ICLR

Toward Principled Flexible Scaling for Self-Gated Neural Activation

Sudong Cai[1,3], Shuyuan Zheng[2*], Bingzhi Chen[3], Shuai Yuan[3], Chuan Xiao[2], Jianbin Qin[4], Bing Wang[1*]

[1]The Hong Kong Polytechnic University; [2]The University of Osaka;
[3]Beijing Institute of Technology, Zhuhai; [4]Shenzhen University

Presenter: Sudong Cai
Contact: {sudong.cai, bing-w.wang}@polyu.edu.hk; zheng@ist.osaka-u.ac.jp

Motivation

Why do modern self-gated activations help CNNs much more than Transformers?

- Recent activation designs improve fitting flexibility by adding content-aware or non-local modulation.
- These ideas often boost CNNs, but much smaller gains appear in Transformer / MetaFormer layers.
- Transformer blocks already encode fine-grained non-local context through token mixing and attention.
- The paper names the resulting conflict inside activation design non-local tension.

How can a gate remain discriminative after informative features have already been enriched by non-local cues?

Core Insight $\phi(\tilde{x}) = \rho(\tilde{x})\,\tilde{x}$

Convergence limitation causes weak gating discrimination

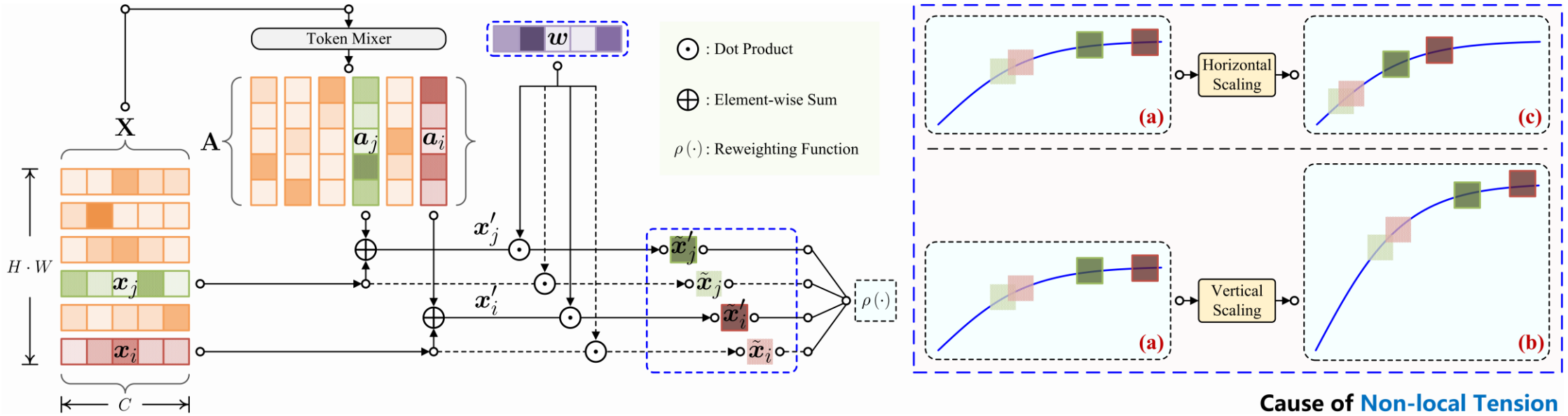

Fig. 1. Non-local tension and the intuition behind FleS.

- Importance score
- Trivially discriminative gating weights
- Convergence limitation

➢ The gate saturates, so important features receive almost identical weights.

➢ This weak discrimination is the key reason behind non-local tension.

