**Title:** Detecting deception: using neural networks to identify deepfake satellite imagery

**Names and UVA IDs:**

* Gabriella (Gabby) Belmarez, gb6ux
* Frederick (Will) Blickle, fwb4cx
* Patrick Corbett, pjc8cq
* Nicholas Keeley, Ngk3pf

**Introduction and Motivation**

Deepfake imagery – artificially generated images that appear real to human viewers – can be used for a variety of purposes, to include adversarial purposes such as propaganda or misinformation campaigns. Forensic tools in use today lack robustness and scalability,[[1]](#endnote-2) especially when confronted with the growing ability of neural networks to quickly generate convincing imagery. A potential area of concern to the national security community involves deepfaking satellite imagery, which has been theorized as an impending possibility. This possibility is no longer theoretical: Zhao, et al., have recently demonstrated the ability of Generative Adversarial Networks (GANs) to produce such deep fake imagery.[[2]](#endnote-3) Tools that help analysts and consumers to identify deepfaked images at scale must be developed and maintained in order to prevent the infusion of misinformation into otherwise reliable data sources.

We hypothesize that deepfake satellite images will possess certain telltale features detectable by a neural network, and that we can use existing neural network architectures to identify deepfake satellite images. We train multiple neural network models on deepfake satellite image detection (some using transfer learning), compare accuracy, and explore the feature activation occurring in each model. Our results show that these models can achieve a high degree of accuracy, in excess of 90% across all models tested, and approaching 100% in our top-performing model.

**Hypothesis**

Deepfake images are improving in quality but there are still some general telltale signs that may help distinguish between real and deepfaked images. We hypothesize that blurriness and a lack of detail in parts of an image may be significant in distinguishing real from fake images, as previously studied deepfake images of people tend to contain areas of blurriness and possess less-than-expected pixel variation.

Deepfake detection is typically viewed as a classification problem. As such, pretrained neural networks designed to classify images could potentially be adapted to recognize deepfake satellite images. These kinds of neural networks possess layers that have typically learned to recognize certain color or shading patterns and geometric primitives – skills which may be helpful in identifying the satellite image features such as blurriness and lack of detail that may be indicative of deepfakes.

However, pretrained neural networks like InceptionV3 are often highly complex and difficult to explain, making training on new geographical datasets difficult and time consuming. We propose use of the models developed by Afchar et al. to reduce training time and model complexity while still maintaining high accuracy.

**Literature Review**

The proliferation of deepfake images across the internet has created an alarming societal problem of separating truth from misinformation.[[3]](#endnote-4) While much of the literature on “deep fakes” has revolved around propagated images of predominant political figures or celebrities[[4]](#endnote-5), [[5]](#endnote-6), a new threat has emerged of potentially fraudulent satellite images. Bo Zhao, et. al., have recently demonstrated the ability of GANs to produce deepfake satellite images.ii While manipulations of maps is not a new phenomenon, companies and governments will need better tools to differentiate deepfakes from real satellite images in a new AI arms race.[[6]](#endnote-7) Zhao and others have been able to detect GAN-generated images with varying levels of success.ii,[[7]](#endnote-8) However, as deepfake detection becomes more advanced, so does the generation. Many existing image classification networks are available and can be leveraged to help solve this problem.[[8]](#endnote-9), [[9]](#endnote-10), [[10]](#endnote-11)

Hsu, et al. have demonstrated a deepfake detection method by passing images through DenseNet to create pairwise information that serve as training inputs into a common fake feature network with a concatenated classification layer. [[11]](#endnote-12) Afchar, et al., have developed a network they dubbed Meso-4 that detects deepfaked faces by examining the mesoscopic level (between microscopic and semantic levels) of input images using four convolutional layers.[[12]](#endnote-13) They have also developed MesoInception-4, which replaces two of the convolutional layers with a module from an Inception model. They propose that these networks work, at least in part, by identifying areas of images lacking in detail relative to the rest of the image: as deepfaked faces tend to be blurry, it makes sense that these networks would hone in on areas likely to have been faked.

