SP-ViT: Learning 2D Spatial Priors for Vision Transformers

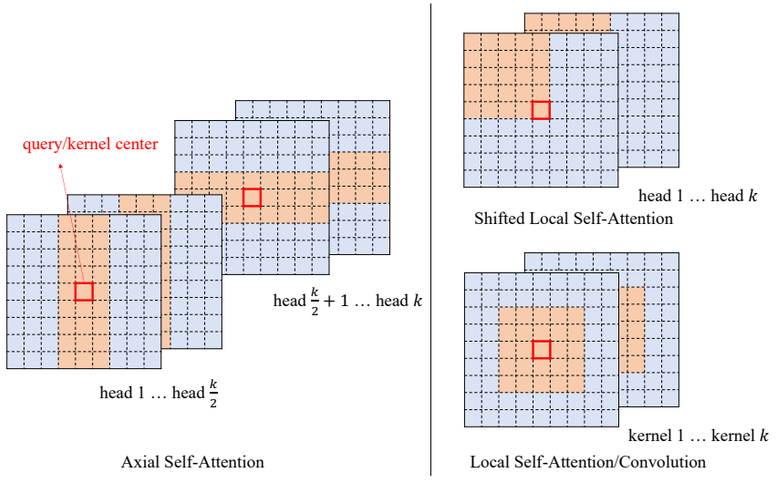
Yuxuan Zhou, Wangmeng Xiang, Chao Li, Biao Wang, Xihan Wei, Lei Zhang, Margret Keuper, Xiansheng Hua

UNIVERSITY OF MANNHEIM · 達摩院 ALIBABA DAMO ACADEMY · max planck institut informatik · UNIVERSITÄT SIEGEN

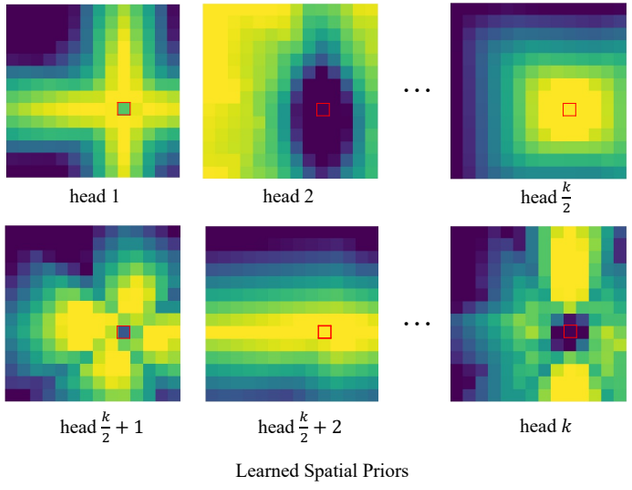
BMVC 2022

Inductive Biases for Vision Transformers

Hand-crafted convolutional inductive biases



Our learned Spatial Priors.



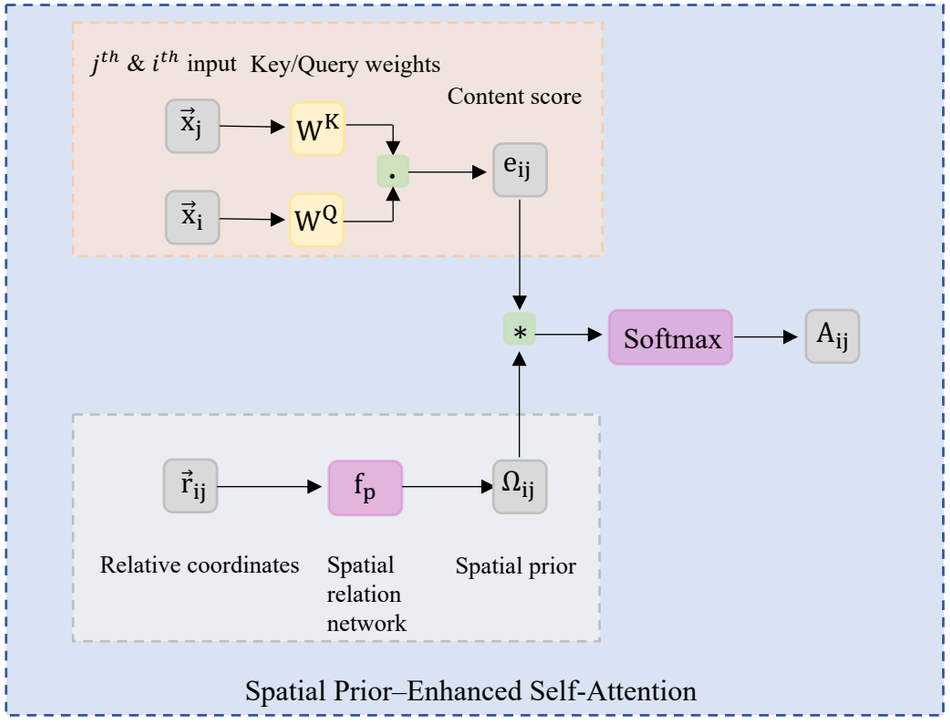
Focus on different spatial relations including local and non-local at each head.

Incorporate Spatial Priors into Self-Attention

Formulation of our SP-SA:

$$A_{ij} = \frac{\exp(e_{ij} \cdot \Omega_{ij})}{\sum_{k=1}^{n} \exp(e_{ik} \cdot \Omega_{ik})}, \quad \text{where} \quad e_{ij} = \frac{(\vec{x}_i^\top \mathbf{W}^Q)(\vec{x}_j^\top \mathbf{W}^K)^\top}{\sqrt{d_z}}, \quad \Omega_{ij} = f_p(\vec{r}_{ij})$$

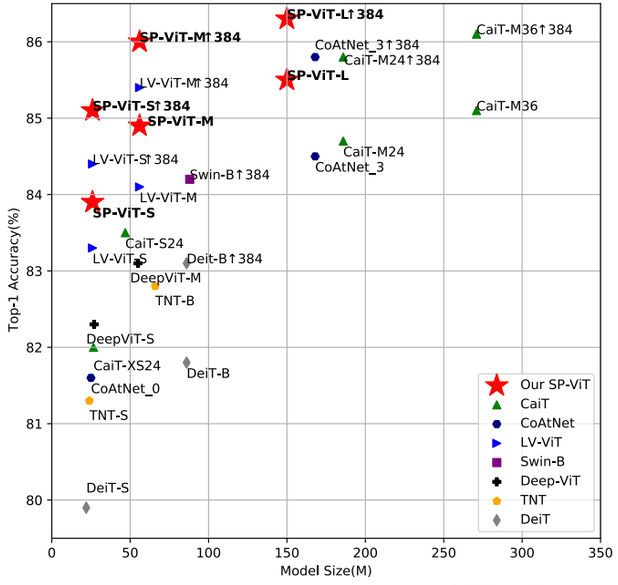
The mapping from 2D relative coordinates to our learnable Spatial Prior is parameterized by a 2-layer MLP.



- We propose a family of inductive biases for ViTs that focus on different types of spatial relations, called Spatial Priors (SP).
- Parameterized with neural networks, SPs are automatically learned during training and incorporated into the vanilla SA.

Experiments

Classification accuracy on ImageNet-1K



Semantic segmentation on ADE20K

Method	mIoU (SS)	P.Acc. (SS)	mIoU (MS)	P.Acc. (SS)
LV-ViT-S	47.9	82.6	48.6	83.1
SP-ViT-S	49.0	83.0	49.8	83.4

Ablation on the first 100 classes of ImageNet-1K

SP-SA	Acc (%)
Shared SP	82.6
Unique SPs per layer	82.1
Unique SPs per layer&head	83.6

Models and code are publicly available:
https://github.com/ZhouYuxuanYX/SP-ViT