INF 550 Homework 2: Large Scale Generation of Falsified Scientific Literature and Detection

Project Team 03
Abouelnaga, Raghda - abouelna@usc.edu - 2735798900
Cherry, Carlin - ccherry@usc.edu - 8211265507
Johnson, Andrew - andrewbj@usc.edu - 9950819736
Lee, Matthew - mdlee@usc.edu - 4356300240
May, David - davidmay@usc.edu - 5801939142

Introduction

In this assignment, we explored various content based tools for generating and detecting falsified media. Using tools like DCGAN and Grover, we were able to generate rough articles with the Bik dataset's biomedical papers as context. We also discussed the facets of the media falsification problem in general.

Generating Falsified Media

DCGAN - Images

We made several choices in developing our training sets which influenced the resulting falsified images. First, we resized all retrieved images to 128x128 to match the parameters of the model. We did this automatically using python, meaning we were indiscriminate about which portions of training images were trimmed / resized. Second, we filtered out all grayscale images to ensure the number of channels for all images matched. This resulted in a set of ~3,300 photos. We further broke this out into three buckets for image training: images of blots (i.e.m western blots), images of tissue samples, and images of fluorescently stained cells. We did this to 1. Speed up the model by using smaller training sets, and 2. To ensure models were trained with relatively similar datasets (ie, not mixing barg graphs with tissue samples).

The resulting fake blots were quite believable. Their color, gradient and clear lanes are difficult to distinguish from real western blots without very close analysis. From a distance, the falsified tissue samples look similar to real images, predominantly due to a consistent color schema and distribution (mixes of the purple and green stains generally seen on tissue samples). Up close, however, it is clear that the shapes contained within the images do not mirror those of true cell structures. It is not surprising that the model was unable to reproduce these patterns as they are not homologous across images in the training set. The least believable images were those coming from the fluorescent cell staining batch. While it was able to generate images with similar color patterns (neon reds, oranges and greens), the patterns were not believable. The falsified set contained a lot of perpendicular lines, likely caused by the image training set containing images with multiple data pieces, separated by white space. Better data could have been generated by manually cropping the training data to contain a single image.

Examples

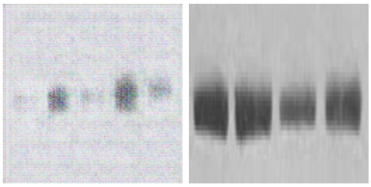

Figure 1: Fake (left) vs Real (right) Blot

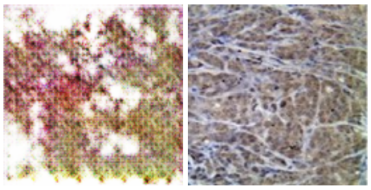

Figure 2: Fake (left) vs Real (right) Tissue

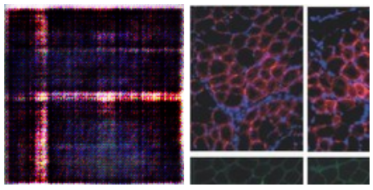

Figure 3: Fake (left) vs Real (right) Fluorescent Stain

Figure 1 depicts a Fake Blot (left) vs Real Blot (right): In the fake image, clear "lanes" can be seen and the color and gradient is quite similar to that of an actual western blot. Similar to real plots, some lanes are darker and more distinct than others.

Figure 2 compares a Fake Tissue (left) vs Real Tissue (right): While the mix of colors is fairly similar, close inspection reveals distinct tissue patterns in the image on the right while the generated image is more pixelated.

Figure 3 looks at a Fake Fluorescent Stain (left) vs Real Fluorescent Stain (right): Again, the generator was able to develop consistent colors, but the patterns clearly did not match those of the real images. Many contained perpendicular color patterns, likely stemming from the presences of training images containing multiple images.

