Paper ID 837

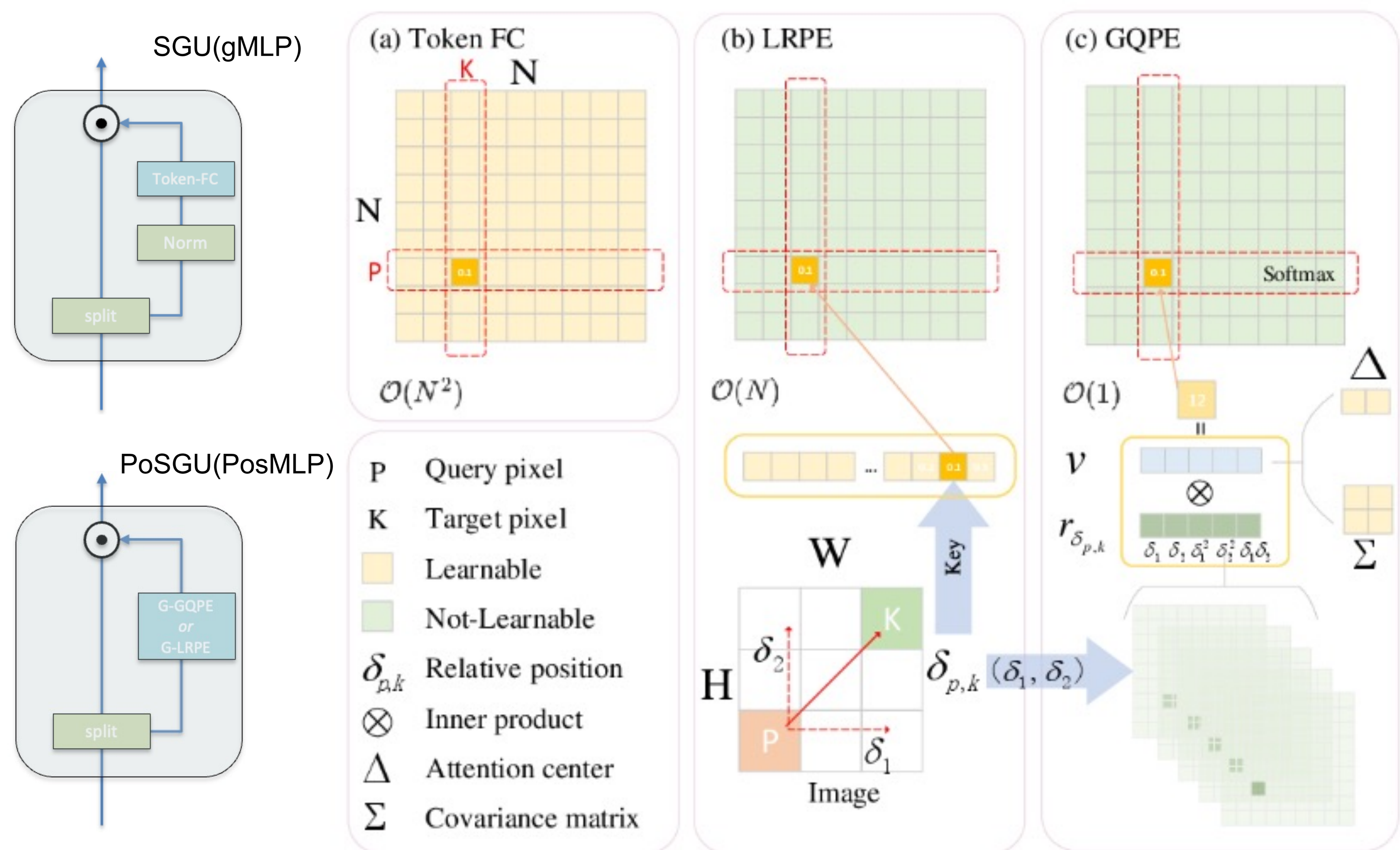
Parameterization of Cross-Token Relations with Relative Positional Encoding for Vision MLP

