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

Abstract

This paper provides a user study to explore how crowdsourcing can be used to detect deepfake media. We quantify our exploration with Amazon Mechanical Turk crowd workers and conducting surveys and Human Intelligent Tasks (HITs). Workers are asked to distinguish between real and fake media to the best of their knowledge and are asked to fill out a brief demographic survey. We then conduct further analysis to understand how well do the crowdworkers perform and recognize patterns, if any. We then conclude our results and user study with the quantified data and any recommendations for future research involving deepfake detection using human computing element.

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 among 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). Such convincing media can be easily spread in social media to be readily available to unsuspecting citizens. One of the more infamous examples is from President Obama and his video posted by Vincent where Jordan Peele shares the grim reality and the ease of creating a deepfake video that can make politicians such as the former president of The United States say offensive and inciting statements, as seen in Figure 1 (Vincent 2018). Furthermore, deepfakes can get more detailed as more data is available to train the deep learning algorithms. President Obama's video was created after feeding the algorithm fifty-six hours of sample recording (Update 2018) indicating that a large number of sample recordings or input can result in more convincing media. This is especially a problem with celebrities and politicians who make numerous public appearances, thereby having more samples available publicly.

Figure 1: Screenshot captured from the infamous Obama deepfake video where the former president says aggravating words against President Trump, leaving the audience with a eerie reality that they cannot trust everything that they see and hear on the internet. The screenshot is captured by Fagan, with Business Insider (Fagan 2018).

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.

This paper will focus on conducting a user study in the hopes to gain more information to increase deepfake detection using crowd sourcing. The experiment crowdworkers will be assigned from Amazon Mechanical Turk (AMT) through the use of Amazon Sagemaker Ground Truth (GT). The overall budget of the project will be constrained under 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. Account credentials were provided through Amazon Web Services (AWS) Edu-

cate services by Amazon [1].

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: Does the crowd detect deepfaked videos better than deepfaked images?

RQ2: Does the crowd reason differently for images and videos?

RQ3: What features does the crowd use to aid in deepfake detection?

RQ4: What demographic of crowd has the highest accuracy?

RQ5: Does performing well on a CRT test imply a higher accuracy at identifying deepfakes?

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 that influence the crowd when it comes to deepfake detection and mitigate/exploit them as need be.

The overall timeline of the project is from November, 2020 through the first week of December, 2020.

2 Related Works

There has been extensive academic work to understand and mitigate the spread of deepfakes: from the creation of surveys to the creation of new algorithmic approaches in order to incorporate Artificial Intelligence in the combat of spreading misinformation through media.

Some previous work in the field, such as Korshunov and Marcel in their work compare the accuracy between humans and machines when it comes to deepfake videos (Korshunov and Marcel 2020), that aims to find how well human subjects are able to detect deepfake videos. The research, however, 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 and pooled from a controlled environment.

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. The survey goes through different modes of detection of deep fake media, including physiological indicators. However, they do not conduct quantified experiments of any human detection and features that enable them to see if the media is a deepfake or not.

Moreover, Yang, Li, and Lyu explored the creation of deepfake images to detect inconsistent head poses as indicators of fake image. The authors trained Support Vector Machine (SVM) classifiers that detect the error that is posed when AI creates an image by splicing synthesized face region on an original image. However, the paper quantifies its experiment by studying the detection only through head orientation vectors and attempting to create a classifier (Yang, Li, and Lyu 2018), it does not incorporate the human element when detecting fake close up images that the authors trained their model to do.

Overall, there were multiple academic publications that aimed to mitigate the spread of incorrect information through detection of fake images and media. However, a lot of the research is focused on the creation of better classifiers and automated detectors. We aim to utilize the pool of crowdsourcing workers to divide detection in micro tasks as opposed to using classifiers or AI.

3 Experiment

We plan on conducting heavy data analysis offline through the use of Python 3.8. 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 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.

3.1 Dataset

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)[2], which features 124,000 videos. These videos will either contain deepfakes or real footage 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 considered two methods to create our image dataset. The first method to create our images dataset was by sampling frames from the videos in the DFDC dataset. This will be a more cumbersome process as the frames would need to be chosen carefully and manually to ensure they are not blurry and should contain the person in the center. We decided to instead go by method two. We used the website whichfaceisreal.com that compiles deepfake and real images of people in a game format. The deepfake images are created us-

[1] https://aws.amazon.com/education/awseducate/

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

ing a Generative Adversarial Network (GAN)[3] and are open source. The real images are sampled from the Flickr-Faces-HQ (FFHQ) Dataset[4], which is public domain.

