

Exercise 5: Prune the network to compress structurally

This exercise introduces a new technique: *network projection*. This technique analyzes the covariance of neural excitations on layers of interest and reduces the number of learnable parameters by modifying layers to operate in a projective space. Although the operations of the layer take place in a typically lower rank projective space, the expressivity of the layer remains high as the width—i.e., the number of neural activations—remains unchanged when compared to the original network architecture. This can be used in place of or in addition to pruning and quantization.

```
%% prune the trained network with data
% assume the network to
arrds = arrayDatastore(dataTable, "OutputType","same");
mbqTrain = minibatchqueue(arrds, OutputAsDlarray=1, MiniBatchFormat="SSBC");
npca = neuronPCA(net,mbqTrain,VerbosityLevel="steps");
```

```
Using solver mode "direct".
Computing covariance matrices for activations connected to: "conv_1/in"
Computing eigenvalues and eigenvectors for activations connected to: "conv_1/in"
Computing covariance matrices for activations connected to: "conv_1/out"
Computing eigenvalues and eigenvectors for activations connected to: "conv_1/out"
Computing covariance matrices for activations connected to: "conv_2/in"
Computing eigenvalues and eigenvectors for activations connected to: "conv_2/in"
Computing covariance matrices for activations connected to: "conv_2/out"
Computing eigenvalues and eigenvectors for activations connected to: "conv_2/out"
Computing covariance matrices for activations connected to: "conv_3/in"
Computing eigenvalues and eigenvectors for activations connected to: "conv_3/in"
Computing covariance matrices for activations connected to: "conv_3/out"
Computing eigenvalues and eigenvectors for activations connected to: "conv_3/out"
Computing covariance matrices for activations connected to: "fc/in"
Computing eigenvalues and eigenvectors for activations connected to: "fc/in"
Computing covariance matrices for activations connected to: "fc/out"
Computing eigenvalues and eigenvectors for activations connected to: "fc/out"
neuronPCA analyzed 4 layers: "conv_1","conv_2","conv_3","fc"
```

```
netProjected = compressNetworkUsingProjection(net,npca,
"LearnablesReductionGoal",0.2);
```

```
Compressed network has 22.0% fewer learnable parameters.
Projection compressed 4 layers: "conv_1","conv_2","conv_3","fc"
```

```
%% test the pruned network
scoresPruned = minibatchpredict(netProjected,xTest);
YPruned = onehotdecode(scoresPruned,classNames,2);
confusionchart(label_Test,YPruned);
```

True Class	BreakTime	190		3	34
	Fist	63	69	140	11
	FlatHand	31	3	198	37
	Love	36		1	246
		BreakTime	Fist	FlatHand	Love
		Predicted Class			

```
accuracyPruned = mean(YPruned == label_Test)
```

```
accuracyPruned = 0.6620
```

```
% Export
exportONNXNetwork(netProjected, "net_Pruned.onnx");
```

Try in Dev Cloud

```
web("https://stm32ai-cs.st.com/home");
```