Vision Transformers for Image Recognition

Group Project Proposal

Student Names:

Zhiwei Dong, Yu Ji, Ruofan Lu

Team Name (optional):

Megatron

Project Title:

Vision Transformers for Image Recognition
(Title of the original paper: AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE)

Project Description:

The original paper developed the Vision Transformer (ViT) that is adopted from Transformers proposed by Vaswani et al. and well-used in the NLP world recently. In the paper, ViT is demonstrated to have competitive performance compared to state-of-the-art CNNs. This project will reproduce at least two experiments of the original paper, including training ViT/BiT on different sizes of datasets, training different ViT/BiT models and visualizing how attention works.

Methodology and techniques:

ResNet (BiT), Transformers, Multi-Head Sefl-Attention (MSA), Multiple Linear Perceptron (MLP), Hybrid Architecture

Resources:

Original Paper:
https://arxiv.org/abs/2010.11929

ViT:
https://github.com/lucidrains/vit-pytorch

Big Transfer (BiT):
https://github.com/google-research/big_transfer

ImageNet Large Scale Visual Recognition Challenge (ILSVRC):
https://www.image-net.org/challenges/LSVRC/

ImageNet Real:
https://www.tensorflow.org/datasets/catalog/imagenet2012_real

CIFAR10 and CIFAR100:
https://www.cs.toronto.edu/~kriz/cifar.html

Oxford-IIIT Pets:
https://www.robots.ox.ac.uk/~vgg/data/pets/

Oxford Flowers-102:
https://www.robots.ox.ac.uk/~vgg/data/flowers/102/

VTAB:
https://github.com/google-research/task_adaptation

Datasets:

ImageNet, ImageNet ReaL, CIFAR10, CIFAR100, Oxford-IIIT Pets, Oxford Flowers-102, VTAB