1.answer:

MLP does have some flaws that make it challenging for some computer vision tasks. For particular, spatial MLP has a fixed input dimension, that impedes its ability to use for subsequent computer vision tasks like object detection and semantic segmentation. The performance of various tasks may be constrained by the MLP's single-stage design. It is also challenging to apply MLP to existing computer vision issues since large consecutive MLPs with fully connected layers demand complex processing and many parameters.

The authors have suggested a hierarchical convolutional MLP for machine vision as a solution to these problems. Convolutional layers and MPL layers are combined in this technique to classify images. It's a lower resource consuming, stage-by-stage procedure that, by learning visual representations for computer vision tasks, can produce competitive outcomes with fewer parameters. Convolutional MLP is modular and is simple to use for tasks like semantic segmentation and object detection.

2.answer:

Convolutional tokenizer is the tokenizer which has three convolutional blocks, with all three of them having batch normalization, ReLU activation & a 3x3 convolution. It has a max pooling layer attached to it as well.

The patch tokenizer was swapped out for a convolutional tokenizer in the publication. It is employed to extract the F1 initial feature map and also to create feature map F2 for speeding up computation and strengthen spatial relationships. They moreover added three Conv-MLP stages, producing the F3 and F4 feature maps. Every Conv-MLP level consists of a number of Conv-MLP blocks, each of which consists of a channel MLP, a depth-wise convolutional layer, and then another channel MLP. Inputs in the block were given Layer Normalization treatment, and they also contained residual connections. Two completely connected layers with a GeLU activation and dropout make up each channel's MLP. The resultant feature map, F4, is next subjected to global average pooling, followed by passage through the classification head. The feature maps F1, F2, F3, and F4 can be utilized to create feature pyramids without restrictions on input size when using ConvMLP for downstream tasks.

3.answer:

The authors tested with are ADE20K benchmark, ImageNet-1k, CIFAR, Flowers-102, and MS COCO datasets. They experimented with datasets from apparently related domains because they all had images.

4.answer:

The authors illustrated how challenging standard MLP is for object detection and semantic segmentation. For particular, it must fix the input dimension in order to encode spatial information using MLP. However, a variety of computer vision applications, such object detection and semantic segmentation, call for inputs with any resolution. Therefore, deploying MLP on these kinds of computer vision applications is highly challenging due to the fixed dimension of its inputs. Additionally, MLP produces predictions based on their feature pyramids for object detection and semantic segmentation. As a result, ViT's single-stage architecture can restrict how well certain tasks are performed. Moreover, it is quite challenging to use MLP to current computer vision problems because huge consecutive MLPs may necessitate intensive processing and far more parameters with larger dimension of hidden layers. The authors found that MLP-big Mixer's variation allowed it to marginally outperform ViT-Base but it was twice as expensive and required complex computation. Additionally, ResMLP seems to be almost 30% more complex and has more parameters than a transformer-based model with comparable performance.

Convolutional layers and MLP layers are combined in the Conv-MLP or hierarchical Convolutional MLP model, which is proposed by authors as a solution to these problems. For computer vision tasks like object recognition and segmentation, this model was able to perform better while requiring less parameters. First, they develop a pure-MLP baseline model by replacing all spatial MLPs with channel MLPs for cross-channel connections in order to remove the input dimension limits on previous MLP-like frameworks. The spatial information interaction is then created by layering a light convolution stage on top of the remaining MLP stages. In addition, convolution layers are used for down sampling.

5.answer:

The authors used the CIFAR-10/CIFAR-100 and Flower The authors used the CIFAR-10/CIFAR-100 and Flower 102 datasets to assess the transfer capability of the Conv-MLP. Every model had been fine-tuned for 50 epochs at a learning rate of 3e 4, weight decay of 5e-2, and 10 warmup and cooldown epochs. Identical augmentations and training script were used as the ImageNet-1k experiments. All images are also downsized to 224into224. As a result, it is clear from this that they needed to adjust lesser parameters and their results were improved. The publication shows that imageNet-1k's top 1% accuracy was 76.8%, while CIFAR-10, CIFAR-100, and Flowers-102 had respective accuracy of 98%, 87.4%, and 99.5%.