General Detection of Image Manipulation

Jacob Hendricks, Chirantana Krishnappa, Santhosh Manohar Gouda Patil, Bharath Varma Kantheti.

Abstract

In today's world, image editing software is widely available, leading to an increase in image manipulation. It's vital to detect and prevent the spread of altered images, especially in journalism, forensics, and social media. We present our results obtained from experimenting with various deep neural network-based methods for identifying manipulated images by training the models on a dataset of original and altered images. We present analysis of six different modern deep learning architectures on our dataset. We also propose a new architecture that combines ELA with ResNeXt to get results surpassing the basic deep learning models. Our results suggest that general detection of image manipulations is possible using deep learning techniques.

1. Introduction

The art of image manipulation has been around for decades, but it was the advent of Adobe Photoshop that popularized it and made it more accessible to the masses. Today, almost every person has access to photo editing software and is capable of modifying original photos. While many of these modifications may be harmless, such as color correction or cropping, some may be more dangerous. This has led to the widespread dissemination of modified images that can misrepresent events, individuals, or organizations, leading to severe consequences. Furthermore, photo editing techniques have become more sophisticated. This makes it harder to tell which photos have been modified. As a result, detecting these false images can require considerable expertise in a given field, making human detection unreliable. Therefore, there is a need for a more efficient, automated, and user-friendly solution to detect image manipulation in real-time.

In this paper, we aim to utilize deep learning techniques to automatically classify images as photoshopped or not photoshopped. Convolutional Neural Networks and Transformers can learn specific features of images that can be used to indicate whether a photo is manipulated. Our idea is that, by providing a probability of manipulation, users can be more cautious of the image content and take appropriate action if necessary.

Our contribution to the field is two-fold:

1. First, we experiment with six popular
2. deep-learning models to get an initial baseline for image classification on the PS-battles dataset [2].
3. Second, we introduce a new architecture which combines the ResNeXt architecture with ELA to produce better results for manipulation detection.

2. Background and Related Work

There has been some previous experiments with automatic image manipulation detection. However, most current methods focus on identifying the specific areas where an image has been edited [1][3][4], rather than image classification. While this approach has some theoretical advantages, it suffers from low accuracy rates in practice.

There has also been some work in detecting whether images of human faces are manipulated [5]. This approach has had reasonable success by using deep learning methods. However, this success is limited to the specialized task of only detecting faces.

By looking at these current methods, we aim to create a more generalized model that can work on a wide range of images and provide the user with information on whether the image is likely to be photoshopped.

3. Dataset

For our experiments, we used the PS-Battles dataset [2]. We used images from r/photshopbattles, a popular community on Reddit where original images are posted and interested users manipulate these images using photoshop as replies. Most of the manipulations are designed to be funny, so a lot of the images are easy for humans to classify. However, there are many images in the dataset that are much more difficult to detect. We chose to use this dataset because we believe that the mix of obvious and subtle manipulations are representative of common image manipulations.

The full dataset is composed of over 100,000 images. It contains 11,142 unmodified images and 90,886 photoshopped images. However, due to computational and time constraints, we were unable to use the full set for our experiments. Instead, we took 5,000 original images and 5,000 photoshopped images to use for the experiments.

4. Methods

To begin, we conducted an analysis of six distinct deep learning models to establish a benchmark for the manipulation detection task. The models we utilized were GoogLeNet [9], ResNet [6], InceptionV3 [8], DenseNet [10], VisionTransformer [7], and ResNeXt [11]. These models were selected as a baseline because they represent state-of-the-art deep learning techniques. Although we experimented with several other popular architectures, we found that these six models yielded the best results.

After the baseline models were established, we set out to build a new architecture that can achieve better results. TO do this, we used Error Level Analysis (ELA), which is a technique used in forensic investigations to identify possible areas of digital image manipulation or tampering [12]. The process involves saving an image at a low quality or compression level, creating various error levels in the image. ELA then compares the error levels of different parts of the image to determine if any areas have significantly

