Fake News Detection and Classification with Multimodal Learning

Stanford University CS224N Final Project

Charles Bai
cbai@stanford.edu

Yiqi Chen
yiqic@stanford.edu

Elaine Liu
yilinliu@stanford.edu

Abstract

In recent years, the prevalence of fake news has increased significantly with the rapid progress in digitization and the rise of social media. It has harmed our society greatly by spreading misinformation and escalating social issues. To combat the spread of misinformation in multiple modalities, we experimented with various new multimodal machine learning models and mutimodal feature fusion techniques to improve the current benchmark on fake news detection with Fakeddit dataset. [1] Our experiments demonstrate the importance of learning associations between the two modalities and aligning visual and text signals in the fake news detection task. Also, learning visually-grounded language understanding has also been proven to be transferable and pretrainable among different vision-and-language tasks.

Mentor: Shikhar Murty

1 Introduction

At the age of digital revolution and information explosion, the spread of fake news online, especially on social media, has become a prevalent problem in society. In recent years, we have all witnessed the great harm of misinformation on climate change, global pandemic, vaccine distribution, racism, and politics, etc. Hence, we aimed to build multimodal machine learning models to detect and categorize online fake news, which usually contains both images and texts.

We are using a new multimodal benchmark dataset, Fakeddit, for fine-grained fake news detection. [2] It contains 1 million well labeled samples of fake news data points sourced from Reddit. The paper presents a baseline model which combines BERT and ResNet to extract visual and textual features separately, and then it combined the features by simply adding, concatenating, taking the maximum and averaging.

Although the baseline results are already quite impressive, we believe more sophisticated visual/language feature fusion strategies and multimodal co-attention learning architecture could capture more semantic interactions/associations between visual and language features that come in pairs in fake news. The understanding of visuals should be conditioned on the text, and vice versa. This belief motivated us to explore several new approaches to this problem including mBert, MuRel, as well as ViLBERT after implementing the baseline model as a benchmark. Although these approaches were initially designed for other vision-and-language tasks, recent work has suggested transferability of learning visiolinguistic feature representations across tasks.

2 Related Work

2.1 VQA

VQA is a dataset that contains open-ended questions about images. It was first proposed in the original VQA paper [3]. The model is given an image and an open-ended questions about the image, and it is asked to answer the given question. Answering the question requires a deep understanding of vision, language and commonsense knowledge [3]. Since then, it became a classic multimodal task and there are lots of interesting work around it. We took inspirations from several VQA approaches and tried to apply them onto our multimodal fake news detection task.

The two tasks might look different at first: one about question answering and the other about classification, but the nature of these two tasks are actually quite similar. Both tasks require the model to understand visual and language signals, and learn the associations between the two modalities, and both require the model to have visually grounded language understanding.

There are also some fundamental differences between the two tasks. In VQA, only one or two parts of the image is relevant to the question. Take the following sample question and image from VQA dataset as an example, the answer only requires the network to focus on a specific object of the image.

Figure 1: Sample from VQA dataset

On the other hand, for fake news detection, classification often requires the network to take the whole image, or multiple objects within the image, into account. The text is also generally longer, noisier and more complicated compared to the questions in VQA dataset.

2.2 Multimodal Fusion

Fusion is one of the main topics in multimodal research. The baseline models in Fakeddit [1] simply concatenate the image and text features extracted. There are many other ways to fuse different modalities to achieve better results. Tensor Fusion Network [4] fuses modalities while learning both intra-modality and inter-modality dynamics end-to-end; FiLM [5] utilizes feature-wise affine transformation based on conditioning information; MAC (Memory, Attention, and Composition) network chains multiple recurrent MAC cell to turn the network into a series of attention-based reasoning steps [6]; MuRel network fuses two modalities by utilizing a chain of MuRel cells to reason both text and visual representations, and it also models the position relations [7].

Note that novel fusion methods might boost the model performance, but the training time may also be significantly longer, especially for the fusion networks in an iterative manner.

3 Approach

3.1 Baseline

Our baseline model architecture is illustrated in Figure 2. In the model, we applied BERT-base model on text data and ResNet-50 model on image data independently to get BERT token embeddings $T \in \mathbb{R}^{512\times768}$ and an $7 \times 7 \times 2048$ image embedding $M \in \mathbb{R}^{7\times7\times2048}$. Then for BERT embeddings, we append another BERT layer, and use the first token embedding as the final text embedding. For image embeddings, we run a max pooling to get a 2048-dimension image embedding, and feed into a

full-connected layer to get the final 768-dimension embedding (same dimension as text embedding). Finally, we concatenate image and text embedding, apply a layer of dropout and directly output prediction through a softmax classifier.

