Convolutional Deep Neural Networks on a GPU

by

Team Incognitos

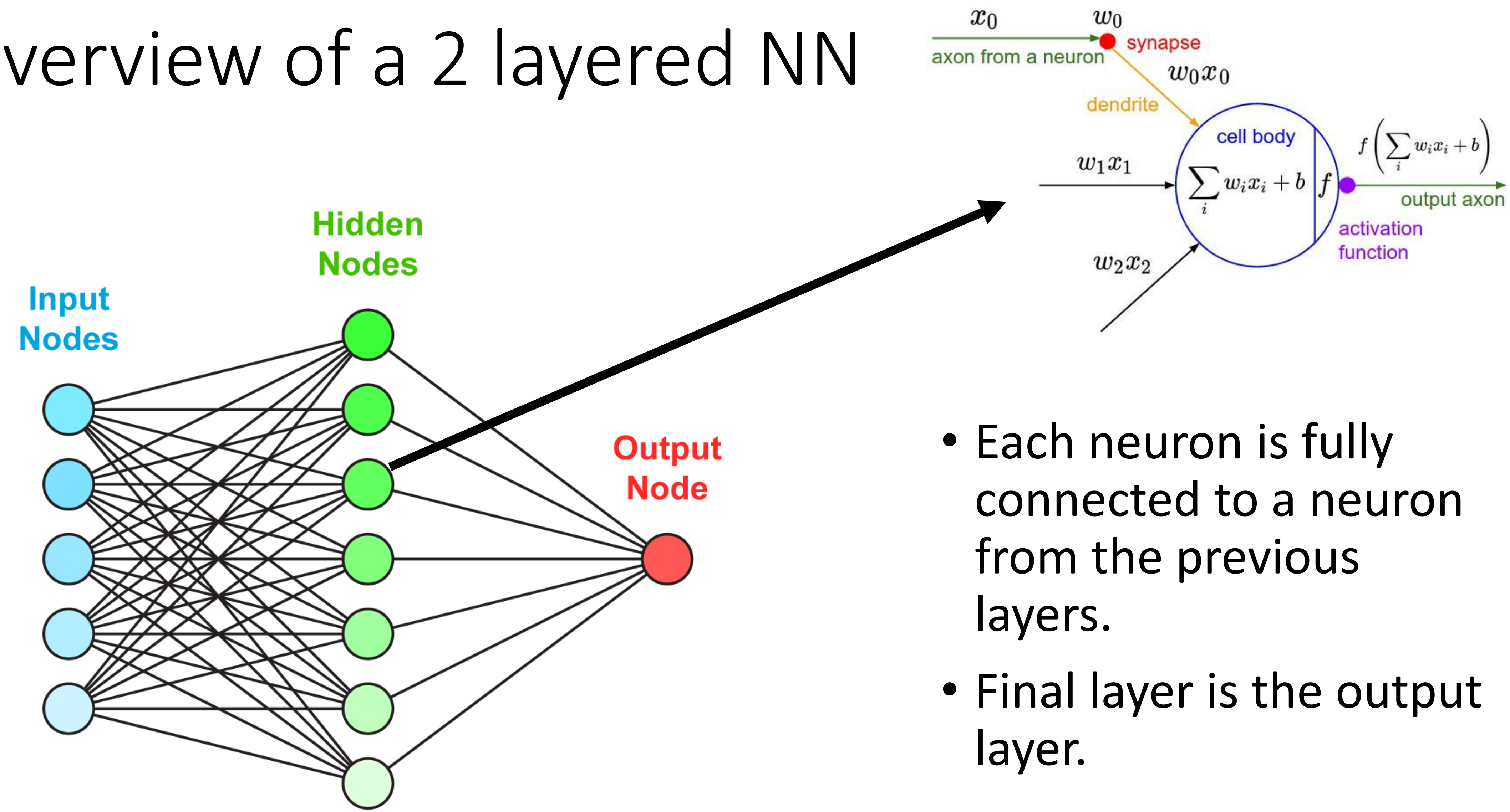
Suhas **P**illai

Siddesh **P**illai

Investigate

- Alex Net Architecture from the paper ImageNet Classification with Deep Convolutional Neural Networks by Alex Krizhevsky (University of Toronto).
- Run 5 convoluted Layers
- 3 Max Pooling Layers
- 2 Fully connected layers with 4096 neurons
- 1 output layer.