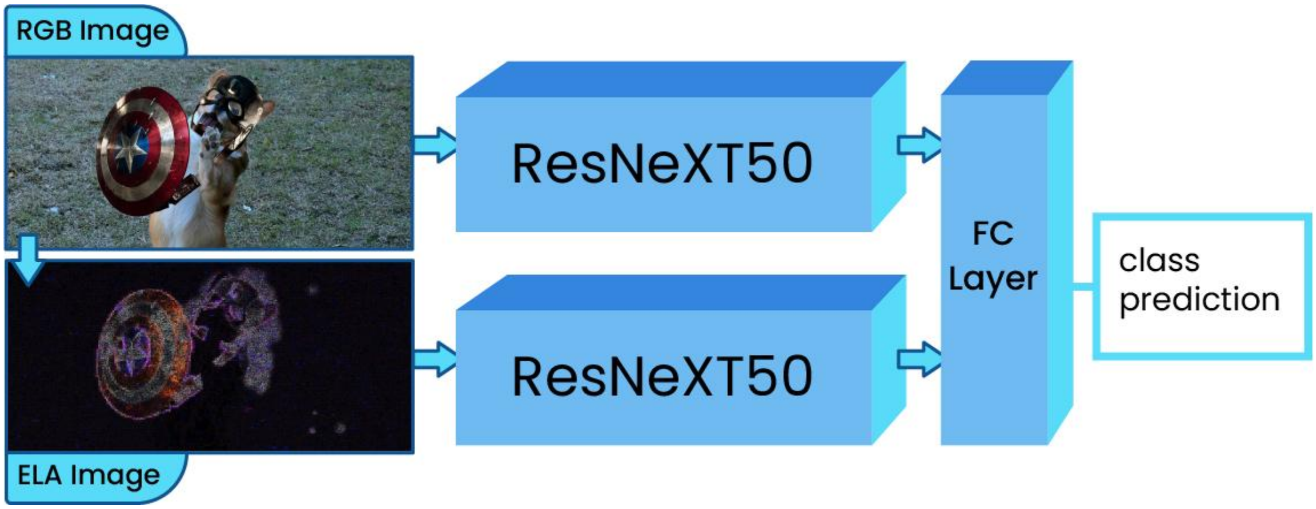


Figure 1: ELA + ResNeXt architecture. The top image is the normal RGB image. The bottom image is the ELA version of the same images. Each of these are fed through separate ResNeXt models and combined at the end to get the final prediction.

different error levels than others, which could indicate potential manipulation or tampering. However, ELA is not completely reliable since several factors, such as different image compression algorithms, can impact error levels in an image. Nonetheless, ELA can be a valuable tool when combined with other deep learning techniques to detect potential image manipulation.

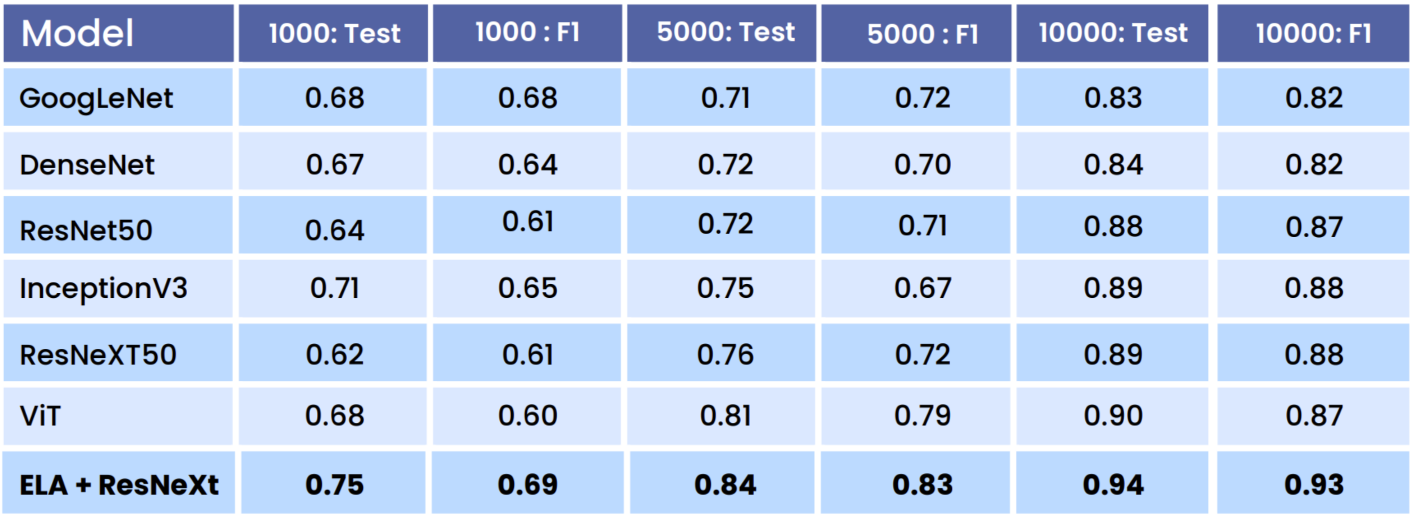
Initially, raw RGB images were trained using state-of-the-art deep learning models, and the Vision Transformer and ResNeXt models achieved the highest accuracies. However, when the vision transformer model was combined with error level analysis, the accuracy did not improve. Therefore, we chose the ResNeXt model to expand on. To preprocess the data, error level analysis (ELA) of the images was generated by compressing them to 90% quality. The raw RGB images and their corresponding ELA images were used as inputs to two separate ResNeXt50 networks, and the features from the last layer of the ResNeXt50 models were combined and forwarded to a linear layer to predict manipulated images. An illustration of this model is shown in FIgure 1.

