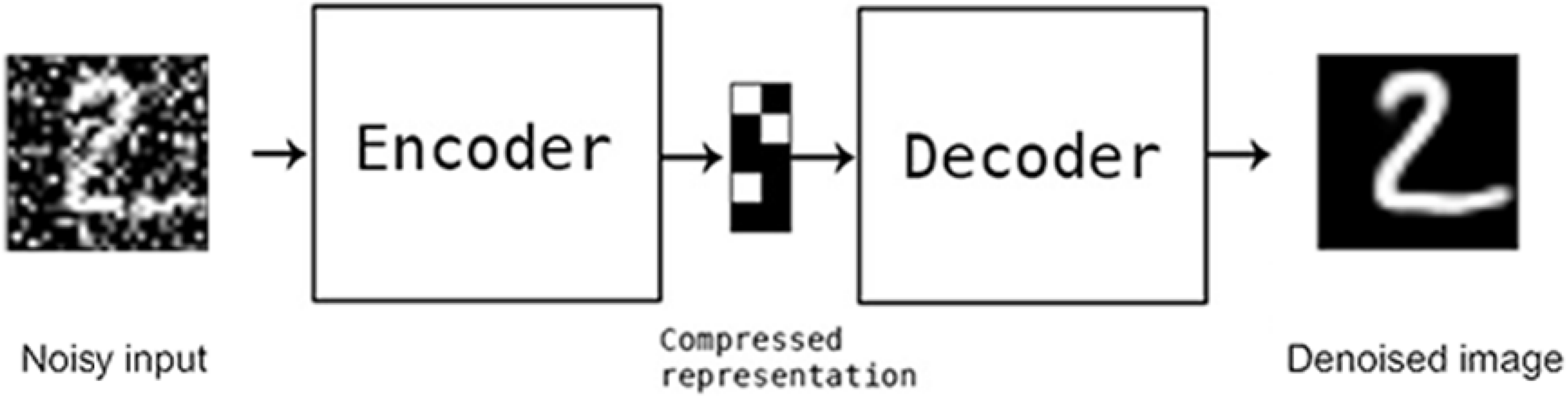
Image Restoration

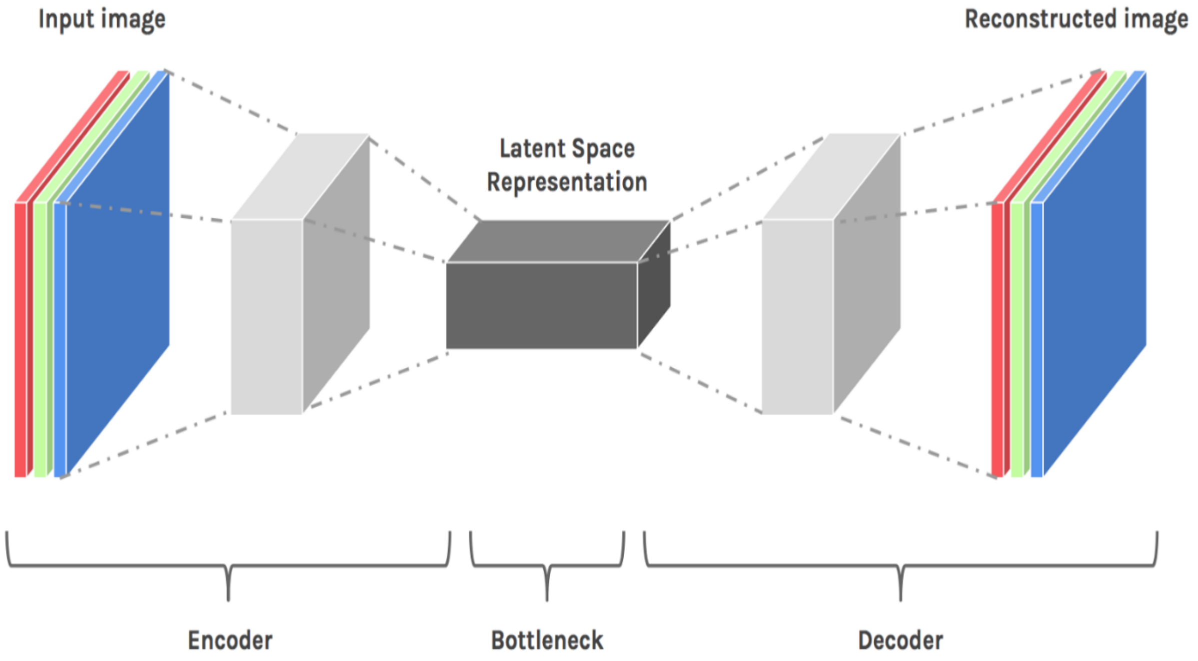
DANIELE ANGIONI - GIACOMO GALLUS

Goal of the project

- Train an autoencoder to filter out noise from input images