Overview of a 2 layered NN



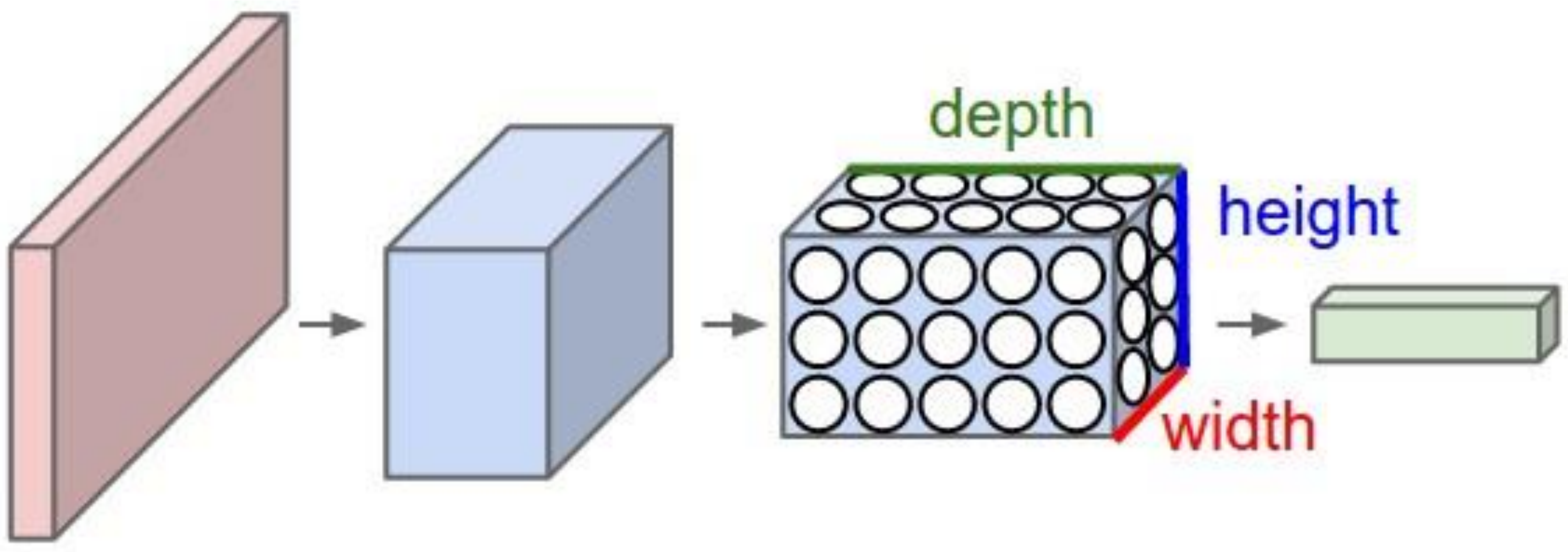
- Each neuron is fully connected to a neuron from the previous layers.
- Final layer is the output layer.

Need for Convolutional Neural Network (CNN)