We chose 20 real videos and 20 deepfake videos randomly from the DFDC dataset that were short in duration and had a file size of lower than 5MB. Additionally, 20 pairs of real and deepfake images were sampled from the game website. With this, we compiled a total of 40 videos and 40 images. To avoid bias in aggregating the images from the website, no pairs of images were skipped regardless of whether it seemed easy or difficult to identify which image in the pair was a deepfake, and the first 20 pairs were all chosen.

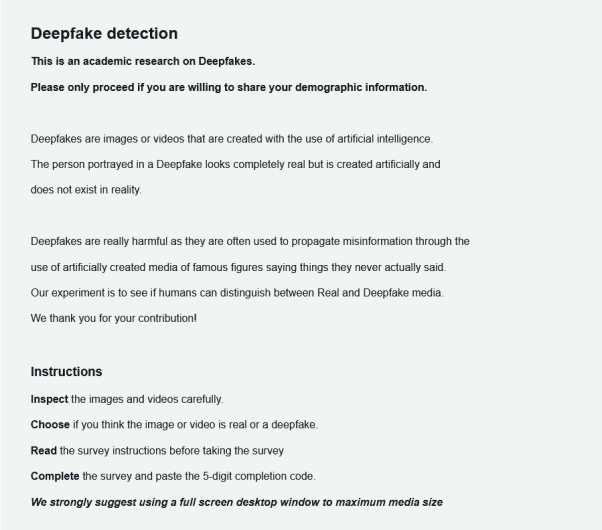

Figure 2: Screenshot of the current HIT instruction design that the workers on AMT will see.

3.2 Task design

Data will be collected by publishing Human Intelligence Tasks (HITs) on Amazon Sagemaker GT, which publishes the tasks to AMT for crowdworkers to annotate. Crowdworkers will be asked to annotate the images and videos as real of fake. We will use the available ground truth extracted from the game website and the DFDC dataset to analyze the responses of the crowdworkers. To understand trends in the crowdworkers performing our HITs, the HITs will contain a batch of media - two images and two videos each. Additionally, each HIT will be annotated by five different crowdworkers. The images and videos selected for each HIT are completely randomized from the dataset and it is possible that a single HIT will have two real images or two deepfake images instead of being balanced with one of each. This is to mitigate correct answers that will corrupt the data due to any potential uniformity bias.

The crowdworkers will also be asked for their reasoning behind why they think the images or videos are real or deepfakes. They will only be asked for this rationale after annotating the two images in the batch, and again after annotating the two videos in the batch. This is to avoid fatigue among the crowdworkers by answering multiple free response questions. It additionally provides us with an opportunity to analyze potential differences in their thinking processes when evaluating images versus evaluating videos. A part of an example HIT can be seen in Figure 2 and Figure 3.

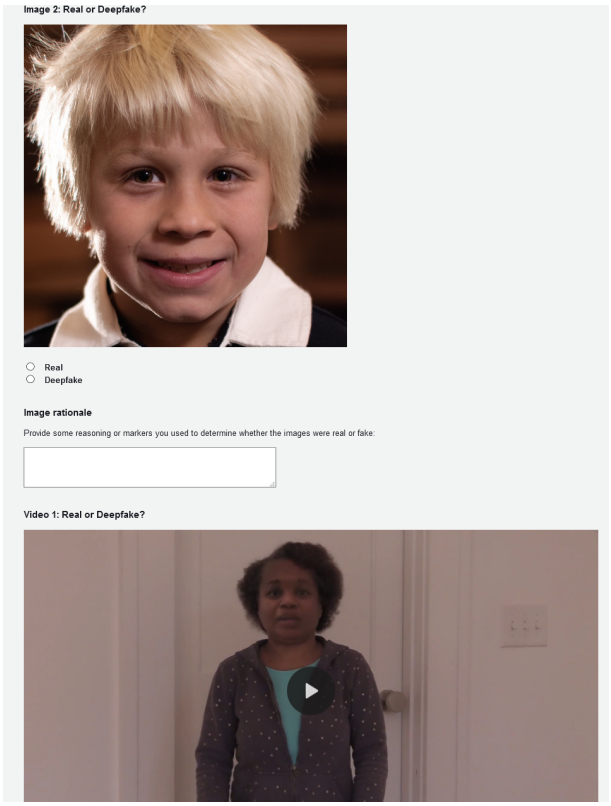

Figure 3: Sample of the HIT that consists of two images and two videos for annotation, two free-form fields to provide rationale for images and videos, and a link to the demographic survey.

