Adversarial Attacks And Interpretability Covid Chestxray Dataset

Cognitive Computing and Artificial Intelligence - UniCT a.a 2019/2020

Raiti Mario O55000434
Nardo Gabriele Salvatore O55000430
Sortino Renato O55000405

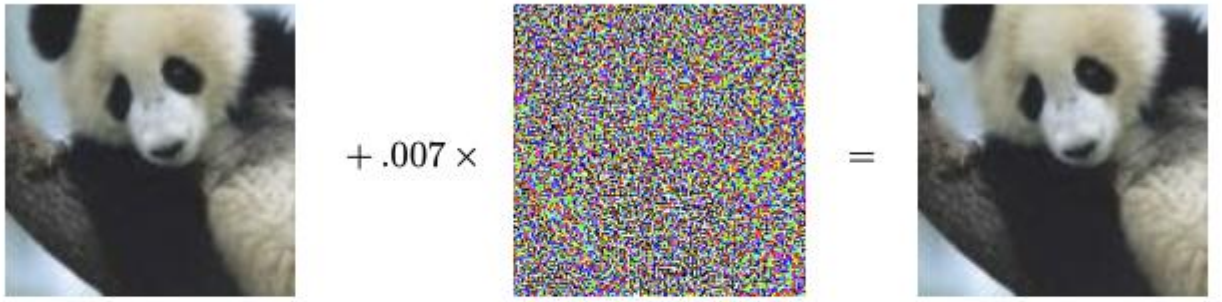
Problem Definition

- The goal of this project is to define a convolutional model that classifies CoViD-19 cases starting from chest x-ray scans.
- In order to make the model more robust to adversarial attacks, the idea is to fine-tune the model with perturbed images after a first phase of training on normal data.

Problem Definition

- To create the adversarial samples, the Fast Gradient Sign Method method has been used
- This method creates an image by adding a small amount to the original image.
- Such quantity is computed multiplying an epsilon (e.g. 0,07) by the sign of the gradient.

Fast Gradient Sign Method



$\boldsymbol{x}$
"panda"
57.7% confidence

$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

$$adv\_x = x + \epsilon * \text{sign}(\nabla_x J(\theta, x, y))$$

Introduction

- The project is available on github at following link : https://github.com/RaiMar96/AdversarialAttacksAndInterpretability-covid-chestxray-
- The Dataset is available on github at following link : https://github.com/ieee8023/covid-chestxray-dataset
- The project has been implemented in python on Jupyter environment, using Google Colab platform.

Dataset Creation

Dataset Creation

- Dataset creation has been realized with the image_dataset class, which takes as input parameters the file paths of CSV file containing the structure of the dataset, and the image folder, the transformations to apply and the relative phase for the subset (Transformations are defined separately for train and test).

Data Transformations

- Data Normalization with mean value 0,5 and variance 0,25
- Resize to 128 pixels
- ToTensor
- Grayscale

Data Augmentation

- Random Rotation of 30 degrees
- RandomCrop of 112
- RandomHorizontalFlip

Dataset Creation

- In the following experiments a subset of the provided dataset has been utilized (images from perspective 'PA'), made of 303 samples.
- After creating the image dataset, the train, validation and test subset are generated using the class torch.Subset, and the relative dataloaders.

Class Distribution

COVID	NO-COVID
201 (66.34 %)	102 (33.66 %)

From Scratch CNN, train and evaluation

From Scratch CNN

- The from scratch model is made of 9 convolutional levels, each of which applies a 3x3 kernel and is followed by ReLU, BatchNormalization.
- Three fully connected layers follow with 1024, 128 and 16 neurons respectively.
- The first two convlutional layers have a stride value of 2. The last two layers are followed by a MaxPooling layer.

Train Details

- Binary Cross Entropy is used as evaluation criterion for loss computation.
- As optimizer, ***Adam*** is used**.**

Model interpretabilty using Integrated Gradients

Model interpretabilty using Integrated Gradients

- Pytorch Captum module is used for model interpretability.
- modelInterpretation function use integrated gradient criterion to define feature meaning after model computations.
- Attributes are plotted trough '*visualize_image_attr*' called on '*visualization*' element of captum.

Adversarial attack FGSM

Adversarial attack FGSM

- The adversarial attack taked into consideration is ***Fast Gradient Sign Method***.
- Fgsm_attack function perform images perturbation that adds noise to the original images, multiplying an epsilon value to the gradient sign of the data. In our experiment, epsilon = 0.025 is used.

Adversarial Training, Fine tuning and evaluation of the model

Adversarial Training, Fine tuning and evaluation of the model

- Same operations performed before, are repeated in adversarial mode.
- A new adversarial image dataset is created; the new loaders are created through concatenation of the original dataset and the adversarial one; to do this, ConcatDataset class is used.
- Model is trained with perturbed images, tuning parameters. Then, evaluation operations are repeated.