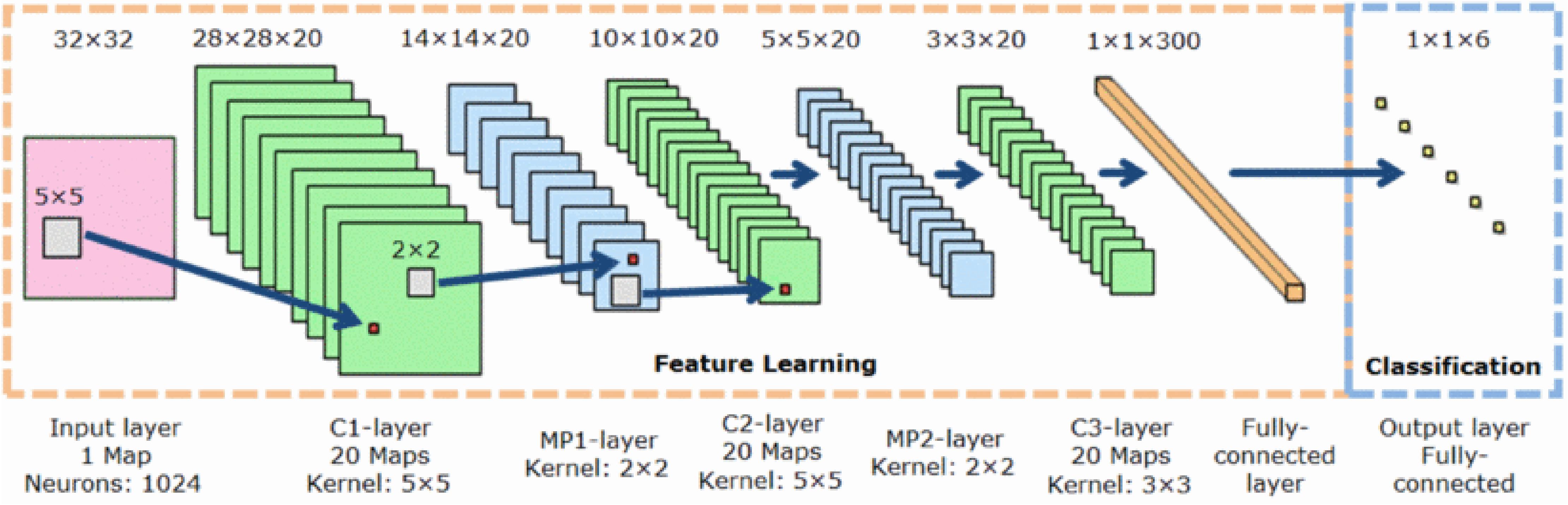
- Regular NN do not scale well to full images.
- E.g. CIFAR-10 image size is 32*32*3 = 3072 weights. Still manageable
- For a more respectable image e.g. 200*200*3 = 120,000 weights.
- Large number of parameters leads to overfitting.

Convolutional Neural Networks

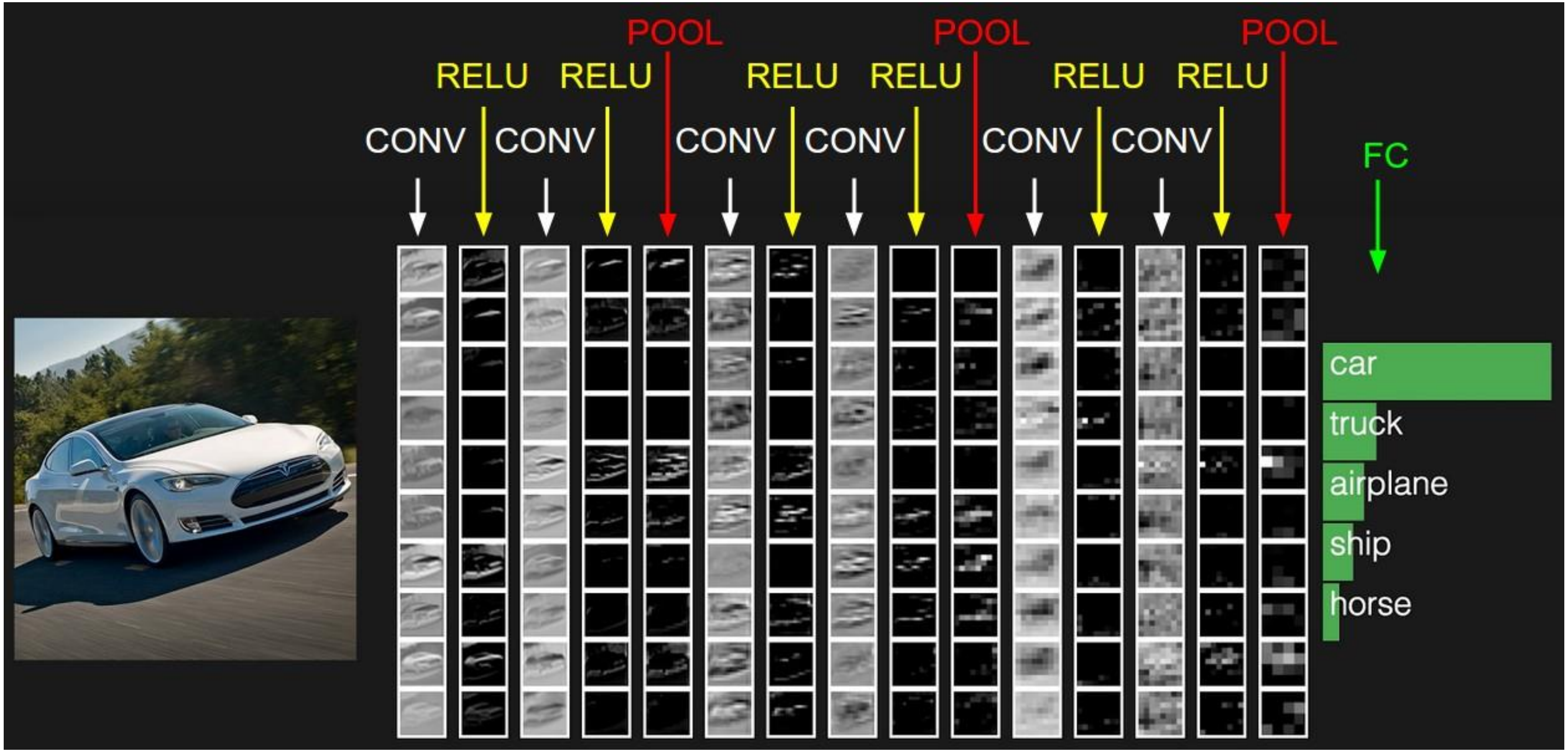
- Unlike, normal NN layers of CNN have neurons arranged in 3 dimensions.



