In [1]:

```
import numpy as np
import matplotlib.pyplot as plt
from torch import nn
import torch
from torchvision.datasets import MNIST
import torchvision.transforms as transforms
from sklearn.metrics import accuracy_score
%matplotlib inline
```

For this homework you will be using `pytorch` and `torchvision` library for neural networks and datasets. You can install them with `pip install torch torchvision` .

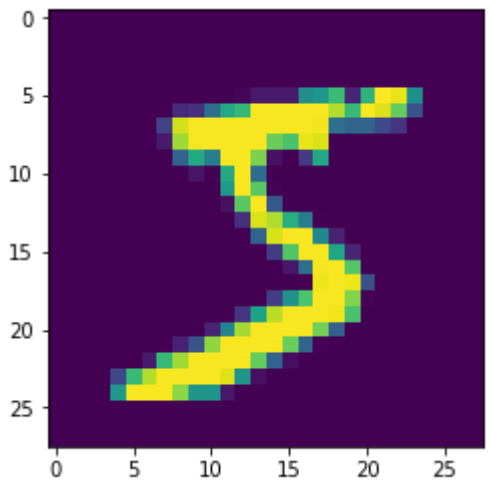
# Question 1 Principal Component Analysis

This problem will guide you through the principal component analysis. You will be using a classical dataset, the MNIST hand written digit dataset.

In [2]:

```
# Load the MNIST dataset
mnist = MNIST('.', download=True)
data = mnist.train_data.numpy()
labels = mnist.train_labels.numpy()
print('shapes:', data.shape, labels.shape)
plt.imshow(data[0])
print('label:', labels[0])
```

```
shapes: (60000, 28, 28) (60000,)
label: 5

/usr/local/lib/python3.7/site-packages/torchvision/datasets/mnist.p
y:53: UserWarning: train_data has been renamed data
  warnings.warn("train_data has been renamed data")
/usr/local/lib/python3.7/site-packages/torchvision/datasets/mnist.p
y:43: UserWarning: train_labels has been renamed targets
  warnings.warn("train_labels has been renamed targets")
```

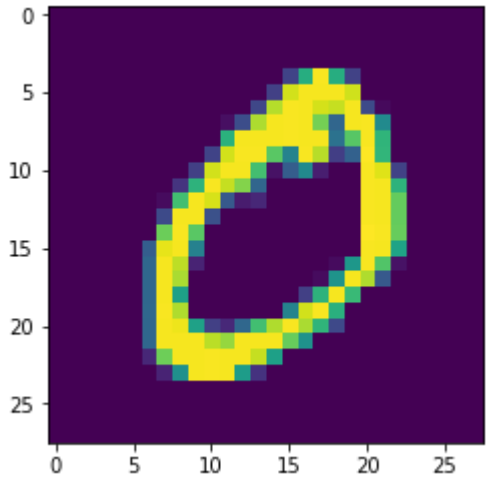


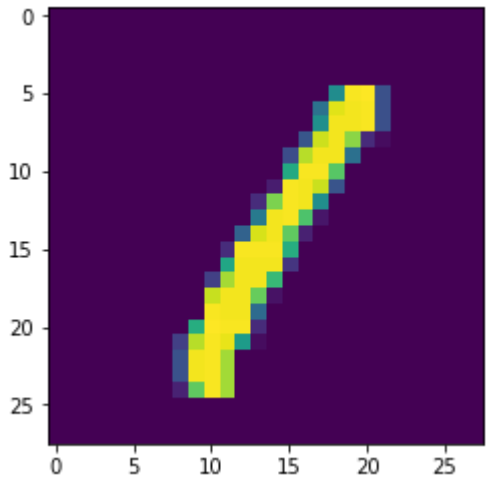
# Question 1.1 Familiarize yourself with the data [5pt]

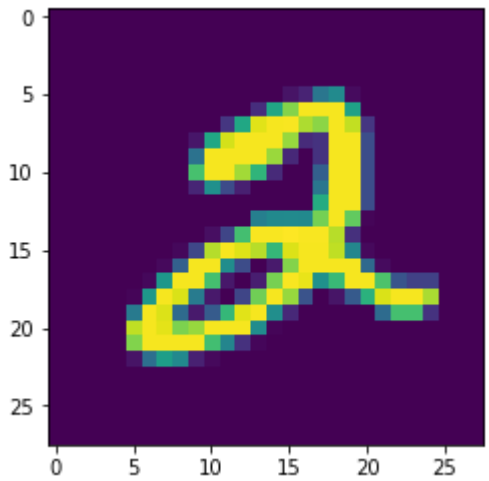
For this task, you will be using the torchvision package that provides the MNIST dataset. For each digit class(0-9), plot 1 image from the class and store those 10 images for each digit class in the array `digit_images` .

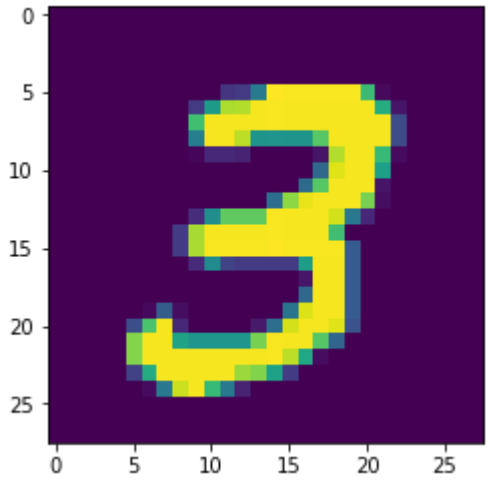
In [3]:

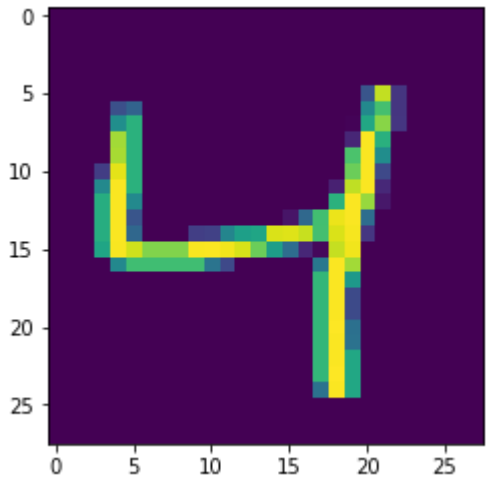
```
digit_images = np.zeros([10, 28, 28])
### YOUR CODE HERE
digit = []
for i in range(len(labels)):
    l = labels[i]
    if l not in digit:
        digit += [l]
        digit_images[l] = data[i]
    if len(digit) == 10:
        break

for i in range(10):
    plt.imshow(digit_images[i])
    plt.show()

### END OF CODE
```

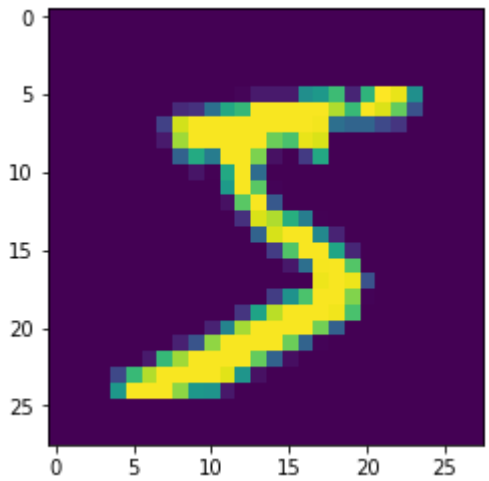

0
5
10
15
20
25
0
5
10
15
20
25

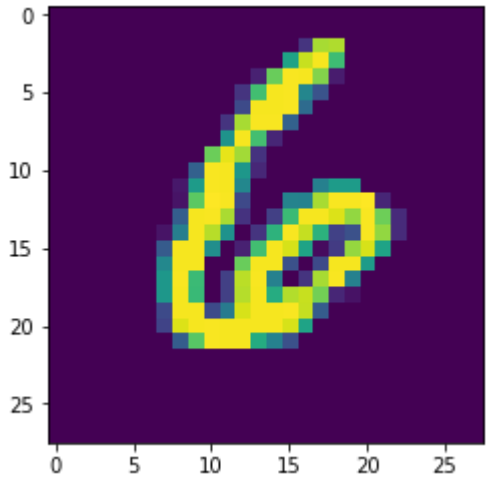

0
5
10
15
20
25
0
5
10
15
20
25

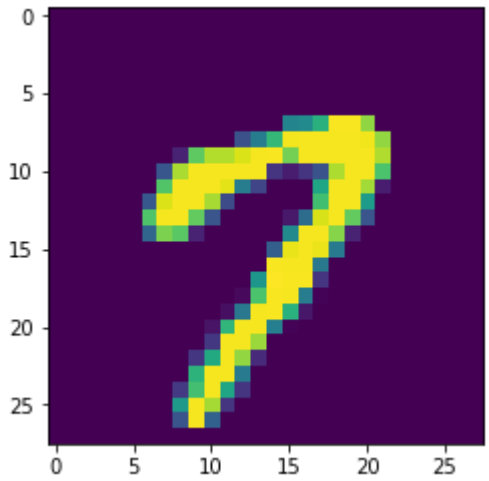

0
5
10
15
20
25
0
5
10
15
20
25

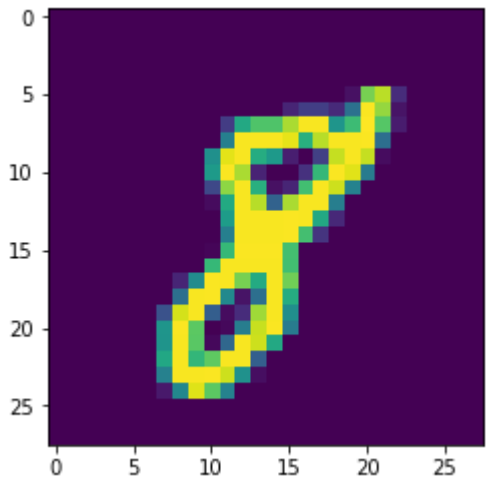

0
5
10
15
20
25
0
5
10
15
20
25

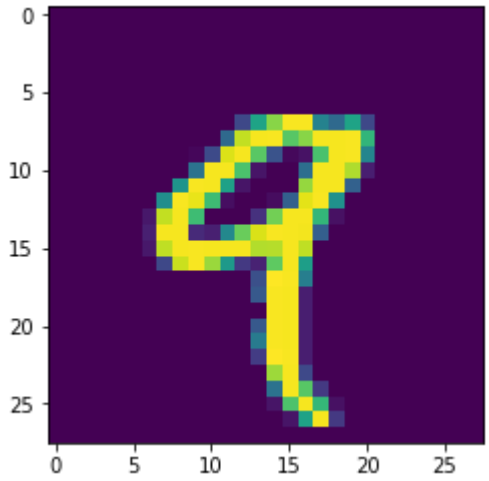

0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25




# Question 1.2 PCA

The following questions will guide you through the PCA algorithm.

## Question 1.2.1 Centering the data [5pt]

For each image, flatten it to a 1-D vector. To perform PCA on the dataset, we first move the data points so they have 0 mean on each dimension. Store the centered data in variable `data_centered` and the mean of each dimension in variable `data_mean` .

In [4]:

```
data_centered = None
data_mean = None
### YOUR CODE HERE
data_flatten = np.array([d.flatten() for d in data])
data_mean = np.mean(data_flatten, axis = 0)
data_centered = data_flatten-data_mean
### END OF CODE
```

## Question 1.2.2 Compute the covariance matrix of the data [5pt]

You need to store the covariance matrix of the data in variable `data_covmat` . You may **not** use numpy.cov

In [5]:

```
data_covmat = None
### YOUR CODE HERE
data_covmat = 1/len(data_centered)*data_centered.T @ data_centered
### END OF CODE
```

## Question 1.2.3 Compute the eigenvalues of the covariance matrix [5pt]

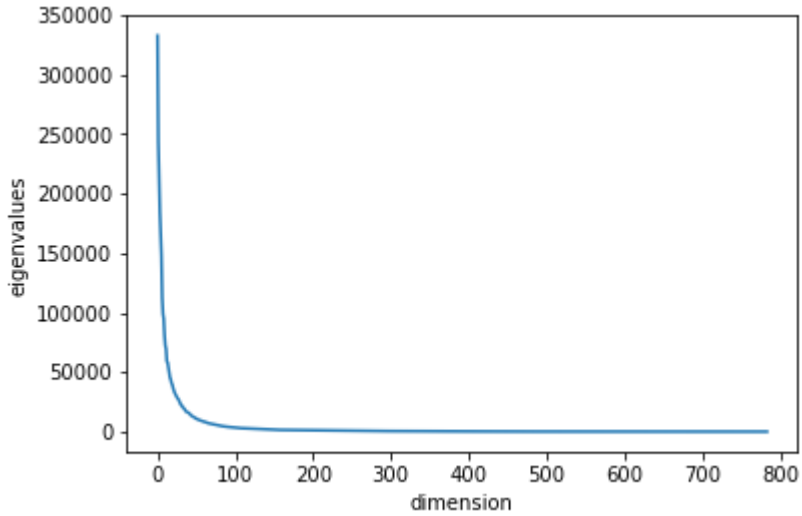
You need to store the eigenvalues of the covariance matrix in variable `covmat_eig` , sorted in descending order. Then you need to plot the eigenvalues with `plt.plot` . You can use any numpy function.

In [6]:

```
covmat_eig = None
### YOUR CODE HERE
evals, evecs = np.linalg.eigh(data_covmat)
idx = np.argsort(evals)[::-1]
evecs = evecs[:,idx]
covmat_eig = evals[idx]
plt.plot(covmat_eig)
plt.xlabel("dimension")
plt.ylabel("eigenvalues")
### END OF CODE
```

Out[6]:

```
Text(0,0.5,'eigenvalues')
```

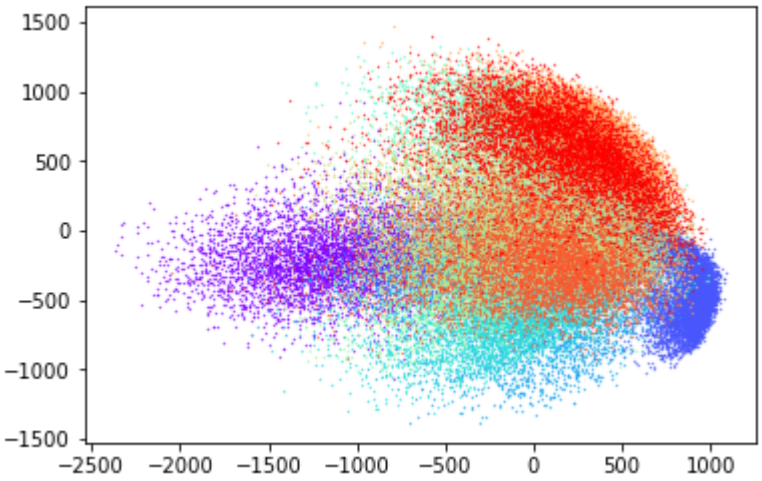



## Question 1.2.4 Project data onto the first 2 principal components [5pt]

Now you need to project the centered data on the 2D space formed by the eigenvectors corresponding to the 2 largest eigenvalues. Create a 2D scatter plot where you need to assign a unique color to each digit class.

In [7]:

```
### YOUR CODE HERE
import matplotlib.cm as cm
data_PCA = evecs[:,:2].T @ data_centered.T
colors = cm.rainbow(np.linspace(0, 1, 10))
for i in range(10):
    l = np.where(labels == i)
    plt.scatter(data_PCA[0][l],data_PCA[1][l], s = 0.1, c = colors[i])
### END OF CODE
```

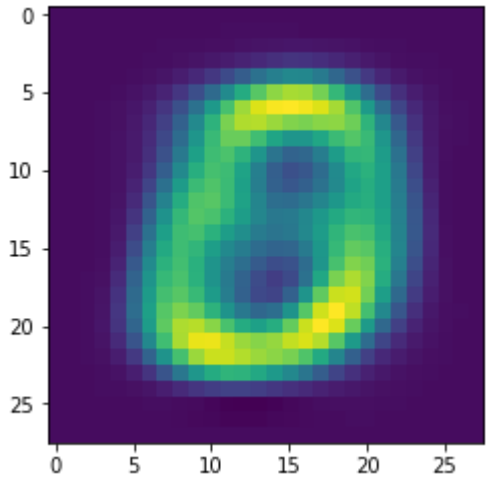


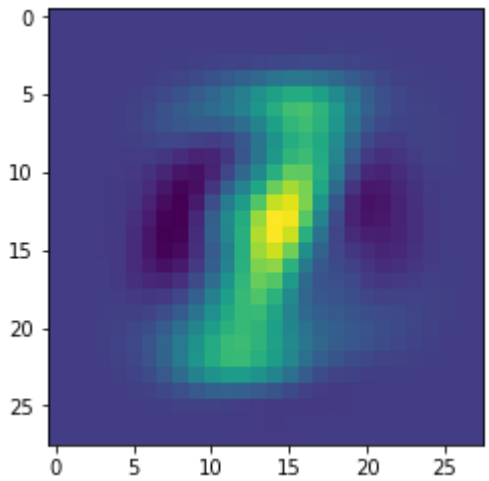

***Question 1.2.5 Unproject data back to high dimensions [10pt]***

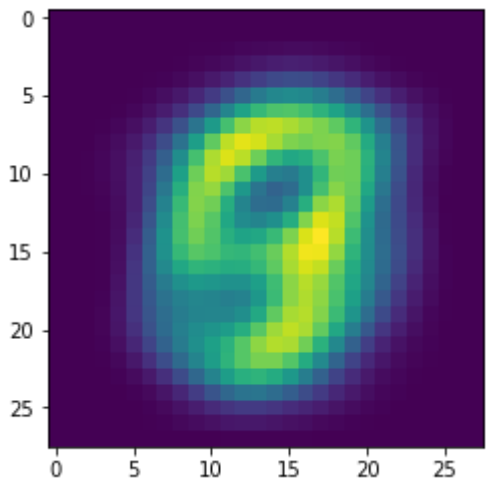
For this question, you need to project the 10 images you plotted in **1.1** on the first 2 principal components, and then unproject the "compressed" 2-D representations back to the original space. Plot the "compressed" digit (the reconstructed digit). Do they look similar to the original images?

