CS229 Project Report: Fake News Detection with Multimodal Machine Learning

Danni Ma
PanAgora Asset Management
dannima@stanford.edu

Kaijun Feng
OmniVision Technologies
kfeng19@stanford.edu

Chuanqi Chen
KLA
cchuanqi@stanford

Abstract

Nowadays the prevalence of fake news online has become a serious social issue with very negative impact societal impact. In this project, we intend to develop novel hybrid text+image models based on the Fakeddit database for 2-way classification of fake news. We implement and compare the classification results using both unimodel to detect just fake text or fake image. And then we use multimodal methods to detect fake text and image together, which further improve the performance compared to that of unimodal approach. The late fusion multimodal method achieves state of the art performance.

1 Introduction

Fake news has been a serious social problem with huge negative influence on both politics and culture. With the rise of social media and other online platforms, individuals can easily broadcast fake news, which poses a threat to the stability of society. Fake news has weakened public trust in governments, technology and business. According to the latest Edelman Trust Barometer 2021, trust in the tech industry fell to new lows in the majority of 27 countries surveyed. In particular, social media is the most distrusted sector with only 37% trust in the US. The extensive spread of fake news can have serious consequence to the society. For example, the misinformation on the cure of COVID may lead to misuse of drug which has not been proved to be safe and effective. As such, research in detecting fake news is of high importance for the society.

2 Related work

Multimodal fake news detection is an active research area. Here we survey some of the most well known work . Granik and Mesyura (2017) developed a naive Bayes classifier to detect fake Facebook news posts. Tan et al. (2020) demonstrated a method to detect neural fake news which are news articles generated by machines. Lu and Li (2020) showed a graph aware co-attention neural network to detect fake news on social media. Some other relevant efforts include using knowledge graph for fact aware language modeling [IV et al. (2019)].

Recently transformer based deep learning architectures have gain more and more attention for multimodal tasks; some prominent works include: Tan and Bansal (2019); Su et al. (2020); Hu and Singh (2021). These models are well suited for fake news detection because of their multimodality and high prediction accuracy. Among them, the pretrained vision-language cross modality framework "LXMERT" shows great potential for applications in fake news detection [Tan and Bansal (2019)].

Table 1: Dataset Statistics

Dataset	# Total Samples	# True Samples	# Fake Samples
All	1,063,106	527,049	628,501
Text Sentences	84,481	42,326	42,155
Multimodal samples	682,996		

3 Dataset and Features

To assist the development of machine learning methods for fake news detection, a number of fake news databases have been published. For instance, Wang (2017) developed a new dataset called LIAR which consists of 12.8k manually labeled short statements for model training. In this project, we uses a new dataset called Fakeddit, introduced by Nakamura et al. (2020), which consists of both images and text. They also constructed hybrid of language representation models and computer vision models for fake news classification.

Refer to table 1, Fakeddit dataset is composed of over 1 millions submission from 22 different subreddits. Approximately 64% of the dataset are multimodal samples which contain both text and images. We use these multimodal samples for our model training and experiment and error analysis.

The text sentences are preprocessed to be all lower cases, tokenizen, prepended with the '[CLS]' token to the start, appended the '[SEP]' token to the end, mapped tokens to their IDs, and then padded or truncated to 'max_length'. For example:

```
Original:  jamie gilt florida mom accidentally shot by yearold son
Tokenized BERT:  ['jamie', 'gil', '##t', 'florida', 'mom', 'accidentally',
'shot', 'by', 'year', '##old', 'son']
Token IDs BERT:  [6175, 13097, 2102, 3516, 3566, 9554, 2915, 2011, 2095, 11614, 2365]
Max sentence length BERT:  614

Original:  do you see the face
Tokenized BERT:  ['do', 'you', 'see', 'the', 'face']
Token IDs BERT:  [2079, 2017, 2156, 1996, 2227]
Max sentence length BERT:  93
```

And the images are transformed to the right dimension to be compatible to the convolutional neural network (CNN) model's input format (for example ResNET's input image dimension is 224x224.)

4 Method

We first implement separate language and vision models for unimodal detection. We then implement two types of multimodal architectures: late-fusion fake news detector and cross-modality encoders, as explained in the following sections.

