TLDR

Using the newly released DMDF_V1, Deepmedia AI endeavored to test the effectiveness of modern deepfake detection techniques on a wide variety of deepfake detection datasets. To accomplish this, we trained a Cross Efficient Visual Image Transformer based upon the architecture released by Davide Coccomini on 8 different deepfake detection datasets, containing over 10,000 fake videos and 6900 real videos. Below contains the first stage results of this machine learning architecture on the first version of DMDF. Though the results are quite promising, we intend to update this deepfake detection effort as DMDF contains to be updated to ensure that our deepfake detectors are being trained on the most sophisticated deepfakes in the wild. For more information, please see the attached white paper as well as the DMDF github page.

Abstract

The work of deepfake detection requires evolution at a rate matching the rapid pace of deepfake generation itself. It is crucially important that deepfake detectors be trained on the highest quality synthetic media available. Implementing the Deep Media Data Files(DMDF), a large and consistently growing collection of deepfake detection data released by Deep Media AI, this project aimed to train a Cross Efficient Visual Image Transformer(CE-VIT) to effectively detect deepfakes across a wide variety of datasets. We modified and trained a CE-VIT network to perform deepfake detection on 8 different deepfake datasets, and have produced this paper to release the results of this experiment. Overall, the detector functioned exceptionally well, with an expected deviance of accuracies between different qualities of deepfake datasets. As DMDF continues to update, and add additional modalities such as voice detection, this document will also be updated.

Methods

For a more robust introduction as to the datasets used in this project, as well as for more details on the size, quality, and robustness of the data, please refer to the DMDF white paper.

For the purposes of this analysis, we modified the CE-VIT architecture to allow for the simultaneous processing of multiple deepfake datasets, and further reports the accuracy of the deepfake detector for each dataset individually. We find this to be crucially important, as the high accuracies of less technologically developed deepfakes have the potential to overshadow the difficulties analyzing advanced deepfakes. By separating results by dataset, we will be able to directly observe the accuracies on the most advanced deepfake datasets.

We performed logging and analysis using tensorboard, and results will be presented through tensorboard logging graphs. As we continue to progress through epochs of training, we will continue to update the results of this tensorboard logging. After only 36 hours or so, we found the results to be sufficiently compelling as to share.

Results

Below includes the tensorboard logging for nearly two days of CE-VIT runtime. Please check back to this github page, as the result logging will be updated quite rapidly.

One important note to mention is that due to the splitting of accuracies by datasets, and the varying size of individual datasets, certain datasets are selected by the network far less than others, some of which significantly less than others. For this reason, we observe certain logging graphs initially reporting accuracies at 100%, but as the training continues they appear to fix themselves. Further work is being done to see if it is possible to perform this kind of individual analysis without these logging inconsistencies.

Discussion

Even from a limited runtime, it is clear that the CE-VIT architecture seems very well adapted to the work of deepfake detection. Moreso, this project emphasized the necessity of deepfake detection tools to be trained on the most advanced deepfake generating techniques. As can be illustrated by the discrepancy between an advanced dataset such as deepfacelab and a more outdated deepfake generator such as DFDC.

To continue advancing our knowledge of how deepfake detectors function against a wide variety of deepfake generation techniques, it will be crucially important to continue testing our detectors against new and advanced deepfakes. For this reason, we intend to continue updating our CE-VIT results as the Deep Media Data Files(DMDF), the largest publicly available deepfake detection resource ever released, continues to add datasets and additional modalities.

An important secondary question of the CE-VIT infrastructure is how it might handle modalities such as voice altered deepfakes. Be on the lookout for the next update of DMDF at its homepage.