Grover - Text

We were able to run Grover in generation mode, but we were ultimately unable to train a Grover model or run discrimination mode, which relies on training a model as well. Full detail on our attempts to get the training mode working are in the README. As a workaround, we used Grover's pre-built "large" model to generate text given 500 contextual articles. These include the 214 real papers from the Bik data, plus 286 "articles" formed by randomly sampling from the 214 (pick a paper at random, pick a few reasonably long sentences at random, then concatenate).

The results were decent, but there were several flaws that limit its believability to both human and machine eyes. When fed an entire article as context, Grover does not automatically format sections in a way that one would expect from a scientific paper–for instance, there is no references section. Some fake papers had blatant artifacts, where nonsensical strings repeated constantly (blue highlight in Figure 4). The generations also had the tendency to overuse capitalization, likely due to the proliferation of acronyms in the contextual input. This led the generations to write entire phrases in capitals (yellow highligh in Figure 4) or to include acronyms with little explanation or connection to the content. Surprisingly, when Grover did form complete sentences, they often seemed coherent. For example, consider these two snippets, one from an article and the other from its generated counterpart:

> Abstract This study investigates the subclinical or embryonic loss of bone cell massess as a consequence of enzymatic inhibitors of proteins found on many PI3KAKT targets. The conclustion that elevating the PI3KAKT pathway in bone cell masses due to overexpression of BMP9 in bone loss of osteosarcoma patients was a promising therapeutic target, albeit a subset study area...
>
> Abstract. Bone morphogenetic proteins (BMPs) are members of the TGF-B superfamily of signaling molecules and have previously been show to be associated with the biological behavior of osteosarcoma. However, to date the effects of molecular mechanisms of BMP-9 on osteosarcoma progression are unknown...

The former snippet is actually the generation; the latter is the original article. Grover only missed subtle

things–like the period after Abstract, the slash in PI3K/AKT, or the hyphen in TFG-8–and otherwise seemed quite believable. Despite these occasions of reasonable content, the lack of consistent style and logical flow throughout many of the generated articles ultimately make them not very convincing overall.

Without the training or discrimination modes, we could not utilize Grover for testing if a paper was fake. However, based on the Grover authors' paper, Grover would in theory excel at identifying its own generations as fake (even in what the authors coin a "zero-shot" setting, where it evaluates generations that it has never seen during training). Furthermore, we used the "large" model to generate rather than "mega"; Grover-mega is much stronger, so it should have no trouble identifying its weaker variant's fakes.

Perhaps fine-tuning Grover would improve the emulations slightly, as we were not able to train a new Grover model on scientific papers. It also might be more believable if we fed specific sections to Grover as context (giving abstracts to generate an abstract; giving references to generate references). We also only ran Grover with the default value of p (a tuning parameter that limits the variance in the text that Grover generates with); the paper mentions that there is a sweet spot for choosing p, and too high or too low would create frequent artifacts.

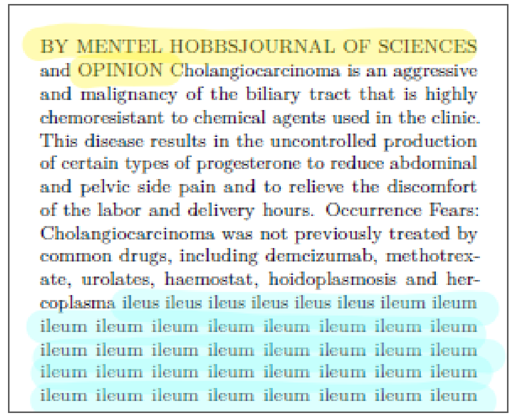
BY MENTEL HOBBSJOURNAL OF SCIENCES and OPINION Cholangiocarcinoma is an aggressive and malignancy of the biliary tract that is highly chemoresistant to chemical agents used in the clinic. This disease results in the uncontrolled production of certain types of progesterone to reduce abdominal and pelvic side pain and to relieve the discomfort of the labor and delivery hours. Occurrence Fears: Cholangiocarcinoma was not previously treated by common drugs, including demcizumab, methotrexate, urolates, haemostat, hoidoplasmosis and hercoplasma ileus ileus ileus ileus ileus ileus ileum

Figure 4: Generated text from Grover; yellow highlight showcases missuse of capitalization; blue highlight shows an artifact of nonsensical text

Thoughts on Falsified Media

Content Based vs Metadata Features

The previous assignment looked at metadata features to detect (or in theory, disguise) falsified media. This assignment dives into content based approaches to generate and detect falsified media.