Along with the annotations, crowdworkers will also be asked for their demographic information and to answer three CRT questions. This data will be used in the analysis of the annotations. This data is not asked directly in the HIT. This is so that crowdworkers do not have to fill the demographic information multiple times if they choose to perform multiple HITs. Instead, the method of asking for the demographic information and the CRT responses is through a Qualtrics survey. A link to the survey will be available in the HIT. The survey provides a "Survey Completion Code" at the end of the survey which the crowdworkers will be asked to paste in their HIT response. If a crowdworker performs multiple HITs, they can use the same completion code each time. The survey responses will be combined with the annotation responses during analysis with the help of the completion code. A portion of the survey can be seen in Figure 4.

As part of the user study, filtering bad data will be an important precursor to the analysis of the data. This bad data

[3]https://thispersondoesnotexist.com/

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

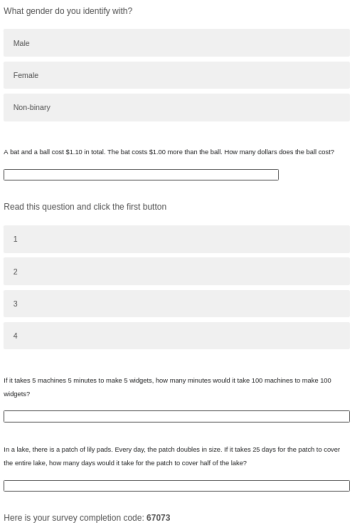

Figure 4: Screenshot of a portion of the survey that the workers on AMT will fill out. The HIT consists of three demographic questions, three CRT questions, and an attention check. The survey completion code can be seen at the bottom.

may be intentional or unintentional (outliers). One way of knowing whether crowdworkers are annotating the dataset diligently is by the use of attention checks. Attention checks will help filter out inattentive crowdworkers, but malicious workers will still be hard to catch. An attention check will be added to the survey instead of the HIT itself. This is to prevent spam demographic information. Additionally, the input field asking for the survey completion code in the HIT will act as another attention check in itself. This is because it is not a completely obvious step and if the crowdworker does not properly read the instructions, they might not realize to save the code from the survey.

Another way of filtering bad data is through the use of honeypot questions. However, considering the complexity of the task, honeypots will be difficult to create/choose from out dataset and may not be a very effective metric to filter data. The deepfake game website provides some useful insights about how to detect deepfakes[5]. According to the website, some dead giveaways are water splotches and bad backgrounds. Maybe images and videos with these issues can be carefully handpicked from the datasets to be chosen as honeypots. However, observing whether the crowdworker accuracy and inter-annotator agreement is better with or without these filters, especially honeypots, will be an important part of the analysis.

3.3 Deployment of tasks

To ensure that we do not overuse the worker pool, the HITs will be released in batches. 50% of the batches will be released in the working hours of the United States to cater to the worker pool here. To ensure that all our work is seen by the US population, the second batch will be delayed by at least 5 business days in US and then published in the working hours of India to ensure that we capture the largest pool of workers by catering to India and US. Although there is a risk that some Indian workers will see the tasks deployed for US workers due to latency or because the tasks are "in queue", this method will ensure that the task is viewed by a diverse audience and has the opportunity to be completed with the most minimal overlap for a study that is bound to take a few weeks without violating any privacy issues and filtering workers through some personally identifiable information.

3.4 Analysis of data

With the data collected, several different analyses will be conducted to answer the research questions presented in section 1. The data collected will be in the post-processed form from Amazon Sagemaker GT, and in CSV format from the Qualtrics survey. For each HIT, data will include four annotations (two images, two videos), two rationales, and a survey completion code. The survey responses will first be matched with their appropriate completed HITs, and surveys with completion codes that do not have associated HITs will be discarded. The data will then be filtered through the attention checks and potential honeypots.