- Test its robustness against existing adversarial attacks using the library SecML



What is an Autoencoder?



Convolutional Encoder-Decoder architecture

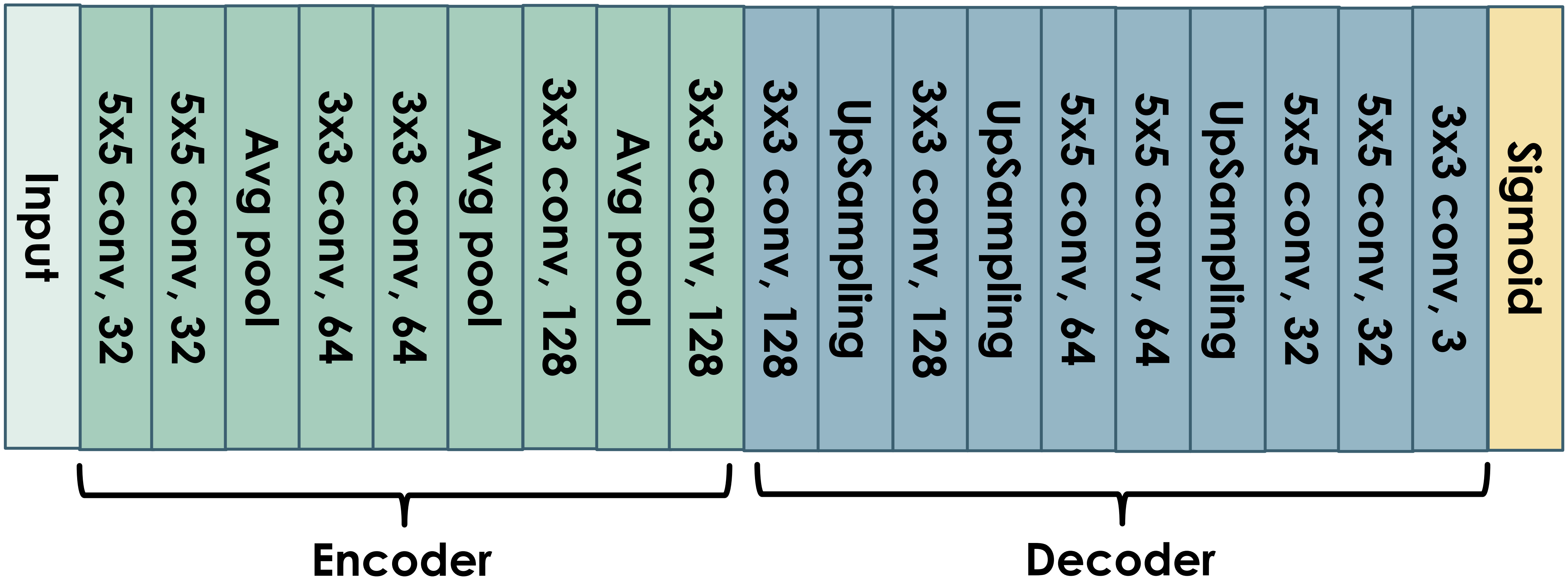
Model capable of compressing data into a lower dimensional feature space