5. Experiments

To make up for the smaller set of data as mentioned above, we ran tests using 10%, 50%, and 100% of the smaller data in order to see training patterns. That way we can make inferences on how the models would perform using the full 10,208 images from the original data.

For each of the six deep learning architectures that were chosen, we used the pretrained models built into TorchVision. The weights we used were trained from the ImageNet dataset.

All of our testing was done on Google Colab using PyTorch. 80% of the data was used for training and 20% was used for testing. We trained the model for 20 epochs and tested the model on our testing set. The results for these experiments are illustrated in the following section.



Model	1000: Test	1000 : F1	5000: Test	5000 : F1	10000: Test	10000: F1
GoogLeNet	0.68	0.68	0.71	0.72	0.83	0.82
DenseNet	0.67	0.64	0.72	0.70	0.84	0.82
ResNet50	0.64	0.61	0.72	0.71	0.88	0.87
InceptionV3	0.71	0.65	0.75	0.67	0.89	0.88
ResNeXT50	0.62	0.61	0.76	0.72	0.89	0.88
ViT	0.68	0.60	0.81	0.79	0.90	0.87
ELA + ResNeXt	**0.75**	**0.69**	**0.84**	**0.83**	**0.94**	**0.93**

Table 1: Results from each of the models that we used. The leftmost column is the model used to get the results. The next two columns are the results using 10% of our collected data. The two columns after that are the results from using 50% of the data. The final two columns are the results from using 100% of our data.

6. Results

After conducting experiments on the dataset using various models, we have identified the best performing model, and their training accuracy, testing accuracy and F1 scores in correspondence with the size of the dataset. We used 1000 images as our base size and 10000 images as the maximum dataset.

As seen on the table above, the accuracies increase as the size of the dataset increases no matter the model used. We decided to include the models that were showing promising results (above 80% accuracy) for 10,000 images. Empirical results show that ELA + ResNeXt to be our best performing model that has an accuracy of 94% and an F1-score of 93%. This does perform much better than even the best performing models such as ViT or normal ResNeXt. However, every one of the models tested achieved results of at least 83%. Table 1 shows the rest of these results.

It is also important to note that these are the results without using all of the available data from PS-battles. As mentioned before, we did not have the time or computational power to work with the entire dataset. However, we can look at the pattern of how the model improves by using increasing amounts of data to speculate on what the results would be when trained on the entire dataset. These smaller subsets of the data indicate that the model would theoretically be able to work even better on more data since each of them do improve over time.

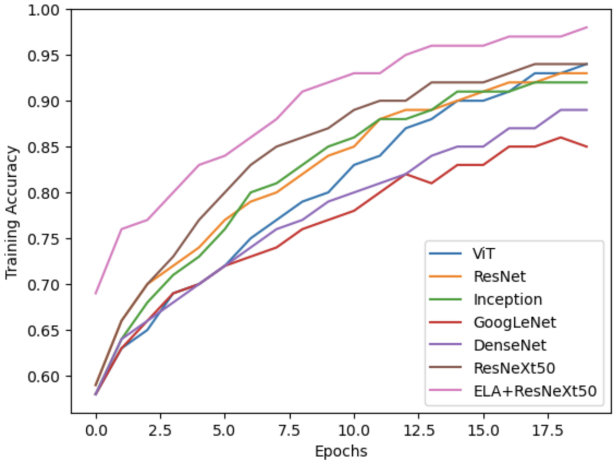


Figure 2. Plot of Training accuracy vs Epoch

We also plotted the training accuracies of each model over the course of multiple epochs in Figure 2. This enabled us to observe and evaluate their learning progress. This approach enabled us to determine how well the