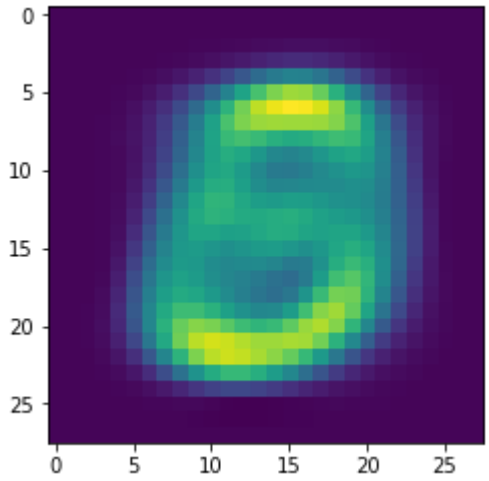
In [8]:

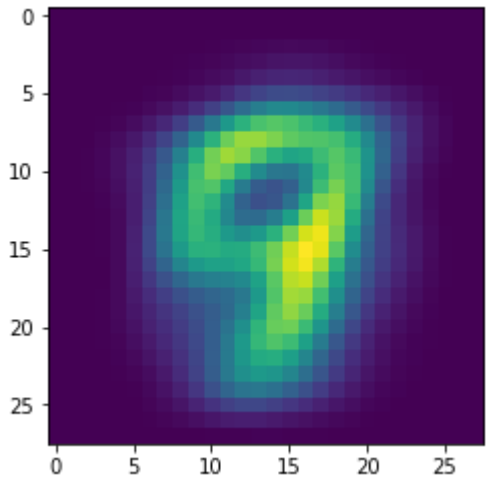
```
### YOUR CODE HERE
for i in digit_images:
    proj = evecs[:,:2].T @ (i.flatten() - data_mean)
    unproj = np.linalg.pinv(evecs[:,:2].T) @ proj + data_mean
    plt.imshow(unproj.reshape((28,28)))
    plt.show()
### END OF CODE
```

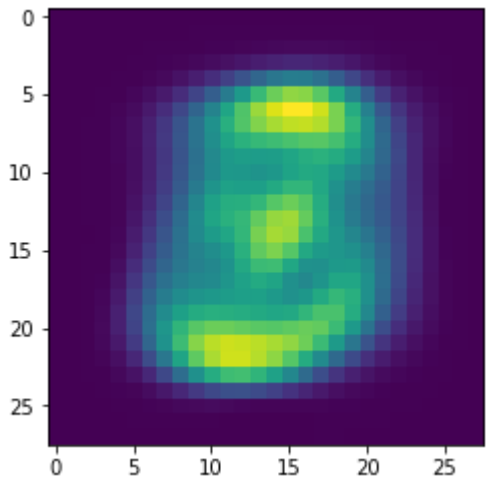

0
5
10
15
20
25
0
5
10
15
20
25

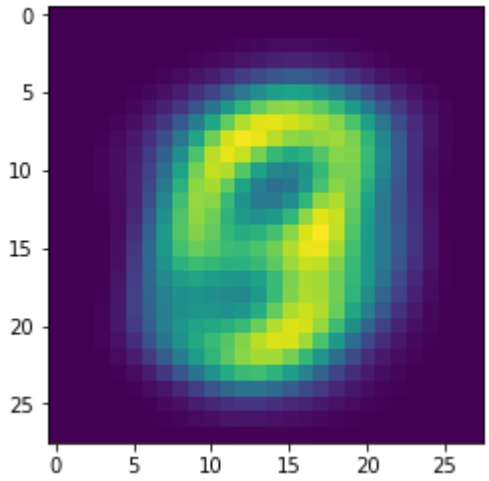

0
5
10
15
20
25
0
5
10
15
20
25

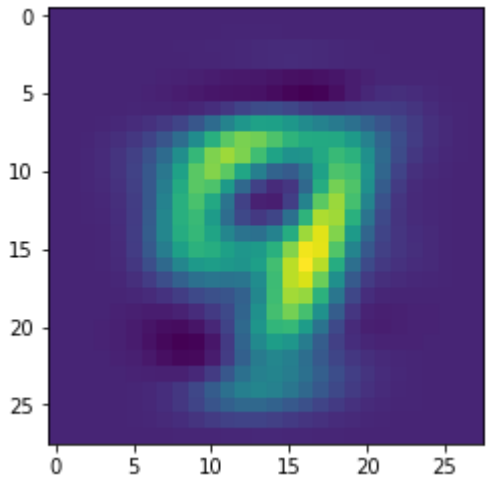

0
5
10
15
20
25
0
5
10
15
20
25

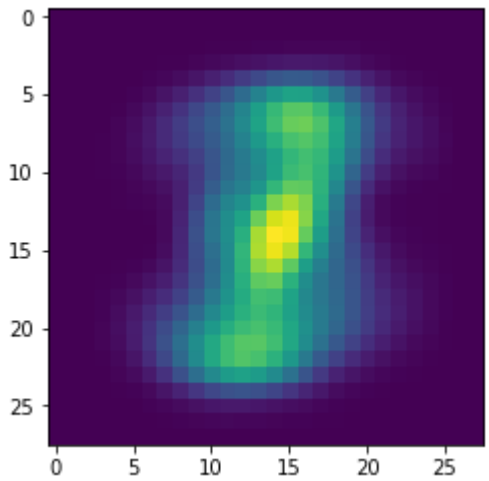

0
5
10
15
20
25
0
5
10
15
20
25

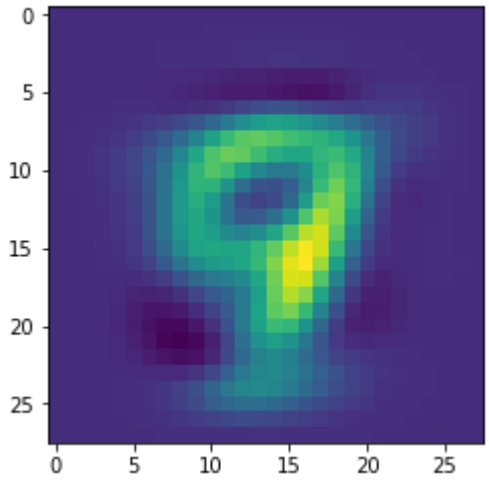

0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25



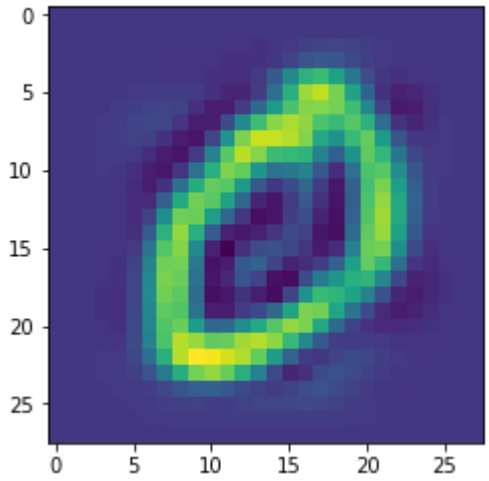

It is a little bit similar to the original images but more blurred and hard to recognize because when doing PCA(project to eigenvectors corresponding to the 2 largest eigenvalues), we cannot avoid the loss of some important information.

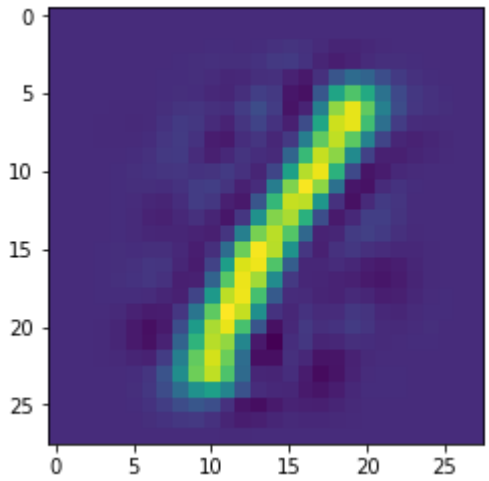
## Question 1.2.6 Choose a better low dimension space. [5pt]

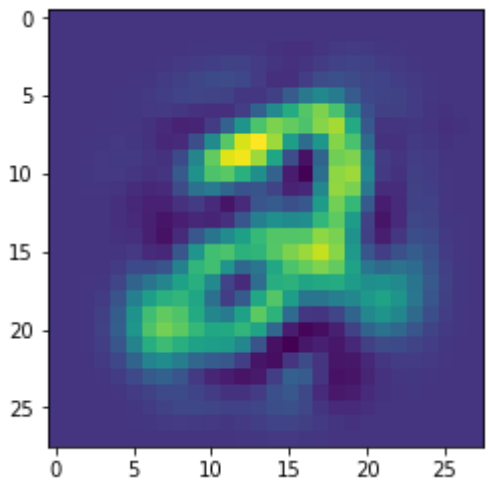
Do the previous problem with more dimensions (e.g. 3, 5, 10, 20, 50, 100). You only need to show results for one of them. Answer the following questinos. How many dimensions are required to represent the digits reasonably well? How are your results related to **question 1.2.3**?

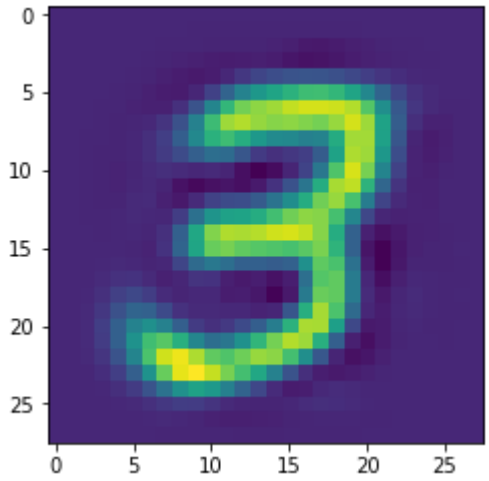
In [9]:

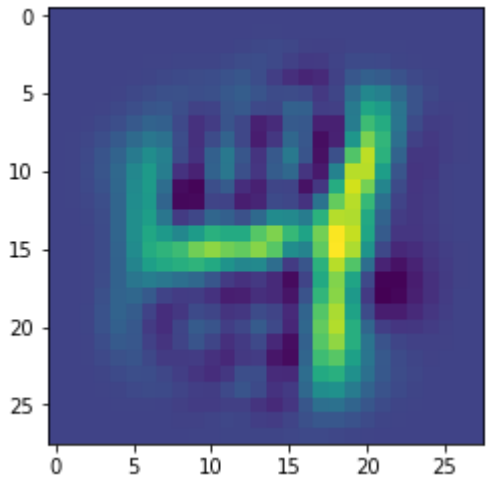
```
### YOUR CODE HERE
for i in digit_images:
    proj = evecs[:,:50].T @ (i.flatten() - data_mean)
    unproj = np.linalg.pinv(evecs[:,:50].T) @ proj + data_mean
    plt.imshow(unproj.reshape((28,28)))
    plt.show()
### END OF CODE
```

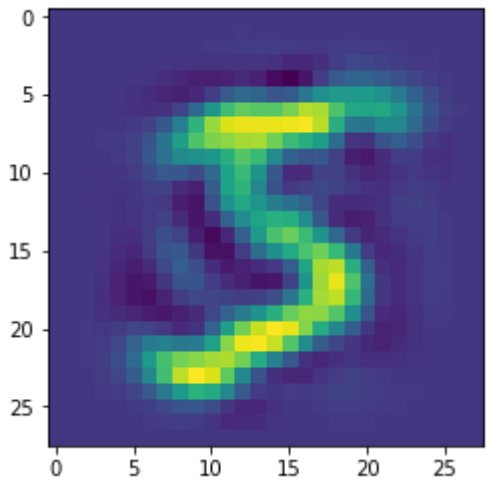

0
5
10
15
20
25
0
5
10
15
20
25

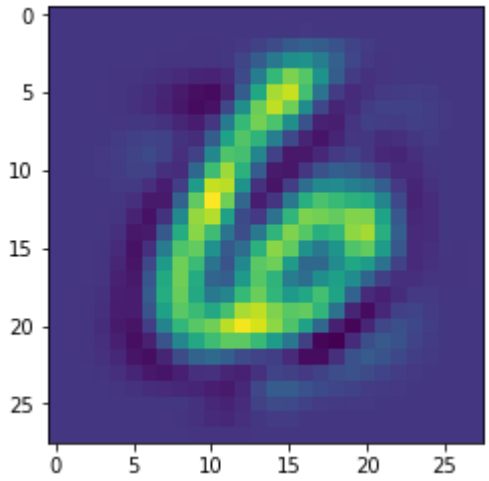

0
5
10
15
20
25
0
5
10
15
20
25

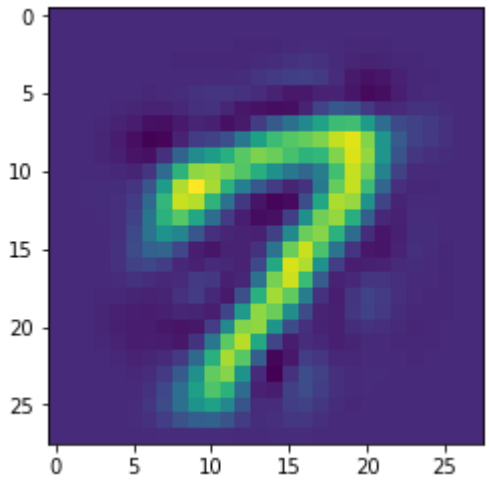

0
5
10
15
20
25
0
5
10
15
20
25

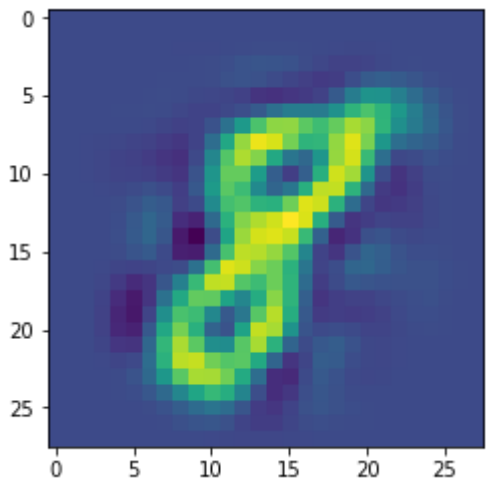

0
5
10
15
20
25
0
5
10
15
20
25

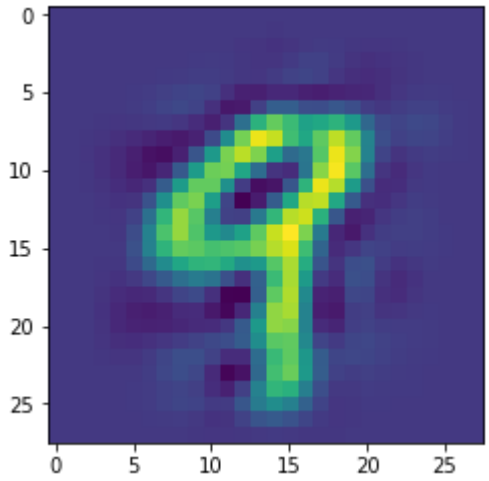

0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25




(Your explanation)
50 dimensions are required to represent the digits reasonably well.

From the plot in question 1.2.3, we can see the evigenvalues drop to 0 at around 50 dimension, which means the rest dimensions only have little effect on making the reconstructed digits look better.

# Question 1.3 Harris Corner and PCA [10pt]

Recall Harris corner detector algorithm:

1. Compute $x$ and $y$ derivatives ($I_x, I_y$) of an image
2. Compute products of derivatives ($I_x^2, I_y^2, I_{xy}$) at each pixel
3. Compute matrix $M$ at each pixel, where $$ M(x_0,y*0)* = *sum*{x,y} w(x-x_0,y-y_0) \begin{bmatrix}

```
I_{x}^2 & I_{x}I_{y} \\
I_{x}I_{y} & I_{y}^2
```

\end{bmatrix} $$ Here, we set weight $w(x,y)$ to be a box filter of size $3 \times 3$ (the box is placed centered at $(x_0, y_0)$).

In this problem, you need to show that Harris Corner detector is really just principal component analysis in the gradient space. Your explanation should answer the following quesions.

1. As we know, PCA is performed on data points. What are the data points in Harris corner detector when we think of it as a PCA?
2. What is the covariance matrix used in Harris corner detector and why it is a covariance matrix?
3. What are the principal components in Harris corner detector?
4. Briefly explain how principal components imply "cornerness".

(Your proof here)
1. The data points are x and y derivatives ($I_x, I_y$) of an image. As we are taking graidents ,we can assume $(I_x, I_y)$ as centered data points.

2. $M$ is the covariance matrix used in Harris corner detector. As $w(x, y)$ is the box filter which is similar to give us an average around the neighbor pixels, we can denote is as $\frac{1}{N}$ , N is the number of pixels of the images. Then $M$ can be represented as $\frac{1}{N}X^TX$ where $X = [I_x, I_y]$ for all pixels $(x, y)$ of the image.

3. The principal components in Harris corner detector are the eigenvectors of matrix $M$ corresponding to the 2 most siginificant eigenvalues.(Actually M has 2 eigenvalues in total)

4. If the 2 eigenvalues corrsponding to the principal components are very close to each other and are large, then it is corner at $(x_0, y_0)$.
If the 2 eigenvalues corrsponding to the principal components have a great difference between them, then it is edge at $(x_0, y_0)$.

If the 2 eigenvalues corrsponding to the principal components are small, then it is flat at $(x_0, y_0)$.

# Question 2 KNN, Softmax Regression

In [10]:

```
train_dataset = MNIST(root='.', train=True, transform=transforms.ToTensor, downl
oad=True)
test_dataset = MNIST('.', train=False, transform=transforms.ToTensor())
train_X = train_dataset.data.numpy()  # training data, uint8 type to reduce memo
ry and comparison cost
train_y = train_dataset.targets.numpy()  # training label
test_X = test_dataset.data.numpy() # testing data, uint8 to reduce memory and co
mparison cost
test_y = test_dataset.targets.numpy()  # testing label
```

In [11]:

```
train_X = train_X.reshape((train_X.shape[0], -1))
test_X = test_X.reshape((test_X.shape[0], -1))
```

# Question 2.1 K-Nearest Neighbor [10pt]

In this problem you will be implementing the KNN classifier. Fill in the functions in the starter code below. You are are allowed to use `scipy.spatial.KDTree` and `scipy.stats.mode` (in case of a tie, pick any one). Please avoid `sklearn.neighbors.KDTree` as it appears extremely slow. You are **not** allowed to use a library KNN function that directly solves the problem.

If you do not know what a KD-tree is, please read the documentation for `scipy.spatial.KDTree` to understand how you can use it.