**Methods and Data Sources**

For our training and test dataset we used the open source satellite images (real and fake) provided by Bo Zhao via figshare at: <https://figshare.com/articles/dataset/Fake_Satellite_Imagery/12197655/2?file=22424628.> [[13]](#footnote-2)

This dataset contains real and faked satellite images. The real satellite images are of Beijing, Tacoma, and Seattle, and are coupled with GAN-generated fake images that appear similar to these regions. As such, these images are focused primarily on metropolitan areas. The dataset contains roughly 4,000 real and 4,000 fake satellite images.

We explore the use of four neural network models to identify deep faked satellite imagery: InceptionV3, Meso-1, Meso-4, and MesoInception-4. As this is ultimately an image classification challenge, we felt applying transfer learning to the ImageNet-trained InceptionV3 on Zhao’s deepfaked satellite imagery may prove fruitful. Similarly, given that the Meso-series networks have been trained to detect deepfake videos of people, it seemed reasonable that we could use them to detect deepfaked satellite images.

We first implemented the Meso1, Meso4, and MesoInception4 architectures proposed by Afchar, et al. to detect deepfaked human faces – only this time, trained and assessed on satellite images. We then compared those results to that of a simple InceptionV3 model, a recent iteration of GoogLeNet, to evaluate the extent to which transfer learning could distinguish between deepfake and real satellite images. We utilized transfer learning to identify deepfaked satellite images via the InceptionV3 network which was pretrained on the ImageNet database.

We then searched for the ‘fingerprints’ of the generative (deepfake) process within the resultant images by exploring the activations within the neural networks themselves. We hypothesized that at least some filters within the layers of the network would prove useful in labeling an image as real or fake, and we attempted to identify those filters to produce findings similar to Afchar (i.e., what features, specifically, seem to drive the classification).

**Analysis and Interpretation**

*Model Accuracy*

Results for deepfake detection via our pretrained models (below) suggest our hypothesis was correct: transfer learning from pretrained models can be used to reliably detect deepfake satellite images.

Figure 1. Model Accuracy

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Train Accuracy | Validation Accuracy | Test Accuracy |
| InceptionV3 | 99.37% | 100% | 99.4% |
| Meso1 | 91.11% | 91.10% | 92.00% |
| Meso4 | 90.97% | 95.12% | 94.87% |
| MesoInception4 | 92.56% | 91.07% | 91.65% |

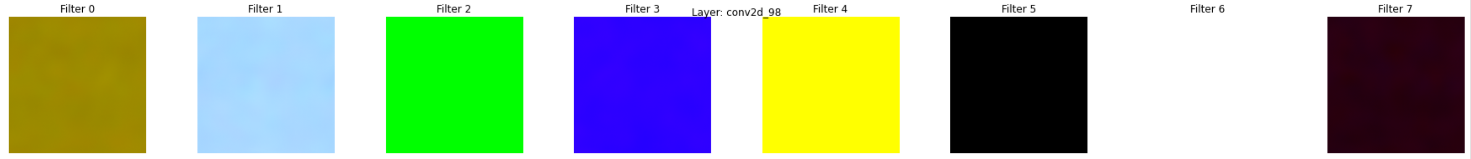
The best model performance clearly comes from the pre-trained InceptionV3 model which we transferred to our classification task. It is not surprising that the model pretrained on ImageNet shows more optimal performance than the Meso1/Meso4/MesoInception4 models, which are trained from scratch using only the few thousand satellite images available in our dataset. Even with this disadvantage, however, our other models are performing surprisingly well. Meso4, in particular, exhibits very high accuracy even with this limited training set, supporting the benefits of this model.