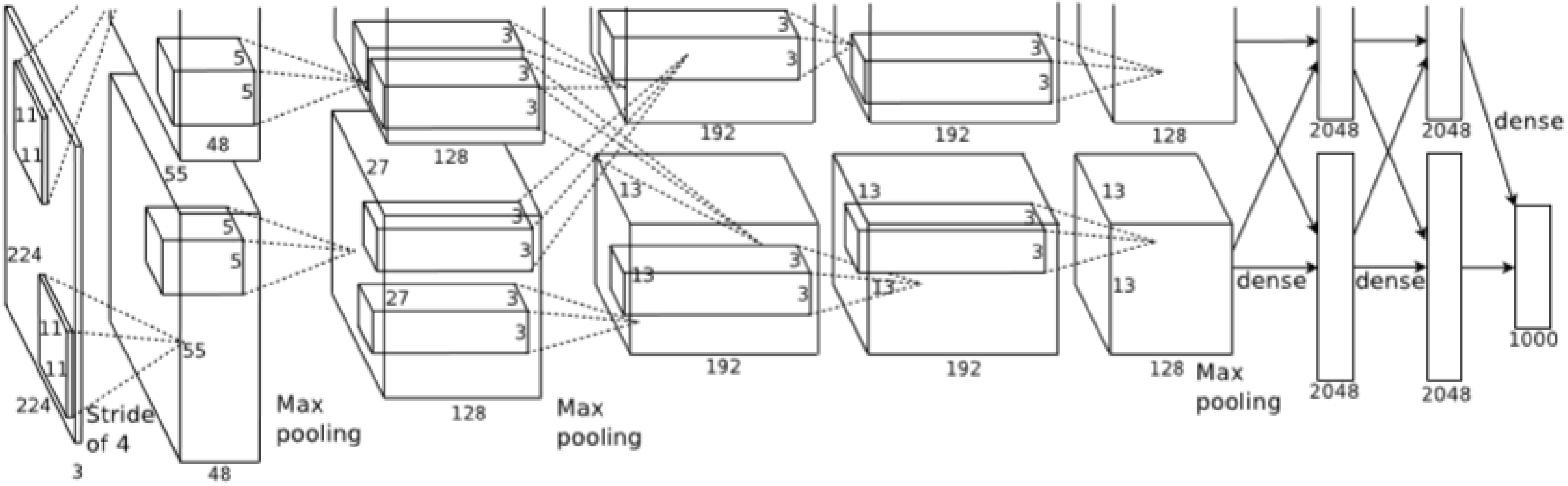
Convolutional Neural Net Architecture



Visualizing 3D volume



ImageNet Parallelization(Alex Net)



