Proposal: Determining Crowd's Ability to Distinguish Deepfakes in Images and Videos

Suhas Dara
Department of Computer Science
The University of Texas at Austin
suhasdara@utexas.edu

Aditya Tyagi
Department of Electrical and Computer Engineering
The University of Texas at Austin
adityatyagi6498@utexas.edu

1 Introduction

The overall theme of the class has revolved around fake news and integrity detection. It comes as no surprise that exploitation and weaponization of social media to spread misinformation is a major issue in The United States. When discussing the idea of "fake news", Figueira and Oliveira refers to ideas of news articles that are tagged with catchy headlines and inaccurate or distorted information (Figueira and Oliveira 2017). It should be noted that fake news is just another issue in the current issues of misinformation and there is an ample amount of research going on to mitigate the spread as well as stop it all together.

Although there is a plethora of academic research available when attempting to stop fake news, misinformation spreading through deepfakes comes as a new challenge: deepfake images are face swapped images that can now be done through readily available GPUs. The ease of access to such technology incentives people to create deepfakes for entertainment purposes, as well as for target attacks to spread misinformation about certain individuals and institutions (Dolhansky et al. 2020). To combat such issues, there has been extensive academic research with datasets created by institutions such as Facebook to understand how to identify deepfake media as well as create algorithms and new ways to mitigate the spread of misinformation through deepfakes.

2 Summary

This section will divide up the proposal to meet the rubric requirements

2.1 What

We plan on conducting a user study in the hopes to gain more information to increase deep fake detection using crowd sourcing. For the project, we will be using Amazon Mechanical Turk (AMT) crowdworkers. We shall also have a combined budget USD 300, given to us as class credit, to ensure that we are able to conduct comprehensive data collection and pay the workers an appropriate amount as well.

2.2 Why

With our curiosity in deepfake detection, we want to create a user study that will ensure that future research work is able to tap into crowdsourcing as one of the options to successfully detect deepfakes. Though there has been previous work done in this field, such as Korshunov and Marcel in their work to compare the accuracy between humans and machines when it comes to deep fake videos (Korshunov and Marcel 2020), the research does not focus on the different factors that influence the human detection and how to fine-tune them for best results in accuracy, but rather focused on how the accuracy of machine detection compares to human detection. The study also does not focus on crowdworkers but rather 60 participants who were not randomized as well. Further, there have been surveys such as Nguyen et al. that aim to understand the key features that a machine model would need to lookout for in order to conclude if an image or a video is fabricated. However, they do not take account of any human detection and features that enable them to see if the media is a deepfake or not.

We aim to pose some research questions that enable us to conduct an exhaustive user study that would help future academic interests as well:

RQ1: How well can the crowd detect images and videos?

RQ2: What demographic of crowd has the highest accuracy?

RQ3: Can the crowd reason through their annotations?

RQ4: What features does the crowd use to aid in deepfake detection?

Throughout the project, we aim to answer these questions in an effort to create an exhaustive user study that would enable researchers to understand the factors of influence for the crowd when it comes to deepfake detection and mitigate/exploit them as need be.

2.3 How

With our background in Computer Science and Engineering, we plan on conducting heavy data analysis offline through the use of Python 3.7. From the crowdsourcing perspective, we plan to use our limit of USD 300 to create task designs that are aimed to gather basic demographic information from the workers, as well as their annotations. We plan to use techniques such as honeypots and attention checks to ensure that we filter out bad data points as well as calculate

inter-annotator agreement data to ensure that if we observe any outliers in the data points, we are either able to explain the cause or are able to replace it. We also plan on tapping into different dimensions of a crowdsourcing task, as described by Sakamoto et al. in an effort to effectively communicate with the workers, as requesters. This would ensure that we are able to use the worker pool efficiently by creating tasks that suit them and convey our expectations well (Sakamoto et al. 2011). The task will be iterated over to ensure that all instructions and other aspects of the task design are clear. We also plan to release versions of the task at different times to ensure that we attract majority workers from The United States and India, AMT's biggest population of workers (McAllister Byun, Halpin, and Szeredi 2015). This should also ensure that we don't overuse the worker population by publishing multiple tasks over a short period of time.

2.4 When

The overall timeline of the project will be from November, 2020 through the first week of December, 2020. Though the milestones will be further expanded on, the timeline will mimic the rest of the course timeline for Fall 2020 semester.

3 Milestones

Dataset aggregation / sampling (November 1, 2020) For this project, we will require two different datasets. The first dataset will consist of videos sampled from the Facebook Deepfake Detection Challenge Dataset (DFDC)[1], which features 124,000 videos. These videos will be contain deepfakes and real videos of people. We will utilize these videos to analyze the capabilities of crowdworkers in detecting deepfakes from videos. The second dataset will consist of deepfake and real images of people. There is currently a lack of image datasets for deepfakes, which is the reason we will be sampling our own dataset for deepfake images. We have two methods to create our image dataset. The first method is using the website whichfaceisreal.com that compiles deepfake and real images of people in a game format. The deepfake images are created using a Generative Adversarial Network (GAN)[2] and are open source. The real images are sampled from the Flickr-Faces-HQ (FFHQ) Dataset[3], which is public domain. Another method to create our images dataset will be by sampling frames from the videos in the DFDC dataset. This will be a more cumbersome process as the frames chosen should not be blurry and should contain the person in the center.