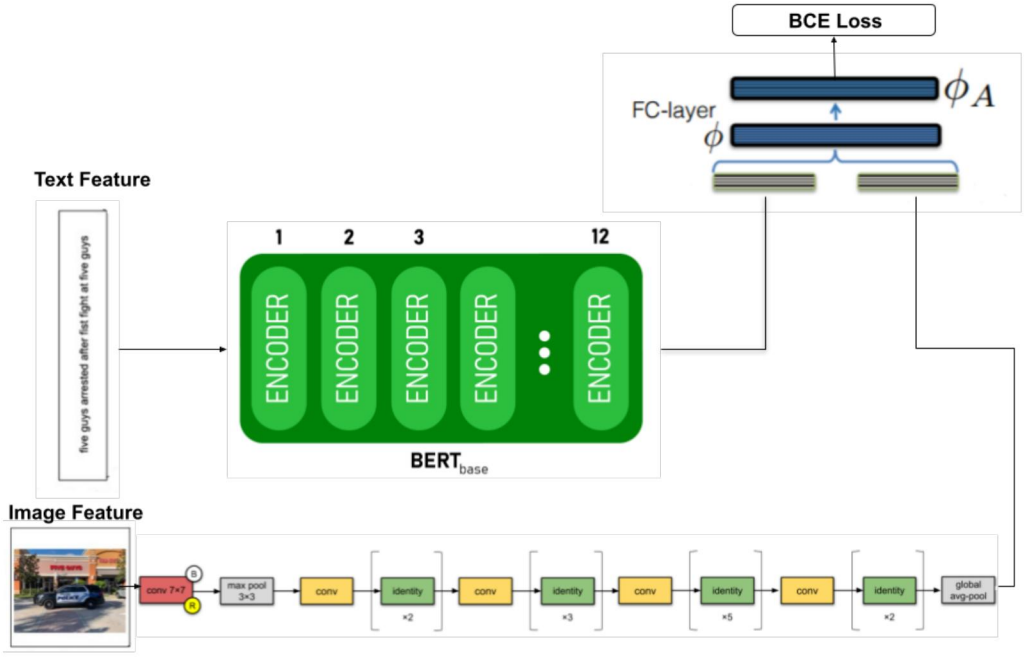
4.1 Fake news detection using images with deep learning methods

For the image data we implement and experiment with some of the well known convolutional neural network (CNN) architectures, including ResNet50, ResNet18 [He et al. (2016)], VGG16 [Simonyan and Zisserman (2015)]. We either train from scratch or fine tune pretrained models to obtain the feature vector which will later be used for feature combination.

4.2 Fake news detection using text with NLP methods

To detect fake text data, we implement and experiment with several latest NLP methods, including BERT[Devlin et al. (2019)], GPT-2 [Brown et al. (2020)], etc.

Figure 1: Model Architecture



Multimodal model for integrating text and image data for 2-way classification (extensible to 3-way and 6-way). Text sentences are fed through NLP model (such as BERT) while images are fed into CNN model (such as ResNET). The output of these two models are concatenated and fed into the hidden FC-layer, which is tuned to combine both output of NLP and CNN model optimally to make the best prediction.

4.3 Multimodal and Learning Cross-Modality Encoder Representations from Transformers

Based on Figure 1, to combine both text and image features for multimodal classification, we combine the image and text feature vectors and subsequently pass it through a fully connected layer with 1, 2 or 3 output features for 2 way or 3 way classification.

To further improve the multi-modal performance, we implemented multi-modal learning, LXMERT: Learning Cross-Modality Encoder Representations from Transformers [Tan and Bansal (2019)] to learning vision-and-language cross-modality representations.

5 Experiments/Results/Discussion

5.1 Fake Text Detection using NLP Unimodal Results

In this task to detect fake text news, we evaluate the model performance by grouping a 2-way binary class label {0: True, 1: False} of text sentences, and evaluate the prediction by accuracy.

We implement multiple NLP models: BERT[Devlin et al. (2019)], DistilBERT [Sanh et al. (2020)], Transformer-XL[Dai et al. (2019)], GPT2-XL [Brown et al. (2020)]. All these models achieve substantial state-of-the-art results on many classification tasks.

Refer to Table 2 for results comparison of these NLP models, we observe that the bert-base-uncased in Table 2 performs the best, it performs better than the original baseline result published in [Nakamura et al.].

The BERT[Devlin et al. (2019)] surprisingly performs better than any other latest much bigger NLP models as it might be more fitful to this tasks with good balances of model expressiveness without over-fitting the data.

Table 2: Fake Text Detection Benchmark Comparison Results

Method (NLP Task)	**Metrics**		
	Parameters	**Validation**	mcc
bert-base-uncased[1]	110M	0.89	0.767
bert-large-uncased	340M	0.501	
distilbert-base-uncased	66M	0.50	
gpt2-xl	1558M	0.73	0.482
Transformer-XL	257M	0.80	0.599

[1] Best NLP model to detect fake news.

5.2 Fake Image Detection using CNN Unimodal Results

Similar to fake text data detection, in this task to detect fake image data, we evaluate the model performance by grouping a 2-way binary class label {0: True, 1: False} of image data, and then evaluate the prediction by accuracy.

We implement and experiment with several of the well known convolutional neural network (CNN) architectures, including ResNet (ResNet50 and ResNet18) [He et al. (2016)], VGG16 [Simonyan and Zisserman (2015)]. The models are either trained from scratch or tuned from pre-trained models. Referring to Table 3, ResNet50 [He et al. (2016)] performs the best to detect fake image data.

Then we conduct fully retrain the ResNet50 model [He et al. (2016)] and the fully retrained model achieves big performance improvement over the fine tuned one. And it performs better than the same ResNet50 model [He et al. (2016)] as reported by original paper published in [Nakamura et al.].