*Activations*

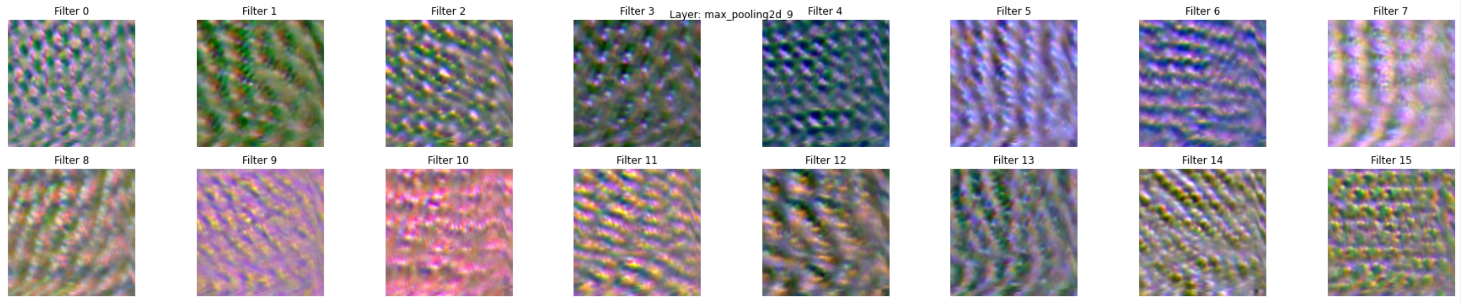
In addition to accuracy, we wished to analyze the activations behind our models to get some sense of what kinds of information lead to a satellite image being identified as real. To do this, we produced activation maximizations for the filters in all layers of each of our models. This allows us to examine differences across models, as well as understand what features emerge in deeper layers compared to shallower layers.

Below are the activation maximizations from each model, with a shallow layer and deep layer presented for each.