To determine crowdworker accuracy in general, the annotations will be aggregated in different formats. An overall accuracy metric will be computed to evaluate the general performance of crowdworkers at detecting deepfakes, regardless of the media type. To understand individual worker accuracy, the accuracy of annotations will be calculated based on the survey completion codes and then averaged. To answer RQ1, the next accuracy metric will aggregate annotations based on each type of media. This is to determine whether crowdworkers perform better at detecting deepfaked images or deepfaked videos. We hypothesize that the crowd is better at detecting deepfake videos than deepfake images because of additional motion information being available.

To answer RQ2, the free-response rationales of the crowdworkers will be analyzed manually. In the responses to the HITs, separate rationales are collected for the crowdworkers' choices for images and videos. We expect a lot of these rationales to be potentially empty or unusable, but from the legitimate responses, we hypothesize that the rationales for videos will contain a lot more responses that emphasize fake human motion, than the human figure themselves. To answer RQ3, once again the free-response rationales will be analyzed. Although, it is not explicitly asked to provide any specific features that are utilized in making their decisions, it is likely that this type of information will be implicitly stated as part of their reasoning.

To answer RQ4, the annotations will be once again aggregated in different formats. In the survey, we collect four different demographic data - age, race, gender, and education. The annotations from the HITs will be grouped according to different demographic categories to see if the crowdworker accuracy is affected by any of these demographic data. We expect age to play some role in being able to better detect

[5]https://www.whichfaceisreal.com/learn.html

deepfakes, with younger people performing better. We also expect groups with better education to perform better as it they probably have more honed reasoning skills through the academic process. However, we do not expect race and gender to have any impact on the ability to detect deepfakes.

To answer RQ5, the CRT responses collected from the survey will be utilized. The CRT contains the standard three cognitive ability question as described by Toplak, West, and Stanovich (2011.) The annotations will be grouped based on the score received on the CRT test (between 0 and 3), and the accuracy will be calculated. An issue with CRT is its variability with our demographic factors. As shown by Welsh, Burns, and Delfabbro, gender and level of education did effect the results of the CRT, although age does not (2013.) This means some of the variability in the accuracy of the CRT groups will be contributed by the differences in the demographics. However, we still expect that a higher CRT score implies a higher accuracy at detecting deepfakes, as the higher score implies better reasoning skills.

Appendix

*Group Contributions

Sections 1, 2, 3, 3.3, and Abstract are written by Aditya Tyagi. Sections 3.1, 3.2, 3.4, and Appendix are written by Suhas Dara. Both authors also did minor contributions to sections other than the ones mentioned.

*Milestones changes

Designing HITs and data collection (November 14, 2020 $\rightarrow$ November 21, 2020) Our data collection phase is a little pushed back. This is because we spent some time iterating over the design of our HIT and the survey to add more components as required. We have been thinking ahead from the perspective of analysis as well, and that's what delayed our response collection phase. We will roll out our HITs over the course of the next week and collect responses.

Filtering bad data (November 22, 2020 $\rightarrow$ November 25, 2020) As a result of the previous push-back, we will also push back our bad data filtering milestone. We will start writing code next week as we roll out our HITs.

Other milestones Other milestones will remain the same to be on track for completing the project by the end of the course.

*Risks changes

There are no changes to the potential risks we will face during the remainder of the project.

References

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

Fagan, K. 2018. A viral video that appeared to show obama calling trump a 'dips—' shows a disturbing new trend called 'deepfakes'.

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.

Toplak, M. E.; West, R. F.; and Stanovich, K. E. 2011. The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & cognition* 39(7):1275.

Update, M. 2018. Are deepfakes the new fakenews?

Vincent, J. 2018. Watch jordan peele use ai to make barack obama deliver a psa about fake news.

Welsh, M.; Burns, N.; and Delfabbro, P. 2013. The cognitive reflection test: How much more than numerical ability? In *Proceedings of the Annual Meeting of the Cognitive Science society*, volume 35.

Yang, X.; Li, Y.; and Lyu, S. 2018. Exposing deep fakes using inconsistent head poses.