Table 3: Fake Image Detection Benchmark Comparison Results

Method (CNN Task)	**Metrics**		
	Parameters	training	**Validation**
resnet50 (Fully Retrain) [1]	23M	0.8475	0.8418
resnet50 (Fine Tune)	23M	0.6846	0.7058
resnet18 (Fine Tune)	11M	0.6434	0.6637
VGG16 (Fine Tune)	138M	0.6464	0.6590

[1] Best CNN model to detect fake images.

For all experiments using CNN models to detect fake images data, we tuned the hyperparameters on the validation dataset. For the optimizer, we used Adam [Kingma and Ba (2017)] and tested three learning rate values: 1e-2, 1e-3, 1e-4, which 1e-4 achieves the best result. For each model, we specified a maximum of 20 epochs and an early stopping callback to halt training if the validation accuracy decreased.

5.3 Multimodal Fake News Detection Results: Late Fusion

The late fusion topology can be roughly depicted by Fig. 1. Note that our implementation is different from that of Nakamura et al. (2020) in the following aspects:

- Instead of using BERT and Resnet as fixed feature extractor, we retrained BERT on the full Fakeddit dataset and ResNet50 from scratch on the multimodal dataset. We then use them as feature extractor to train the classification layer.
- For combination, instead of combining feature vectors of the same size, we concatenate the 768 dimension BERT output and 2048 dimension ResNet50 feature vector to feed into the classification layer

- For 2-way classification, our final layer has an output feature size of 1 with the binary cross entropy loss, so as to reduce model redundancy and improve robustness.

For late fusion, we choose the best two models, i.e. BERT base uncased and ResNet50 as the feature extractor. Because the abundance of data, we trained the linear classifier for 1 epoch and we used Adam [Kingma and Ba (2017)] optimizer and learning rate of 1e-4. The result is listed in Table 4. The obtained test accuracy 0.9217 is much better than the 0.8909 accuracy reported in Nakamura et al. (2020).

Table 4: Multimodal Fake News Detection Benchmark Results

Method (Multimodal Task)	**Metrics**				
	training	validation	test	mcc	F1
Bert-base-uncased + ResNet	0.9403	0.9206	0.9217	0.8374	0.9003

6 Conclusion/Future Work

In conclusion, we have successfully implemented multiple NLP models to detect fake news, implemented and experimented multiple CNN models to detect fake images, and we have combined above two methods to do multimodal fake news detection to achieve state of the art performance.

The NLP model fine tune training takes about average 2 hours per model. Furthermore it's very time consuming to train the CNN models to detect fake images, averaging 21+ hours to fine tune a model with the multimodal dataset. We attempted the implementation of multi-modal learning, LXMERT: Learning Cross-Modality Encoder Representations from Transformers [Tan and Bansal (2019)] but don't have time to do any fine tune and experiments. In the summer, we plan to use the 3-way labels (in combination of image and text) with multi-modal learning, LXMERT [Tan and Bansal (2019)] to see if we can further improve prediction accuracy of Fake News detection.

Furthermore, we can explore multi-modal's capability to do multi-task such as additional hate speech detection [Gomez et al. (2019)]. Multi-tasks methods as outlined in paper [Hu and Singh (2021)] can be applied with additional hate speech dataset MMHS150K to train the multi-modal to detect fake news and hate speech at the same time. We can study the effects of such multi-task cross model training to see if these multi-tasks can benefit each other or undermine each other's performance.

7 Contributions and Code

The team members equally contribute to this project. The code repository is located at github: `https://github.com/chuanqichen/FakeNewsDetection`.

References

Granik, M.; Mesyura, V. *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)* **2017**, 900–903.

Tan, R.; Plummer, B. A.; Saenko, K. Detecting Cross-Modal Inconsistency to Defend Against Neural Fake News. 2020.

Lu, Y.-J.; Li, C.-T. GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media. 2020.

IV, R. L. L.; Liu, N. F.; Peters, M. E.; Gardner, M.; Singh, S. Barack's Wife Hillary: Using Knowledge-Graphs for Fact-Aware Language Modeling. 2019.

Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. 2020.

Hu, R.; Singh, A. UniT: Multimodal Multitask Learning with a Unified Transformer. 2021.

Wang, W. Y. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* **2017**, 422–426.

Nakamura, K.; Levy, S.; Wang, W. Y. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)* **2020**, 6149–6157.

He, K.; Zhang, X.; Ren, S.; Sun, J. *Proceedings of the IEEE conference on computer vision and pattern recognition 2016* **2016**, 770–778.

Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations. 2015.

Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019.

Brown, T. B. et al. Language Models are Few-Shot Learners. 2020.

Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2020.

Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q. V.; Salakhutdinov, R. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context. 2019.

Nakamura, K.; Levy, S.; Wang, W. Y. Fakeddit. `https://github.com/entitize/Fakeddit`.

Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. 2017.

Gomez, R.; Gibert, J.; Gomez, L.; Karatzas, D. Exploring Hate Speech Detection in Multimodal Publications. 2019.