DS 6050: Deep Learning Final Project Report

Utilizing convolutional neural networks to detect fake satellite imagery

MacKenzye Leroy (zuf9mc), Edwin Purcell (jzd6af), Stephen Wheztel (sjw5ke)

May 3, 2022

Introduction & Background

Deepfakes, or the technology that generates realistic looking simulated imagery, has been a growing concern among ethicists and law enforcement since the technology first gained notoriety in 2017 (Brooks et al, 2021). Individuals, both private and public, face the threat of videos and images being doctored to show them in false contexts and perhaps performing actions harmful to their reputation (Hosanagar 2021). The ability to create realistic false videos and images with faces is largely accomplished through the use of convolutional neural networks (CNNs) acting as autoencoders to superimpose someone's facial structure onto another (Afchar et al, 2018). The use of CNNs has been expanded to other ethically dubious purposes, creating a new concern for intelligence communities around the world: the ability of adversaries to falsify intelligence and generate convincing imagery for the purposes of deception and propaganda (Tucker 2019).

One way this concern is manifesting itself is in the realm of fake satellite imagery. Novel imagery generation via artificial intelligence has been rapidly developing in recent years. Some of these impressive advances have come in less nefarious (though still ethically dubious) forms, such as NVIDIA's generation of hyper-realistic novel faces through their generative adversarial networks (GANs) (Tangerman 2018). GANs generate fake imagery by simultaneously training a generator and discriminator CNN. The generator network is trained to take in random noise and produce an image, which is judged to be authentic or false by the discriminator network. The signal generated by the discriminator's judgment allows both the generator and discriminator to optimize their parameters to better accomplish their respective tasks.

A well designed and properly trained GAN will eventually see performance converge to a point where the images created by the generator are good enough that the discriminator cannot tell whether they are real or fake and is unable to provide a nuanced enough signal to further improve performance. This same GAN technology is now being leveraged to create falsified satellite imagery (Eckart 2021). The generation of hyper-realistic false satellite imagery has the potential to mask geo-political intentions, create pretexts for war, or create contexts for other large intelligence miscalculations.

This paper demonstrates the ability for a sufficiently complex CNN to detect the authenticity of satellite imagery generated by a GAN even when the network was not a part of the GAN training that produced the images. A successful demonstration on available GAN generated images

nods to the potential for fake satellite imagery detection. Even with much more sophisticated test imagery this study points to the potential for CNNs to provide informative feedback pointing to which images are the most dubious and warrant special consideration in an intelligence setting.

Data Description

To explore how well we could differentiate between authentic and fake satellite images, we first needed a sufficient number of both types of these images to train a model on. We used a publicly accessible dataset created by Bo Zhao (2021). This dataset contained authentic satellite images from Beijing, Tacoma, and Seattle as well as a set of fake satellite images generated from those originals. In sum, there were 4032 authentic images (1008 from Beijing, 1008 from Seattle, 2016 from Tacoma) and 4032 fake images in the dataset (2016 modeled after the Beijing images, and 2016 modeled after the Seattle images.)

Data Filtering

In this data set, and in real intelligence settings, not all false imagery is going to be of the same quality. In this particular dataset there were hundreds of images that were obviously fake, even to the naked eye. We did not want to include these images in an honest assessment of our network's performance and did not want to waste model parameters by training them to detect these low-quality images. Luckily, many of these images were quite similar to one another. We developed an algorithm that cycled through each image in our dataset, flattened the tensor representation of the image, and compared the cosine distance between this vector and vector representations of several obviously false images that we had found. If the similarity, as measured by the cosine distance, met a preset threshold, the image in question was removed from the dataset. This resulted in hundreds of images being removed that would have otherwise tainted both our training and evaluation processes (Figure 1).

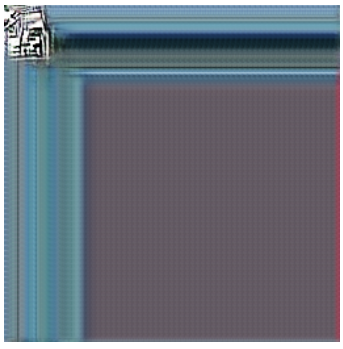

Figure 1: Example of a very obvious fake image that was filtered out. There were hundreds of nearly identical copies of this image.

In a more sophisticated environment, it would be possible to replace this simple filtering process with a CNN trained to detect images that are obviously false before passing images that did not

meet a certainty threshold to a more sophisticated GAN which would be trained on more realistic imagery. This would allow our high-level CNN to learn features of an obvious fake while the lower-level GAN would be free to learn more nuanced features that discriminate authentic from false images.