- **Encoder** : compresses the data
- **Decoder** : reconstructs the data

Dataset

- **CIFAR10** : 50000 samples for the train set and 10000 for the test set
- A **gaussian noise** with zero mean and standard deviation equal to 0.05 has been added to dataset

Autoencoder's architecture



Input

5x5 conv, 32

5x5 conv, 32

Avg pool

3x3 conv, 64

3x3 conv, 64

Avg pool

3x3 conv, 128

Avg pool

3x3 conv, 128

3x3 conv, 128

UpSampling

3x3 conv, 128

UpSampling

5x5 conv, 64

5x5 conv, 64

UpSampling

5x5 conv, 32

5x5 conv, 32

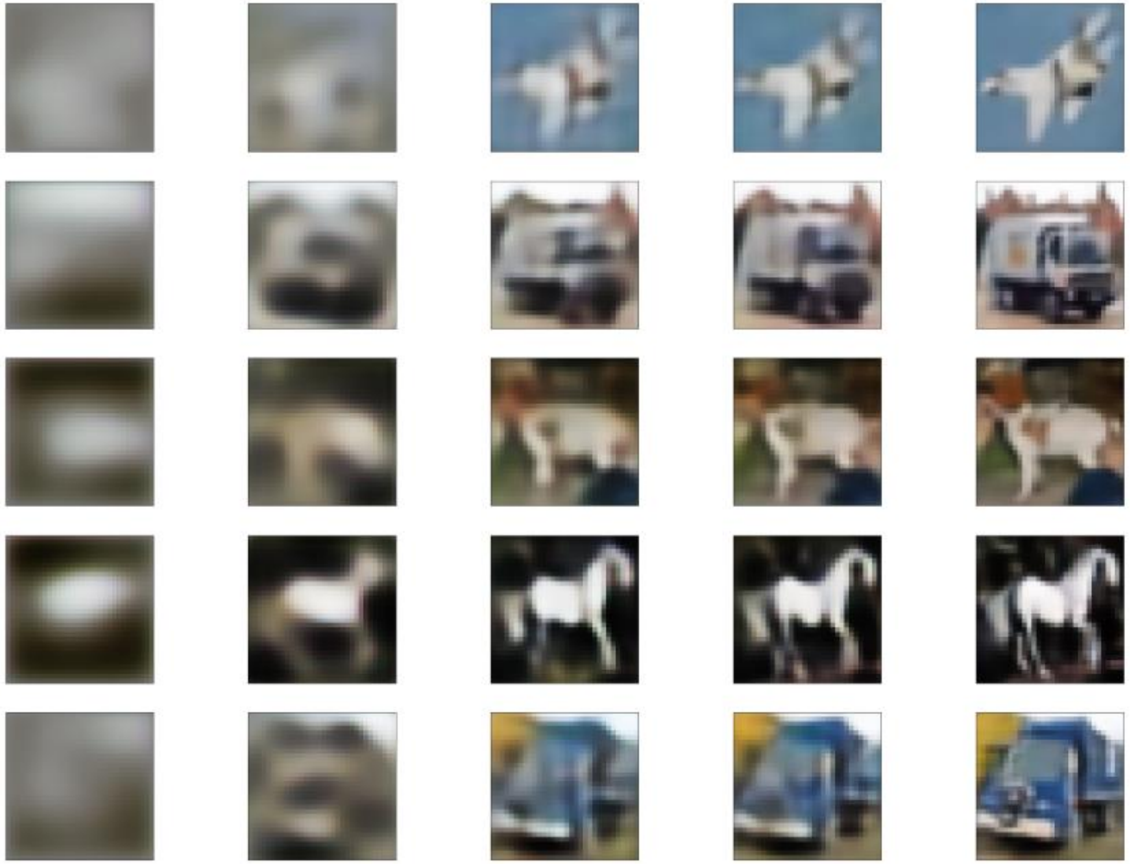
3x3 conv, 3

Sigmoid

Encoder

Decoder

Training

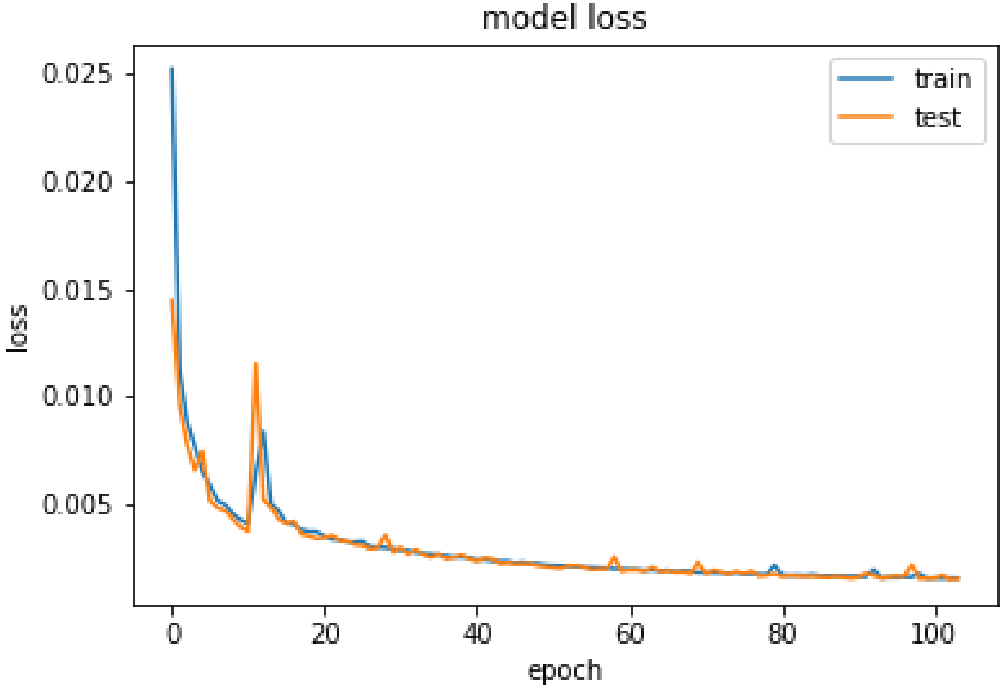


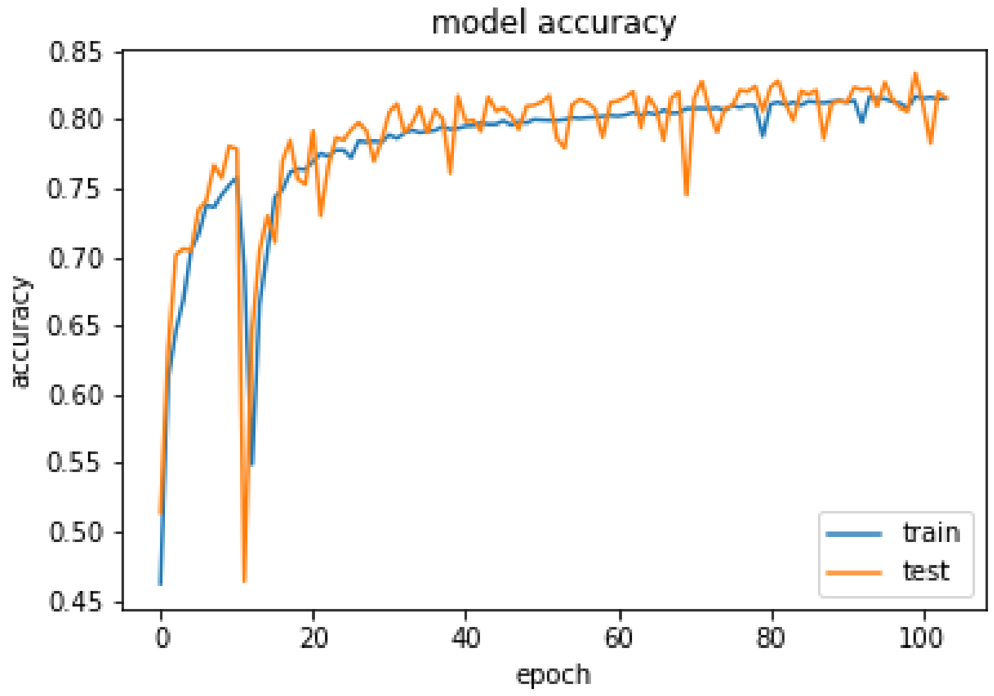
- A **Mean Squared Error Loss** between the output of the autoencoder and the original images has been definined