Note: if you run into memory issues or neighbor queries run for more than 10 minutes, you are allowed to reduce the data size, and explain what you have done to the training data.

In [12]:

```
from scipy.spatial import KDTree
from scipy.stats import mode
```

In [13]:

```
class KNNClassifier:
    def __init__(self, num_neighbors):
        """
        construct the classifier
        Args:
            num_centers: number of neighbors
        """
        ### YOU CODE HERE
        self.num_neighbors = num_neighbors
        ### END OF CODE

    def fit(self, X, y):
        """
        train KNN classifier
        Args:
            X: training data, numpy array with shape (Nxk) where N is number of
 data points, k is number of features
            y: training labels, numpy array with shape (N)
        """
        ### YOU CODE HERE
#         kdtree  = KDTree(X, leafsize=100)
        self.train = X
        self.target = y

        ### END OF CODE
        return self

    def predict(self, X):
        """
        predict labels
        Args:
            X: testing data, numpy array with shape (Mxk) where M is number of d
ata points, k is number of features
        Return:
            y: predicted labels, numpy array with shape (N)
        """
        pred = None

        ### YOU CODE HERE
        pred = []
        kdtree = KDTree(self.train, leafsize=100)
        for x in X:
            d, knn_i = kdtree.query(x, self.num_neighbors)
            knn_t = self.target[knn_i]
            m, c = mode(knn_t,axis = None)
            pred += [m]
        pred = np.array(pred)
        ### END OF CODE
        return pred
```

In [14]:

```
from sklearn.metrics import accuracy_score
knn = KNNClassifier(3).fit(train_X, train_y)
pred_y = knn.predict(test_X)
print('KNN accuracy:', accuracy_score(test_y, pred_y))
```

```
/usr/local/lib/python3.7/site-packages/scipy/spatial/kdtree.py:388:
RuntimeWarning: overflow encountered in ubyte_scalars
  sd[node.split_dim] = np.abs(node.split-x[node.split_dim])**p


KNN accuracy: 0.5135
```

# Question 2.2 Softmax Regression

In this problm, you will be implementing the softmax regression(multi-class logistic regression). Here is a brief recap of several important concepts. In the following explanation, I will use $x$ for data vector, $y'$ for ground truth label, and $y$ for predicted label.

Suppose we have a problem where we need to classify data points into $m$ classes.

1. Softmax function $S$ normalize a vector to have sum 1. (it turns any vector into a probability distribution)
$$S(x) = [\frac{e^{x_1}}{\sum_{j=1}^{m} e^{x_j}}, \frac{e^{x_2}}{\sum_{j=1}^{m} e^{x_j}}, \dots, \frac{e^{x_m}}{\sum_{j=1}^{m} e^{x_j}}]$$
2. Cross entropy loss $J$ is the multiclass logistic regression loss.
$$J(y', y) = -\sum_{i=1}^{m} y'_i \log y_i$$
where $y'$ is the one-hot ground truth label and $y$ is the predicted label distribution.
3. Softmax regression is the following optimization problem.
$$\min_{W,b} \sum_{(X,y')\in\{\text{training set}\}} J(y', S(Wx+b))$$
where $W$ has shape $(m \times k)$ where $k$ is the number of features in a data point; $b$ is a $m$ dimensional vector.
4. This objective is optimized with gradient descent. Let
$$L = \sum_{(x,y')\in\{\text{training set}\}} J(y', S(Wx+b))$$
Update $W$ and $b$ with $\frac{\partial L}{\partial W}$ and $\frac{\partial L}{\partial b}$.

## Question 2.2.1 Compute the gradients [10pt]

In this question, you need to do the following:

1. Compute the gradient $\frac{\partial J}{\partial y}$. i.e. compute
$$\frac{\partial J}{\partial y_i}$$
Express it in terms of $y_i'$ and $y_i$.
2. Let $u = Wx + b$, $y_i = S_i(u)$ Compute
$$\frac{\partial y_i}{\partial u_j}$$
Express it in terms of $y_i, y_j$ and $\delta_{ij}$, where
$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$
3. Compute
$$\frac{\partial J}{\partial W_{jk}} \text{ and } \frac{\partial J}{\partial b_j}$$
Express them in terms of $y_j, y_j', x_k$. Explain your results in an intuitive way. Hint: the results should have a very simple form that makes sense intuitively.
4. Compute
$$\frac{\partial J}{\partial W}$$
in the matrix form. It should be a matrix with the same shape as $W$, and entry $jk$ is $\frac{\partial J}{\partial W_{jk}}$. Similarly, compute
$$\frac{\partial J}{\partial b}$$

(Your proof here)

1.

$$\frac{\partial J}{\partial y_i} = -\frac{\partial \sum_{k=1}^{m} y'_k \log(y_k)}{\partial y_i} = -\sum_{k=1}^{m} y'_k \frac{\partial \log(y_k)}{\partial y_i} = -\frac{y'_i}{y_i}$$

2.

$$\frac{\partial y_i}{\partial u_j} = \frac{\partial S_i(u)}{\partial u_j} = \frac{\partial \frac{e^{u_i}}{\sum_{k=1}^{m} e^{u_k}}}{\partial u_j} = \frac{\frac{\partial e^{u_i}}{\partial u_j} \sum_{k=1}^{m} e^{u_k} - e^{u_i} e^{u_j}}{(\sum_{k=1}^{m} e^{u_k})^2} = \frac{\delta_{ij} e^{u_i} \sum_{k=1}^{m} e^{u_k} - e^{u_i} e^{u_j}}{(\sum_{k=1}^{m} e^{u_k})^2} = \frac{e^{u_i}}{\sum_{k=1}^{m} e^{u_k}} \frac{\delta_{ij} \sum_{k=1}^{m} e^{u_k} - e^{u_j}}{\sum_{k=1}^{m} e^{u_k}} = y_i(\delta_{ij} - y_j)$$

3.

Since y' is one hot encoding, $\sum_{i=1}^{m} y'_i = 1$

$$\frac{\partial J}{\partial W_{jk}} = \frac{\partial J}{\partial y} \frac{\partial y}{\partial u_j} \frac{\partial u_j}{\partial W_{jk}} = -\sum_{i=1}^{m} \frac{y'_i}{y_i} y_i(\delta_{ij} - y_j)x_k = -\sum_{i=1}^{m} y'_i(\delta_{ij} - y_j)x_k = -(\sum_{i=1}^{m} y'_i \delta_{ij} x_k - \sum_{i=1}^{m} y'_i y_j x_k) =$$

$$\frac{\partial J}{\partial b_j} = \frac{\partial J}{\partial y} \frac{\partial y}{\partial u_j} \frac{\partial u_j}{\partial b_j} = -\sum_{i=1}^{m} \frac{y'_i}{y_i} y_i(\delta_{ij} - y_j)1 = -\sum_{i=1}^{m} y'_i(\delta_{ij} - y_j) = -(\sum_{i=1}^{m} y'_i \delta_{ij} - \sum_{i=1}^{m} y'_i y_j) = -y'_j + y_j$$

Explanation: $W_{jk}x_k$ comprise the jth component of our predicted label $y_j$, and cross entropy $J$ depends on the difference of our predicted label $y$ and the truth label $y'$, then $\frac{\partial J}{\partial W_{jk}}$ should just depends on both $y_j - y'_j$ and $x_k$ , which results in $x_k(y_j - y'_j)$. Similarly, $b_j$ (a constant) comprise the jth component of predicted label $y_j$, and $J$ depends on $(y - y')$, so $\frac{\partial J}{\partial b_j}$ should just depends on $(y_j - y'_j) * 1$, which is $(y_j - y'_j)$.

4.

$$\frac{\partial J}{\partial W} = \begin{bmatrix} x_0(y_0 - y'_0) & x_1(y_0 - y'_0) & \cdots & x_k(y_0 - y'_0) \\ x_0(y_1 - y'_1) & x_1(y_1 - y'_1) & \cdots & x_k(y_1 - y'_1) \\ x_0(y_2 - y'_2) & x_1(y_2 - y'_2) & \cdots & x_k(y_2 - y'_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_0(y_m - y'_m) & x_1(y_m - y'_m) & \cdots & x_k(y_m - y'_m) \end{bmatrix} \qquad \frac{\partial J}{\partial b} = \begin{bmatrix} y_0 - y'_0 \\ y_1 - y'_1 \\ y_2 - y'_2 \\ \vdots \\ y_m - y'_m \end{bmatrix}$$

## Question 2.2.2 Stochastic Gradient Descent [10pt]

In gradient descent algorithm, we update $W$ and $b$ with $\partial L/\partial W$ and $\partial L/\partial b$. However, this requires the gradient w.r.t. the whole dataset. Computing such gradient is very slow. Instead, we can update the weights with per-data gradient. This is known as the SGD algorithm, which runs much faster. You need to take the following steps.

1. Implement softmax function $S$. We need to take special care in this function since $e^x$ tends to overflow easily. However, we observe that $S(x) = S(x - m)$ for any constant vector $m$. We can stabilize softmax using $S(x) = S(x - \max(x))$.
2. Implement function `J` (loss) and `dJ` (loss gradient). Note: `J` is not required to run the algorithm, but you may want to implement it for debug purposes.
3. Implement the SGD algorithm.
4. Run the algorithm for 20 epochs (each epoch iterates the whole data set once) with learning rate `1e-3` and report accuracy on test set. You may use `sklearn.metrics.accuracy_score`. You need to achieve accuracy > 90%. You are allowed to experiment with different epoch numbers and learning rates (even learning rate decay) to achieve this accuracy, but they are not required.

You may use print (or progress bar packages) to track the training progress since it might take several minutes.

In [15]:

```
train_dataset = MNIST(root='.', train=True, transform=transforms.ToTensor, downl
oad=True)
test_dataset = MNIST('.', train=False, transform=transforms.ToTensor())
train_X = train_dataset.data.numpy() / 255.  # normalize data to 0-1
train_y = train_dataset.targets.numpy()  # training label
test_X = test_dataset.data.numpy() / 255.  # normalize data to 0-1
test_y = test_dataset.targets.numpy()  # testing label
train_X = train_X.reshape((train_X.shape[0], -1))  # flatten the image
test_X = test_X.reshape((test_X.shape[0], -1))  # flatten the image
```

In [16]:

```
def softmax(x):
    """
    softmax function
    Args:
        x: a 1-d numpy array
    Return:
        results of softmax(x)
    """
    ### YOUR CODE HERE
    m = len(x)
    max_x = np.array([np.max(x)]*m)
    softmax = np.exp(x-max_x)/np.sum(np.exp(x-max_x))
    return softmax
    ### END OF CODE

def J(W, b, y_true, x):
    """
    Softmax Loss function
    Args:
        W: weights (num_classes x num_features)
        b: bias (num_classes)
        y_true: ground truth 1-hot label (num_classes)
        x: input data
    Return:
        J(y', y)
    """
    ### YOUR CODE HERE
    num_classes, num_features = W.shape
    u = W @ x + b
    y = softmax(u)
    J = - y_true @ np.log(y)
    return J
    ### END OF CODE

def dJ(W, b, y_true, x):
    """
    Softmax Loss gradient
    Args:
        W: weights (num_classes x num_features)
        b: bias (num_features)
        y_true: ground truth 1-hot label (num_classes)
        x: input data (num_features)
    Return:
        (dW, db): gradient w.r.t. W and b
    """
    ### YOUR CODE HERE
    num_classes, num_features = W.shape
    u = W @ x + b
    y = softmax(u)
    dW = (y - y_true)[:,None] @ x[None,:]
    db = (y - y_true)
    return (dW,db)
    ### END OF CODE
```

In [17]:

```
from tqdm import tqdm_notebook
```

In [18]:

```
def SGD(f, df, Xs, ys, n_classes=10, lr=1e-3, max_epoch=20):
    """
    Args:
        f: function to optimize
        df: the gradient of the function
        Xs: input data, numpy array with shape (num_data x num_features)
        ys: true label, numpy array with shape (num_data x num_classes)
        lr: learning rate
        max_epoch: maximum epochs to run SGD
    Return:
        optimal weights and biases
    """
    N, m = Xs.shape
    W = np.random.rand(n_classes, m) - 0.5  # you do not need to change random i
nitialization
    b = np.random.rand(n_classes) - 0.5

    ### YOUR CODE HERE
    for epoch in range(max_epoch):
        for n in range(N):
            delta_W, delta_b = df(W, b, ys[n], Xs[n])
            W = W - lr*delta_W
            b = b - lr*delta_b

    ### END OF CODE
    return W, b
```

In [19]:

```
train_y_onehot = np.zeros((train_y.shape[0], 10))
train_y_onehot[np.arange(len(train_y)), train_y] = 1
W, b = SGD(J, dJ, train_X, train_y_onehot, 10,  max_epoch=20)
accuracy_score(test_y, np.argmax(test_X @ W.T + b, axis=1))
```

Out[19]:

```
0.919
```

# Question 3 Convolutional Neural Networks

This question requires you to use the PyTorch framework for neural network training. You will not need GPU to train the networks for this problem.

The following is a code sample for training a simple multi-layer perceptron neural network using PyTorch. Running it should give you about 98% testing accuracy.

Since network training takes long, I recommend installing the tqdm package for progress tracking.

In [20]:

```
from tqdm import tqdm_notebook
```

In [21]:

```
train_dataset = MNIST(root='.', train=True, transform=transforms.ToTensor(), dow
nload=True)
test_dataset = MNIST('.', train=False, transform=transforms.ToTensor())
```

In [22]:

```
class MLP(nn.Module):
    def __init__(self, input_size, hidden_size, num_classes):
        """init function builds the required layers"""
        super(MLP, self).__init__()  # This line is always required
        # Hidden layer
        self.layer1 = nn.Linear(input_size, hidden_size)
        # activation
        self.relu = nn.ReLU()
        # output layer
        self.layer2 = nn.Linear(hidden_size, num_classes)

    def forward(self, x):
        """forward function describes how input tensor is transformed to output
 tensor"""
        # flatten the input from (Nx1x28x28) to (Nx784)
        torch.flatten(x, 1)
        x = self.layer1(x)
        x = self.relu(x)
        x = self.layer2(x)
        # Note we do not need softmax layer, since this layer is included in the
CrossEntropyLoss provided by torch
        return x
```

In [23]:

```
model = MLP(784, 1024, 10)
model
```

Out[23]:

```
MLP(
  (layer1): Linear(in_features=784, out_features=1024, bias=True)
  (relu): ReLU()
  (layer2): Linear(in_features=1024, out_features=10, bias=True)
)
```

In [24]:

```
opts = {
    'lr': 5e-4,
    'epochs': 5,
    'batch_size': 64
}
```

In [25]:

```
optimizer = torch.optim.Adam(model.parameters(), opts['lr'])  # Adam is a much b
etter optimizer compared to SGD
criterion = torch.nn.CrossEntropyLoss()  # loss function
train_loader = torch.utils.data.DataLoader(dataset=train_dataset, batch_size=opt
s['batch_size'], shuffle=True)
test_loader = torch.utils.data.DataLoader(dataset=test_dataset, batch_size=opts[
'batch_size'], shuffle=True)
```

In [26]:

```
for epoch in range(opts['epochs']):
    train_loss = []
    for i, (data, labels) in tqdm_notebook(enumerate(train_loader), total=len(tr
ain_loader)):
        # reshape data
        data = data.reshape([-1, 784])
        # pass data through network
        outputs = model(data)
        loss = criterion(outputs, labels)
        optimizer.zero_grad()  # Important! Otherwise the optimizer will accumul
ate gradients from previous runs!
        loss.backward()
        optimizer.step()
        train_loss.append(loss.item())
    test_loss = []
    test_accuracy = []
    for i, (data, labels) in enumerate(test_loader):
        # reshape data
        data = data.reshape([-1, 784])
        # pass data through network
        outputs = model(data)
        _, predicted = torch.max(outputs.data, 1)
        loss = criterion(outputs, labels)
        test_loss.append(loss.item())
        test_accuracy.append((predicted == labels).sum().item() / predicted.size
(0))
    print('epoch: {}, train loss: {}, test loss: {}, test accuracy: {}'.format(e
poch, np.mean(train_loss), np.mean(test_loss), np.mean(test_accuracy)))
```

```
/usr/local/lib/python3.7/site-packages/ipykernel_launcher.py:3: Tqdm
DeprecationWarning: This function will be removed in tqdm==5.0.0
Please use `tqdm.notebook.tqdm` instead of `tqdm.tqdm_notebook`
  This is separate from the ipykernel package so we can avoid doing
imports until


epoch: 0, train loss: 0.287885642002847, test loss: 0.14951454556197
12, test accuracy: 0.956906847133758


epoch: 1, train loss: 0.11411322941527025, test loss: 0.094398773892
24281, test accuracy: 0.9713375796178344


epoch: 2, train loss: 0.07259026515349222, test loss: 0.077498968767
74966, test accuracy: 0.9766122611464968


epoch: 3, train loss: 0.05160615372700867, test loss: 0.067946167959
3393, test accuracy: 0.9793988853503185


epoch: 4, train loss: 0.0377031400864785, test loss: 0.0626822245428
8726, test accuracy: 0.9807921974522293
```

# Question 3.1 Implementing CNN [15pt]

You need to implement a convolutional neural network for the same task as above. You may find the PyTorch documentation helpful. https://pytorch.org/docs/stable/nn.html (https://pytorch.org/docs/stable/nn.html)

We provide a working network structure below. You can adjust the network size and training options for better performance, but a correct implementation of the provided network should give you the required accuracy. For convolutional layers, (conv MxM, N) means the layer has kernel size $M$ by $M$ and $N$ output channels; for pooling layers, (maxpool MxM) means doing max pooling with kernel size $M$ by $M$.

(conv 5x5, 32) -> (relu) -> (maxpool 2x2) -> (conv 5x5, 64) -> (relu) -> (maxpool 2x2) -> (flatten) -> (linear 10) -> (output)

For full score, you need to achieve 99% testing accuracy. Also, plot the hand-written digits that your network got wrong.