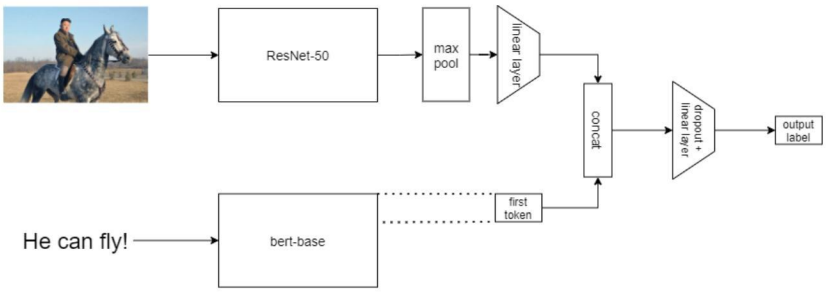


Figure 2: Illustration of Baseline Architecture

3.2 mBERT

MBERT (multimodal BERT) is a model proposed in [8] for sentiment categorization task on image + task data. The architecture is illustrated in Figure 3. It is an extension of the baseline model, which allows richer interactions between text and image outputs. The image embedding generation process is identical to the baseline model with output $m \in \mathbb{R}^{768}$. For text input, after applying the same pre-trained BERT model to get embeddings $T \in \mathbb{R}^{512\times768}$ and one additional BERT layer to get $T' \in \mathbb{R}^{512\times768}$, we concatenate m and T' on the embedding dimension to get a combined embedding $E \in \mathbb{R}^{513\times768}$. E is fed into another BERT layer such that attention can be applied on image and text and the same time. At the end we pull the transformed image embedding and first token of text embedding to generate final output through dropout and a softmax classifier.

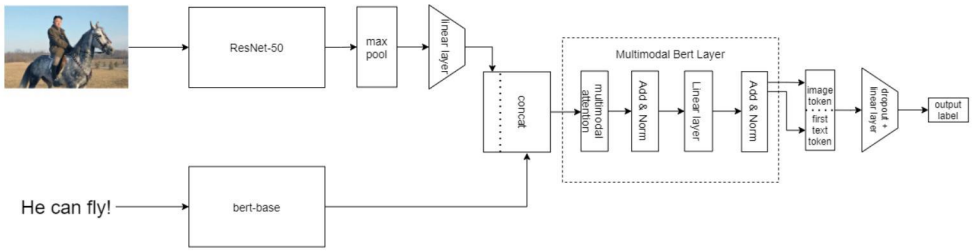


Figure 3: Illustration of mBert Architecture

3.3 MuRel

MuRel Network is a multimodal relational network originally developed for VQA task. The building block of MuRel Network is MuRel cell (see Figure 4), which allows rich interactions between the text/question and specific image regions.

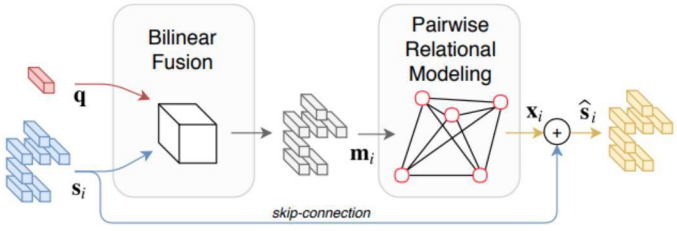


Figure 4: Illustration of MuRel cell [7]

Rather than using attention, in which each image region is only represented and weighted by scalars, MuRel cell represents each image region as a vector. A MuRel cell first joins text (represented as $\boldsymbol{q}$) with each image region embedding (represented as $\boldsymbol{s_i}$) through bilinear fusion (see Equation 1).

$$\boldsymbol{m_i} = B(\boldsymbol{s_i}, \boldsymbol{q}; \Theta) \tag{1}$$

Then the resulting representation $\boldsymbol{m_i}$ as well as each region's spatial representation $\boldsymbol{b_i} = [x, y, w, h]$ are passed through a pairwise modeling block (see Equation 2).

$$\boldsymbol{m_i} = B(\boldsymbol{b_i}, \boldsymbol{b_j}; \Theta_b) + B(\boldsymbol{m_i}, \boldsymbol{m_j}; \Theta_m) \tag{2}$$

Finally, a skip connection of image region embedding (represented as $\boldsymbol{s_i}$) is added to form the final output of the cell. Chaining multiple MuRel cells in an iterative manner, followed by max pooling and residual block forms a MuRel Network (see Figure 5).

