Evolution of Image Classification

Model Architecture

Sakura

2025/06/12
bili_sakura@zju.edu.cn

Background: Image Classification with Deep Learning

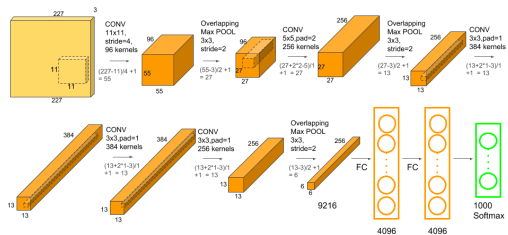
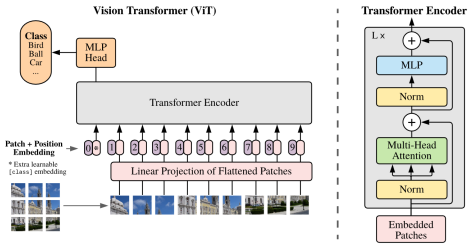


Figure: Left: AlexNet on ILSVRC-2010 (Berg, Deng, and Fei-Fei, 2010) Right: Architecture of AlexNet (Krizhevsky, Sutskever, and Hinton, 2012).

Berg, et al. Large Scale Visual Recognition Challenge 2010, 2010.
Krizhevsky, et al. ImageNet Classification with Deep Convolutional Neural Networks, NeurIPS, 2012.

Architecture Evolution of Image Classification

- **2012: AlexNet**, **2016: ResNet**
- **2021: ViT**
  (Dosovitskiy et al., 2021)
- **2021: Swin Transformer**
  (Liu et al., 2021)
- **2021: CLIP-ViT**
  (Radford et al., 2021)
- **2022: MAE-ViT**
  (He et al., 2022)
- **2022: CoCa-ViT**
  (Yu et al., 2022)



Overview of Vision Transformer (Dosovitskiy et al., 2021).

Dosovitskiy, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, ICLR, 2021.
Liu, et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows, ICCV, 2021.
Radford, et al. Learning Transferable Visual Models From Natural Language Supervision, ICML, 2021.
He, et al. Masked Autoencoders Are Scalable Vision Learners, CVPR, 2022.
Yu, et al. CoCa: Contrastive Captioners Are Image-Text Foundation Models. TMLR. 2022.