In [27]:

```
class CNN(nn.Module):
    def __init__(self, input_size, num_classes):
        """
        init convolution and activation layers
        Args:
            input_size: (1,28,28)
            num_classes: 10
        """
        super(CNN, self).__init__()
        ### YOUR CODE HERE

        self.num_classes = num_classes
        self.conv1 = nn.Conv2d(input_size[0], 32, (5,5))
        self.relu = nn.ReLU()
        self.maxpool = nn.MaxPool2d((2,2))
        self.conv2 = nn.Conv2d(32 ,64, (5,5))
        self.flatten = nn.Flatten()
        self.linear = nn.Linear(1024,num_classes)
        ### END OF CODE


    def forward(self, x):
        """
        forward function describes how input tensor is transformed to output ten
sor
        Args:
            x: (Nx1x28x28) tensor
        """
        ### YOUR CODE HERE
#         torch.flatten(x, 1)
        x = self.conv1(x)
        x = self.relu(x)
        x = self.maxpool(x)

        x = self.conv2(x)
        x = self.relu(x)
        x = self.maxpool(x)

        x = self.flatten(x)
        x = self.linear(x)

        ### END OF CODE
        return x
```

In [28]:

```
model = CNN((1, 28, 28), 10)
model
```

Out[28]:

```
CNN(
  (conv1): Conv2d(1, 32, kernel_size=(5, 5), stride=(1, 1))
  (relu): ReLU()
  (maxpool): MaxPool2d(kernel_size=(2, 2), stride=(2, 2), padding=0,
dilation=1, ceil_mode=False)
  (conv2): Conv2d(32, 64, kernel_size=(5, 5), stride=(1, 1))
  (flatten): Flatten()
  (linear): Linear(in_features=1024, out_features=10, bias=True)
)
```

In [29]:

```
### You may (and should) change these
opts = {
    'lr': 1e-3,
    'epochs': 20,
    'batch_size': 64
}

### if you cannot get 99% with SGD, Adam optimizer can help you
optimizer = torch.optim.Adam(model.parameters(), opts['lr'])
```

In [30]:

```
criterion = torch.nn.CrossEntropyLoss()  # loss function
train_loader = torch.utils.data.DataLoader(dataset=train_dataset, batch_size=opt
s['batch_size'], shuffle=True)
test_loader = torch.utils.data.DataLoader(dataset=test_dataset, batch_size=opts[
'batch_size'], shuffle=True)
```

In [31]:

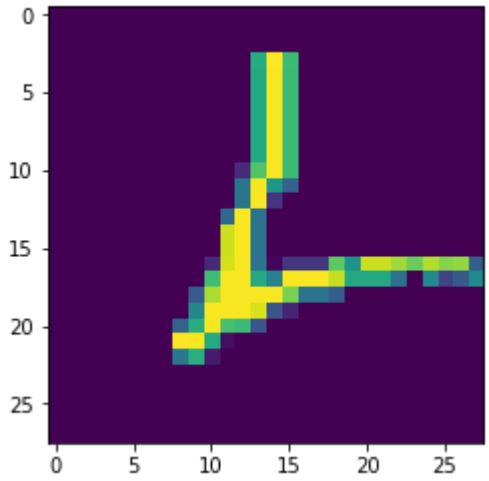
```
for epoch in range(opts['epochs']):
    train_loss = []
    for i, (data, labels) in tqdm_notebook(enumerate(train_loader), total=len(tr
ain_loader)):
        # pass data through network
        outputs = model(data)
        loss = criterion(outputs, labels)
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()
        train_loss.append(loss.item())
    test_loss = []
    test_accuracy = []
    error = []
    n = 0
    for i, (data, labels) in enumerate(test_loader):
        # pass data through network
        outputs = model(data)
        _, predicted = torch.max(outputs.data, 1)
        loss = criterion(outputs, labels)
        test_loss.append(loss.item())
        test_accuracy.append((predicted == labels).sum().item() / predicted.size
(0))
        error_idx = np.where(predicted != labels)[0]
        if len(error_idx) > 0:
            n += len(error_idx)
            error += [data[e][0] for e in error_idx]
    print('epoch: {}, train loss: {}, test loss: {}, test accuracy: {}'.format(e
poch, np.mean(train_loss), np.mean(test_loss), np.mean(test_accuracy)))
```

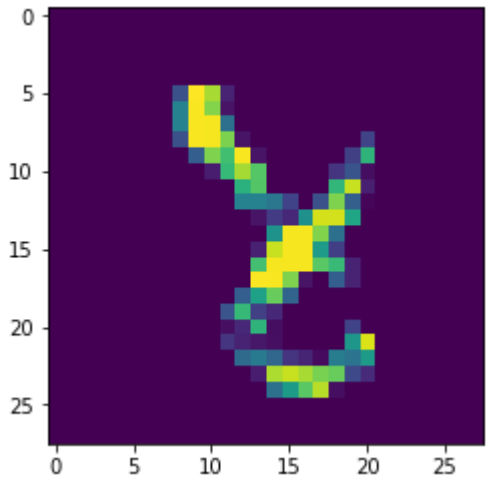
```
/usr/local/lib/python3.7/site-packages/ipykernel_launcher.py:3: TqdmDeprecationWarning: This function will be removed in tqdm==5.0.0
Please use `tqdm.notebook.tqdm` instead of `tqdm.tqdm_notebook`
  This is separate from the ipykernel package so we can avoid doing imports until


epoch: 0, train loss: 0.1647081941116486, test loss: 0.05560216813045702, test accuracy: 0.9829816878980892


epoch: 1, train loss: 0.04915811687170256, test loss: 0.034321922234965455, test accuracy: 0.9887539808917197


epoch: 2, train loss: 0.036046947290471046, test loss: 0.03015227345000822, test accuracy: 0.9894506369426752


epoch: 3, train loss: 0.026050088006236937, test loss: 0.030086378129807176, test accuracy: 0.9891520700636943


epoch: 4, train loss: 0.021646494667149033, test loss: 0.027626315975579457, test accuracy: 0.990843949044586


epoch: 5, train loss: 0.01745964811417584, test loss: 0.02762153877351718, test accuracy: 0.9911425159235668


epoch: 6, train loss: 0.013336070074946116, test loss: 0.034684459422982413, test accuracy: 0.9886544585987261


epoch: 7, train loss: 0.01221570678499624, test loss: 0.02526637025367117, test accuracy: 0.991640127388535


epoch: 8, train loss: 0.009255607786699155, test loss: 0.031224525596110993, test accuracy: 0.9902468152866242


epoch: 9, train loss: 0.007962019026544169, test loss: 0.023879671207774389, test accuracy: 0.992734872611465


epoch: 10, train loss: 0.00638781518490486, test loss: 0.031575230005562989, test accuracy: 0.9911425159235668


epoch: 11, train loss: 0.007015494154095348, test loss: 0.026892773140607864, test accuracy: 0.993531050955414


epoch: 12, train loss: 0.004517363090417081, test loss: 0.036284373136983698228, test accuracy: 0.9902468152866242


epoch: 13, train loss: 0.0048642926511034395, test loss: 0.03263570872647048, test accuracy: 0.9920382165605095


epoch: 14, train loss: 0.003729217232173704, test loss: 0.035585670377224137, test accuracy: 0.9923367834394905


epoch: 15, train loss: 0.003733043346276719, test loss: 0.041614705691677606, test accuracy: 0.9914410828025477
```

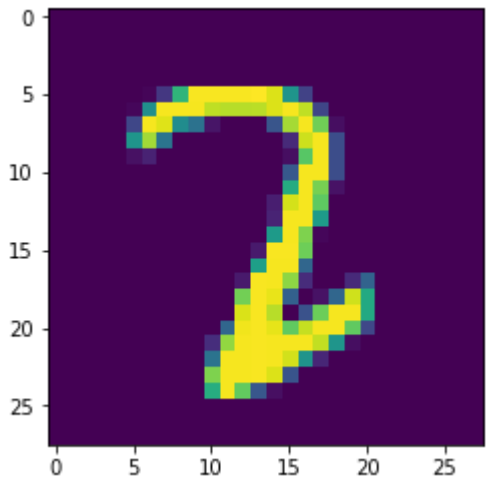
```
epoch: 16, train loss: 0.0043042451297435275, test loss: 0.045198301
260771664, test accuracy: 0.9898487261146497


epoch: 17, train loss: 0.004447008825030113, test loss: 0.0394310957
32562316, test accuracy: 0.9906449044585988


epoch: 18, train loss: 0.0036576191682695346, test loss: 0.041780356
533230596, test accuracy: 0.9917396496815286


epoch: 19, train loss: 0.0032267633709250052, test loss: 0.034179794
27574317, test accuracy: 0.9930334394904459
```

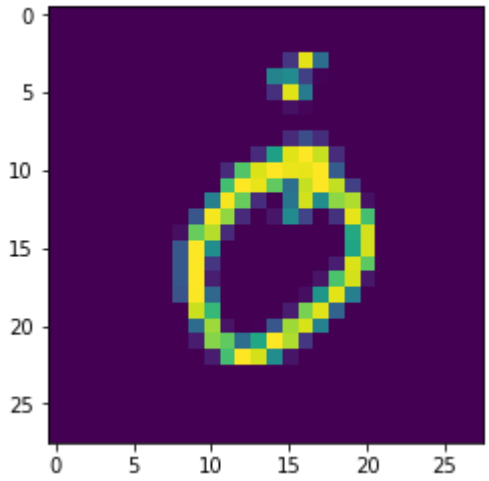
In [32]:

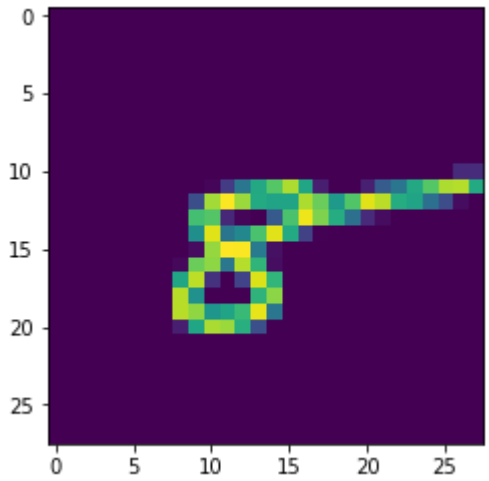
```
for e in error:
    plt.imshow(e)
    plt.show()
```

