Assignment 2 Final Report

Team A++: Ziyue Chen, Ying Wang, Cheng Shi, Zichao Wang, Ziquan Chen, Wenting Shang

Abstract: This assignment is focused on practicing content and image extraction and then using those features for generation and detection of falsified media. Tika was used to extract the content of existing Bik papers, which were put into the Grover model for discrimination and generation of fake paper contents, and we found that the model were able to detect falsified media. Celebrity faces were also used to generate fake faces and inserted into papers, and Tika Docker was used to add captions for them. All the features were combined, and we were able to create fake scientific papers successfully.

1. Introduction

In this report, we will mainly discuss our observations during the process of generating and detecting falsified data using Grover. We would like to briefly mention other processes in this assignment as well such as extracting text using Tika, generating a falsified face using DCGAN, generating captions for the fake face images using Tika Dockers, and putting things together using LaTeX.

2. Tika Content Extraction

First of all, by extracting dois and pmcids from Pubmed websites, we downloaded the papers from Pubmed and other sources. To use the discriminator for the Bik papers and to generate new fake papers from them, we began with extracting the content from the 214 Bik papers using Tika. The package was very easy to use, as we could simply import the list of papers, ask the package to parse the entire documents, and load the content section into a json file. The contents seemed to be accurate with no obvious errors, so we moved on to use them as the training set to generate fake papers and for discrimination using Grover.

3. Grover Text Generation

While having the text we extracted from the original paper, we use Fake Name Generator to generate 500 fake names as well as fake domains. Then we also generated some random dates, and took the first 1000 words as the summary. After compiling all the needed information into a .jsonl file, we use Grover to generate the text.

Surprisingly, the generated text was readable and some of the texts were quite believable to us, but Grover discriminator was able to detect most generated content as falsified media. At the same time, Grover also recognized most of the original Bik papers as falsified media, so we thought adding ancillary features would not help to make media falsification solvale. And we noticed that the grover discriminator only takes domain, date, author, title and article as metadata, which means the newly added features are not considered as metadata in the model, so the discriminator is not able to solve media falsification.

4. Grover Text Discriminator

In this step, we input the contents for both the original Bik papers and the generated fake papers, and both types are regarded as false media in the Grover text discriminator. The average rate of Machine:Human for all papers is 0.74:0.26, and the average rate of Machine:Human for bik papers is 0.75:0.25; thus, there is no big difference on the result even though we know that bik papers are written by humans. We are not sure that if the Grover discriminator is good at discriminating the academic papers for the following reasons. First, Grover discriminator is trained on news instead of academic papers; second, the discriminator always says the papers are made by machines including all humans' papers; third, the bik papers have modified or plagiarized behaviors that may influence the accuracy of the discriminator to some extent. Therefore, we are hard to assert the Grover discriminator did a good job on differentiating true and fake papers.

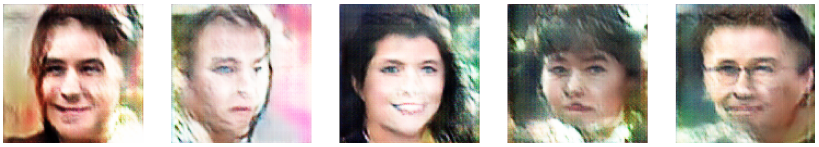
The Grover discriminator used some metadata, such as author, title and published data, as the features to distinguish the machine-made or human-made among texts. However, this tech also contains the whole article content as one of the features. Thus, we regard Grover discriminator as using both ancillary metadata features, and actual content based techniques. In our opinion, the actual content based technique is better than the metadata, because the metadata has limited information about the content when we aim to determine whether the content is written by machine or human. The metadata may include some distractors sometimes, such as the date or the author name, which is hard to be helpful for a text discriminator.

5. DCGAN Face Generation

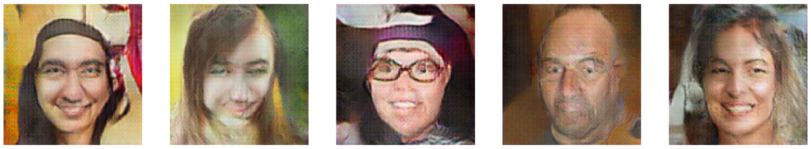
As the professor recommended in the guideline, we used the DCGAN model to generate faces for each of the false authors for our 500 papers. Since the dataset used in the original article is offline and it was difficult for us to upload it to google drive due to its large size. We used the ffhq dataset which is already on google drive so we don't need to upload it again.