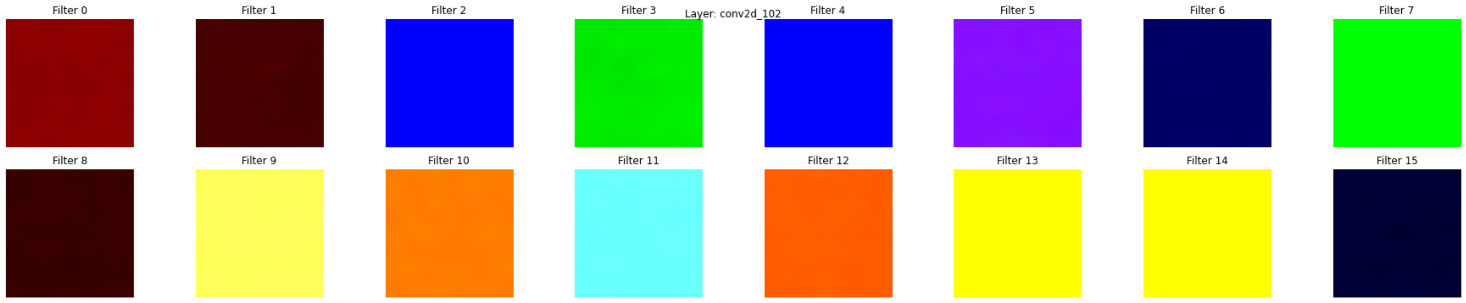
InceptionV3- Shallow



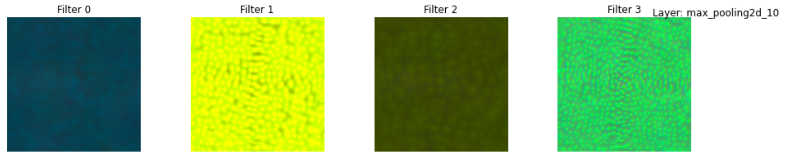
InceptionV3- Deep



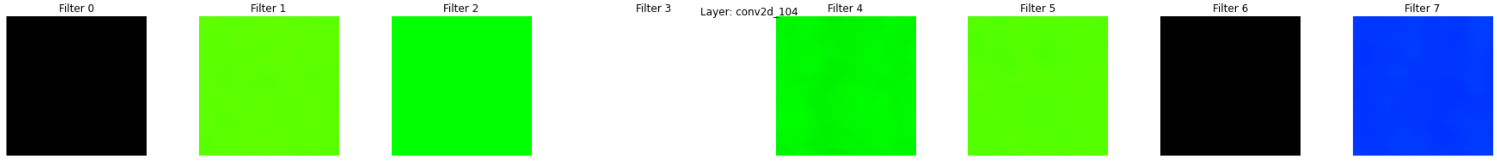
Meso1- Shallow



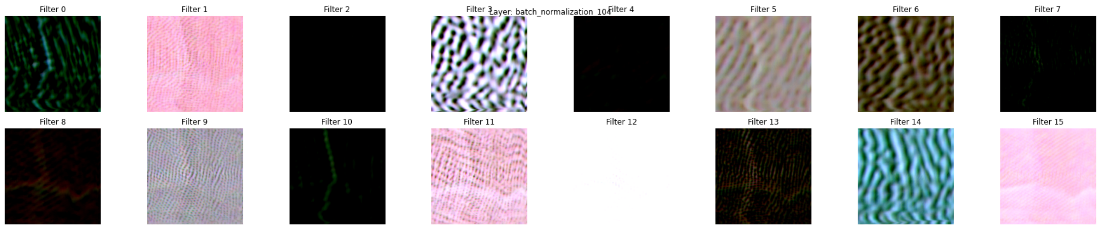
Meso1- Deep



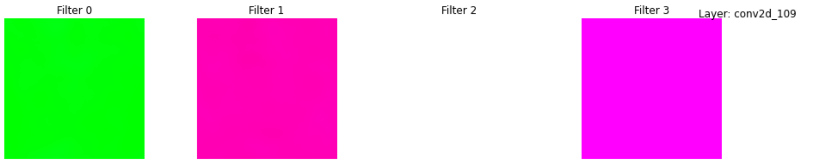
Meso4- Shallow



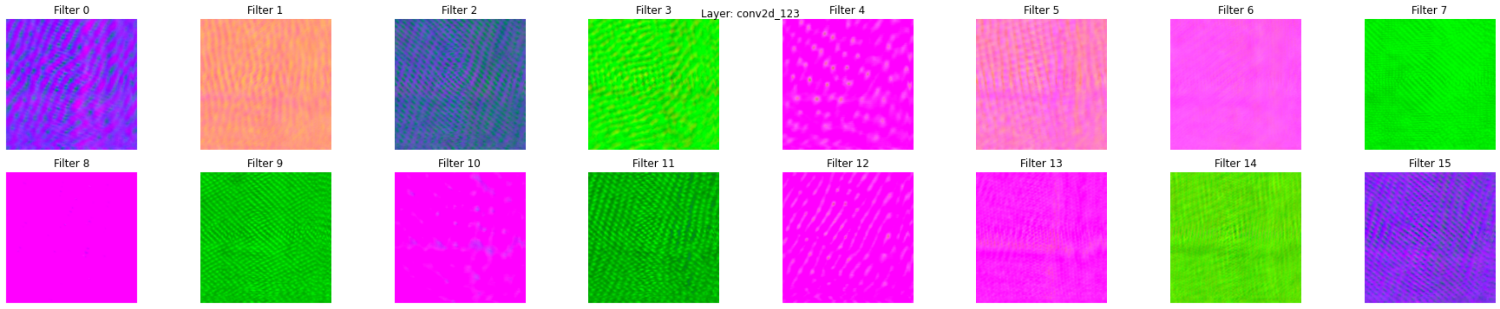
Meso4- Deep



MesoInception4- Shallow



MesoInception4- Deep



The above represent the extreme ends of the filters from our models, but we can infer some conclusions about the features that contribute to the identification of real satellite images and how they relate to our hypotheses. Most notably, the deeper layers do show indications of the types of geometric primitives we discuss above, particularly contour lines and, in some cases, what appear to resemble roads. The first few layers in each model are just activated by different colors, which is unsurprising for shallow layers but indicates that some amount of depth is needed to distinguish the types of features that are truly helpful in separating real images from deepfakes. These observations do seem to support our hypothesis that more detailed geometric features are necessary in distinguishing the fake from real images, and so additional depth in models is beneficial to this objective.

The differences between our models provide some evidence in favor of this hypothesis as well. Our best performing model – the InceptionV3 – as well as our top Meso model, Meso-4, show some more distinct geographic features like contours and road-like shapes compared to the detailed features from the other Meso models. The ability of these models to activate on these features may help to explain the modest improvement in performance of this model relative to our other candidates. A possible explanation for Meso4 outperforming MesoInception4 may be MesoInception-4's comparative use of “skip connections,”[[14]](#endnote-14) potentially causing deeper layers of the model to be interfered with (or impacted by) the shallower modules earlier in the architecture.

**Discussion and Conclusions**

Through our analysis, we discovered that transfer learning could be used to distinguish deepfakes from real satellite images with relative accuracy. This leads us to believe that despite the rapid development of GANs, existing tools are still available to help identify deepfake geographical images.

Beyond this, our results show that the models built from scratch without transfer learning are quite close in performance to a pre-trained model. Given that we achieved these results with a relatively limited number of training examples, we see this as a promising indication for the potential future performance of these models for this classification task if they were to be deployed and maintained with new data and retraining.

