Conv MLP

1. What are the lacking of MLP's for current computer vision tasks? How authors proposed to overcome such process?
   Solution: There are multiple lacking of MLP's for current version of computer.
   a. Fixed dimension inputs, for that reason facing difficulty in downstream tasks, likes object detection and semantic segmentation.
   b. Single stage design do limited performance because of heavy computation connected layer.

   Authors proposed ConvMLP: A hierarchical convolutional MLP for visual recognition. Because of light weight, stage-wise, co-design of convolution layers and MLP.

2. What do you understand by convolutional tokenizer? How it is used with MLP and made CONV MLP?
   Solution: Convolutional tokenizer is a method which was proposed by CCT. It is leading to better performance on smaller datasets training from scratch, with fewer parameters compared with ViT.

   It is used with MLP and made CONV MLP like
   a. extract the initial feature map
   b. educe computation and improve spatial connections
   c. producing feature map
   d. We place 3 conv-MLP stages, generated 2 feature maps
   e. Each Conv-MLP stage includes multiple Conv-MLP blocks
   f. each Conv-MLP block has one channel MLP followed by a depth-wise convolutional layer, succeeded by another channel MLP

3. What are the datasets the authors experimented upon? Did they try some different domain datasets or similar domain datasets? Try to mention all.
   Solution: ImageNet-1k, CIFAR, Flowers-102, MS COCO and ADE20K benchmark are datasets the authors experiment upon.

   They tried some different domain datasets:
   - ImageNet-1k used for training images
   - CIFAR-10/CIFAR-100 and Flowers102 used for evaluate transferring ability of ImageNet pertained ConvMLP variants.
   - MS COCO is a widely-used benchmark for evaluating object detection model.
   - Dataset ADE20K is a widely-used dataset for semantic segmentation, which has 20k images for training and 2k images for evaluating the performance of semantic segmentation models.

4. Object detection and Semantic segmentation are difficult in the traditional MLP depicted by the authors. Explain with example why they are difficult and how Conv-MLP overcome it.
   Solution: Object detection and Semantic segmentation are difficult in the traditional MLP because of arbitrary input size resolutions. Moreover feature pyramids, single-stage design may limit performance on object detection and semantic segmentation.

   Conv-MLP overcome this issue.
   - The convolutional tokenizer is used to extract the initial feature map.
   - It blocks are then utilized to generate successive feature maps. These feature maps can be used to construct feature pyramids with no input size limits for use in downstream tasks overcoming the limitations of traditional MLPs.
   - For reduce computation and improve spatial connections, we follow tokenization with a pure convolutional stage, producing feature map F2 (H8 × W8 × C2 dimensional).

5. Visual representation learned by Conv-MLP can be seamlessly transferred and achieve competitive results with fewer parameters. What are the parameters the authors used to achieve such remarkable results?

   Solution: Parameters are weights and tokenization with a pure convolutional stage which was learned during training. Weight matrices are used for back-propagation process and contribute to the predictive capability of the model.
   Moreover, for reduce computation and improve spatial connections tokenization with a pure convolutional stage, author used tokenization. At each stage the learnable parameters are reduced from hx w x c→ H/2^L x 2/w^L x cx2^L, where L is the level of down convolution. If we consider MLP-Mixer at each MLP block the number of learnable parameters are hxwxc. The hyper-parameters are also tweaked to achieve competitive results.
   Author is used as optimizer in lieu of SGD. For optimum output he used learning rate and momentum.