Image-tampering Classification for Fake News Detection

Veronica Wang, Yingdan Lu {wangh28, yingdan}@stanford.edu

Introduction

Image-tampering is facilitated by powerful image-manipulation software and visual techniques - **Actively used** in creating **fake news**.

Fake news could contain misleading information and convey provocative negative emotions - image manipulation as one type of fake news

Goal: research on CNN algorithms, implement and compare VGG16 and ResNet50 , and apply transfer learning to achieve a model that is specifically designed to detect tampering images.

Data

- **CASIA-v2.0**
  - 12,324 color images in total
  - 7491 authentic and 5123 tampered color images
  - Tampered images: generated with random image-cropping, splicing, post-processing and realistic operations to avoid over-tampering

Authentic pictures

Tampered pictures

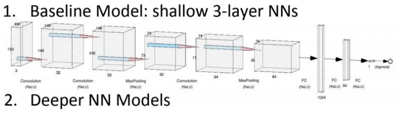
- **Kaggle Fake News Dataset**
  - Images from online fake or biased news
  - Unlabeled data
- **Pre-processing**

CASIA-v2.0	Kaggle
• Convert tif file to jpg image • Resize to 150 x 150 for baseline model and 224 x 224 for improved models • 90: 5: 5 train/valid/test set • Same authentic vs tampered ratio in three datasets	• Sampled 50 pictures with no logos, no overly tampering and not paintings • Resize to 224 x 224 • Human tagged the pictures as tampered/authentic

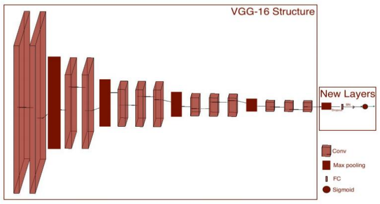
Methods

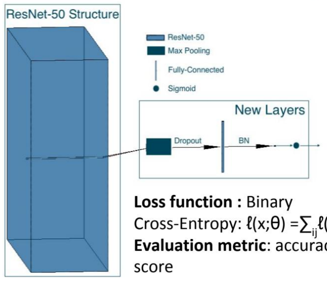
Models

1. Baseline Model: shallow 3-layer NNs



2. Deeper NN Models
   Transfer-learning with VGG-16 and ResNet-50
   - Tried both fine-tuning with pre-trained weights and retraining + customized layers.
   - Hyperparameter tuning on optimizer, learning rates, dropout rate regularizer, and batch size.





Loss function : Binary Cross-Entropy: $\ell(x;\theta) = \sum_{ij} \ell(x_{ij};\theta)$

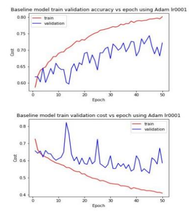
Evaluation metric: accuracy, F1 score

Workflow

1. Model training and enhancing based on the CASIA v2.0 dataset
2. Use the weights from the best model to predict the proportion of manipulated pictures in the sampled Kaggle dataset

Results

Baseline Results

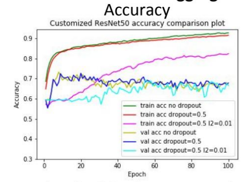


Model Comparison

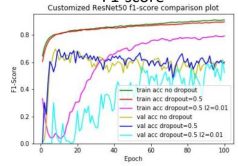
Model	Optimizer	Train Accuracy	Val Accuracy	Train F1	Val F1	Val Loss
ResNet-50 (No dropout)	Adam	0.93	0.66	0.91	0.57	1.38
ResNet-50 (0.5 dropout)	Adam	0.92	0.68	0.90	0.61	1.31
ResNet-50 (0.5 dropout, l2=0.01)	Adam	0.83	0.66	0.80	0.48	0.62
ResNet-50 (pre-trained)	Adam	0.88	0.59	0.85	0.00	6.45
VGG-16 (No 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

Retrained ResNet-50 Aggregated Results

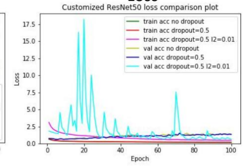
Accuracy



F1-score

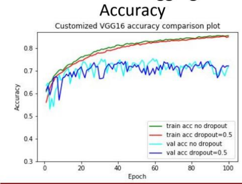


Loss

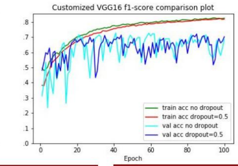


Retrained VGG-16 Aggregated Results

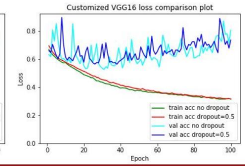
Accuracy



F1-score



Loss



Fake news prediction

- Machine prediction: 50 authentic, no tampered picture.
- Human-tagging: 6 tampered, 46 authentic.
- The model could be improved to better differentiate false positives.

Future Work

- Gather more well-labeled fake news images
- Improve and test the model on more reliable fake news image data

Conclusion

- The retrained VGG-16 model with 0.5 dropout has the best performance
- Overall VGG-16 models performs better than the ResNet-50. Possible reasons are:
  - Later layers in ResNet-50 may weaken the model performance of capturing non-semantic features (edge, corner, sudden change in color shade)
  - Deep NN like ResNet50 tends to overfit very quickly.
- VGG-16 and ResNet-50 results both did not achieve our expectation. This could because of the property of our task (tampered vs not), which is not a typical classification problem.

References

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151.
Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Schudson, M. (2018). The science of fake news. *Science*, *359*(6380), 1094-1096.
Tandoc Jr, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news" A typology of scholarly definitions. *Digital Journalism*, *6*(2), 137-153.
Zubiaga, A., & Ji, H. (2014). Tweet, but verify: epistemic study of information verification on twitter. *Social Network Analysis and Mining*, *4*(1), 163.
Dong, J., Wang, W., & Tan, T. (2013, July). Casia image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing* (pp. 422-426). IEEE.
Gunther, R., Beck, P. A., & Nisbet, E. C. (2018). Fake news did have a significant impact on the vote in the 2016 election: Original full-length version with methodological appendix. *Columbus, OH: Ohio State University. https://u. osu. edu/cnep/files/2015/03/Fake-News-Piecefor-The-Conversation-with-methodological-appendix-11d0ni9. pdf*