Image-tampering Classification for Fake News Detection
(Project Final Report)

Huaxiuyue Wang
Department of Civil Engineering
Stanford University
wangh28@stanford.edu

Yingdan Lu
Department of Communication
Stanford University
yingdan@stanford.edu

Abstract

Growing research body has shown that image-tampering has become a salient strategy for fake news. To address this problem, we researched on ResNet-50 and VGG-16 models, implemented with the pretrained weights for ImageNet, and retrained on the CASIA dataset. We present evaluations of different models and finally predicted the proportion of manipulated images among the sampled fake news images.

1 Introduction

Fake news have drawn many public scrutiny and academic attentions, especially after the 2016 Presidential Election(1)(2). While current studies focus more on the text-based features of fake news, less attention is cast on another important type of fake news – image manipulation. In fact, image tampering, facilitated by powerful software and visual techniques(3), has become more pervasive in online content-creation, and may create a misleading narrative for news reports(4).

Given that image tampering in fake news has not yet been explored systematically, this project aims to research on Convolutional Neural Network (CNN) algorithms, implement and compare several established CNN models, and apply transfer learning to achieve a model designed to accurately classify tampering images and utilize in fake news detection.

Figure 1: Example of the manipulated image in fake news.

2 Datasets

Two dataset are used in this project. The main dataset used in training, development and testing is CASIA ITDE V2.0 dataset, which is popularly used for tampering image detection. This dataset contains 12,324 color images divided into two subsets: the authentic subset contains 7,491 authentic images, and the tampered subset contains 5,123 fake images. The tampered images are generated with random image-cropping, splicing, post-processing and realistic operations to avoid over-tampering.

The second dataset is a relatively small dataset from Kaggle Getting Real about Fake News, only used at test time. This dataset contains images appeared in online fake or biased news (with unknown label). The purpose of adding this small test set is to estimate the proportion of images used in fake news that are indeed tampered images.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

3 Methods

3.1 Data preprocessing

For the **CASIA v2.0** dataset, we first converted all tif images into jpg format and then split the dataset to 90: 5: 5 training, validation and test set. Images from authentic and tampered subsets are divided into training/development/test sets with the same ratio. Also, since the CASIA V2.0 images are with difference sizes, ranging from 240 $\times$ 160 to 900 $\times$ 600 pixels, we re-sized them to have the same size: 150 $\times$ 150 for the baseline model and 224 $\times$ 224 for the further training.

For the **Kaggle** dataset, we downloaded the images and cleaned the pictures with logos, overly tampering and pure paintings. Then, we sampled 50 pictures for the prediction and did human-tagging for comparison. We also resized then to 224 $\times$ 224.

3.2 Models and workflow

We started from a shallow neural network with three layers of CNNs as our baseline model. The architecture of the baseline model is shown in Figure 2.

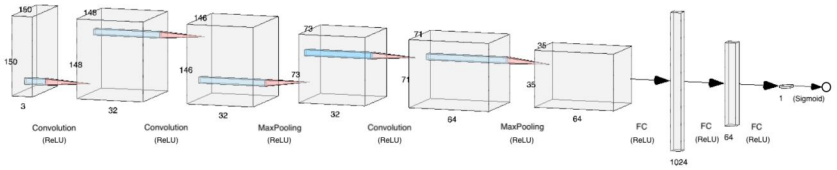
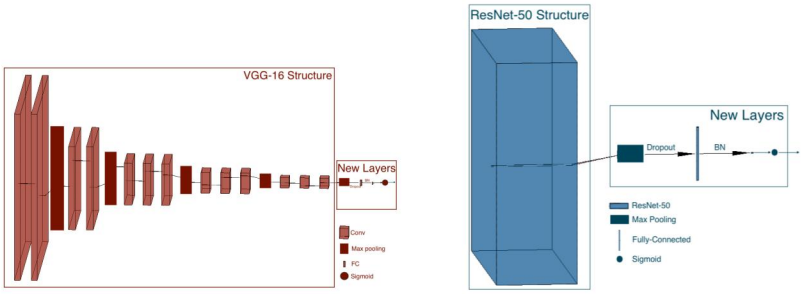


Figure 2: Architecture of the baseline model

Further, we used two deeper-NN architectures for transfer learning. For better feature-learning and given limited computational power and time, we started from using the pretrained weights from **ResNet50**. We also tried **VGG16** for its advantages in image classification. Then, we retrained the models and added our customized layers. We also tried different hyperparameter tuning on optimizer, learning rates, dropout rate regularizer, and batch size. Finally, we used the weights from the best model to predict the proportion of manipulated pictures in the sampled Kaggle dataset.



(a) Architecture of the VGG16 model (b) Architecture of the ResNet50 model

Figure 3: Architectures of improved models

3.3 Loss function

Since our task is a binary classification problem (authentic or fake), we decided to use Binary Cross-Entropy as our loss function: $\ell(x;\theta) = \sum_{ij} \ell'(x_{ij};\theta)$.