Zhicai Wang, Yanbin Hao §, Xingyu Gao §, Hao Zhang, Shuo Wang, Tingting Mu, Xiangnan He

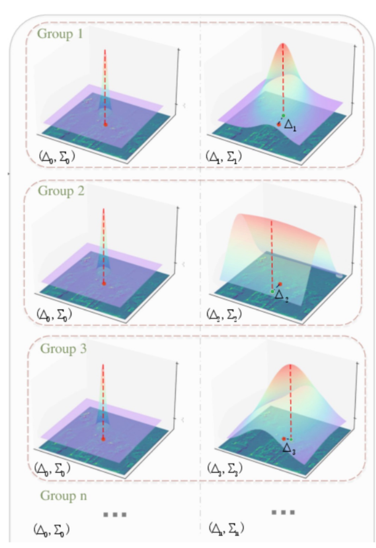
• Motivation & Contributions:

- Vision MLPs like MLP-Mixer, gMLP and et al. are modeling cross-token relations via heavy parameterized token-MLP layers. Inspired by the extensive implementation of the positional encoding method in Transformers, we explicitly integrate the positional prior into gMLP, which is also treated as our baseline, and propose our vision backbone network named PosMLP.
- The single layer of token FC in the SGU of gMLP is weak at capturing the complex spatial interaction. We propose to implement a channel group-wise strategy that assigns each group a individual RPE-based token FC layer to achieve a multi-granular information aggregation.
- The gMLP also suffers from its weak extendability to input resolution and thus the pretrained wights are hard to be transferred into other downstream scenes with flexible input resolution. To reduce the transfer cost, PosMLP is utilizing a window-portioning and convolutional down-sampling architecture.

• Method:



- **LRPE in PoSGU:** Learnable relative positional encoding (LRPE) predefines a learnable weight dictionary in which keys are defined as the relative displacement between two tokens. The weights of the mapping matrix are obtained via an assignment operation based on pairwise displacement, $\mathbf{Z}^{\text{lrpe}} = (\boldsymbol{r}_{\boldsymbol{\delta}}^{\text{lrpe}} + \boldsymbol{b})\text{Norm}(\mathbf{X}^1) \odot \mathbf{X}^2.$
- **GQPE in PoSGU:** Generalized quadratic positional encoding (GQPE) adapts a Non-isotropic Gaussian distribution to realize a continuous weights assignment (unlike the discrete dictionary in LRPE). The mapping matrix is predefined from a second-order function in which the attention center and covariate determine the certain aggregation pattern,



$$\mathbf{Z}^{\text{gqpe}} = (\mathbf{A}^{\text{gqpe}} + \boldsymbol{b})\mathbf{X}^1 \odot \mathbf{X}^2,$$

$$\mathbf{A}_{i,j}^{\text{gqpe}} := \text{Softmax}_j\left(-\frac{1}{2}\left(\boldsymbol{\delta}_{i,j} - \Delta\right)\Sigma^{-1}\left(\boldsymbol{\delta}_{i,j} - \Delta\right)^{\top}\right),$$

Right shows the illustration figure of the aggregation pattern before-and-after training from different groups of the G-GQPE. Due to its **inherit localized positional prior** and **light-weights** property, we take G-GQPE version of PosMLP as the default setting.