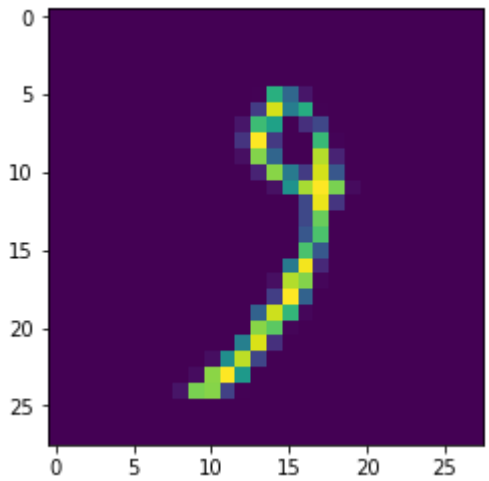

0
5
10
15
20
25
0
5
10
15
20
25

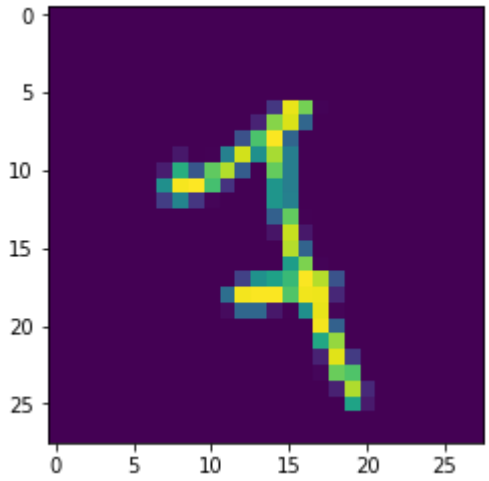

0
5
10
15
20
25
0
5
10
15
20
25

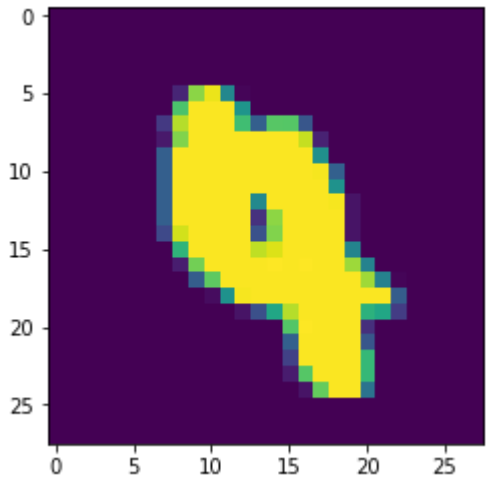

0
5
10
15
20
25
0
5
10
15
20
25

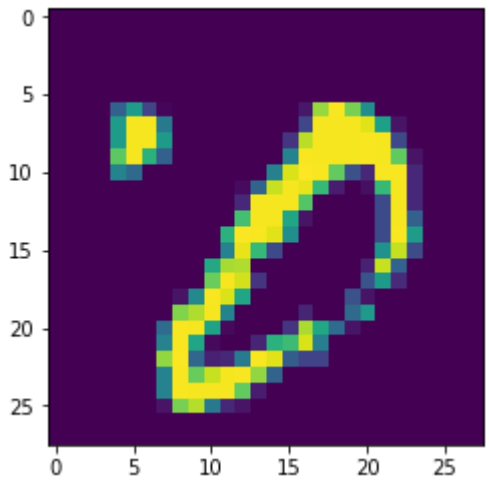

0
5
10
15
20
25
0
5
10
15
20
25

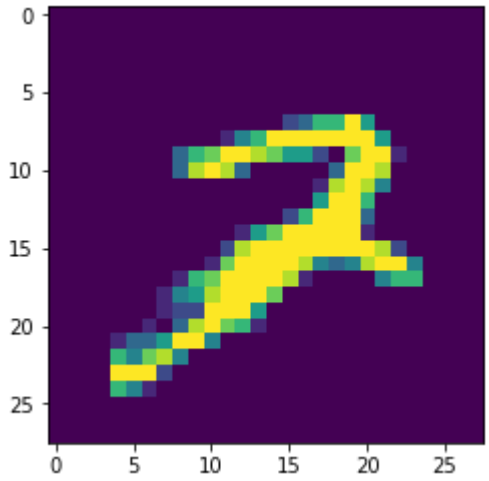

0
5
10
15
20
25
0
5
10
15
20
25

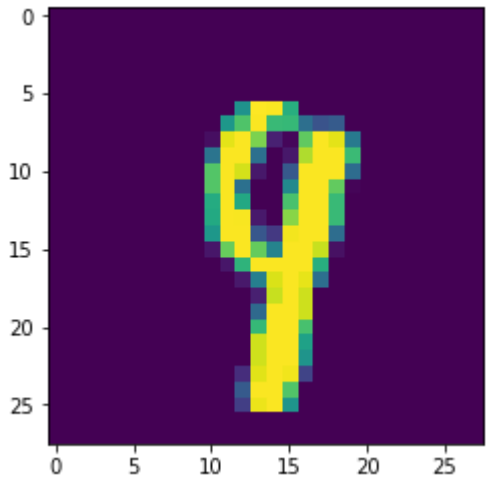

0
5
10
15
20
25
0
5
10
15
20
25

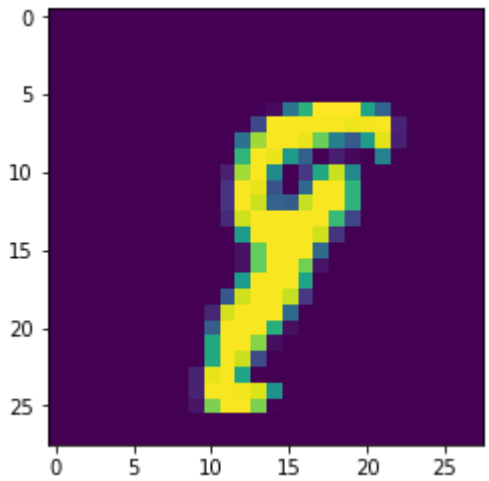

0
5
10
15
20
25
0
5
10
15
20
25

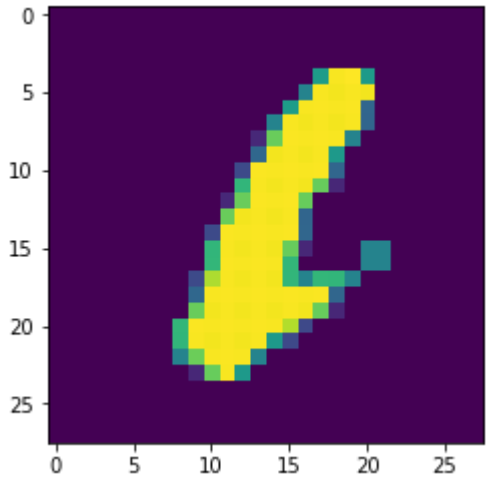

0
5
10
15
20
25
0
5
10
15
20
25

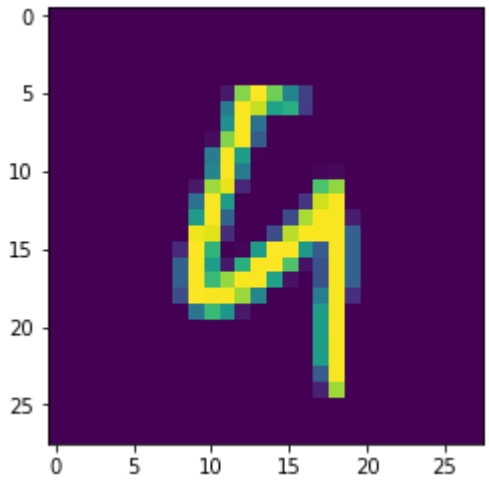

0
5
10
15
20
25
0
5
10
15
20
25

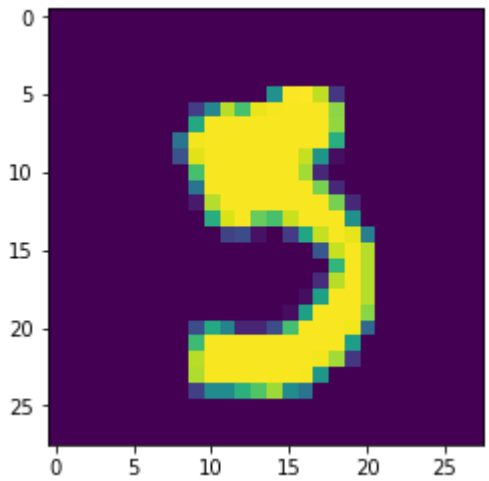

0
5
10
15
20
25
0
5
10
15
20
25

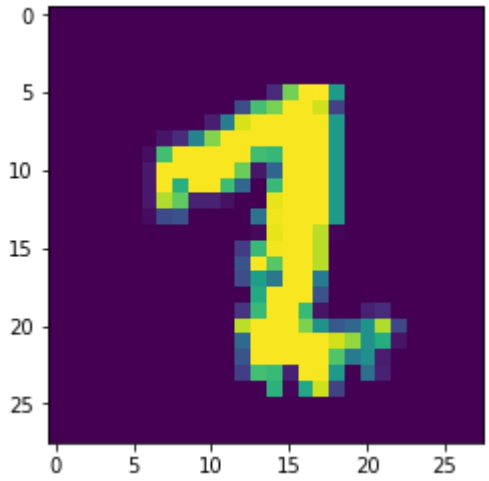

0
5
10
15
20
25
0
5
10
15
20
25

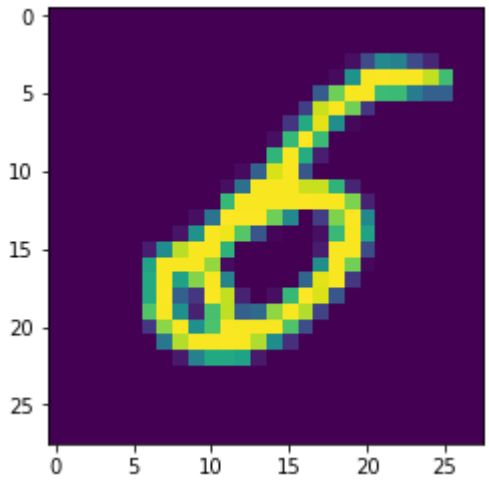

0
5
10
15
20
25
0
5
10
15
20
25

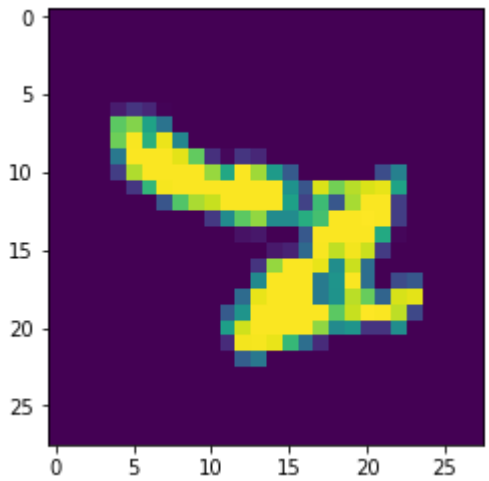

0
5
10
15
20
25
0
5
10
15
20
25

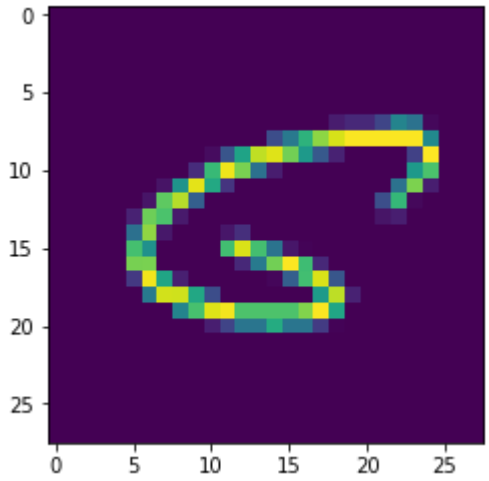

0
5
10
15
20
25
0
5
10
15
20
25

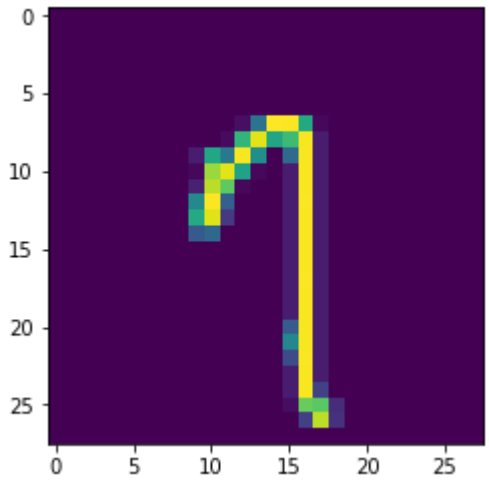

0
5
10
15
20
25
0
5
10
15
20
25

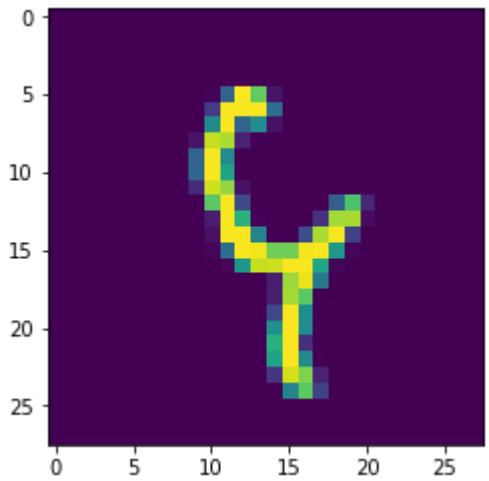

0
5
10
15
20
25
0
5
10
15
20
25

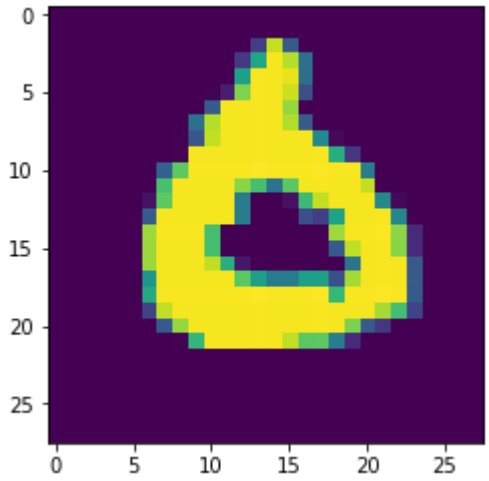

0
5
10
15
20
25
0
5
10
15
20
25

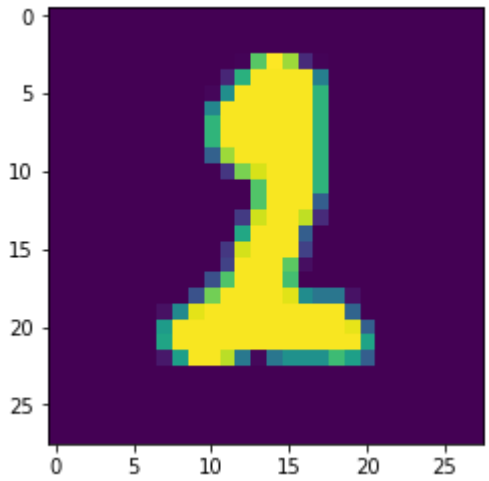

0
5
10
15
20
25
0
5
10
15
20
25

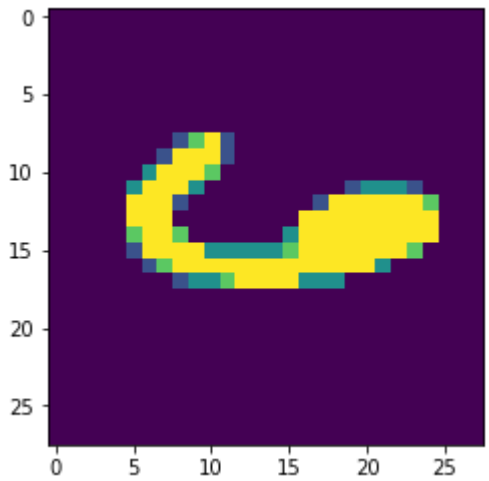

0
5
10
15
20
25
0
5
10
15
20
25

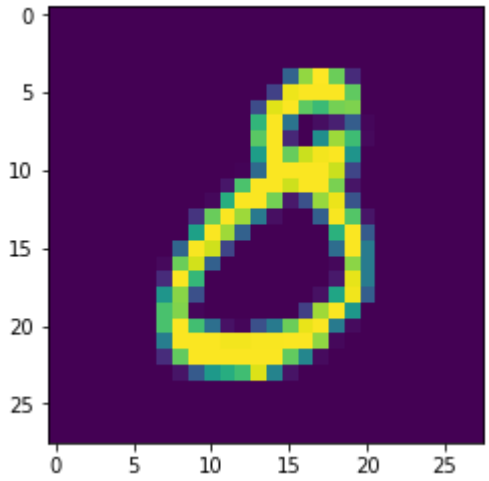

0
5
10
15
20
25
0
5
10
15
20
25

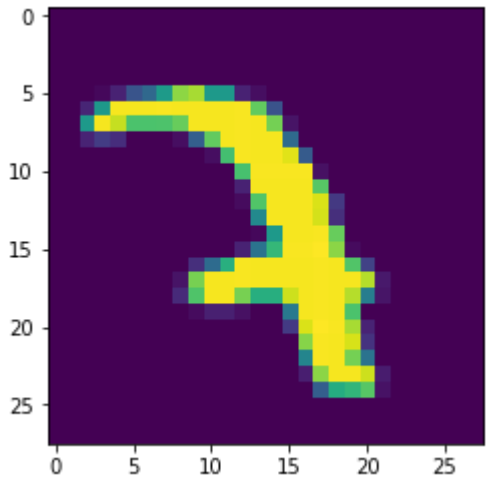

0
5
10
15
20
25
0
5
10
15
20
25

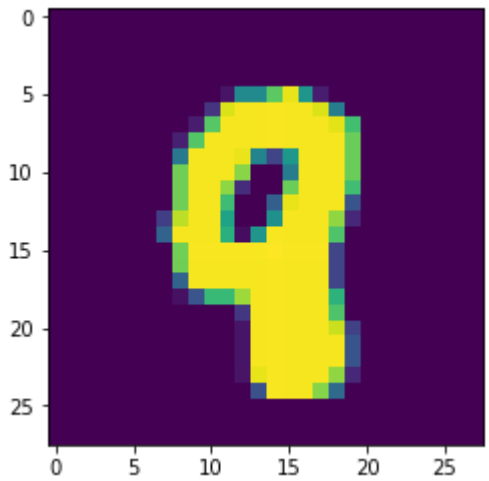

0
5
10
15
20
25
0
5
10
15
20
25

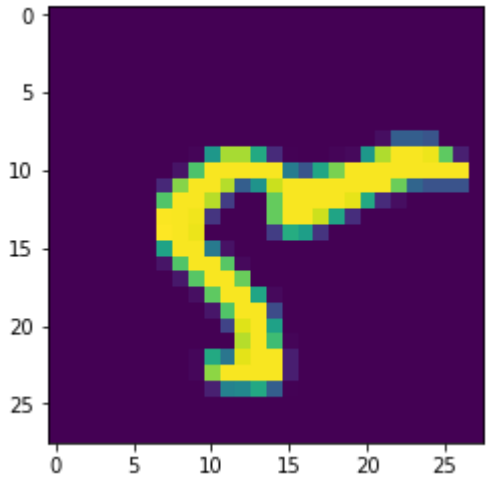

0
5
10
15
20
25
0
5
10
15
20
25

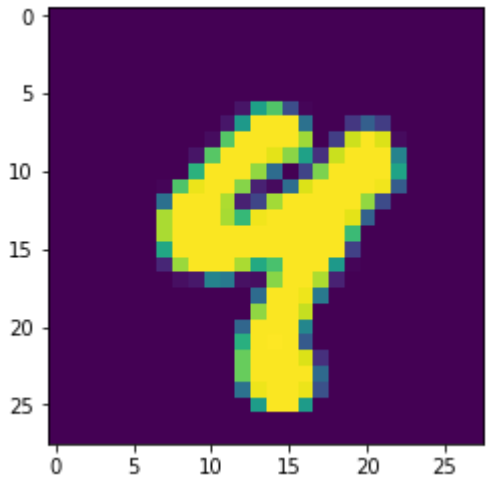

0
5
10
15
20
25
0
5
10
15
20
25

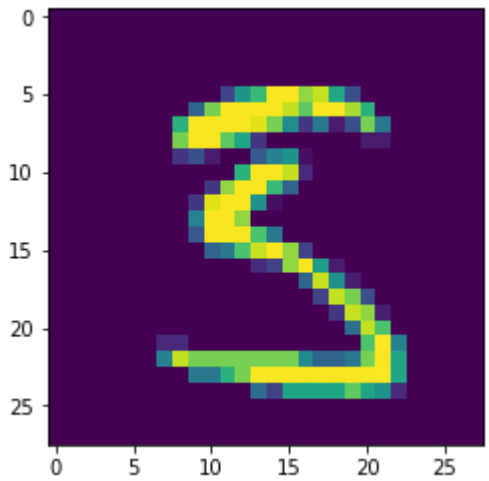

0
5
10
15
20
25
0
5
10
15
20
25