Sequential Algorithm

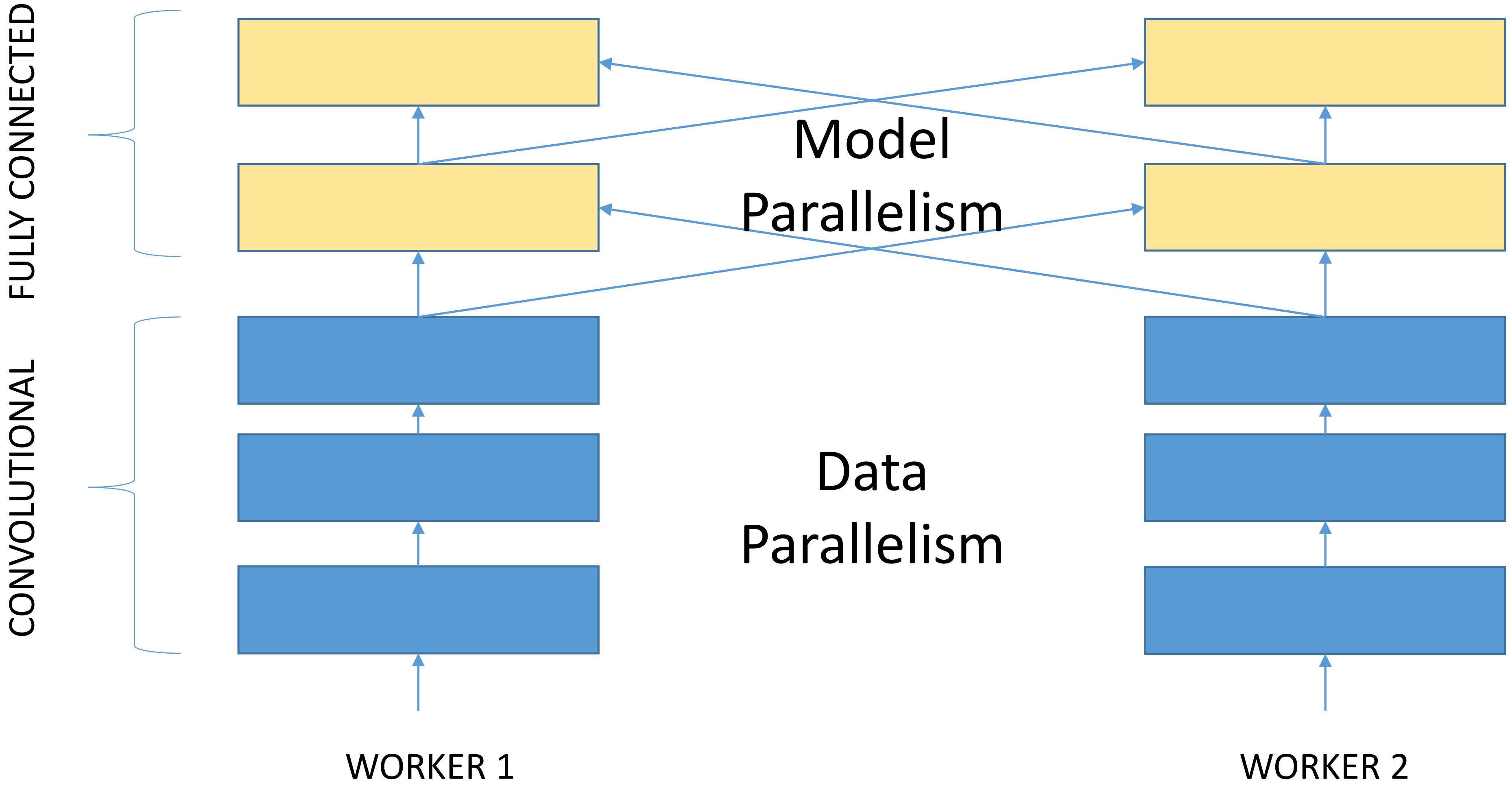
- Simple Convolutional Neural Network on CIFAR-10 images
- Input – 32*32*3, W-32, H-32, RGB
- Convolutional Layer each computing dot product between weights from the neuron and the region they are connected to in the input volume.
- RELU layer will apply element wise activation function.
- Pooling layer will perform down sampling operation around spatial dimensions.
- Fully connected layer will compute class scores resulting in a vector of classes.

Sequential Algorithm(cont’d)

- Parameters of Convolutional/FC layers will be trained using the gradient descent.
- Such that the class scores are consistent with the labels in the training set for each image.
- Training such networks takes a lot of time.

Current Approaches for Parallelizing CNN

(Data and Model Parallelism)



Parallel Algorithm

- Instead of running CNN on one GPU, we will have CNNs running on multiple GPUs.
- Every CNN will run on different batches, eg. 128 images per batch.
- Each CNN will compute values till last convolutional layers.
- Then the computation of fully connected layers will be split among CNNs running on multiple GPUs (i.e Forward and Backward pass computation)
- Need to synchronize weight updates across multiple GPUs.
- Update weights and run for another batch till CNNs converge.

Deeper the better

Thank you

Q.