- **PosMLP**

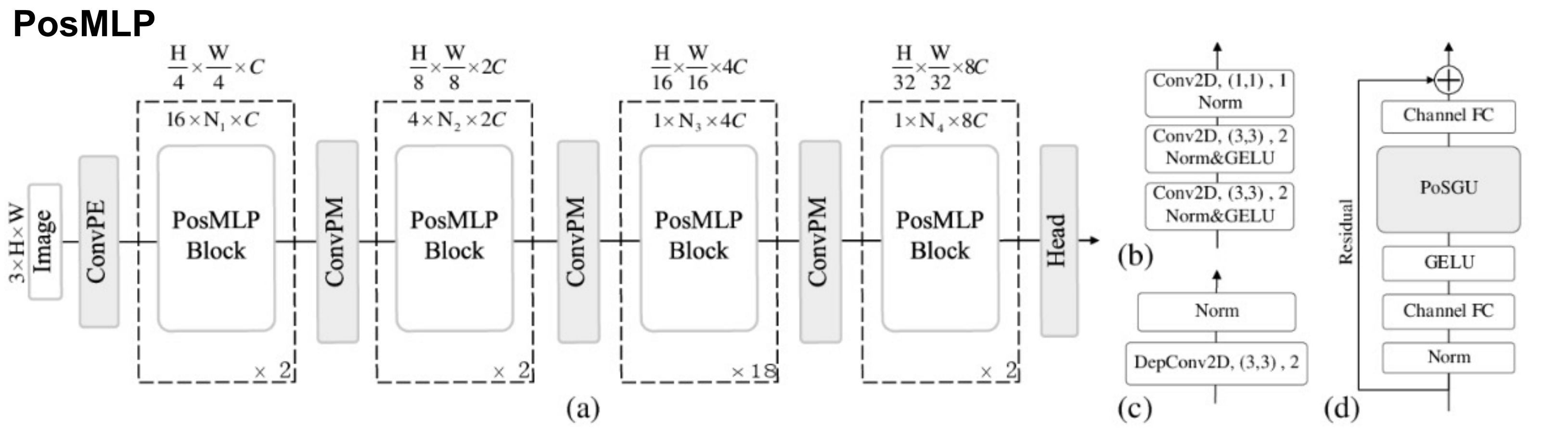


Figure 2: The proposed PosMLP: (a) Overall architecture; (b) Convolutional Patch Embedding block; (c) Convolutional Patch Merging block; (d) Architecture of PosMLP block with PoSGU.

• Experiments

- Ablation study:

1. *SGU VS* PosGU and *gMLP VS PosMLP*

Model	Module		Token-mixing complexity	Extra FLOPs	Top-1 acc.
gMLP	SGU		$O(N^2)$	✗	72.14
	PoSGU	LRPE-M	$O(N^2)$	$O(N^2)$	73.96(+1.82)
		LRPE	$O(N)$	$O(N^2)$	72.44(+0.30)
		GLRPE	$O(N)$	$O(sN^2)$	**74.56**(+2.42)
		GGQPE	$O(1)$	$O(sN^2)$	74.02(+1.88)
PosMLP	SGU		$O(N^2)$	✗	76.33
	PoSGU	LRPE-M	$O(N^2)$	$O(N^2)$	76.95(+0.62)
		LRPE	$O(N)$	$O(N^2)$	76.93(+0.60)
		GLRPE	$O(N)$	$O(sN^2)$	**77.41(+1.08)**
		GGQPE	$O(1)$	$O(sN^2)$	77.40(+1.07)

Table 1. Effectiveness of RPE in gMLP and PosMLP.

2. Main result (image classification)

Method	Input Size	#Param.	FLOPs	Top-1 Acc.
Tiny Models				
RegNetY-4G [41]	224^2	21M	4.0G	80.0
Swin-T [36]	224^2	29M	4.5G	81.3
Nest-T [56]	224^2	17M	5.8G	**81.5**
gMLP-S [35]	224^2	20M	4.5G	79.6
Hire-MLP-S [17]	224^2	33M	4.2G	81.8
ViP-Small/7 [23]	224^2	25M	6.9G	81.5
PosMLP-T	224^2	21M	5.2G	**82.1**
PosMLP-T	384^2	21M	17.7G	**83.0**
Small Models				
RegNetY-8G [41]	224^2	39M	8.0G	81.7
Swin-S [36]	224^2	50M	8.7G	83.0
Nest-S [56]	224^2	38M	10.4G	**83.3**
Mixer-B/16 [45]	224^2	59M	11.7G	76.4
S2-MLP-deep [52]	224^2	51M	9.7G	80.7
ViP-Medium/7 [23]	224^2	55M	16.3G	82.7
Hire-MLP-B [17]	224^2	58M	8.1G	**83.1**
AS-MLP-S[31]	224^2	50M	8.5G	**83.1**
PosMLP-S	224^2	37M	8.7G	83.0
Base Models				
RegNetY-16G [41]	224^2	84M	16.0G	82.9
Swin-B [36]	224^2	88M	15.4G	83.3
Nest-B [56]	224^2	68M	17.9G	**83.8**
gMLP-B [35]	224^2	73M	15.8G	81.6
ViP-Large/7 [23]	224^2	88M	24.3G	83.2
Hire-MLP-L [17]	224^2	96M	13.5G	83.4
PosMLP-B	224^2	82M	18.6G	**83.6**

