Discussion Quiz #11

Ashutosh Tiwari (ashutiwa@iu.edu)

April 18, 2022

1 Papers

1.1 Paper 1

MLPMixer

1.2 Paper 2

CycleMLP

2 Questions

2.1 Describe all the steps of the mixer layer design (shown in Fig. 1). Also describe the architecture of MLP in the same figure. (5 pts)

MLP mixer architecture only includes multiple layers of MLP. These are applied either across spatial dimensions or feature channels. Input to the mixer are the linear image patches, shaped as $patches \times channels$ architecture. This format is consistent throughout the architecture.

First an image is divided into patches which are then converted to a latent representation. This latent representation is of size C and is non overlapping among patches. In other words each patch results in a embedded vector. Then there is this mixer layer which takes these latent embeddings.

This creates a table like structure which is taken as input by the Mixer layer. Each Mixer layer has two components. First part is the token mixer which mixes the patches of the input table by mapping them through a single MLP, which is shared across all channels. Second part is the channel mixer, applies layer normalization and then mixes the spatial dimensions of the input table. Both of these are connected by skip connections. For both of those components, there are single MLPs which are shared on a channel or spatial basis. So in total for every mixer block, there are two MLPs. In case of toekn mixer there is also a transpose at the beginning and end of the mixer.

2.2 Why does the mixer layer use Transpose operation twice inside the layer? (1 pt)

First transpose is for aligning it for patch mixer and second transpose is for aligning it for channel mixer. This is important since there are skip connections at the point of second transpose and so we want to have same dimensionality for both the skip connection and the output of last layer.

2.3 Describe in detail the functioning of channel-mixing and token-mixing MLPs. (3 pts)

Each mixer has these two MLPs. The first MLP is the token mixer which takes the input of the table and maps it through a single MLP. The second MLP is the channel mixer which takes the input of the table and maps it through a single MLP.

Token mixer acts on the columns of the embedding vector and then a shared MLP is applied to it. The output of the token mixer is then transposed and passed to the channel mixer. Channel mixer acts on the rows of the embedding vector and then a shared MLP is applied to it. The output of the channel mixer is then passed to the token mixer of the next mixer. So each mixer has two fully connected layers.

2.4 Describe how larger convolutional kernels perform both feature mixing at a given spatial location and between different spatial locations. (3 pts)

In case of large convolutional kernels, the feature mixing is performed at a given spatial location and between different spatial locations. Local spatial mixing is performed by convolving the input with a kernel and then taking the average/mean or some other operation of the output. This is done for all the channels. Spatial mixing between different locations is done because of the increasing receptive field of the convolutional kernels as the depth increases.

2.5 State True or False: (6 pts)

2.5.1 1x1 convolution can perform feature mixing at a given spatial location.

True

2.5.2 CNNs require pooling to perform feature mixing between different spatial locations.

True, this is required along with convolution operation.

2.5.3 Mixer layer uses the same instance of MLP layer for both channel-mixing and token-mixing.

False, mixer block has two MLPs. One for channel mixing and one for token mixing.

2.5.4 All patches use the same projection matrix to get a desired hidden dimension C.

True.

2.5.5 Computational complexity of the network (paper 1) is quadratic in the number of input patches like ViT models.

False, it is linear.

2.5.6 In ViT models, self-attention layers can perform feature mixing at a given spatial location as well as between different spatial locations but the MLP layers in ViT can only do feature mixing at a given spatial location.

True.

2.6 Describe briefly (based on paper 1): (2 pts)

2.6.1 Positional Invariance

Positional invariance is the property of the model that the model is invariant to the position of the input patches. In MLP mixer this is achieved by using the same MLP for both the token and channel mixer.

2.6.2 Isotropic architecture of models

Isotropic architecture of models is the property of the model that every mixer block has the same architecture, input and output dimensions.

2.6.3 Pyramidal architecture of models

This is the fact that in case of most CNN architectures, deeper layers have more receptive fields and lower resolution than the previous layers. And the channel size increases as the depth increases.

2.6.4 Why dies the mixer model NOT use position embedding?

Authors point out that they did not use position embedding since token mixing MLPs are sensitive to the position of the input patches.

2.7 According to paper 2, why all-MLP models, like MLP-mixer, can not be used for dense prediction tasks? Give at least three reasons. (3 pts)

Authors present three reasons for same:

1. Because all these models are composed of non hierarchical structure, therefore they can't high resolution feature representations.
2. These MLP models can't deal with variable length input.
3. Computational and memory complexity of these models are quadratic with respect to size of input.

2.8 Explain the architecture of CycleMLP block shown in fig. 5 (d). (3 pts)

In this architecture CycleMLP block has two MLPs. One for token mixing and one for channel mixing, very similar to the mixer block. There is no difference in the channel mixing MLP. All changes are in the first token mixing MLP. In spatial projection there are three different size convolutional operations performed on the input in parallel.

2.9 Describe briefly (based on paper 2): (4 pts)

2.9.1 Channel FC

In channel FC layer, all features in a channel dimention for all patches are aggregated using a single MLP. It can handle different input scales but cannot learn spatial context.

2.9.2 Spatial FC

Spatial FC on the other hand is a MLP that can learn spatial context. It has the maximum possible receptive field of the input. It's computational complexity is quadratic with respect to the input size.

2.9.3 Cycle FC

Cycle FC has the global receptive field ver similar to the spatial FC. It's computational complexity is however linear as that of channel FC. It applies weighting matrix along the channel dimension on fixed dimension of the input.

2.9.4 Why Cycle FC can be used for dense prediction?

It can be used for dense prediction tasks because it can learn spatial context using CycleFC with a linear computational complexity.