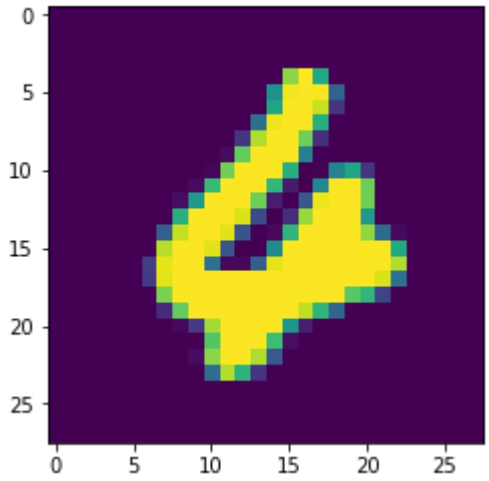

0
5
10
15
20
25
0
5
10
15
20
25

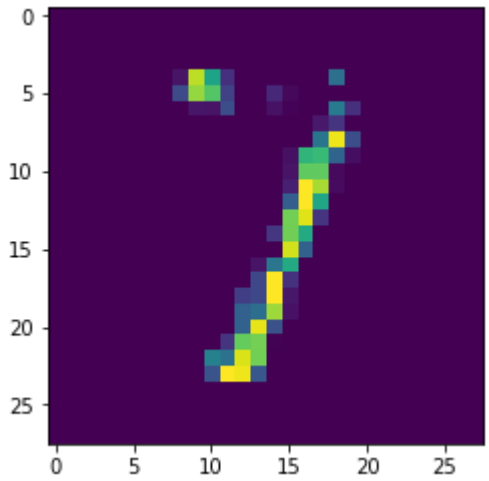

0
5
10
15
20
25
0
5
10
15
20
25

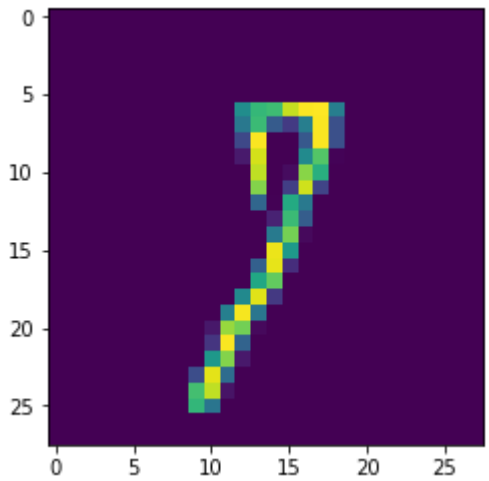

0
5
10
15
20
25
0
5
10
15
20
25

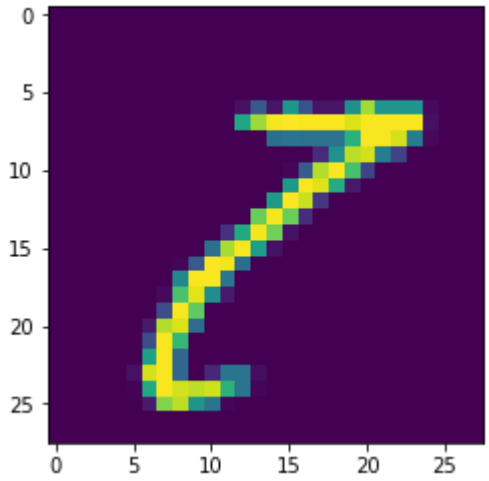

0
5
10
15
20
25
0
5
10
15
20
25

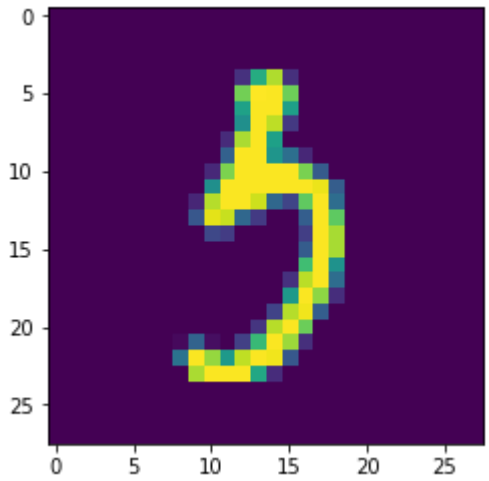

0
5
10
15
20
25
0
5
10
15
20
25

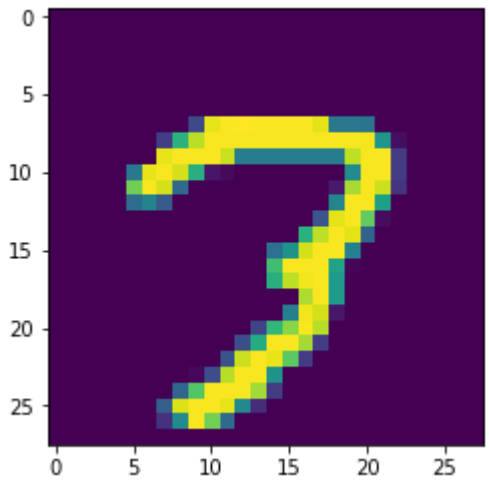

0
5
10
15
20
25
0
5
10
15
20
25

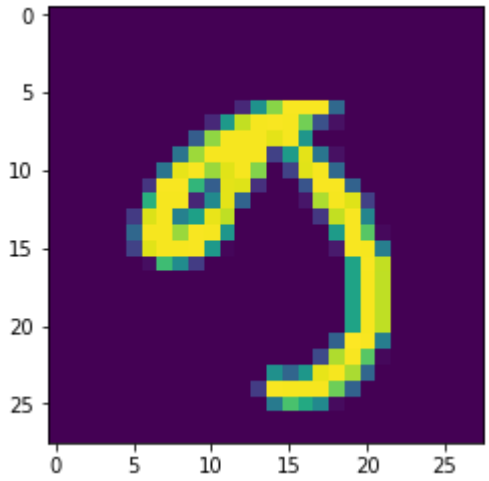

0
5
10
15
20
25
0
5
10
15
20
25

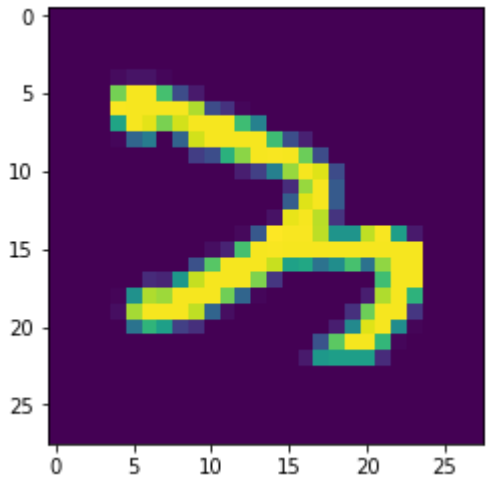

0
5
10
15
20
25
0
5
10
15
20
25

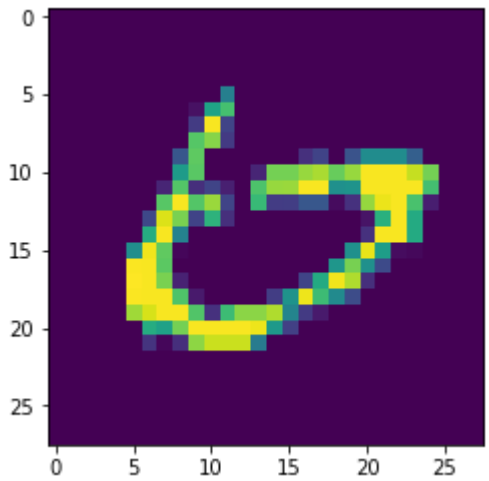

0
5
10
15
20
25
0
5
10
15
20
25

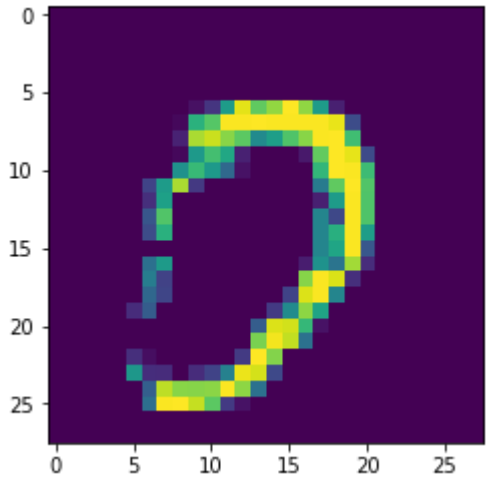

0
5
10
15
20
25
0
5
10
15
20
25

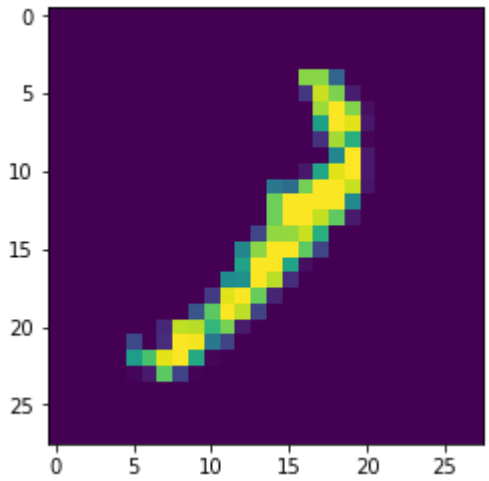

0
5
10
15
20
25
0
5
10
15
20
25

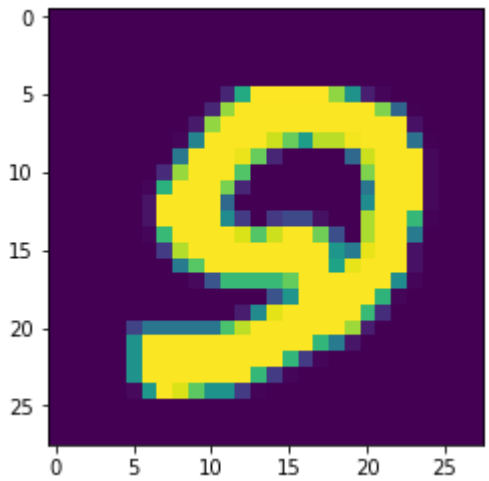

0
5
10
15
20
25
0
5
10
15
20
25

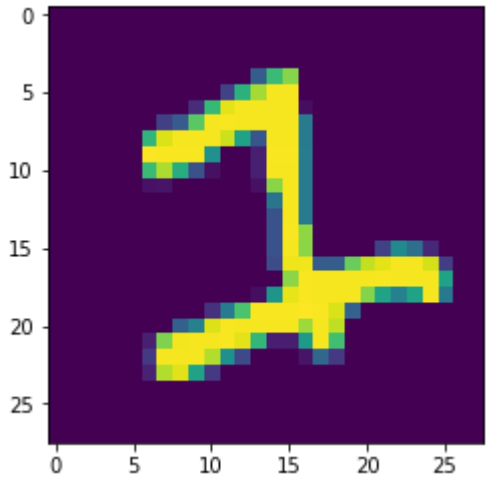

0
5
10
15
20
25
0
5
10
15
20
25

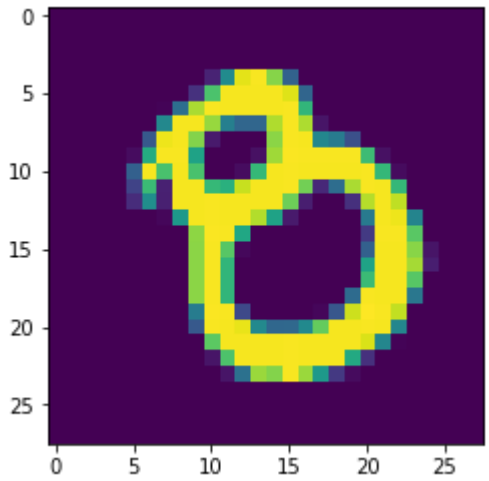

0
5
10
15
20
25
0
5
10
15
20
25

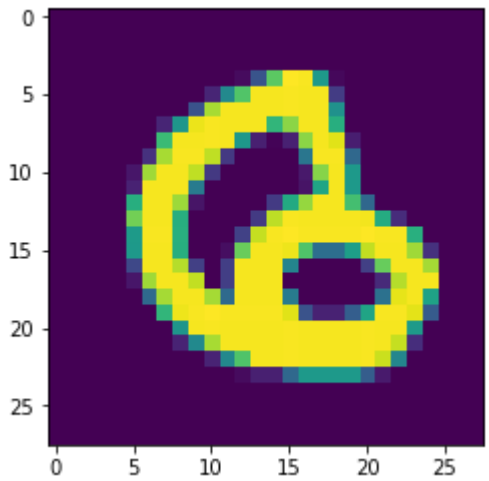

0
5
10
15
20
25
0
5
10
15
20
25

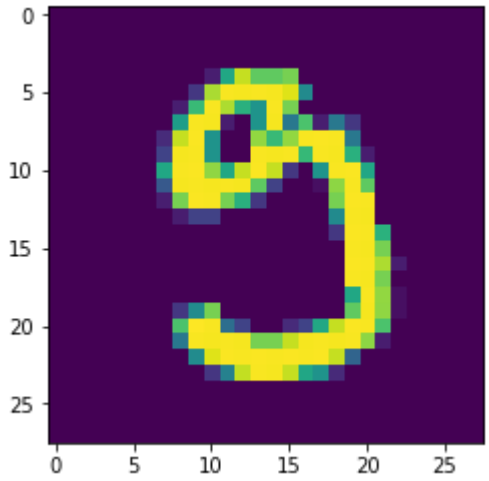

0
5
10
15
20
25
0
5
10
15
20
25

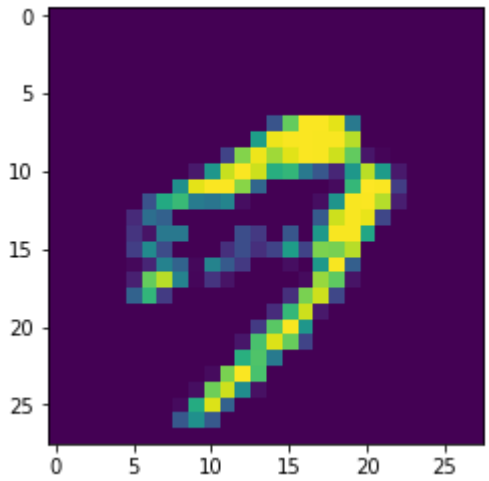

0
5
10
15
20
25
0
5
10
15
20
25

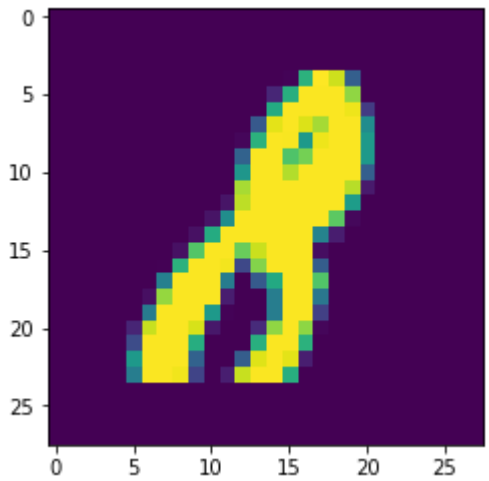

0
5
10
15
20
25
0
5
10
15
20
25

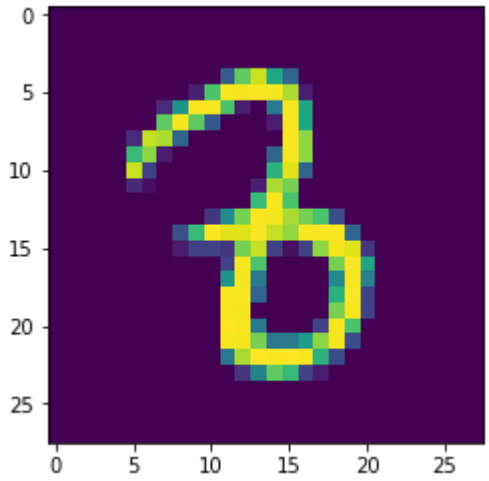

0
5
10
15
20
25
0
5
10
15
20
25

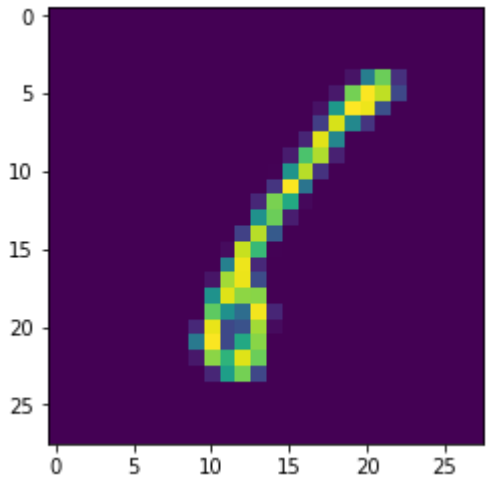

0
5
10
15
20
25
0
5
10
15
20
25

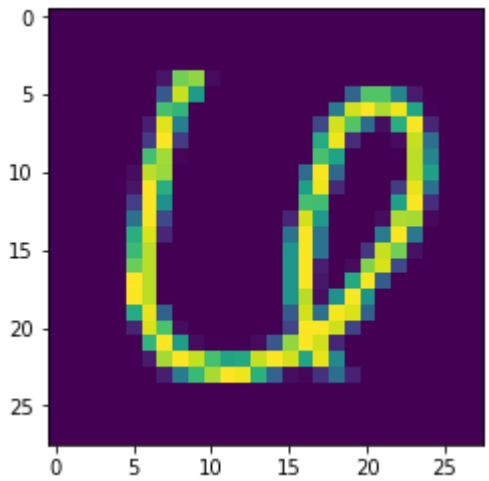

0
5
10
15
20
25
0
5
10
15
20
25

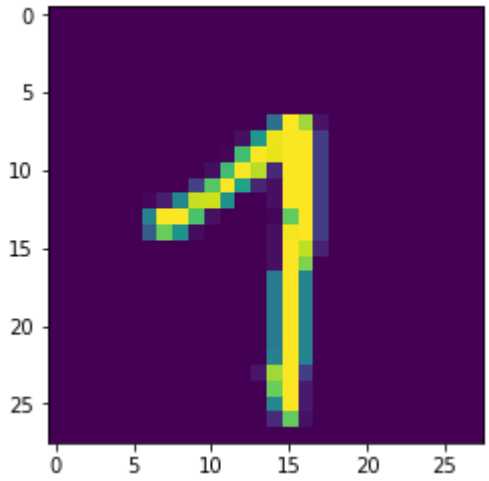

0
5
10
15
20
25
0
5
10
15
20
25

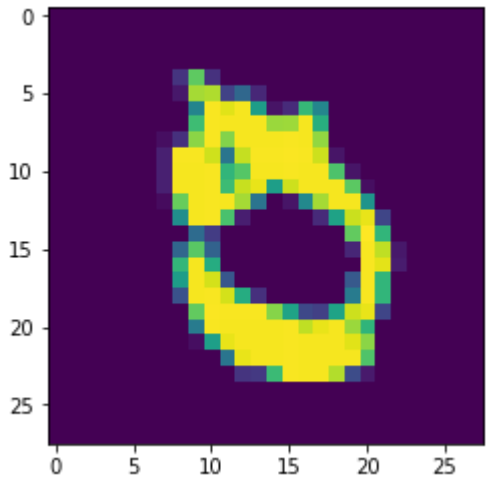

0
5
10
15
20
25
0
5
10
15
20
25

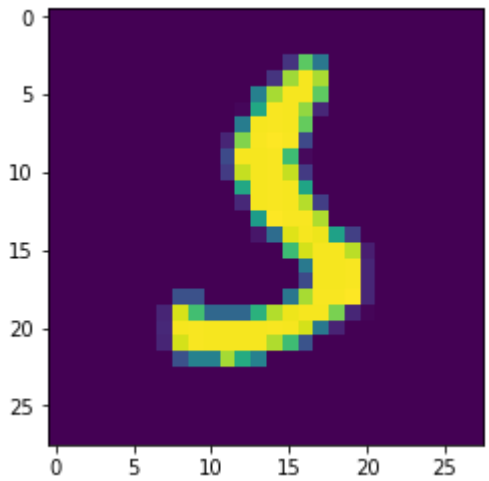

0
5
10
15
20
25
0
5
10
15
20
25

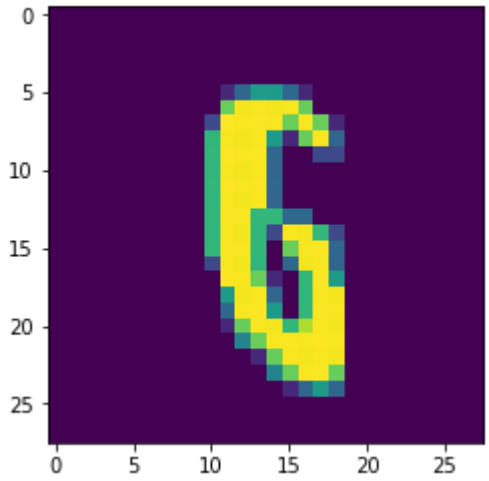

0
5
10
15
20
25
0
5
10
15
20
25

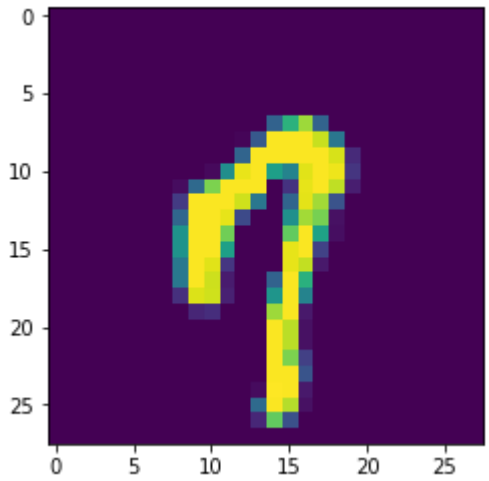