Table 2: Performance comparison of PosMLP variants with the state-of-the-arts such as CNNs, vision transformers and vision MLPs on ImageNet1K dataset

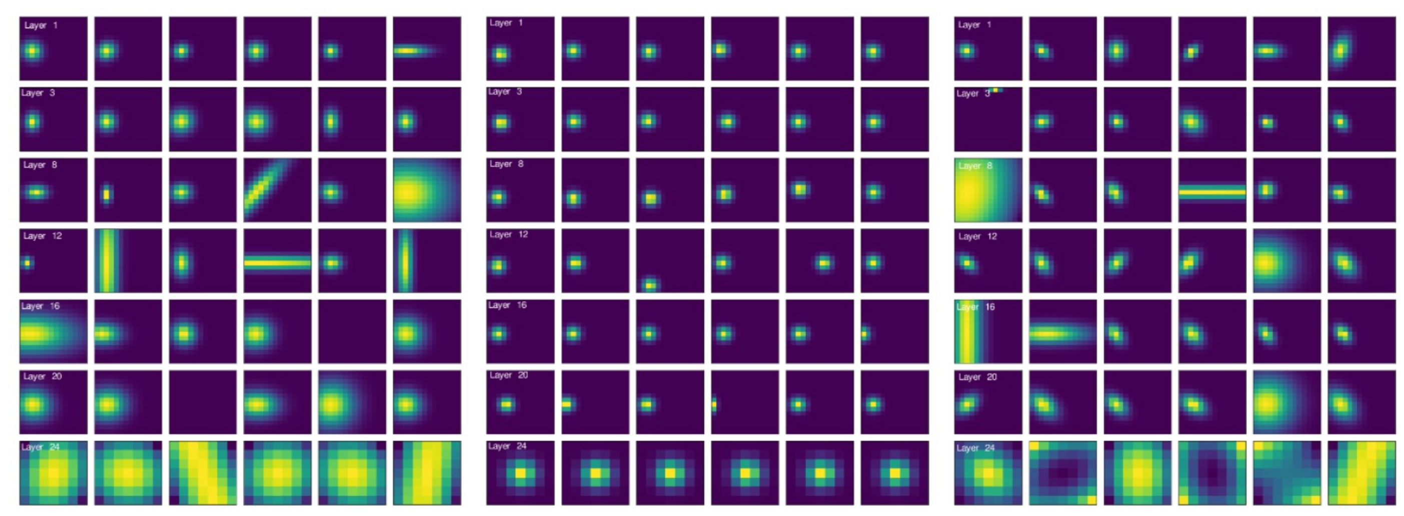
2. Main result (objection detection):