- The model has been trained for **100 epochs** with the **Adam optimizer**

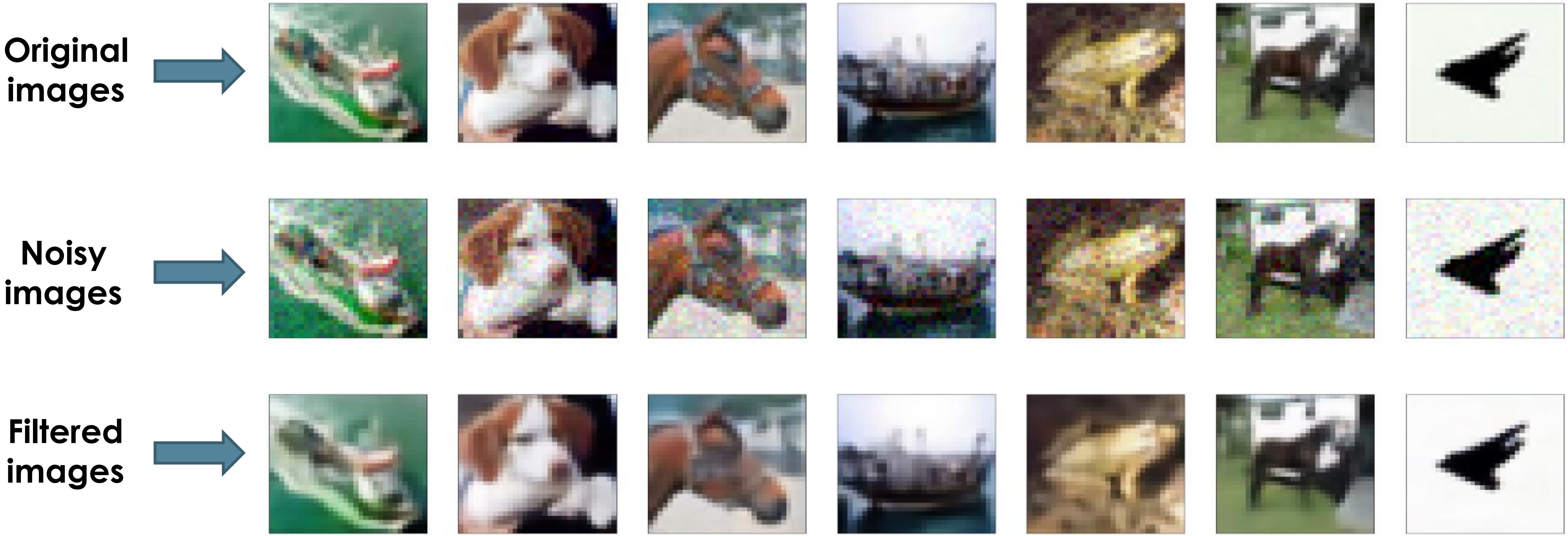
Results of the autoencoder

- **Final Loss** = 0.015
- **Final Accuracy** = 81.5%