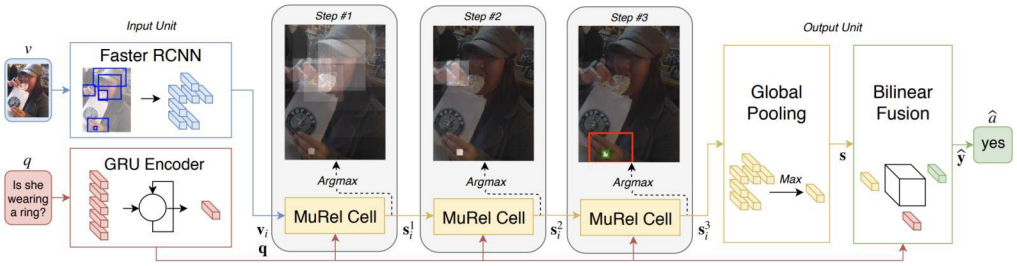


Figure 5: Illustration of MuRel network [7]

This architecture helps the model to gradually focus on the relevant image regions mentioned by the input text over iterations. Faster RCNN and GRU Encoder are used to generate image region embeddings and text embedding in the original paper.

We made some changes on top of the original paper's work. MuRel Network is designed for VQA task, where only a few regions of a image are relevant to the question. The architecture of the MuRel Network focuses on one specific region through the MuRel cells and max pooling. However, for Fake News Detection, the text is generally relevant to the whole image instead. So instead of taking a max pooling, we took a mean pooling instead, so the network can have information on more regions. Instead of using GRU Encoder, we used pre-trained BERT model to get the embedding of the input text. We also added a linear layer before the MuRel cells, to have visual features in the same dimension as the text features.

Since the MuRel Network applies the MuRel cell iteratively, the training speed is significantly longer when we apply this approach to our model. MuRel cell also relies on object detection, which makes it less transferable to other tasks that focus more on understanding the high-level scene rather than specific objects.

3.4 ViLBERT

ViLBERT (Vision-and-Language BERT) is a model for learning task-agnostic joint representations of image content and text. It has achieved state of the art results on multiple visual-and-language tasks. As explained in the original paper[9], it extends BERT architecture to a multi-modal two-stream BERT-like transformer model, processing both visual and textual inputs in separate streams that interact through co-attentional transformer layers (see Figure 6 and 7). Images are preprocessed to generate regional representations, including bounding boxes and regional features are generated with a pretrained object detection model (MaskRCNNn in our case). It also encodes the spatial location of the regions. Regional image features and location features are then projected to the same dimension and summed to form the image embedding. Text tokens are generated from the BERT's tokenizer.

Compared to using a single-stream BERT model, ViLBERT allows different treatments on visual and text inputs, and it enables two modalities to exchange information at different training depth. And the

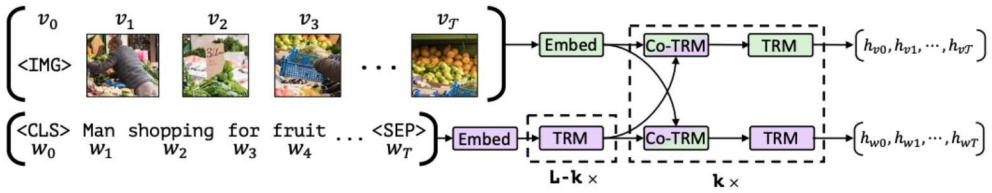


Figure 6: ViLBERT architecture[9]

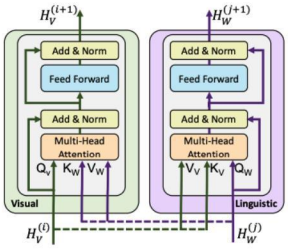


Figure 7: ViLBERT co-attention layers[9]

authors' intuition is that visual features are already high-level and require less transformer layers to learn contextual information.

