ADVANCE Labs - Harassment in Images Detection Lab

1 Lab Overview

In this lab, you will keep learning about how AI/ML can be used to detect societal issues such as cyberbullying. Cyberbullying is bullying performed via electronic means such as mobile/cell phones or the Internet. The objective of this lab is for students to gain practical insights into online harassment such as cyberbullying, and to learn how to develop AI/ML solutions to defend against this problem.

In this lab, students will be given a starter-code. Their task is to follow the instructions provided in the Jupyter notebook, train an AI/ML model on the given dataset, evaluate their model, and deploy the model by testing it on their own samples. In addition to the attacks, students will also be guided to perform hyperparameter tuning to further improve the performance of their detection models. Students will be asked to evaluate whether their tuning effort improves their detection models or not. This lab covers the following topics:

- Detection of a cyberbullying in images
- AI-based classifier models to predict cyberbullying vs. non-cyberbullying in images

Content Warning: This lab contains potentially triggering language and deals with difficult subject material. We minimized showing such language samples in this lab. They do not represent the views of the authors.

2 Lab Environment

This ADVANCE lab has been designed as a Jupyter notebook. ADVANCE labs have been tested on the Google Colab platform. We suggest you to use Google Colab, since it has nearly all software packages preinstalled, is free to use and provides free GPUs. You can also download the Jupyter notebook from the lab website, and run it on your own machine, in which case you will need to install the software packages yourself (you can find the list of packages on the ADVANCE website). However, most of the ADVANCE labs can be conducted on the cloud, and you can follow our instructions to create the lab environment on the cloud.

3 Lab Tasks

3.1 Getting Familiar with Jupyter Notebook

The main objective of this lab is to learn how AI/ML can be used to detect online harassment, such as cyberbullying. Before proceeding to that, let us get familiar with the Jupyter notebook environment.

Jupyter notebooks have a Text area and a Code area. The Text area is where you’ll find instructions and notes about the lab tasks. The Code area is where you’ll write and run code. Packages are installed using `pip`, and need to be preceded with a `!` symbol. Try accessing the lab environment for this task here.

The lab has three areas: one text area and two code areas. Follow the instructions for the three areas, fill the three areas with the instructed content and add a screenshot to your report.

3.2 Cyberbullying Detection

In this lab, you will develop AI to detect cyberbullying. You will use a dataset of real world images to train your AI model, evaluate the performance of a pre trained AI classifier model and check the result with one random instance from the dataset. You can access the lab by clicking here.

Approach towards analysing the cyber bullying in images in a dataset, there are three steps: (i) Understand and identify the factors related to cyberbullying in images. (ii)Extract those factors from images. (iii) Examine the usage of those factors in classifier models.

3.2.1 Datasets Selection

In this lab, we provide three dataset: auxes dataset, poses dataset and images.You can download the model and dataset as per the lab instruction.

In this lab, you will be using the auxes,poses and image data set which consist real world cyberbullying images. Run all the code of the lab. Report the validation sets and check with one random instance from the validation dataset.

3.2.2 Load Dataset

Follow the instructions in the text areas and run the subsequent codes to load your data from a predefined class, as follows. Here is a sample from the lab:

```
class PosesDataset(Dataset)
```

This predefined class to check cyberbullying and non cyberbullying images from the dataset, and include the generated tokens in your report. You can add a code block to run your code.

Generate a valid set from poses and auxes dataset by following the lab instruction. Here is a sample from the lab:

```
valid_set = PosesDataset(’cyberbullying_data/cyberbullying_data_splits_clean/test/’, ’
    cyberbullying_data/cyberbullying_poses/test/’, ’cyberbullying_data/
    cyberbullying_data_auxes/test/’)
```

3.2.3 Load pre trained AI classifier model

After you have loaded the dataset, the next task is to load a pre-trained AI classifier model. Follow the lab instruction to load the AI model. Any other classifier model which can be used here?

```
# load vgg16 pre-trained model
orig = models.vgg16(pretrained = True)
```

3.2.4 Generate the detection results from the validation dataset

Now it is time to run your pre-trained model on a validation dataset. Recall that we have already partitioned the dataset into train, validation and test sets. Run your model on the validation data and report your results here. Use lab instruction:

```
with torch.no_grad():
    print('Val loss is: {:.3f}'.format((sum(running_loss) / len(running_loss)).item())
    )
    print('The accuracy for validation dataset is: {}%'.format((correct / total) *
    100))
```

3.3 Check with one random instance from validation dataset

After generating results from validation dataset, now you need to random select an instance in validation dataset using,

```
random_index = random.randrange(len(valid_set))
instance = valid_set[random_index]
```

Plot the image for results as per the lab instructions.

3.3.1 Check the prediction

Now after generating the image randomly from the validation dataset, you need to check if the prediction is correct or not.

Follow the lab instruction to check prediction.

```
print("The AI prediction for this image is: {}, which is {}!".format(annot_label,
    comparision))
```

Report the results here.

3.3.2 Discussion

We trained a real world cyberbullying dataset of images using a classifier model, why are we using the classifier model?

4 Submission instruction

You need to submit a detailed lab report, with screenshots, to describe what you have done and what you have observed. You also need to provide explanation to the observations that are interesting or surprising. Please also list important code snippets followed by explanation. Simply attaching code without any explanation will not receive credits.