# **Computer Vision_DiscussionQuiz_12**

1. Separate modules for spatial and channel processing are used in the SPACH system. All architectures can attain superior results at a modest level, according to the author's experiments using the SPACH framework. When the network size grows larger, however, they exhibit diverse behaviors. He developed two hybrid models employing convolution and Transformer modules based on their findings. When a sweet spot between generalization capability and model capacity is reached, the performance of these simple hybrid models is already on par with SOTA models with advanced architectural designs, according to varied network performance behavior at different scales with this framework.

2. The following are the four conclusions presented in article 1:

   a. At a relatively low computational cost, both the Transformer and MLP structures gain from local modeling. The absence of local modeling is shown by the superscription.

   b. When weight sharing is used to avoid over-fitting, MLP models perform significantly better.

   c. In the MLP-MS model, using shared weights delays the development of over-fitting symptoms even further. As a result, we conclude that MLP-based models are still viable.

   d. Conv-MS outperforms Trans-MS in terms of generalization because it gets a greater test accuracy at the same training loss before the model saturates.

3. When trying to construct a stand-alone transformer network for vision tasks, the following are the two key challenges:

   a. Transformer controls a set of tokens. Transformers, on the other hand, do not work with natural tokens. A picture, comparable to words in natural language, exists.

   b. Unlike Images, the Transformer structure ignores location and considers all tokens identically. Images, on the other hand, have a fantastic local structure. Because the picture data has strong localization, this becomes difficult.

4. The channel mixing function is concerned with channel information fusion. Here we need to employ MLP and add normalization and residual connection. Whereas the spatial mixing function is concerned with aggregating context information from many spatial places. The spatial mixing feature is essential for implementing various architectural designs. With the use of structures like as MLPs, convolution, and self-attention, the author was able to construct three new unique architectures.

5. Spatial mixing feature is needed to facilitate different architectural designs. Convolution, self-attention, and MLP are used to create three structures. Normalization and residual connection are two typical components. Because channel mixing will be handled individually in following phases, the convolution structure is achieved via a 3x3 depth-wise convolution. In the original design, there is a positional embedding module for the Transformer structure. However, according to current research, absolute positional embedding destroys translation variance, which is incompatible with pictures.

6. Overfitting is a problem with MLP-based models. There are two factors that are discussed in paper-1 that helped in reducing the over-fitting problem. They are as follows:

   When weight sharing is used to reduce over-fitting, MLP models perform significantly better.

   Multi-stage structure also played major role in solving the over-fitting problem and help to resolve this. As per paper 1, both weight sharing and multi-stage structure are the two factors that played major role in reducing over-fitting issue.

7. Translation equivariance is the most crucial, as it is a desired quality for jobs like object identification. Convolutional Neural Networks are also naturally efficient since the calculations are shared when employed in a sliding-window fashion.

   Translation-invariance characteristics are included by ViT architecture. To enable successful training of deeper ViT networks, a Layer Scale method is developed. It's also been revealed that various class-attention layers constructed on top of the ViT network perform better than class embedding.

8. a. Macro Design: The model is focused with image processing at the start of the network computation. The resnet-4 model is compatible with object detection tasks, and the swin-T model does the same computation in 1:1:3:1, but the swin-T 1:1:9:1 compute ratio is utilized to accommodate bigger datasets.

   b. Inverted BottleNeck: The MLp block's hidden dimensions are quadruple times greater than the swin-T input size.

   c. ResNeXTiFy: The capacity of the model is boosted by widening the kernels. The ResNet employs grouped 3x3 convolution, in which the convolution filters are divided into groups. The performance is improved by extending the breadth of the Resnet to match the swin-T.

   d. Micro-Designing: Activation functions such as altering Relu to Gelu, fewer activation functions, and less normalization layers are among the architectural differences for models.

e. Kernal Sizes: Swin-T will have a non-local receptive field that is global owing to the increase in receptive field. As a result, increasing the kernel size improves accuracy.

9.

a. Local Modeling: The convolution model can achieve similar performance as a Transformer model in our SPACH framework using just light-weight depth-wise convolutions. Local modeling is both effective and necessary. A considerable performance improvement is realized with insignificant parameters by adding a local modeling bypass in both MLP and Transformer architectures, and FLOPs rise.

b. Hybrid Models: In paper-1, the author developed two scale-dependent hybrid models based on convolution and Transformer layers. The performance of these simple hybrid models is already on par with SOTA models with advanced architecture designs when a sweet spot between generalization capability and model capacity is achieved, as shown in paper-1.

c. Architectural difference between single-stage and multi-stage SPACH frameworks: In single-stage SPACH frameworks, the picture is followed by Patch Embedding, N mixing blocks, Global Average Pooling layer, and Linear Classifier to the class. Patch Embedding is followed by N1, N2,....Ni mixing blocks where the stage level is the stage level in multi-stage SPACH frameworks. And the class is followed by the Global Average Pooling layer and the Linear Classifier. In paper 1, it is demonstrated that multi-stage models outperform single-stage models.

d. Inductive bias: The inductive bias is a collection of assumptions that the learner makes in order to anticipate outputs from unknown inputs. Convolutional Neural Networks have multiple inductive biases built in. This makes them extremely versatile of Machine Learning and Computer vision related tasks. In these inductive biases, the tasks of Object detection is done by a inductive-bias. Which makes it most crucial one. It is the Translation equivariance.

e. Isotropic Model Design: There are no down sampling layers in isotropic model structures, and these are not dependent on the direction. Therefore feature resolutions remain constant at all depths.

Note: https://en.wikipedia.org/wiki/Inductive_bias

Papers:

A Battle of Network Structures: An Empirical Study of CNN, Transformer, and MLP

A ConvNet for the 2020s