1. **What are the lacking of MLP's for current computer vision tasks ? How authors proposed to overcome such process**

MLP's performance is limited when used to encode spatial information because it must fix the dimension of the input. Furthermore, because MLPs use feature pyramids to predict, single stage design and adherence to ViT may limit their performance. Furthermore, large consecutive MLPs have a high computational burden and more parameters, including a high dimension of hidden layers, and large consecutive MLPs may not perform as well as MLPs for current computer vision tasks due to the high computational burden and high dimensional parameter.

To solve the problem, the author proposes ConvMLP, a Hierarchical Convolutional MLP backbone for visual identification. To free other MLPs from input dimension constraints, the authors first converted all spatial MLPs into channel MLPs and built a baseline model using only MLPs. A thin convolution stage was layered on top of the remaining MLP stages to create the spatial information interaction. Down sampling was also accomplished using convolution layers. They then implemented a simple 3\*3 depthwise convolution between the two channel MLPs in each MLP block to improve spatial connections in MLP stages. Convolution and MLP layers work together to create the ConvMLP image categorization model. ConvMLP was able to overcome this procedure and correct MLP's flaw as a result.

1. **What do you understand by convolutional tokenizer? How it is used with MLP and made CONV MLP**

A convolutional tokenizer is a type of tokenizer that aids in the extraction of the first feature map while also reducing computation and improving spatial connections.

The author employs a convolutional tokenizer in the paper, which is made up of three convolutional blocks, each of which is made up of blocks. Conv-MLP block is a depth-wise convolution layer created in one MLP block using a convolution tokenizer between two channel MLPs. It is a 3x3 convolution layer with the same channel as the two channel MLPs. Because the convolution tokenizer includes both a convolution block and a convolution layer, the convolution layer compensates for the lack of spatial MLP removal by significantly improving performance with few parameters.

1. **What are the datasets the authors experimented upon? Did they try some different domain datasets or similar domain datasets. Try to mention all.**

The ImageNet-1k, CIFAR-10/CIFAR-100, Flowers-102, MS COCO, and ADE20K benchmark datasets were used in the author's experiments.

The authors used the same domain for all of the datasets because they all contain image type information that can be used to further classify images.

1. **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.**

Object detection and semantic segmentation are difficult tasks for traditional MLP because they require inputs with fixed dimensions and arbitrary resolutions.

Fixing the input dimension, for example, which is difficult in classic MLP, is required when using MLP to encode spatial information. With its large variation, MLP-Mixer was able to slightly outperform ViTBase in terms of computation.

Similar to a comparable transformer-based model, ResMLP is more complex and has about 30% more parameters. The authors created ConvMLP, a Hierarchical Convolutional MLP backbone for visual recognition, to overcome the limitations of standard MLP. For the purpose of releasing input dimension constraints in other MLPs, the authors initially constructed a simple pure-MLP model by replacing all spatial MLPs with channel MLPs. They added a light convolution stage on top of the existing MLP stages to enable spatial information interaction in addition to using convolution layers for down sampling.

Then, to improve spatial connections in MLP stages, they added a simple 3\*3 depth wise convolution between the two channel MLPs in each MLP block, dubbed it a Conv-MLP block. Convolution layers and MLP layers are co-designed to create the ConvMLP model for image classification. Experiments on different datasets revealed that ConvMLP is superior in terms of both performance and model size. Thus, ConvMLP overcame the challenges because it is scalable and can be used for downstream tasks such as object detection and semantic segmentation.

1. **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?**

The authors used the CIFAR-10/CIFAR-100 and Flowers-102 datasets to assess the transferability of the pretrained Conv-MLP. Each model was fine-tuned for 50 epochs at a learning rate of 3e-4 (using a cosine scheduler), weight decay of 5e-2, and 10 warmup and cooldown epochs. Because they used the same training script, the same augmentations were used in the ImageNet-1k trials. Every image has also been reduced in size to 224224 pixels. We can see that they had to retrain fewer parameters, which resulted in better outcomes. According to the report, ImageNet-1k had a top 1% accuracy of 76.8%, while CIFAR-10, CIFAR-100, and Flowers-102 had top 1% accuracy of 98%, 87.4%, and 99.5%, respectively.