(A) Download the MNIST dataset (or simply load it from keras.datasets) and use keras to train, (ii) a convolutional neural network to classify hand-written digits. Remember to normalize your data.

Once you feel satisfied with your achieved results save your models to disk for later use. Call a MNIST-simple-generator $f(w) = x$, a function which given an input w outputs a 28x28 grayscale image. You can think of a MNIST-simple-generator as a neural network with one input neuron, always equal to 1, as a function of its parameters. In this exercise you will build MNIST-simple-generators to construct adversarial examples for your models from (A).

To understand what an adversarial example is, we restrain our attention to MNIST. For most digits in the dataset, your developed model is going to output the correct label with a high confidence. Given such a digit (that your model classifies correctly with high confidence) an adversarial example is constructed by applying small changes to the digit image (so small that they may be imperceptible to a human), which lead to our model miss-classifying the image. In fact, we can construct adversarial examples for any such digit and make our model predict for a given input whichever class we want. A lot of research has been done on building robust optimization methods that are not prone to adversarial example attacks.

(B) Choose your 5 favorite digits and take a sample for each from the MNIST dataset. Index these samples by i. Define a MNIST-simple-generator f_{ij} as a keras model, constructing adversarial examples around instance i for class j. Compile your model to use the Adam optimizer and the mean squared error loss. Define also the composite model $C \circ f$ (C is the model you built in (A)), making sure that before you compile you disable training for all parameters in C. (*Hint*: You may want to use the Keras functional API). For each sample you picked and for each class 0-10, train f_{ij} in two steps, repeated as many times as necessary. First train f_{ij} with target equal to sample i. Then, train the composite model with target equal the one-hot encoding of j. Plot your adversarial examples together with the original image on a grid.

(C) Repeat (A)-(B) on the CIFAR10 dataset, this time selecting only one image and compute adversarial examples for each class.

(D) (Bonus) Construct a new model which is robust to adversarial examples constructed in the above described way. Carefully explain your approach and validate on MNIST and CIFAR10.