Data Augmentation

Before fitting our models, we augmented our dataset by running our training set through a pipeline that randomly flipped, rotated, zoomed, or changed the contrast of each image. We ran our original data through this data augmentation pipeline twice and concatenated the augmented sets to the original dataset tripling the size of our training dataset.

Methods

Model

Our CNN architecture consisted of 5 convolutional layers each using soft-plus activation, a 3x3 kernel, and each followed by a batch normalization and a dropout layer. Same padding was used in each convolutional layer. The specific dimensions of each layer and the number of filters at each level can be seen in Figure 2. Training was carried out using an Adam optimizer with a learning rate of 0.003 and categorical cross entropy was used as the loss function. A 20% validation set was used to ensure that the model did not overfit, and training was carried out for 30 epochs before convergence at the lowest validation loss.

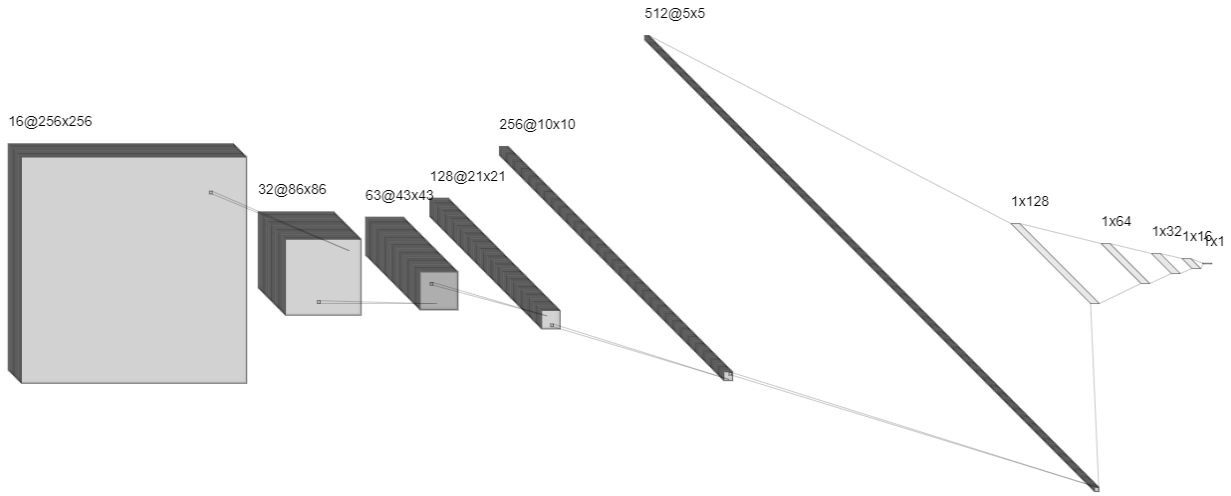

Figure 2: Diagram of the implemented convolutional neural network

The use of the prefiltered data ensured that the training was not tainted with obviously fake images that could possibly make the model suboptimal at detecting more nuanced features while also ensuring that our accuracy metrics gave us a relevant idea of model performance.

Visualizing The Model

After training was completed, we produced saliency maps on a sample of our images to try and interpret the results of the model. These saliency maps showed us which pixels had large responsibilities for the ultimate output of the model (Figures 3 and 4). Examining these maps can often times grant insight into how the model is carrying out its work of discrimination. It can be difficult to distinguish which exact features the model is identifying for each image and class, but it seems that roads in fake images tend to be highlighted along with a few other pixelated looking patterns. This could mean that for our particular model, more convincing roads could beat our classification.