Results of the autoencoder



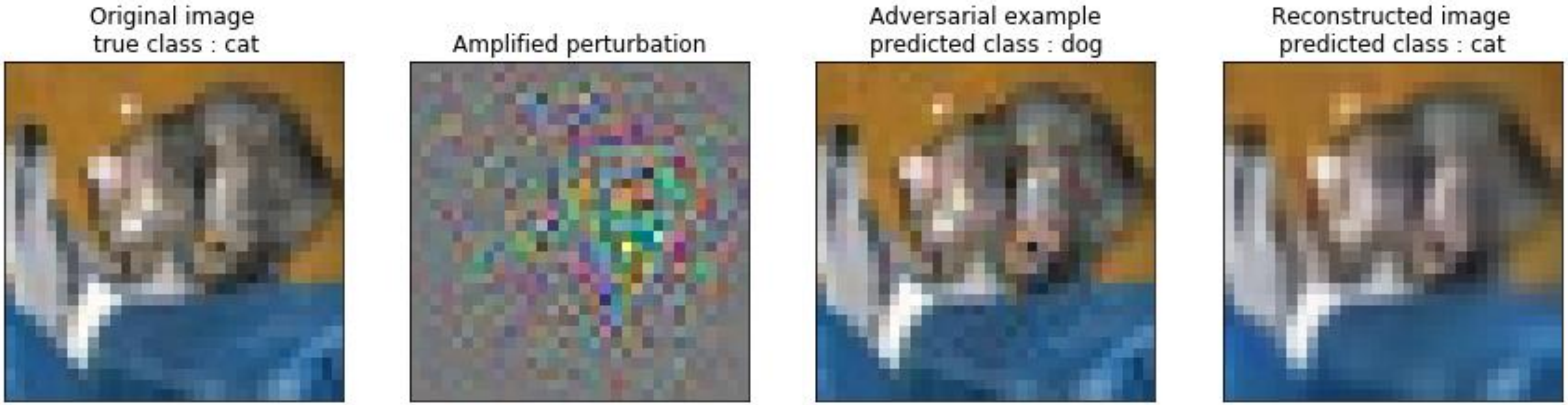
Original images

Noisy images

Filtered images

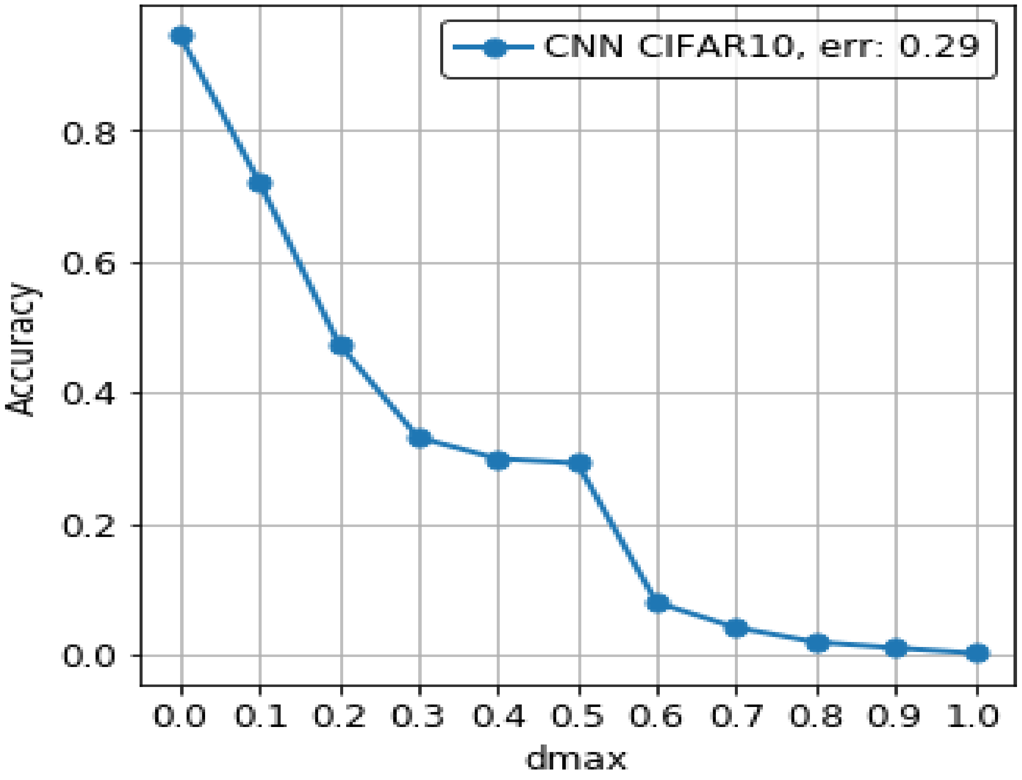
Evasion Attacks on CIFAR10 dataset

- Projected Gradient Descent with Bisect Line Search