Designing HITs and data collection (November 14, 2020) We will be collecting data by publishing Human Intelligence Tasks (HITs) on Amazon Sagemaker Ground Truth (which will publish the tasks to AMT for crowdworkers to annotate.) The crowdworkers will be asked to perform annotations on our datasets. The images/videos to be annotated will be batched together in small batches for the crowdworkers. Along with each batch of data to be annotated, crowdworkers will also be asked for demographic information. For videos, they will potentially also be asked for any supporting proof and/or rationale for why they believe a certain video is either a deepfake or not. We will also add honeypots and attention check questions to the HITs to ensure for good data quality. However, honeypots will be difficult to create for this particular task because of the complexity of the task being high in the first place.

Status report (November 17, 2020) In the three weeks, we hope to accomplish sampling our dataset, designing the HITs for our user study, and collecting the annotated data from the crowdworkers. In the status report, we will explain the method chosen for building our datasets, our thought process with the HIT design, and the raw data that we collected from the crowdworkers with basic statistics. This would be the first half of our final report.

Filtering bad data (November 22, 2020) As part of the user study, filtering bad data will be an important precursor to the analysis of the data. This bad data may be intentional or unintentional (outliers). The attention checks that will be added to the HITs will be one way of knowing whether crowdworkers are annotating the dataset diligently. Attention checks will help filter out inattentive crowdworkers, but malicious workers will still be hard to catch. Additionally, considering the complexity of the task, honeypots may not be a very effective metric to filter data. We will have to observe whether the crowdworker accuracy and inter-annotator agreement is better with or without filters utilizing honeypots.

Analysis of data visualization (November 29, 2020) Along with the filtration of bad data, our preliminary analyses will begin. The analyses will include the accuracy and the inter-annotator agreement of the annotated labels for both the videos dataset and the images dataset and a comparison of where the workers performed better. Additionally, we will utilize the demographics data to understand which groups of workers (age, race) perform better at identifying deepfakes. The last part of our analysis will be to dissect the crowdworkers' rationales and proofs and see common patterns of why crowdworkers think a particular image/video is a deepfake or not. All these analyses will then be converted into appropriate statistical representations for visualizing the outcomes of our user study.

Presentation (December 1, 2020) We aim to have the majority of our analyses complete by the presentation day so that we have visual representations available to share with the rest. We will share our methodology and results.

Final report (December 7, 2020) We plan to re-purpose our status report to create our final report. Other than the methods utilized in our user study, the final paper will include all our analyses and our data visualizations as well. The final paper will note our contributions to the field of study and also the potential for other research that can build upon the research to be conducted by us.

[1] https://ai.facebook.com/datasets/dfdc/

[2] https://thispersondoesnotexist.com/

[3] https://github.com/NVlabs/ffhq-dataset

4 Risks

Time constraints The problem that we have chosen to tackle is a difficult to accomplish because of the complexity of the task that we are studying and the partial dependence on crowdworkers to be able to perform the task. So, one risk that we face is definitely the time constraints in place to finish the project before the end of the semester. Additionally, our later milestones are dependent on the designing HITs and data collection milestone, so any discrepancies in this stage will push our remaining milestones back.

Receiving bad data Our entire user study is dependent on the annotations that we collect from the crowdworkers. The study is at risk if we receive a lot of bad data whether it is intentional or unintentional. This is because the task at hand is difficult. If we lose a lot of data through filtration, our results will no longer have sufficient weight or meaning. Combined with the time constraints, if we do not spend sufficient time designing our HITs to provide the highest clarity and engagement for the crowdworkers, then the risk of bad data is much higher.

Overusing worker pool Another problem that we may face during our user study is potentially overusing the crowdworker pool on AMT. This is because Amazon Sagemaker GT does not have options to limit the number of HITs a single crowdworker can perform unlike AMT directly. If a single worker performs multiple HITs, we will lose data on the demographic end. We will have to chose an appropriate batch size of images/videos from our datasets when creating the HITs. Also, we will have to price the HITs to slightly discourage a single worker performing multiple HITs. Another way we plan to tackle the issue is by staggering the release of batches to ensure we collect data from US workers as well as India works to further diversify our demographics. Lastly, we need to inform the workers to only perform a single HIT. However, none of these methods can fully guarantee a diverse crowdworker pool.

Exhausting funds A last problem that we may face is running out of funds. We have been collectively allocated USD 300 in funds on Amazon Sagemaker GT. If we run out of funds because of going overboard with testing or because of inappropriately pricing our HITs, we may have to end our study early, or pitch in funds from outside the funds already allocated.

Appendix

Sections 1 and 2 are written by Aditya Tyagi. Sections 3 and 4 are written by Suhas Dara.

References

Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; and Ferrer, C. C. 2020. The deepfake detection challenge dataset.

Figueira, , and Oliveira, L. 2017. The current state of fake news: challenges and opportunities. *Procedia Computer Science* 121:817–825.

Korshunov, P., and Marcel, S. 2020. Deepfake detection: humans vs. machines.

McAllister Byun, T.; Halpin, P. F.; and Szeredi, D. 2015. Online crowdsourcing for efficient rating of speech: a validation study.

Nguyen, T. T.; Nguyen, C. M.; Nguyen, D. T.; Nguyen, D. T.; and Nahavandi, S. 2019. Deep learning for deepfakes creation and detection: A survey.

Sakamoto, Y.; Tanaka, Y.; Yu, L.; and Nickerson, J. V. 2011. The crowdsourcing design space. In Schmorrow, D. D., and Fidopiastis, C. M., eds., *Foundations of Augmented Cognition. Directing the Future of Adaptive Systems*, 346–355. Berlin, Heidelberg: Springer Berlin Heidelberg.