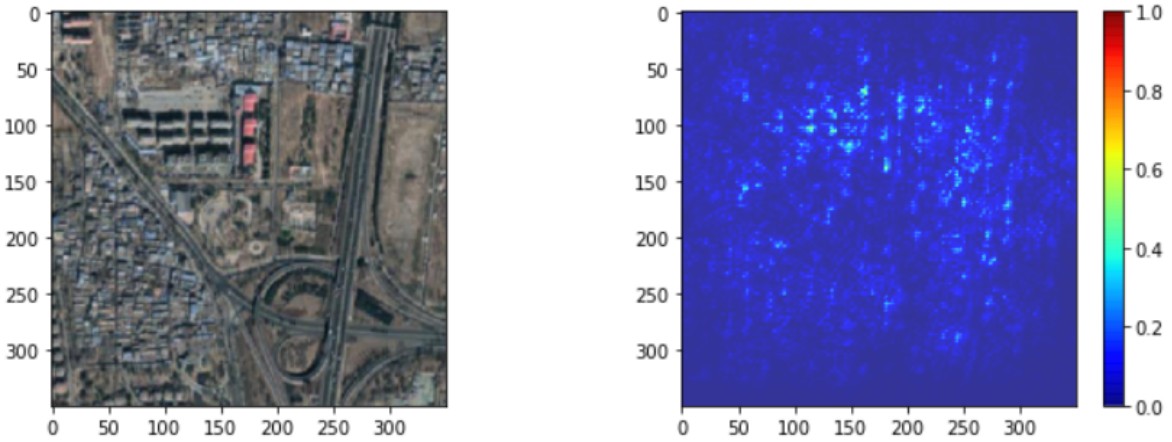

Figure 3: Real image saliency map

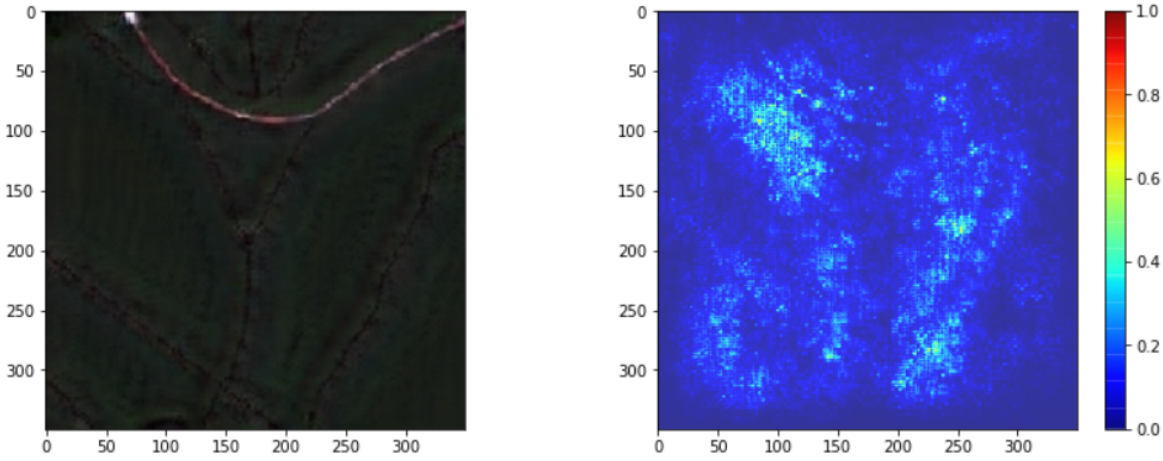

Figure 4: Fake image saliency map

The high-level convolutional filter weights were also examined to try and understand which types of features or shapes the model found to be informative when creating its outputs (Figure 5). These filters are not always significant to the user although we can see from the image we selected to look at, the model is identifying the building in the upper right hand corner. It is unclear from looking at the lower levels of the CNN what patterns are being identified.

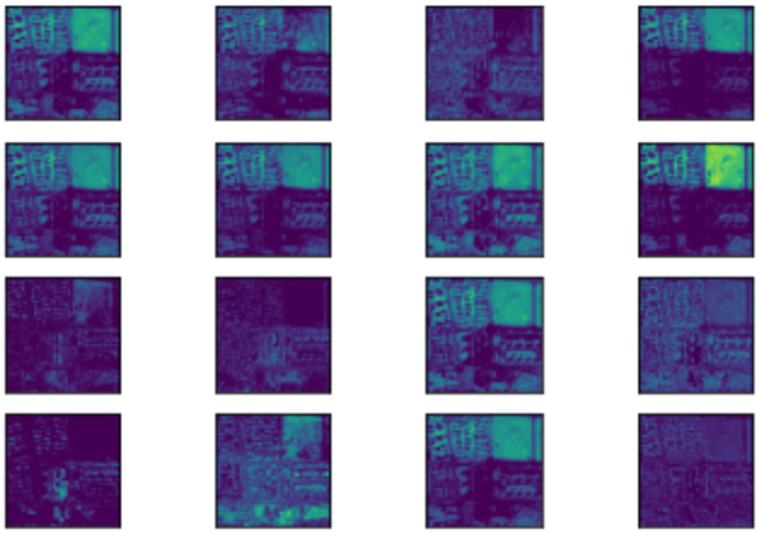

Figure 5: Visualization of the first CNN layer

Results

After training our model for 30 Epochs, we were able to achieve greater than 96% accuracy on our training set. More impressively, we were also able to achieve greater than 96% accuracy on our 20% held out validation set. Our final accuracy figures came in at 96.6% on our training set and 96.5% on our validation set, showing a strong potential for generalizability for our model. Further, our model was not only accurate because of a lucky threshold choice, but because it was truly differentiating between fake and authentic images boasting an area under the ROC curve of greater than .99 on both the training and test set and a very strong Precision Recall Curve (Figure 6). As noted in our sections above on the saliency maps we created, the model does in fact seem to be honing in on some tell-tale patterns that indicate whether an image is fake or not.