Further, with activation maximization, we are able to get some sense of the fingerprints that distinguish real satellite images from deepfake images. Detailed geographic features emerge here, suggesting that the ability of these models to distinguish real images may come from the greater level of terrain or topographic detail that the real images show compared to the deepfakes in our set.

An important implication of this observation is that as new deepfakes continue to improve in quality and level of detail, these kinds of features may not always be able to distinguish real from fake. Indeed, our models were tested exclusively on fakes generated by one kind of GAN. However, existing studies on deepfake imagery suggest an emerging trend to find “common” features between different types of GANs, to account for the introduction of new GANs in the future.[[15]](#endnote-15) This is a very real ongoing concern that needs to be considered in the use of these models. That said, our results do suggest that the deep learning architectures tested here are an effective tool for this task, and future work in this space could include expanding these models even further to try and learn more detailed geographic features that newly generated deepfakes have yet to identify. This demonstrates why examining the model output with tools such as activation maximization is an important step beyond model accuracy: maintaining knowledge of what information deepfakes have not yet picked up on – and whether this information change as more deepfakes are generated – can provide further insight into where vulnerabilities are in separating the two.

As with the deepfake problem in general, there is an ongoing back-and-forth between generative networks improving the quality of fakes and the development of models that can continue to separate real from fake. The models that work well today may not work well tomorrow, but the work presented here and, in the literature, does provide a good level of assurance that deep learning can be deployed as a tool in guarding against misinformation and adversarial attacks.

1. Media Forensics (MediFor) Program, Defense Advanced Research Projects Agency: <https://www.darpa.mil/program/media-forensics> [↑](#endnote-ref-2)
2. Bo Zhao, Shaozeng Zhang, Chunxue Xu, Yifan Sun & Chengbin Deng (2021) Deep fake geography? When geospatial data encounter Artificial Intelligence, Cartography and Geographic Information Science, 48:4, 338-352, DOI: 10.1080/15230406.2021.1910075 [↑](#endnote-ref-3)
3. Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107(6), 1753–1819. <https://doi.org/10.2139/ssrn.3213954> [↑](#endnote-ref-4)
4. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. In K. Lee, D. Forsyth, M. Pollefeys, & X. Tang (Eds.), 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 1–11). [↑](#endnote-ref-5)
5. Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. In 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW) (pp. 83–92). IEEE. [↑](#endnote-ref-6)
6. Tucker, P. (2019). The newest AI-enabled weapon: ‘deep-faking’ photos of the earth. Defense One. <https://www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/> [↑](#endnote-ref-7)
7. Wang, S.-Y., Wang, O., Zhang, R., Owens, A., & Efros, A. A. (2020). CNN-generated images are surprisingly easy to spot … for now. In T. Boult, G. Medioni, & R. Zabih (Eds.), 2020 IEEE/CVF conference on computer vision and pattern recognition (pp. 8692–8701). IEEE. <https://doi.org/10.1109/cvpr42600.2020.00872> [↑](#endnote-ref-8)
8. A. Krizhevsky, “One weird trick for parallelizing convolutional neural networks”. arXiv, 2014. [↑](#endnote-ref-9)
9. Szegedy, Christian, et al. “Going Deeper with Convolutions.” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015. Crossref, <https://doi.org/10.1109/cvpr.2015.7298594>. [↑](#endnote-ref-10)
10. He, Kaiming, et al. “Deep Residual Learning for Image Recognition.” CoRR, vol. abs/1512.03385, 2015, <http://arxiv.org/abs/1512.03385>. [↑](#endnote-ref-11)
11. Hsu, Chih-Chung, Yi-Xiu Zhuang, and Chia-Yen Lee. 2020. "Deep Fake Image Detection Based on Pairwise Learning" Applied Sciences 10, no. 1: 370. <https://doi.org/10.3390/app10010370> [↑](#endnote-ref-12)
12. D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630761. [↑](#endnote-ref-13)
13. Thanks to Dr. Stephen Baek for pointing us towards this dataset. [↑](#footnote-ref-2)
14. He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian. <https://arxiv.org/abs/1512.03385> [↑](#endnote-ref-14)
15. Zhao, 2021. [↑](#endnote-ref-15)