Skip-Attention: Improving Vision Transformers by Paying Less Attention

Shashanka Venkataramanan*, Amir Ghodrati*, Yuki M. Asano, Fatih Porikli, Amirhossein Habibian

Qualcomm AI research

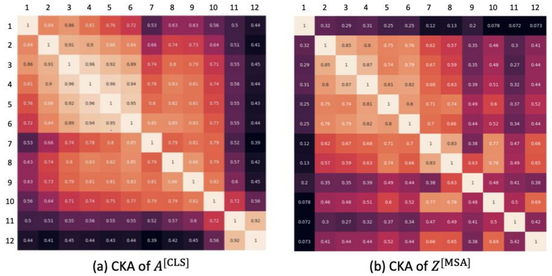
ICLR

Improving efficiency in ViTs

- Transformers are state-of-the-art in most image and video tasks.
- ViTs are **computationally expensive.**
  - Compute and memory grows **quadratically** (N x N). 14x14 for image classification to 128x128 =16K for image denoising.
- How to improve the efficiency of vision transformers?

CKA analysis

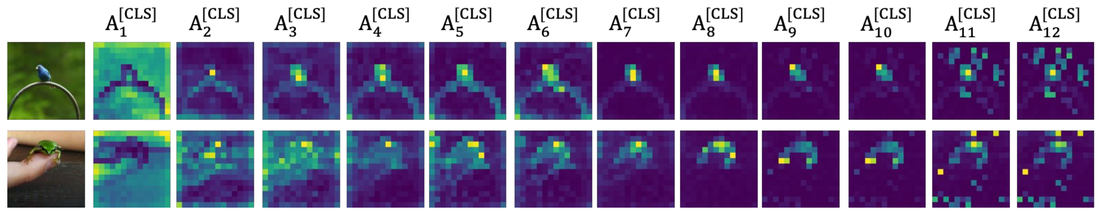
- **High correlation** between attention matrices across layers.
- Same for output of MSA blocks (features Z).



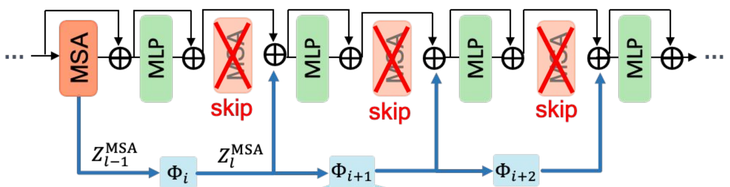
(a) CKA of $A^{[CLS]}$

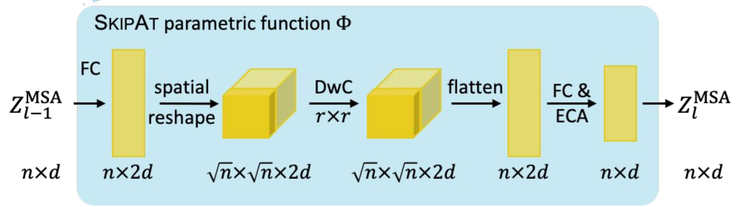
(b) CKA of $Z^{[MSA]}$

We leverage the **high correlation** across MSA blocks to improve **efficiency** of ViTs



Approximating MSA with SkipAT parametric function

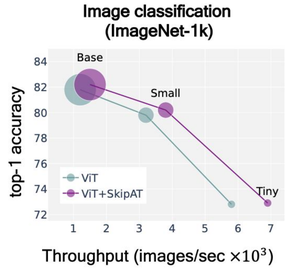


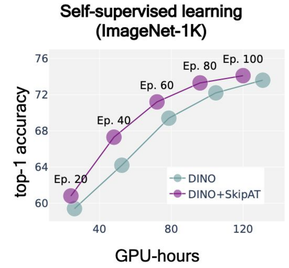


$$Z_l \leftarrow Z_l^{\text{MSA}} + Z_{l-1},$$
$$Z_l \leftarrow \text{MLP}(Z_l) + Z_l.$$

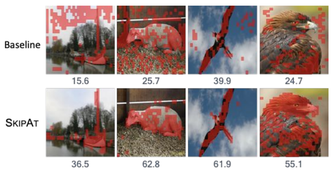
$$Z_l \leftarrow \Phi(Z_{l-1}^{\text{MSA}}) + Z_{l-1}$$
$$Z_l \leftarrow \text{MLP}(Z_l) + Z_l$$

Image classification

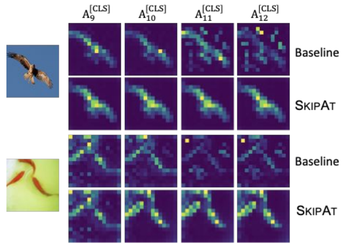




Unsupervised Object Discovery



Method	Jaccard↑	CorLoc↑
ViT-T [13]	32.2	39.5
ViT-T + SkAt	**38.0**	**41.5**
ViT-S [13]	29.0	40.6
ViT-S + SkAt	**34.0**	**41.2**
ViT-B [13]	33.6	36.4
ViT-B + SkAt	**36.8**	**37.2**



Additional Downstream Tasks