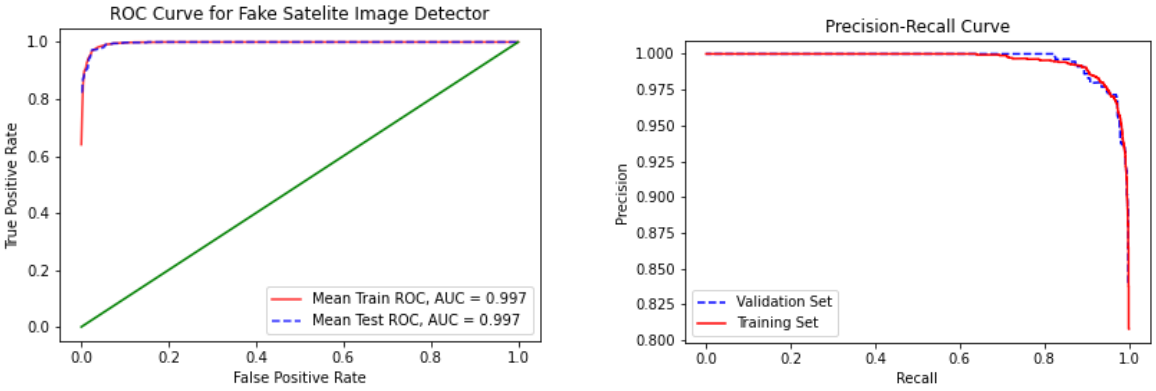

Figure 6: ROC Curve (Left) & Precision Recall Curve (Right) for Convolutional Neural Network Implementation

Discussion

While we were very happy with our final result, we are not certain how generalizable the model would be. On one hand, a .1 percent difference between our training and test accuracy and some promising points to a model with potential for generalizability, but having fake images from only one type of GAN may mean this model would not generalize to fakes created by other GANs. Our next goal would be to supplement our original training data with a range of high quality fakes possibly by creating a set of our own GANs. We attempted this but running a GAN is extremely computationally expensive and time consuming. We successfully created a generator and discriminator that takes as input random noise and attempts to create a satellite image. However, after running the training process for this GAN for dozens of hours, we still had not created any convincing imagery.

Conclusion

To conclude, being able to correctly identify GAN generated images from authentic satellite images is not only an interesting learning exercise, but also has important implications for national security. Using a publicly accessible dataset, we were able to successfully classify satellite images as real or fake even after filtering out images which were clearly fake with greater than 96% accuracy. We did so with a convolutional neural network featuring five layers of convolution and four hidden dense layers. We used visualization techniques like saliency mapping and convolutional layer visualization to better understand our model and what features might be important for identifying fake satellite images. While this research signified a good start, we believe that to ensure generalizability our model would benefit from more data and better deepfakes to learn from. To address this, we attempted to create a GAN that simulates realistic satellite images without success. Future work on this research would start with aggregating a much more diverse set of fake satellite imagery while also working to produce our own generated images via a GAN.

References

Hosanagar, K. (2021, December 7). Deepfake Technology Is Now a Threat to Everyone. What Do We Do? *The Wall Street Journal*. Retrieved May 3, 2022, from https://www.wsj.com/articles/deepfake-technology-is-now-a-threat-to-everyone-what-do-we-do-11638887121.

Eckart, K. (2021, April 21). *A growing problem of 'Deepfake geography': How ai falsifies satellite images*. UW News. Retrieved May 3, 2022, from https://www.washington.edu/news/2021/04/21/a-growing-problem-of-deepfake-geography-how-ai-falsifies-satellite-images/

Brooks, T., G., P., Heatley, J., J., J., Kim, S., M., S., Parks, S., Reardon, M., Rohrbacher, H., Sahin, B., S., S., S., J., T., O., & V., R. (2021). *Increasing threat of deepfake identities - dhs.gov*. Department of Homeland Security. Retrieved May 3, 2022, from https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf

Tucker, P. (2021, April 13). *The newest AI-enabled weapon: 'deep-faking' photos of the Earth*. Defense One. Retrieved May 3, 2022, from https://www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/

Tangermann, V. (2018, December 13). *Look at these incredibly realistic faces generated by a neural network*. Futurism. Retrieved May 3, 2022, from https://futurism.com/incredibly-realistic-faces-generated-neural-network

Zhao, B. (2021). Dataset containing over 8000 fake and authentic satellite images. https://figshare.com/articles/dataset/Fake_Satellite_Imagery/12197655.

Github Repository

Link: https://github.com/eddy3223/Deepfakes