Backbone	#Param.	Mask R-CNN 1×						#Param.	RetinaNet 1×					
		AP	AP_{50}	AP_{75}	AP^m	AP^m_{50}	AP^m_{75}		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
ResNet50[20]	44.2M	38.0	58.6	41.4	34.4	55.1	36.7	37.7M	36.3	55.3	38.6	19.3	40.0	48.8
PVT-Small[49]	44.1M	40.4	62.9	43.8	37.8	60.1	40.3	34.2M	40.4	61.3	43.0	25.0	42.9	55.7
CycleMLP-B2[5]	46.5M	41.7	63.6	45.8	38.2	60.4	41.0	36.5M	40.9	61.8	43.4	23.4	44.7	53.4
PosMLP-T(ours)	40.5M	41.6	64.1	45.6	38.4	61.1	41.0	31.1M	41.9	63.2	44.7	25.1	45.7	55.6
ResNet101[20]	63.2M	40.4	61.1	44.2	36.4	57.7	38.8	56.7M	38.5	57.8	41.2	21.4	42.6	51.1
PVT-Medium[49]	63.9M	42.0	64.4	[illegible]	39.0	61.6	42.1	53.9M	41.9	63.1	44.3	25.0	44.9	57.6
CycleMLP-B3[5]	58.0M	43.4	65.0	47.7	39.5	62.0	42.4	48.1M	42.5	63.2	45.3	25.2	45.5	56.2
PosMLP-S(ours)	56.1M	43.2	65.5	47.4	39.4	62.5	42.1	47.3M	42.4	63.6	45.1	26.5	45.7	56.3

Table 3: Performance comparison with state-of-the-arts on object detection using COCO2017 dataset.

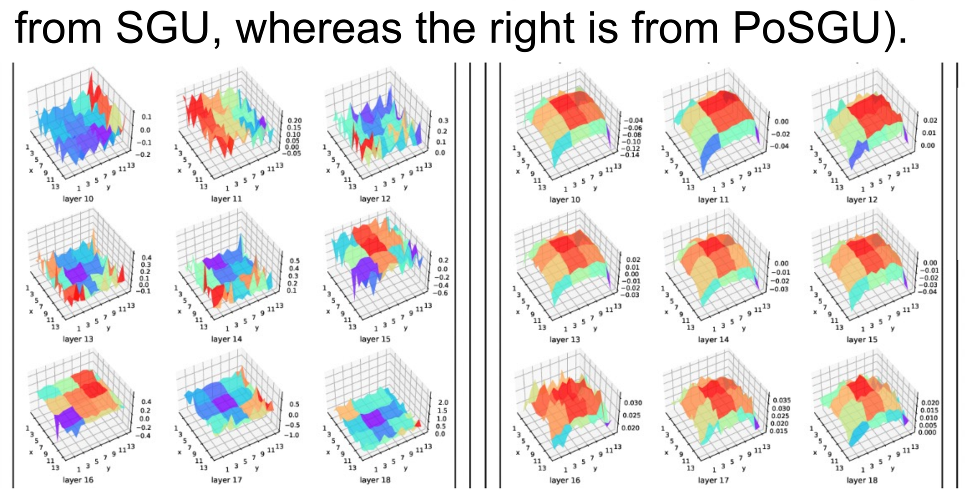
• Visualization

- The covariate matrix determines the aggregation pattern (the mapping weights of a query token that is reshaped in the feature map size).



(a) $\Sigma = \Gamma\Gamma^{\top}$ **(b)** $\Sigma = \alpha\mathbf{I}$ **(c)** $\Sigma = \Gamma$

- The bias term in PoSGU may reveal the absolute information which explains why we do not need absolute positional encoding (APE) in PosMLP (left is the bias map from SGU, whereas the right is from PoSGU).



Bias b	✗	✗	✓	✓
APE	✗	✓	✗	✓
Top-1 acc.	76.8	77.3	77.4*	77.3

Table 3: Ablation study on the relationship between bias term and absolute positional encoding.

• Conclusion and Resources

- Conclusion:

1. Relative positional prior is beneficial for the training of vision MLP and a careful calibration could reduce the model complexity and boost its modeling capability.
2. Group-wise operation is a nearly free-lunch operation in cross-token relation modeling.

- Contact & Resources:

wangzc@mail.ustc.edu.cn
haoyanbin@hotmail.com

GitHub

PDF Adobe