Whether the metadata-feature-based methods from the first assignment could discern false media or not depends on the analyzer of the dataset. To a machine, our ancillary features from assignment 1 would likely not have been able to discern between real and false papers. We can manipulate fake features to imitate real ones in such a way that a machine could not discern between them by sampling from existing distributions of real numeric variables in our dataset to generate fake values in line with the real ones. We could also manipulate non-numeric variables like choosing from the same sample of universities, dates of publication, journal, topic of publication, and other features. Distance measure calculations then would disguise features such that none are suspicious in clustering. To a human, though, our associated ancillary features from assignment 1 might have been able to determine false papers. Humans could do this through deep industry knowledge (for example, a natural sciences professor might know offhand that a particular biology paper was not published in the PNOS in 2011 and therefore must be fake), through claim-verification techniques like a Google search, or through other cues.

Content based techniques have the advantage of learning underlying patterns that are not as easily exploitable as a distance metric. Through a human lens, tools like Grover and DCGAN are already at a point of producing fakes with alarming legitimacy when run in the contexts that they are trained on (e.g. news articles for Grover). However, if run without proper training or context, the generations may not rate very highly to humans; based on our generations from this assignment, the content has some semblance to the context, but the style and overall presentation are off the mark. On the machine side, media falsification/detection is not necessarily solvable, as it is only a matter of time until an improved model counters the fakes or fools the detector. The authors of the Grover paper liken this situation to an escalating "arms race," where verification and generation are constantly improving to beat the other.

All things considered, both content-based methods and metadata augmentation have their respective roles in

falsifying media. The content-based tools are superior for generating the content itself, but properly falsified metadata is important to disguise the content's claims.

Augmenting Generation

The generations from our tests are far from perfect; there is certainly much room for improvement and further experimentation.

One could draw on a database of authors or citations, such as Microsoft Academic, to improve text generations for fake abstracts and reference sections. Perhaps it would be more credible to impersonate real researchers, or at least train a Grover model to generate names from a real person's students or connections. To further obfuscate attempts to unravel the paper trail behind these fake papers, one could create fake LinkedIn profiles, research lab webpages, or degree certifications. These measures still may not fool industry experts or academics, but it might be enough to counter basic attempts by the masses to verify these papers.

A limitation from our observations was the lack of a large training corpus for biomedical text and images. One might scrape the articles and findings from several biomedical conferences en-masse to gather samples text/images to train the models on. In theory, with enough data, DCGAN could generate biomedical images as well as it does faces; Grover could generate scientific prose as well as it does news articles. One might draw from existing data repositories of medical images such as from FlowRepository (https://flowrepository.org) or the Whole Slide Imaging Repository (https://digitalpathologyassociation.org/whole-slide-imaging-repository) to further add training images. Building a large training set of images would also assist in object identification to generate believable captions.

To further augment a training corpus, one could stream in the data from actual biomedical experiments. Having concrete observations to draw from–whether these are simply sampled, or falsely generated themselves–can improve the authenticity of false images. Ideally, instead of generating images based on geometric patterns, a biomedical image generator might use real observations to simulate results. Perhaps such images would fare better in fooling expert eyes, as these images mimic biological structures rather than geometric ones.

Conclusion

Overall, the tasks of this assignment were interesting but challenging. Each required a good amount of background reading and setup to complete, and there were many loose ends that proved difficult. However, by experiencing a few of the many generation frameworks, we have taken our first steps in the "arms race" that the Grover authors liken media falsification to. Like participants in an arms race, we must continuously build our understanding of media falsification.