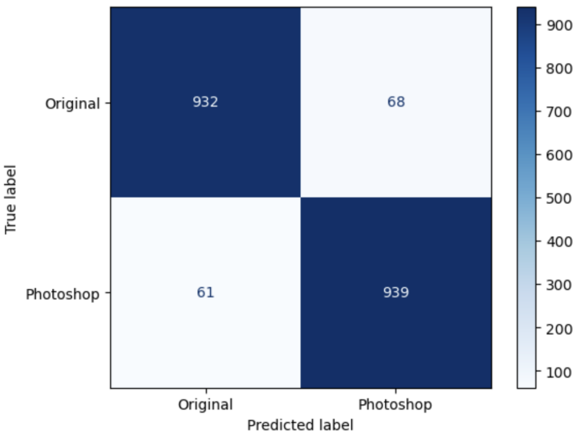


Figure 3. Confusion Matrix

models were adapting to the training data and to assess their potential for generalization.

Through this comparative analysis, we identified that the ELA+ResNeXT50 model consistently outperformed the other models in terms of training accuracy. The noticeable gap between the ELA+ResNeXT50 model and the other models on the plot highlights its superior ability to learn from the training data and suggests a higher likelihood of achieving better performance on unseen data.

In addition, the graphical representation allowed us to spot any stagnation points during the training process. These points were crucial for understanding when the models' learning progress slows down or comes to a halt. Identifying such instances helped in fine-tuning the training process, and selecting the most appropriate model for the given task.

We used a confusion matrix (Figure 3) for evaluating a classification model's performance. This allowed us to present a summary of its true and false predictions. In the context of our ResNeXT50 model combined with ELA, the confusion matrix reveals that the model achieves a balanced distribution of false positives and false negatives. This balanced prediction performance is indicative of the model's ability to maintain a consistent error rate across both positive and negative classes. In other words, it does not display any significant bias towards predicting one class over the other. This is an important aspect of model evaluation, as it ensures that the model provides a fair and reliable representation of the data and is less likely to produce skewed predictions.

7. Conclusion

Overall, our trials with several deep neural network-based approaches for detecting altered photographs produced encouraging results. Our research sought to detect and limit the dissemination of manipulated photos, particularly in fields such as journalism, forensics, and social media, where image manipulation may be highly detrimental. Our models were trained on a dataset of original and changed photos, and their performance was assessed using a variety of measures.

The ELA+ResNXet model, which attained the maximum accuracy of 91%, was our most successful model. This approach uses Error Level Analysis (ELA) as a preprocessing phase to identify parts of a picture that may have been manipulated, followed by a deep neural network based on ResNeXt to categorize the images. The ELA stage aids in identifying regions that may have been altered, which may then be used by the ResNeXt model to appropriately categorize the picture.

Our research highlights the utility of deep neural network-based approaches for detecting picture tampering and lays the groundwork for future research in this field. These models can help to confirm the authenticity of photographs and avoid the spread of misleading or false information. Overall, our findings imply that universal detection of picture modifications is possible, and we expect that our work will help to build more effective approaches for recognizing modified photos.

All code can be found at https://github.iu.edu/jbhendri/CV-Project

Acknowledgement

We would like to thank Professor Crandall for his work in teaching this course - CSCI B-657. We also thank Vibhas Vats for his work teaching the deep learning models of this class. Without their contribution, we would not have been able to achieve such a remarkable learning experience.

References

[1] Bappy, Jawadul H., et al. "Hybrid lstm and encoder–decoder architecture for detection of image forgeries." IEEE Transactions on Image Processing 28.7 (2019):3286-3300.

[2] Heller, Silvan, Luca Rossetto, and Heiko Schuldt. "The ps-battles dataset-an image collection for image manipulation detection." arXiv preprint arXiv:1804.04866 (2018).

[3] Liu, Xiaohong, et al. "PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization." IEEE Transactions on Circuits and Systems for Video Technology 32.11 (2022): 7505-7517.

[4] Zhou, Peng, et al. "Learning rich features for image manipulation detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[5] Neves, Joao C., et al. "Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection." IEEE Journal of Selected Topics in Signal Processing 14.5 (2020): 1038-1048.

[6] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[7] Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv preprint arXiv:2010.11929 (2020).

[8] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[9] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

[10] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[11]Xie, Saining, et al. "Aggregated residual transformations for deep neural networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[12] Abd Warif, Nor Bakiah, et al. "An evaluation of Error Level Analysis in image forensics." 2015 5th IEEE international conference on system engineering and technology (ICSET). IEEE, 2015.