3.4 Evaluation Metric

Accuracy:

$$\frac{TruePositive + TrueNegative}{TruePositive + FalsePositive + FalseNegative + TrueNegative} \tag{1}$$

F1:

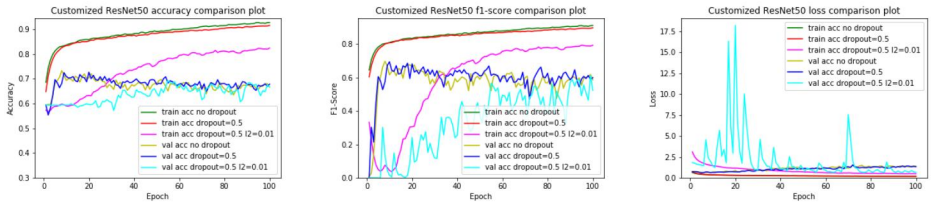
$$\frac{2 * (Precision * Recall)}{Precision + Recall} \tag{2}$$

4 Results

4.1 Model Comparison

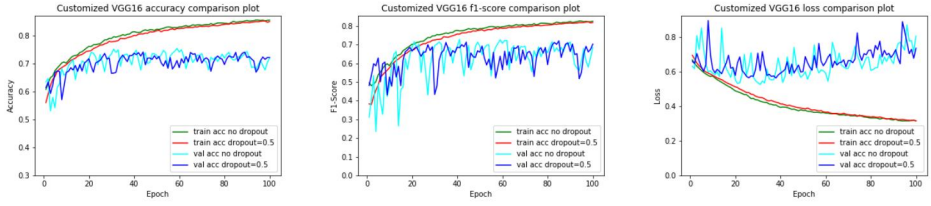
Model	Optimizer	Train Accuracy	Val Accuracy	Train F1	Val F1	Val Loss
ResNet-50 (No dropout)	Adam	0.93	0.66	0.91	0.57	1.38
ResNet-50 (0.5 dropout)	Adam	0.92	0.68	0.90	0.61	1.31
ResNet-50 (0.5 dropout, l2=0.01)	Adam	0.83	0.66	0.80	0.48	0.62
ResNet-50 (pre-trained)	Adam	0.88	0.59	0.85	0.00	6.45
VGG-16 (No dropout)	Adam	0.86	0.73	0.83	0.71	0.84
VGG-16 (0.5 dropout)	**Adam**	**0.85**	**0.72**	**0.82**	**0.72**	**0.76**
VGG-16 (pre-trained)	Adam	0.88	0.67	0.85	0.61	1.29
Baseline Model	Adam	0.82	0.74	0.79	0.74	0.59

Figure 4: Comparison Table for All Models



(a) Accuracy (b) F1-score (c) Loss

Figure 5: Customized ResNet50 Aggregated Results



(a) Accuracy (b) F1-score (c) Loss

Figure 6: Customized VGG-16 Aggregated Results

The baseline model eventually achieves a 0.71 validation accuracy, and 0.62 loss. The performance of the baseline model is not good enough no matter how we tune the parameters.

To improve the baseline model, we further researched on the famous CNN models, i.e. ResNet50 and VGG16, on top of which, we implemented transfer learning to build improved models. As seen

from the comparison table, the customized VGG-16 model with 0.5 dropout has the best performance among all the attempted models, which achieves a 0.82 validation accuracy, 0.71 validation f1 score and 0.76 validation loss.

4.2 Fake news image-tampering prediction

Finally, by using the weight of the optimized model (VGG-16, retrained, 0.5 dropout), we predicted the proportion of tampered images in the sampled Kaggle dataset. 50 images are predicted as authentic while 0 image is predicted as tampered. The human-coding results are 44 authentic vs 6 tampered.

5 Conclusion and Future work

In this paper, we explored different possibilities of detecting image-tampering, and trained a best model to be potentially used in identifying image manipulation in news reports. Overall, we found that VGG-16 performs better than ResNet-50, and we believe that this result is reasonable because:

1. Non-semantic features (like edge, corner, sudden change in color shade) are mostly captured by early layers, and later layers will loss the location information, and thus later layers in ResNet-50 may weaken the model performance of capturing non-semantic features used in image tampering.
2. Deep NN like ResNet50 tends to overfit more quickly.

However, although the deeper network models do work better than the baseline model, both VGG-16 and ResNet-50 results did not achieve our expectation. This could be resulted from the difficulty of our task and dataset. For example, the tampered areas are more subtle and usually only small proportions of the original image, which increases the difficulties in image classification. In other words, the task in this paper might not be a typical classification problem that VGGNet and ResNet are good at. Lastly, as the results of testing on real fake news data show, more work should be done to decrease the false positives at the same time.

For the next steps, what we can start from is to gather more well-labeled fake news images and improve and test the model on more reliable fake news image data. If the model still not perform well, more research should be done to find or train better models to better solve the problem.

References

[1] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[2] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[3] E. C. Tandoc Jr, Z. W. Lim, and R. Ling, "Defining "fake news" a typology of scholarly definitions," *Digital Journalism*, vol. 6, no. 2, pp. 137–153, 2018.

[4] A. Zubiaga and H. Ji, "Tweet, but verify: epistemic study of information verification on twitter," *Social Network Analysis and Mining*, vol. 4, no. 1, p. 163, 2014.