Method $\phi(\tilde{x}) = \kappa_{ve}\rho(\kappa_{ho}\tilde{x})\,\tilde{x}$

FleS: adaptive vertical and horizontal scaling

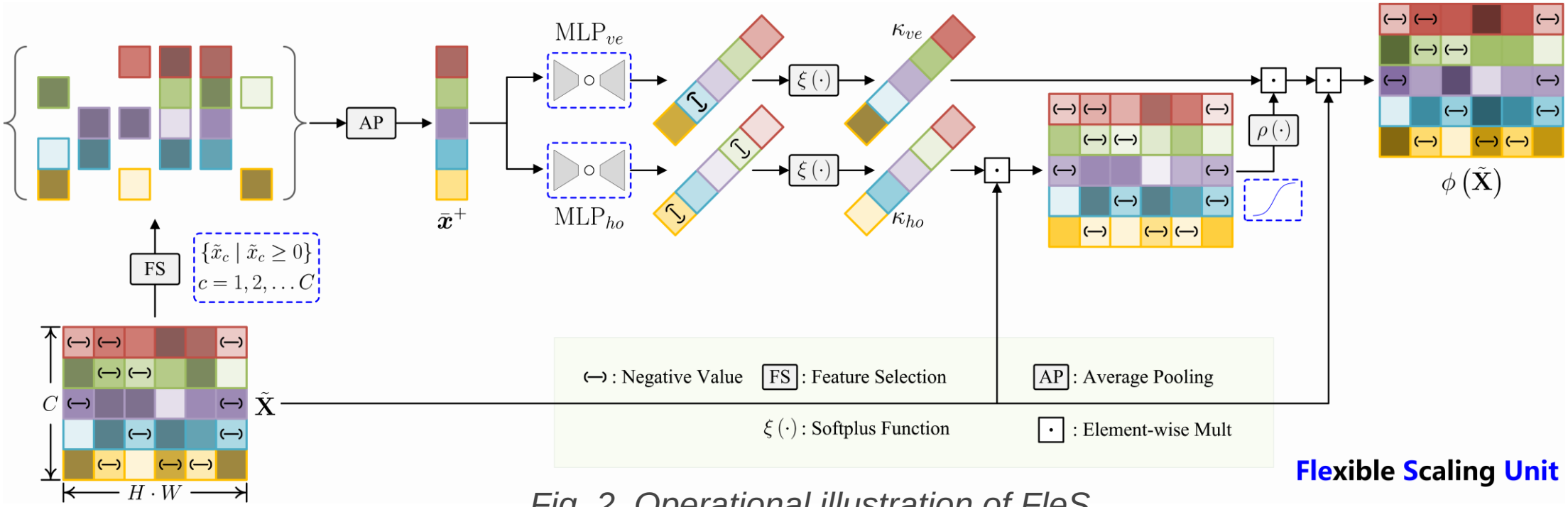

Fig. 2. Operational illustration of FleS.

- Vertical scaling adjusts the activation range
- Horizontal scaling adjusts the steepness
- Coefficients are predicted from positive channel statistics via lightweight MLPs

➢ FleS keeps the gate discriminative by adaptively rescaling both its bound and its steepness.

Results I

Initial evidence and main ImageNet performance

Table 1: Evaluation of FleS-Proto on ImageNet dataset (Deng et al., 2009).

Backbone	Activation	#Shuffle	#Params.	FLOPs	Top-1(%)↑
Swin-Micro	GELU	—	21.1M	2.6G	78.7
	FleS-Proto	—	21.1M	2.6G	**85.2**
		✓			**77.3**
Swin-Base	GELU	—	87.7M	15.1G	83.5

* The Swin-Micro (Liu et al., 2021) backbone is applied, where FleS activation function is compared with the GELU (Hendrycks & Gimpel, 2016) baseline.

Table 2: Comparison of different activation functions on ImageNet (Deng et al., 2009) with **(left)** Swin-Min (Liu et al., 2021) (Swin-[1, 1, 1, 1]) and **(right)** PoolFormer-S12 (Yu et al., 2022) backbones.

Backbone	Swin-Min (Liu et al., 2021)			PoolFormer-S12 (Yu et al., 2022)		
#Epochs		120			300	
Activation	#Params.	FLOPs	Top-1 (%)↑	#Params.	FLOPs	Top-1 (%)↑
GELU (Hendrycks et al., 2016)	11.8M	1.6G	68.7	11.9M	1.8G	77.2
ReLU (Nair & Hinton, 2010)	11.8M	1.6G	68.1	11.9M	1.8G	76.6
SiLU (Elfwing et al., 2018)	11.8M	1.6G	68.9	11.9M	1.8G	77.0
Mish (Misra, 2020)	11.8M	1.6G	68.6	11.9M	1.8G	77.1
Pserf (Biswas et al., 2022a)	11.8M	1.6G	69.0	11.9M	1.8G	NaN
SMU (Biswas et al., 2022b)	11.8M	1.6G	68.9	11.9M	1.8G	77.3
IIEU (Cai, 2023)	13.4M	1.6G	69.5	14.3M	1.8G	78.6
AdaS (Cai, 2024a)	13.7M	1.7G	69.7	15.1M	1.9G	78.2
StarReLU (Yu et al., 2024)	11.8M	1.6G	69.1	11.9M	1.8G	76.8
Meta-ACON (Ma et al., 2021)	13.4M	1.6G	68.3	14.3M	1.8G	78.0
FleS (Ours)	13.0M	1.6G	**71.4**	13.8M	1.8G	**79.4**
FleS-AdaS	14.1M	1.7G	**73.0**	—	—	—

* All competing methods are trained from scratch following the same recipe outlined in ***Implementation details***. "#Epochs" denotes the epochs of training; "NaN" denotes failed training; The baselines use GELU activation.

- FleS-Proto shows that clean channel-wise statistics can dramatically improve Transformer activations.