- Maximum euclidean distance perturbation of 1.0

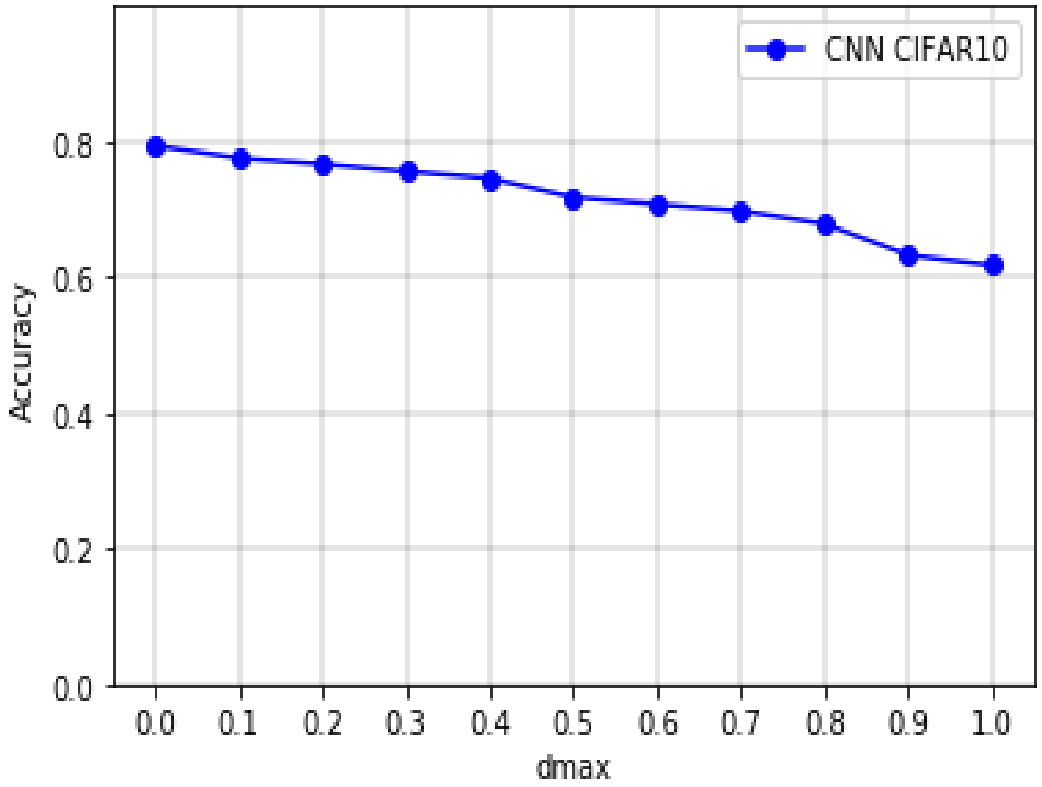


Security Evaluation

SEC before autoencoding



SEC after autoencoding


CNN CIFAR10
Accuracy
dmax