0
5
10
15
20
25
0
5
10
15
20
25

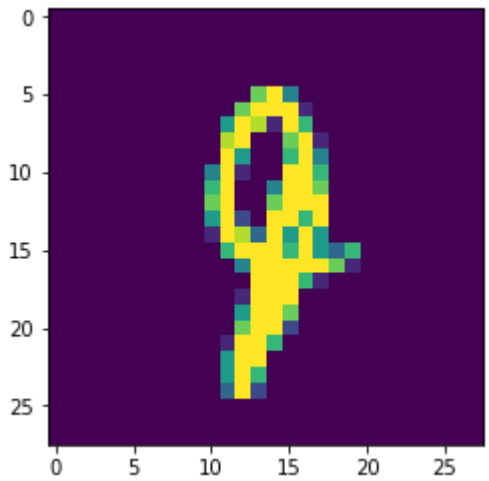

0
5
10
15
20
25
0
5
10
15
20
25

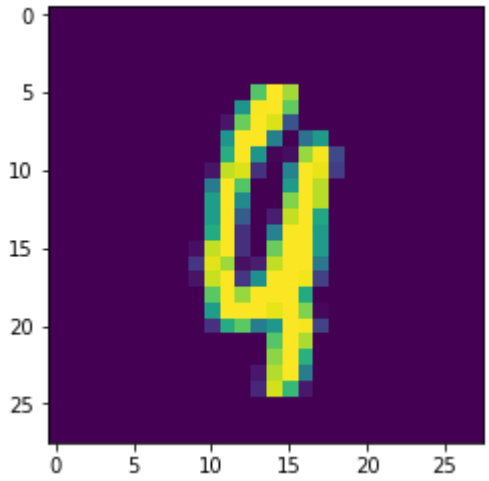

0
5
10
15
20
25
0
5
10
15
20
25

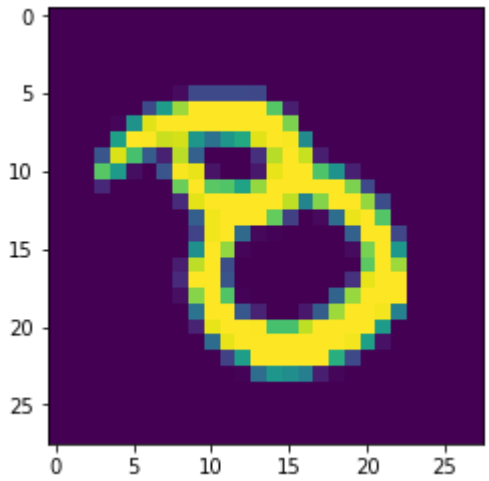

0
5
10
15
20
25
0
5
10
15
20
25

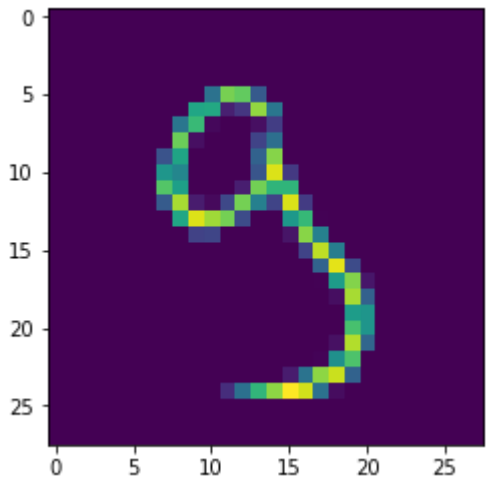

0
5
10
15
20
25
0
5
10
15
20
25

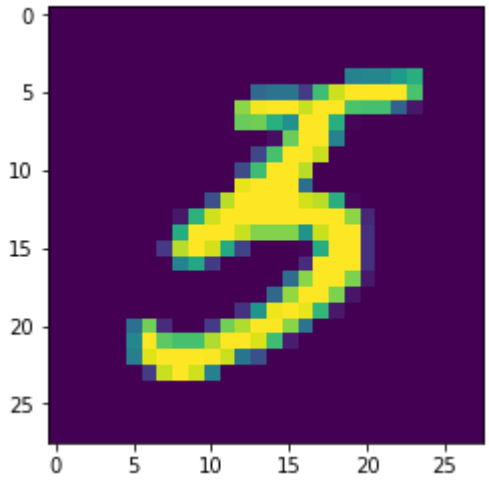

0
5
10
15
20
25
0
5
10
15
20
25

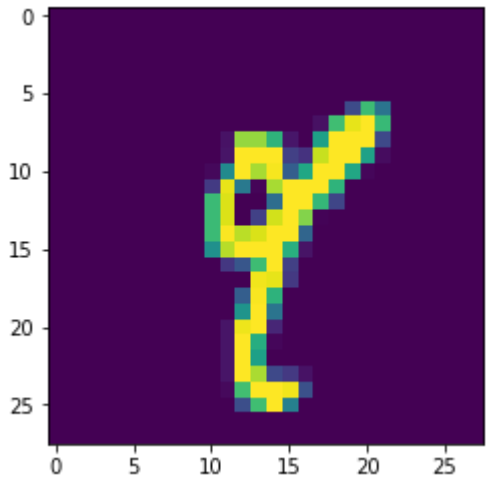

0
5
10
15
20
25
0
5
10
15
20
25

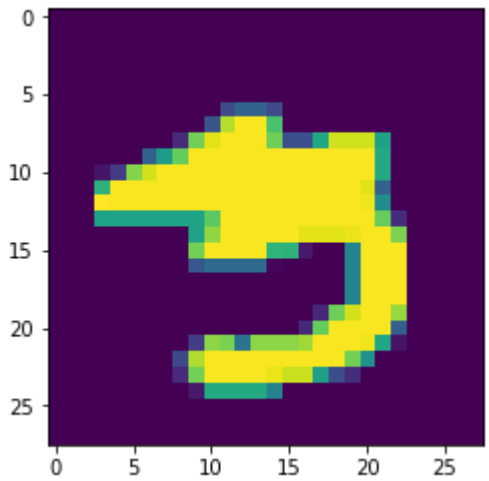

0
5
10
15
20
25
0
5
10
15
20
25

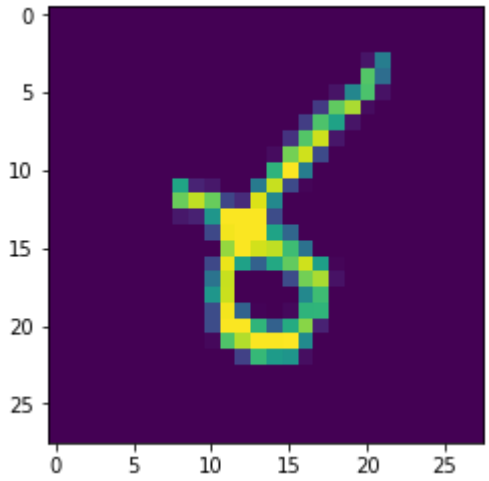

0
5
10
15
20
25
0
5
10
15
20
25

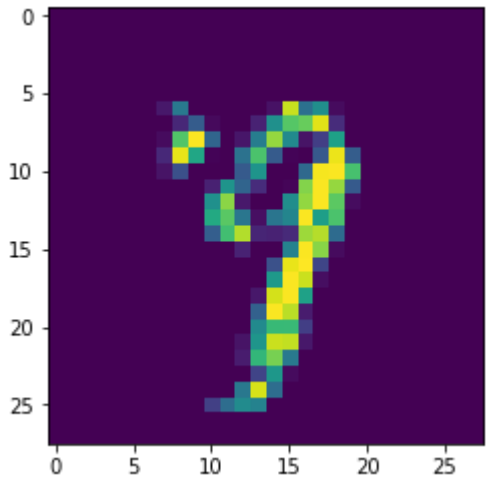

0
5
10
15
20
25
0
5
10
15
20
25

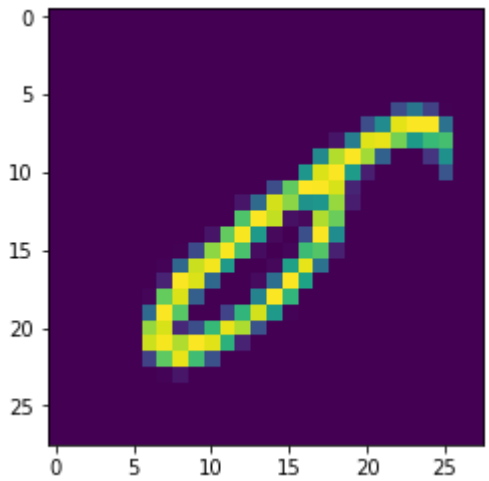

0
5
10
15
20
25
0
5
10
15
20
25

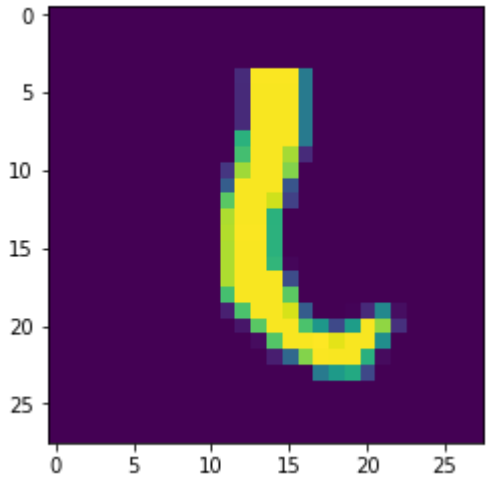

0
5
10
15
20
25
0
5
10
15
20
25

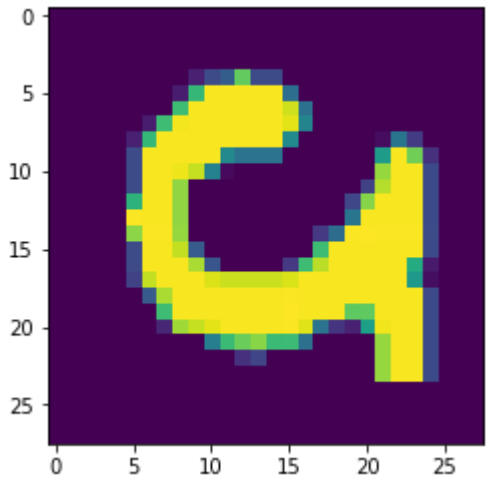

0
5
10
15
20
25
0
5
10
15
20
25

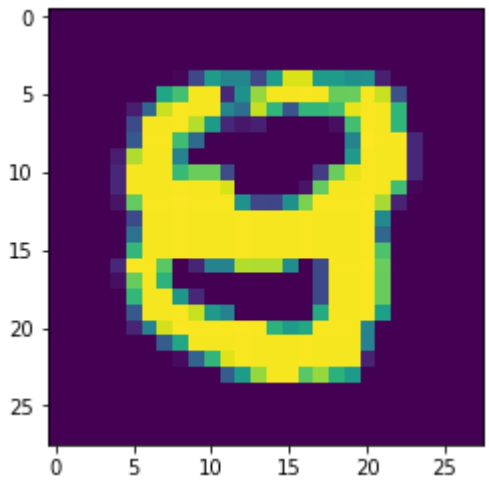

0
5
10
15
20
25
0
5
10
15
20
25

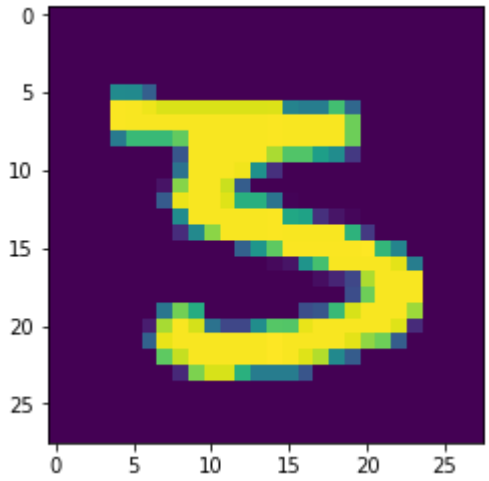

0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25




**Don't forget plotting the digits that the network got wrong.**

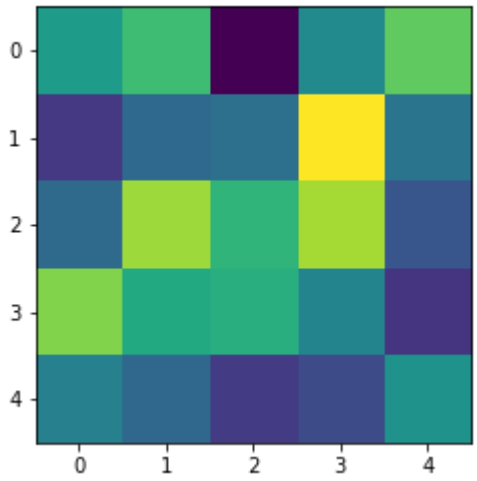
## Question 3.2 Kernel weights visualization [5pt]

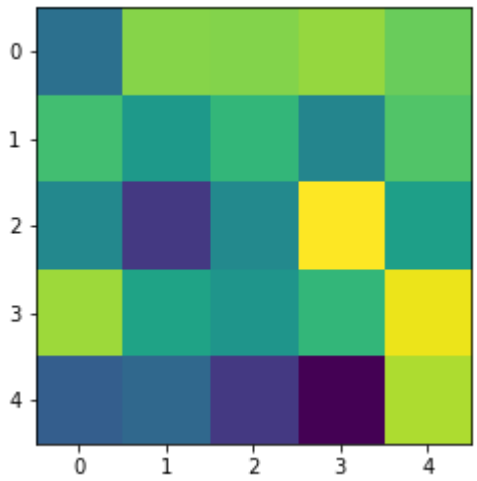
For this question, you need to visualize the kernel weights for your first convolutional layer. Suppose you have 5x5 kernels with 32 output channels. You will plot 32 5x5 images.

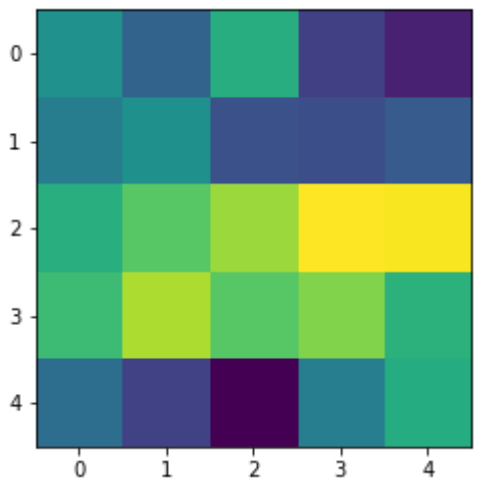
hint: You might need to look at PyTorch documentation (or play with the PyTorch model) to figure out how to get the weights.

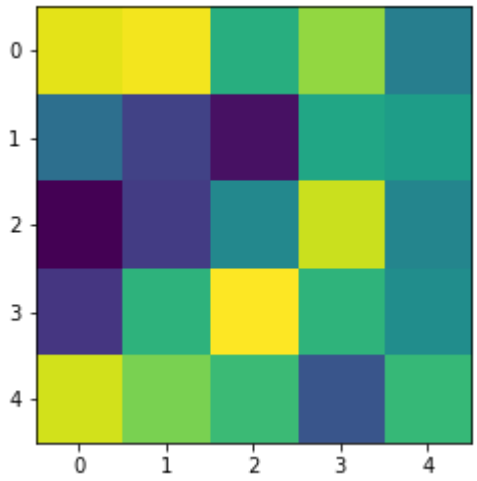
In [33]:

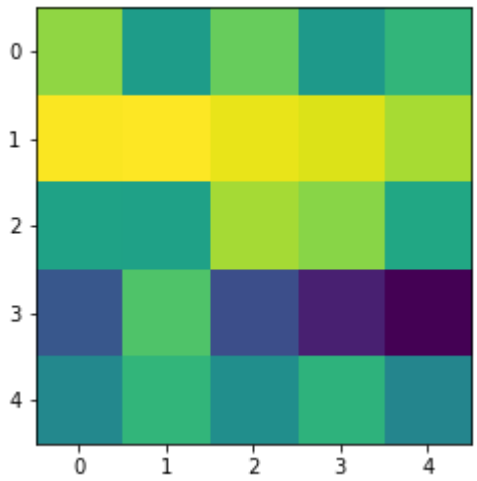
```
### YOUR CODE HERE
for w in model.conv1.weight:
    plt.imshow(w[0].detach())
    plt.show()
### END OF CODE
```