Results

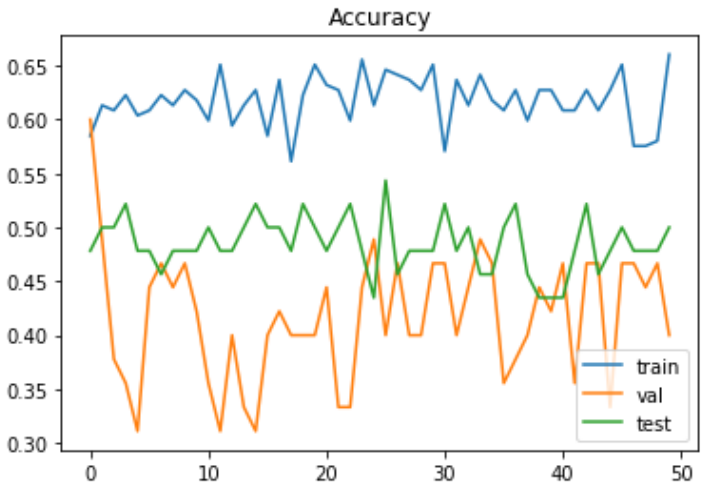
Setup

For test, the following parameters are used:

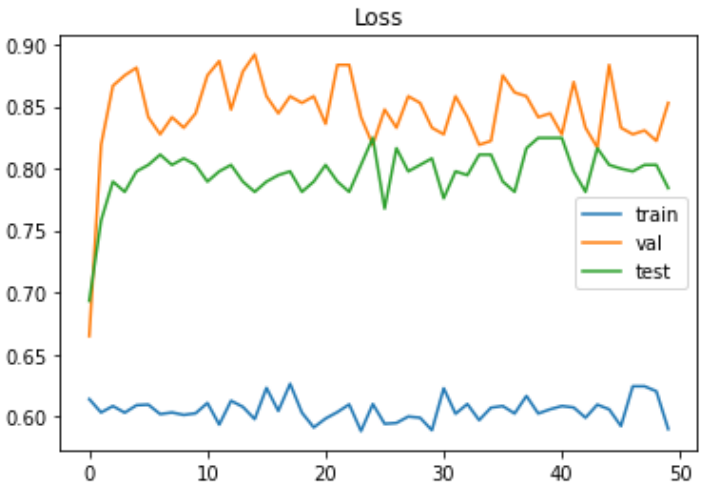
- Batch size = 8;
- Learning rate = 10^-4
- Num of epochs= 50

From Scratch Model Train Results

Accuracy

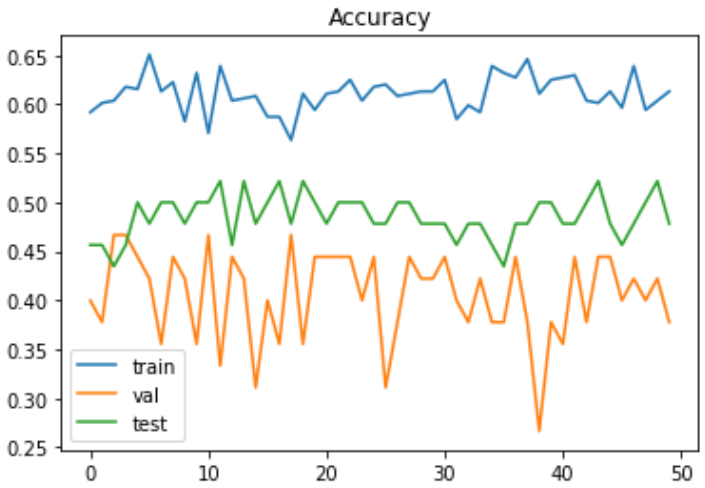


Loss

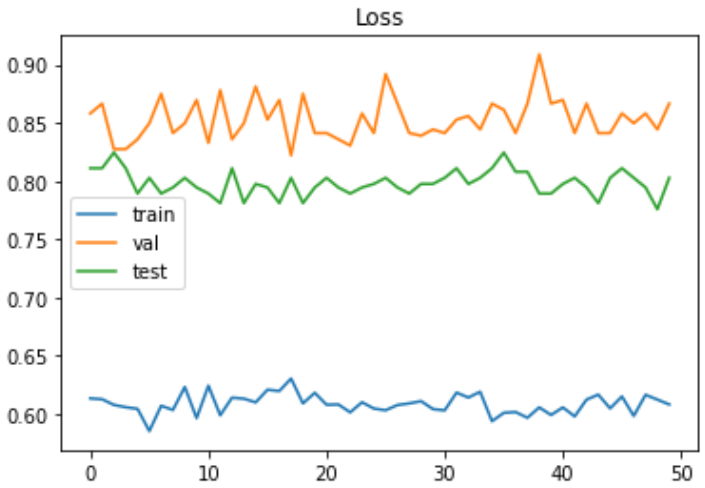


Adversarial Fine Tuned Model Train Results

Accuracy



Loss



Confusion Matrix

True values / Predicted values	COVID	NO-COVID
COVID	93	50
NO-COVID	108	52

Precision = TP/ TP + FP = 93 / 143 = 65.03 %

Recall = TP / TP + FN = 93 / 201 = 46.26 %

From Scratch Model Interpretability Results

Image

Interpretation

Adversarial Fine Tuned Model Interpretability Results

Image

Interpretation

Results Comparison

- Interpretability results are evaluated based on the outcome of the interpretability function. It underlines the differences through the two model's result: if the difference matrix shows values that are zeros (or close to zeros), it means that the evaluated features that have been used in the decision processes, were approximatelly the same.