- On Swin-Micro, FleS-Proto improves Top-1 from 78.7% to 85.2% in the non-shuffle setting, but drops to 77.3% when the batch is shuffled.
- This motivates the practical FleS design for realistic settings.
- Practical FleS consistently outperforms popular and SOTA activations on Swin-Min and PoolFormer-S12, reaching 71.4% and 79.4% Top-1, respectively.

Results II

Scalability and ablation studies

Table 3: **(Left)** Comparison of the FleS-enhanced and vanilla GELU Swin-M(icro) (*i.e.*, Swin-[1, 2, 2, 2]), Swin-T, and ViT-B/16 (Dosovitskiy et al., 2021) models on ImageNet (Deng et al., 2009). **(Right)** Comparison of different activation functions on ImageNet using ResNet-50 backbone.

Activation	Backbone	#Params.	FLOPs	Top-1(%)↑
GELU	Swin-M	21.1M	2.6G	78.7
SiLU		21.1M	2.6G	78.6
SMU		21.1M	2.6G	78.8
Mt-ACON		24.2M	2.6G	78.9
FleS		23.5M	2.6G	**80.3**
GELU	Swin-T	28.3M	4.4G	81.3
SiLU		28.3M	4.4G	81.4
SMU		28.3M	4.4G	81.4
Mt-ACON		32.7M	4.4G	81.5
FleS		31.7M	4.4G	**82.3**
GELU	ViT-B/16	86.6M	16.9G	79.7
FleS		97.4M	16.9G	**80.7**

Activation	Backbone	#Params.	FLOPs	Top-1(%)↑
ReLU	ResNet-50	25.6M	4.1G	77.2
+SE-Net		28.1M	4.1G	77.8
PReLU		25.6M	4.1G	77.1
PWLU		N/A	N/A	77.8
SMU		25.6M	4.1G	77.5
SMU-1		25.6M	4.1G	76.9
FReLU		25.7M	4.0G	77.6
DY-ReLU		27.6M	N/A	77.2
ACON-C		25.6M	3.9G	76.8
Mt-ACON		25.8M	3.9G	78.0
IIEU		25.6M	4.2G	79.7
AdaS		25.6M	4.1G	79.9
FleS		28.1M	4.1G	**80.1**

* Note: FleS with κ_{ve} and κ_{ho} omitted is equivalent to SiLU (Elfwing et al., 2018).

Table 4: Ablation studies on **(left)** w/ or w/o the channel effective mean intensities $\{\bar{x}_c^+\}$ for modeling FleS coefficients; and **(right)** mining statistical cues within positive feature elements for FleS.

Activation	Backbone	#Params.	FLOPs	Top-1(%)↑
GELU		11.8M	1.6G	68.7
FleS-DG	Swin-Min	11.8M	1.6G	69.1
FleS		13.0M	1.6G	**71.4**

* "FleS-DG" denotes the FleS variant omitting $\{\bar{x}_c^+\}$ in generating scaling coefficients.

Activation	Backbone	#Params.	FLOPs	Top-1(%)↑
GELU		11.8M	1.6G	68.7
FleS-P&N	Swin-Min	13.0M	1.6G	69.8
FleS		13.0M	1.6G	**71.4**

* The FleS variant "FleS-P&N" averages positive and negative features for calculation of channel indicators.

- Removing channel statistics or averaging responses both reduces accuracy, confirming the importance of sign-aware flexible scaling.

- FleS scales well across architectures and model sizes, improving Swin-M&T, ViT-B/16, and ResNet-50.
- It raises Top-1 from 78.7% to 80.3% on Swin-Micro, 81.3% to 82.3% on Swin-T.
- Ablation studies show that channel indicators and positive-only statistics are crucial.
- *Comparative evaluations on the GLUE benchmark with BERT can be found in Appendix C.*

Conclusion

Takeaways

- We formalize the convergence limitation in self-gated activations and show that it leads to the overlooked but critical non-local tension in modern neural networks.
- We introduce FleS as an empirical validation of our gray-decision-making-inspired interpretation of neural activations, and further as a principled remedy to non-local tension challenge.
- Extensive experiments across vision and NLP benchmarks demonstrate the effectiveness, generalizability, and extensibility of FleS for interpretable activation modeling.

Thanks!

Toward Principled Flexible Scaling for Self-Gated Neural Activation

Sudong Cai[1,3], Shuyuan Zheng[2*], Bingzhi Chen[3], Shuai Yuan[3], Chuan Xiao[2], Jianbin Qin[4], Bing Wang[1*]

[1]The Hong Kong Polytechnic University; [2]The University of Osaka;
[3]Beijing Institute of Technology, Zhuhai; [4]Shenzhen University