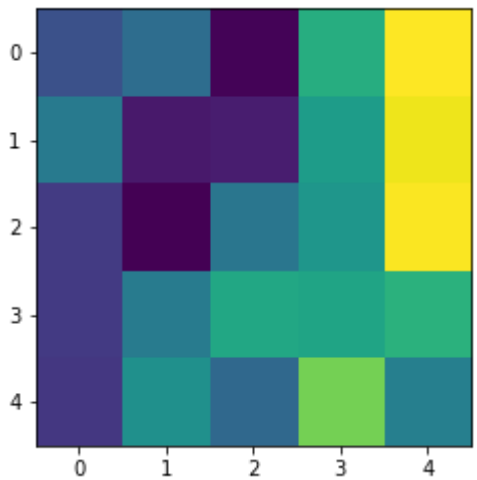

0
1
2
3
4
0
1
2
3
4

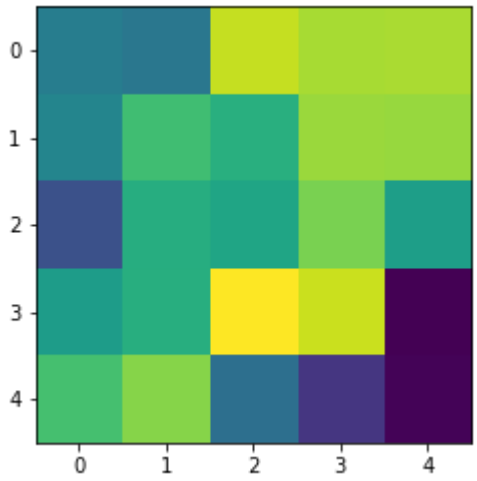

0
1
2
3
4
0
1
2
3
4

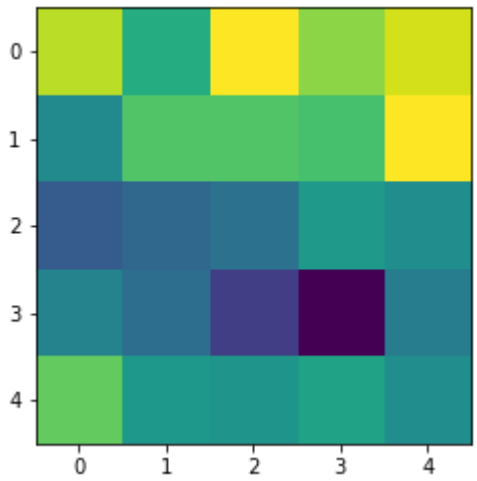

0
1
2
3
4
0
1
2
3
4

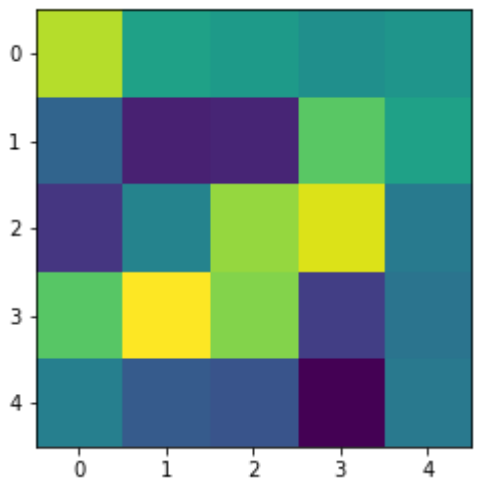

0
1
2
3
4
0
1
2
3
4

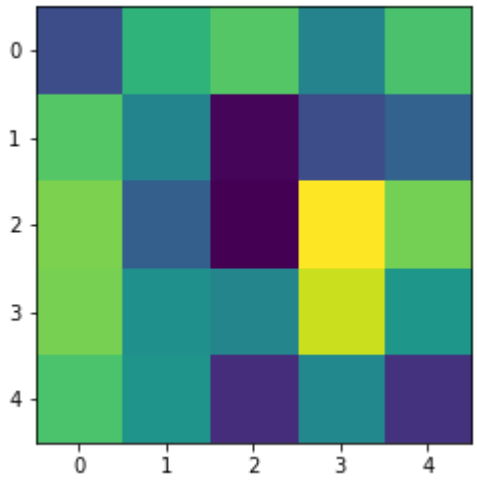

0
1
2
3
4
0
1
2
3
4

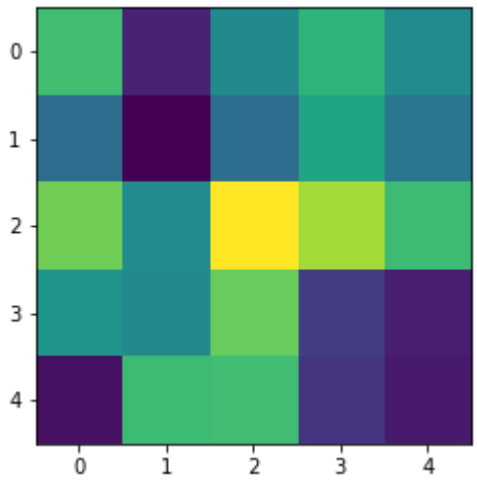

0
1
2
3
4
0
1
2
3
4

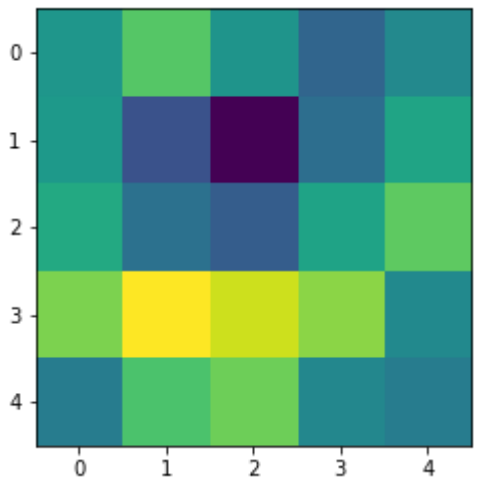

0
1
2
3
4
0
1
2
3
4

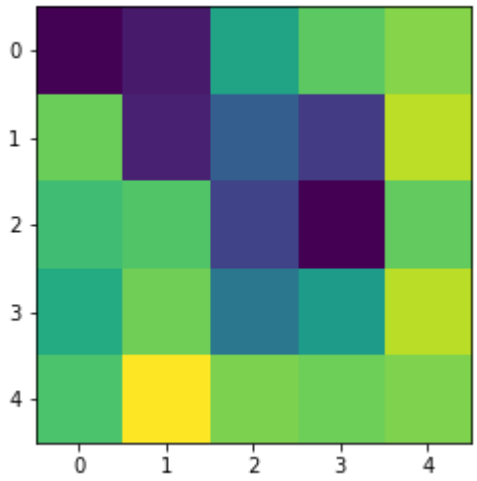

0
1
2
3
4
0
1
2
3
4

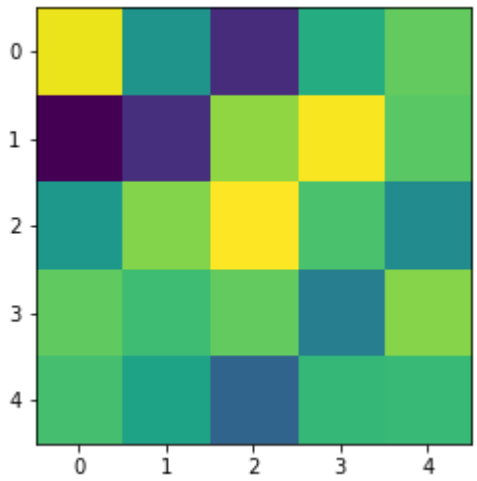

0
1
2
3
4
0
1
2
3
4

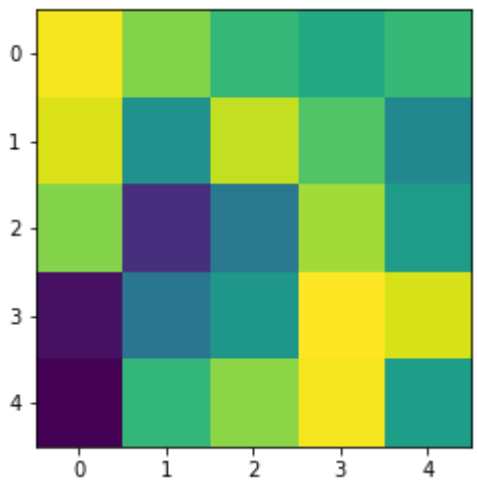

0
1
2
3
4
0
1
2
3
4

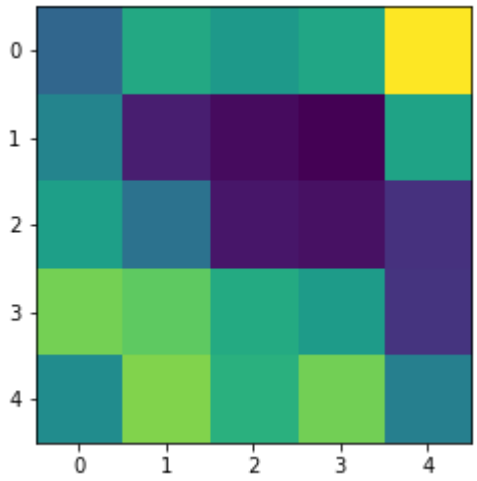

0
1
2
3
4
0
1
2
3
4

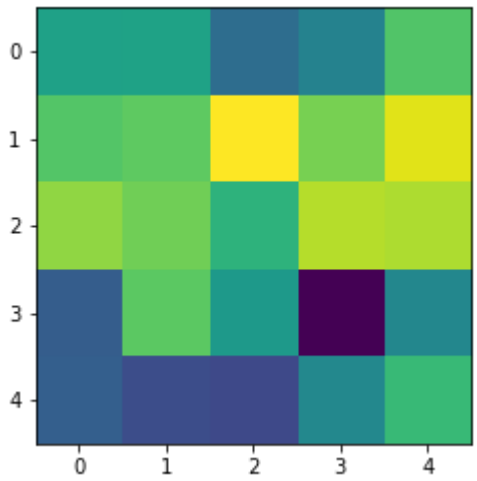

0
1
2
3
4
0
1
2
3
4

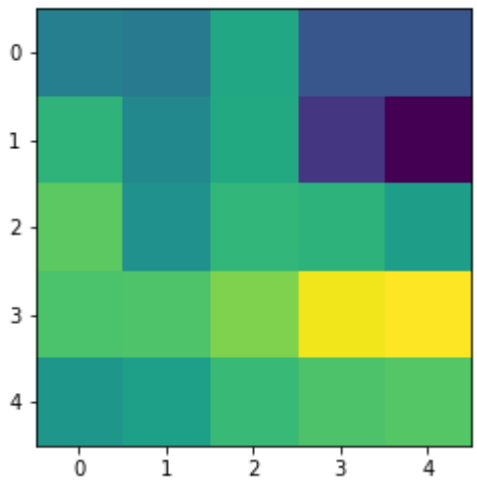

0
1
2
3
4
0
1
2
3
4

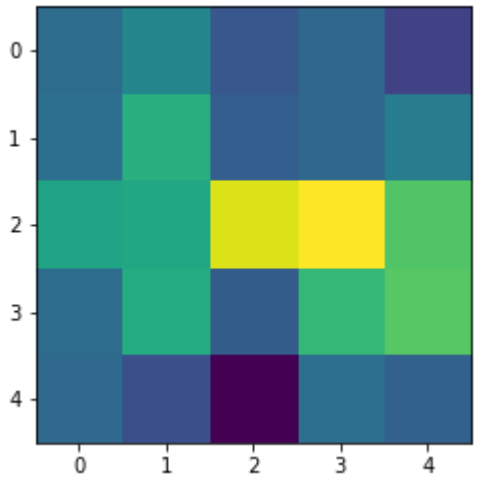

0
1
2
3
4
0
1
2
3
4

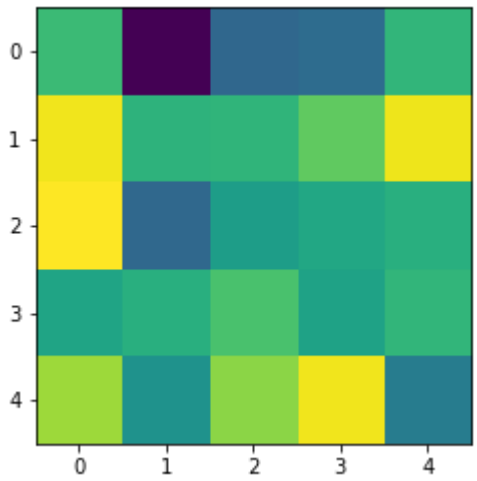

0
1
2
3
4
0
1
2
3
4

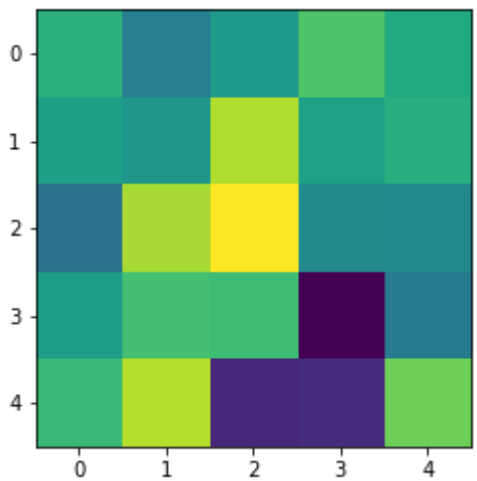

0
1
2
3
4
0
1
2
3
4

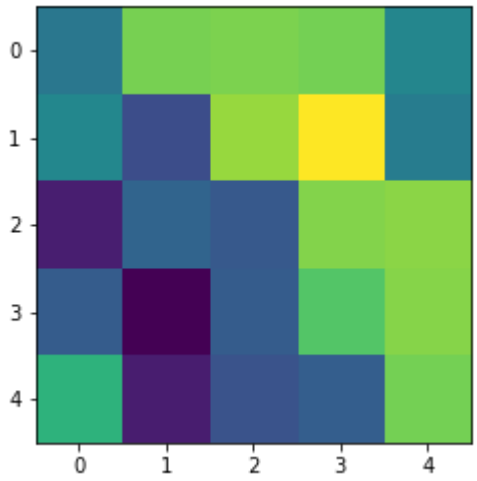

0
1
2
3
4
0
1
2
3
4

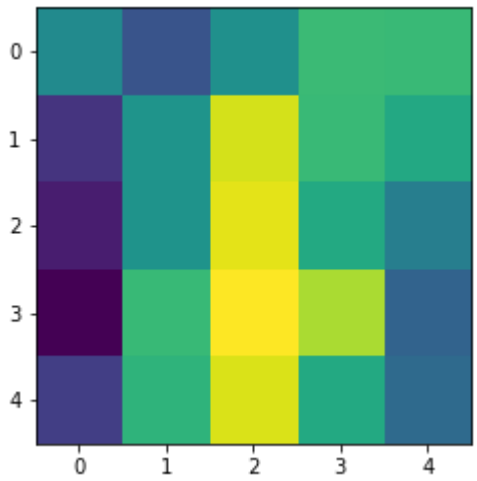

0
1
2
3
4
0
1
2
3
4

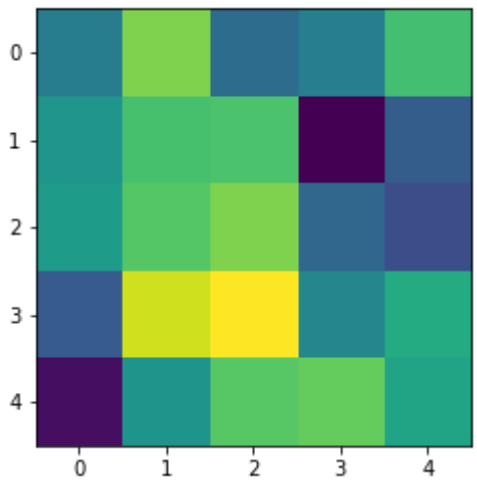

0
1
2
3
4
0
1
2
3
4

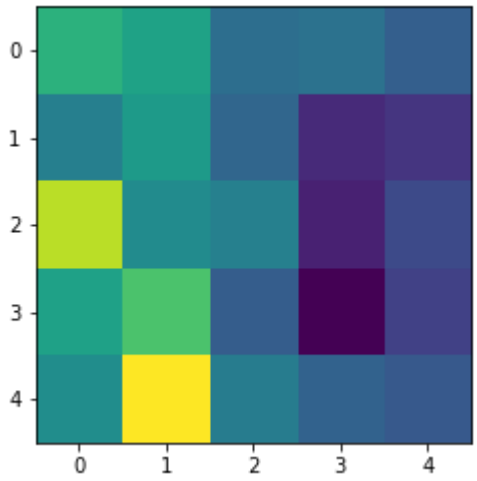

0
1
2
3
4
0
1
2
3
4

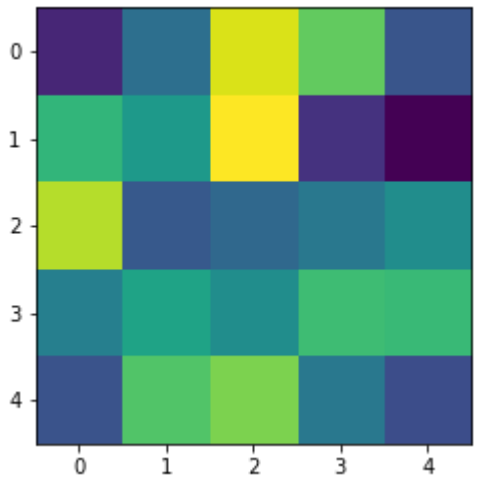

0
1
2
3
4
0
1
2
3
4

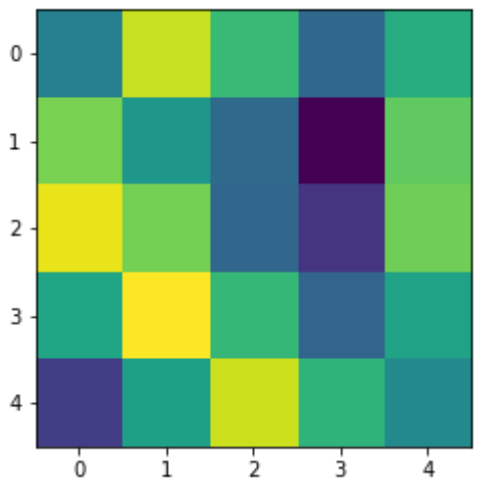

0
1
2
3
4
0
1
2
3
4

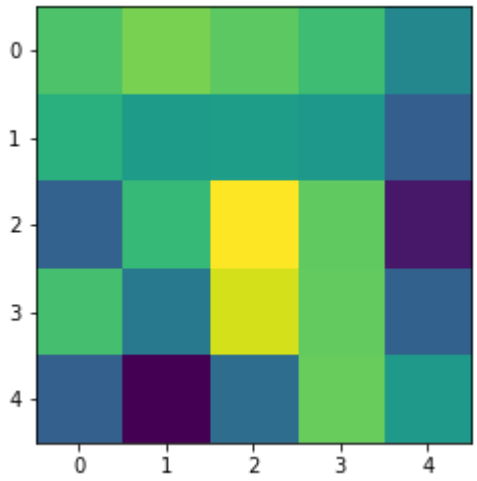

0
1
2
3
4
0
1
2
3
4

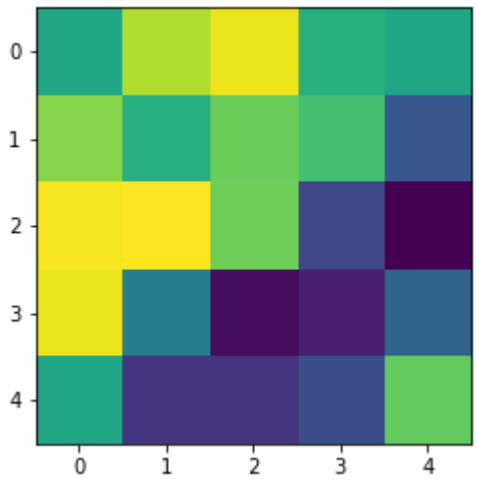

0
1
2
3
4
0
1
2
3
4

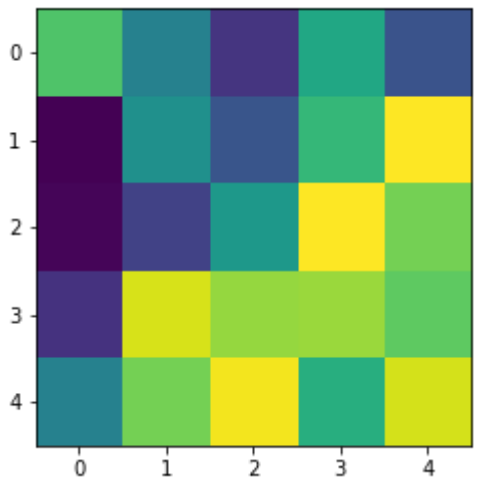

0
1
2
3
4
0
1
2
3
4

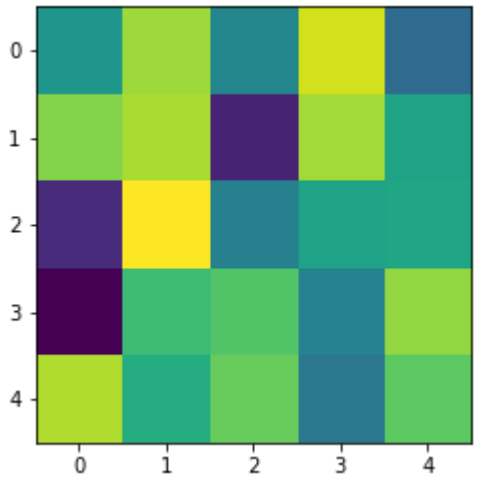

0
1
2
3
4
0
1
2
3
4

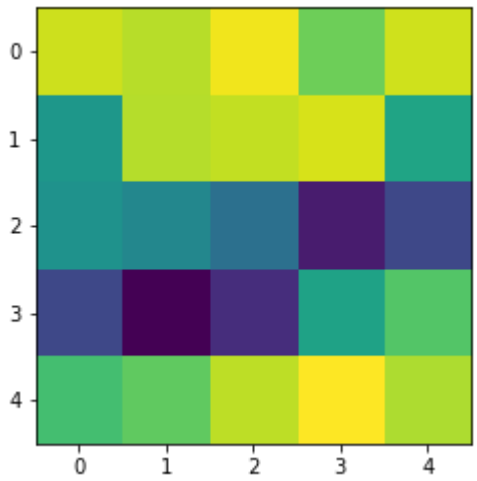

0
1
2
3
4
0
1
2
3
4


0
1
2
3
4
0
1
2
3
4


0
1
2
3
4
0
1
2
3
4


0
1
2
3
4
0
1
2
3
4




In [ ]: