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

Inductive Biases for Vision Transformers

Hand-crafted convolutional inductive biases

Incorporate Spatial Priors into Self-Attention

Formulation of our SP-SA:

\[ A_{ij} = \frac{\exp(e_{ij} \cdot \Omega_{ij})}{\sum_{k=1}^{n} \exp(e_{ik} \cdot \Omega_{ik})} \]

where \( e_{ij} = \frac{(\bar{x}_i^T W^Q)(\bar{x}_j^T W^K)^T}{\sqrt{d_c}} \), \( \Omega_{ij} = f_p(\bar{r}_{ij}) \)

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

Experiments

Classification accuracy on ImageNet-1K

Ablation on the first 100 classes of ImageNet-1K

| 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 |

Semantic segmentation on ADE20K

Models and code are publicly available:

https://github.com/ZhouYuxuanYX/SP-ViT