By training 70k celebrity faces, we got over 500 new authors' faces. We set the papermeter "Epochs = 30", and we expected the bigger epochs number, the face more clear, and the loss function more converged. Also, "Samples_to_Show = 500" so every epoch will generate 500 new faces. From the figures of Epoch 8, 20, 28, we could see that the picture quality of epoch 20 improves largely from epoch 8, and there is no big improvement from epoch 20 to epoch 28. However, even though we spent over 10 hours running 30 epochs, given the low resolution and the twisted look of the faces, these faces might not be believable enough.

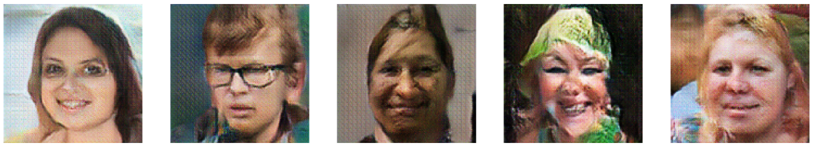
Epoch 8 Photos



Epoch 20 Photos



Epoch 28 Photos



Also, since the dataset we used contains considerable variation in terms of age, ethnicity and image background, there may be some pictures for authors that do not seem to align with age. After consideration, we decided to drop off these people whose faces seemed too young to be an author, such as babies'. And we dropped off pictures that are very vague as well.

To conclude, the falsified images seem to have a low resolution and were a bit blurry since the very blurry ones were dropped, but fortunately we were still able to recognize faces. Additionally, some of them even seemed very similar to some celebrities, which was reasonable as the training data includes celebrity faces. If we were to increase the training set, we might be able to obtain more reliable results.

6. Docker Image Caption

To generate captions for our fake face images, we used the Tika Dockers package for Image Captioning and Object. First, after installing docker successfully, we used the "build" and "run" commands for Docker to start the environment. Then, with a list of image urls for the fake images, we used the requests command to obtain the result for each image and extracted the first "sentence" attribute to be the caption as it had the highest confidence. In the end, the captions were exported for generating fake papers. Through the whole process, we felt the package was really straightforward and easy to understand, such that as long as we have urls for the images, the captions could be obtained. However, it also seemed that the requests took a rather long time to return responses, thus making the process slightly time-consuming. Also, we inspected the captions, and we found that some of them didn't make sense, which might be caused by the low resolution of the fake faces. Therefore, we realized that, if we had images with better resolutions, this object identification and captioning package could theoretically function at its best.

7. LaTeX and PDFs Generation

Our last step was to put things all together including title, author name, generated text, falsified images, and generated image captions by using pylatex in python to generate LaTeX files. In the python file, we load different inputs from previously generated files such as .jsonl files, .png files and .csv files. We noticed some of the titles were including some special characters that caused errors so we preprocessed them by replacing them with an underscore. We also wrote functions to format the pdf by setting up sections in LaTeX and filling the text with inputted generated text. Finally, by using generate_tex command and pdflatex command, we generated 500+ full fake papers of both .tex files and .pdf files.

8. Additional Thoughts

To generate a more believable-looking paper, one of the ways is to add references and citations into the paper, such as adding square brackets and numbers at the end of some random sentences to pretend as in-text citations. In addition, improving the resolution and “Epoch” value when generating the fake images would apparently be more effective in adding more realism to the fake papers to “cheat” a journal.

Thinking more broadly, other types of datasets could have been used to generate the falsified contents as well, such as the video type and the audio type from the MIME types we learned from the lecture. Even though videos and audios are unlikely to appear directly in a paper, they can still be one of the sources of citations in academic papers. One of the most famous techniques these years is the DeepFake technology. It can fake either a person’s voice or their face by using a neural network and with a large collection of videos as training data. This could be a next level of further study after we’ve learned how to generate the text and image types of fake contents in this assignment.

9. Conclusions

As mentioned in all the section above, we had successfully implemented the download, extraction, generation, and discrimination tasks by using Tika, Docker, Grover, and DCGAN techniques. Some of the packages were easy to use as only a few lines of codes were needed, and some required us to consult the instructors for help and to develop our own understanding. We were especially aware of the impact of training data size for text and face generations, such that having a large size took more time but also tended to give better outcomes, so it is important to take all factors into consideration for our future practices of such tasks. Furthermore, the developing AI techniques are bringing many benefits, but it has also made it effortless to generate falsified media that can mislead the audiences, which connects back to the unintended consequences idea. Therefore, we should alway remind ourselves to make our own judgment.

Overall, this learning experience about media falsification for text and images, and image captioning/object identification was relatively impressive. Despite the time-consuming study of the packages, seeing the results was a satisfying moment for us.