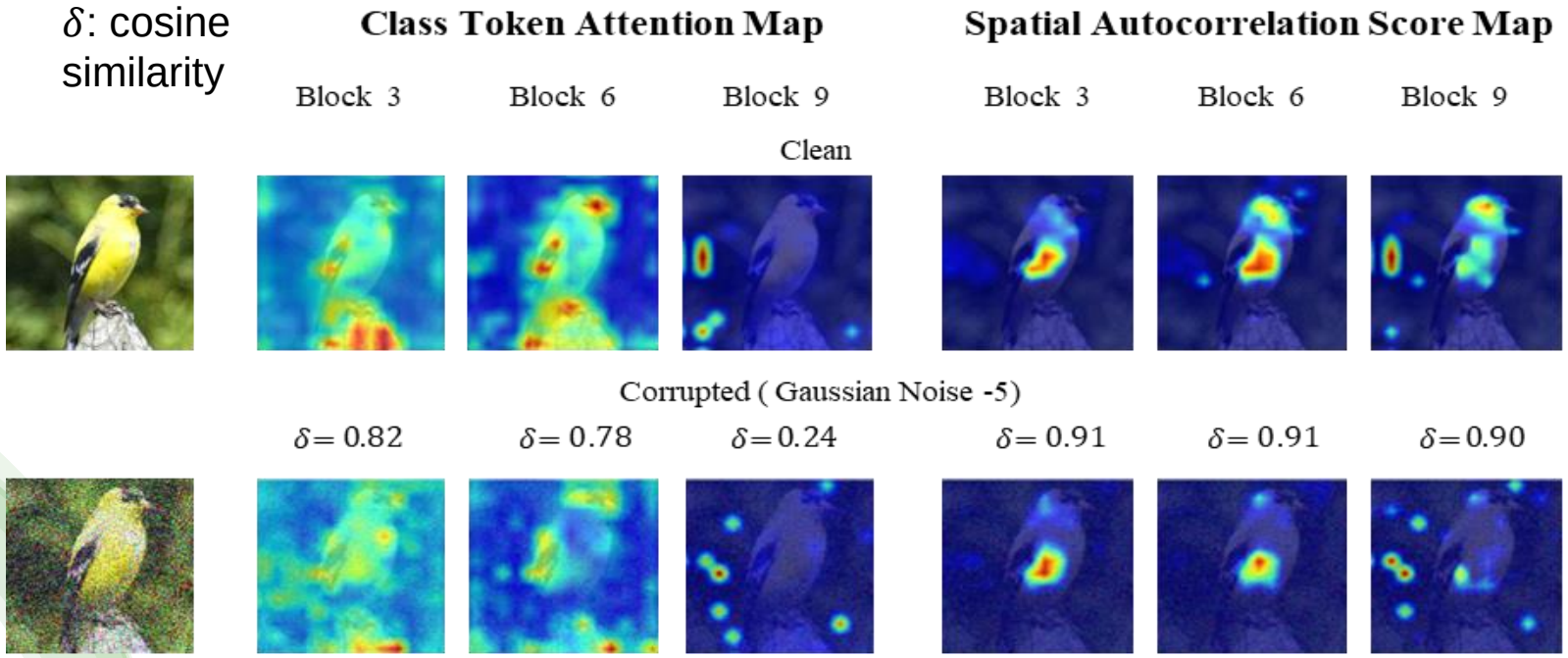
Griffith UNIVERSITY
Queensland, Australia

Analysis



Cosine similarity between the clean and corrupted versions of the class token attention map and spatial autocorrelation scores across different blocks of SATA-B

δ: cosine similarity



Griffith UNIVERSITY
Queensland, Australia

Thank You

https://github.com/nick-nikzad/SATA