This novel architecture has allowed ViLBERT to learn semantic alignment/association between visual and language features through pretraining. The authors of VilBERT pre-trained on the Conceptual Captions dataset using two training objectives: masked multi-modal learning and image-text alignment prediction. In their subsequent paper, they pre-trained the model using multi-task training on 12 established vision-and-language tasks, including visual question answering, visual commonsense reasoning, referring expressions, and caption-based image retrieval, etc [10].

Inspired by the novel architecture and strong results, we believe ViLBERT is applicable in the multimodal fake news detection/categorization task. Through fine-tuning on our dataset, it will learn visually grounded language understanding in the fake news context to help categorize the news content. Using the multi-task pretrained model, we added a linear classification layer on top of the elementwise product of image and text representations to predict a score for each of the six news categories. The final prediction is a softmax over these four scores and is trained under a cross-entropy loss.

4 Experiments

4.1 Data

We are using the large-scale multimodal fake news "Fakeddit" dataset, which consists of 1 million samples from multiple categories of fake news sourced from Reddit. We are only using the a subset of the data with both text and images The samples are labeled with 2-way, 3-way, and 6-way classification categories.[2] The public multimodal training set consists of 564,000 samples; validation set consists of 59,341 samples; public test set consists of 59,342 samples. In particular, we will focus on 6-way classification for this project. Please see Figure 8 for their semantic meanings and examples.

4.2 Evaluation method

We evaluate our results on the test data set by calculating percentage of text/image pairs the model is able to correctly classify for 6-way classifications on public test set.

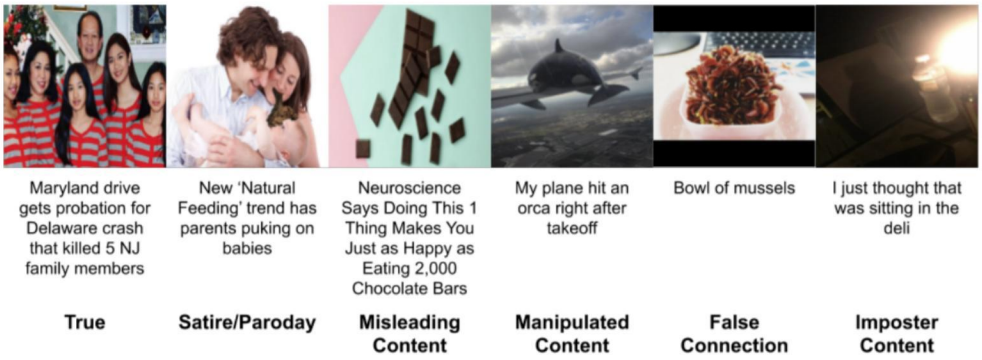


Figure 8: Dataset examples with 6-way classification labels

4.3 Implementation Codebase

Our main codebase is built on top of implementation provided by TomBert [8]. Also, we referenced GitHub repo "12-in-1: Multi-Task Vision and Language Representation Learning" from facebookresearch [10] for ViLBERT implementation and pretrained models, and "murel.bootstrap.pytorch " from Cadene [7] for MuRel implementation.

4.4 Results

We present our 6-way classification accuracy on the test set containing 50,000 samples in Figure 9. We set 10^{-5} with linear warmup as learning rate and use Adam optimizer. Per-model experiment configuration is also in Figure 9.

Model	Layers trained	Training data size	Number of epochs	Accuracy on test set
Baseline	2 BERT layers unfrozen	560,000	3	86.86%
mBERT	2 BERT layers unfrozen	560,000	3	88.35%
MuRel	2 BERT layers unfrozen	560,000	3	65.96%
Baseline	2 BERT layers unfrozen	60,000	1	43.34%
mBERT	2 BERT layers unfrozen	60,000	1	45.68%
ViLBERT	E2E	60,000	1	86.67%

Figure 9: Result Table

We trained all models except for ViLBERT, and some of the MuRel models on full dataset with 3 epochs. We can see that mBERT performs better than baseline model due to the richer image and text interaction.

Although we only fine-tuned the pretrained ViLBERT on a subset of the data due to limited time and resource, prediction accuracy on the randomly selected 10,000 test samples is already high especially comparing to baseline and mBERT model trained under the same condition. This proves that the capability to learn semantic association between visual and language is transferable among different tasks. The pretrained multi-task model is very powerful in aligning image and text signals.

Our MuREL model has slow training speed compared to other models and we tried to train multiple versions of it with different configurations. All of the models are underfit and only one of them showed meaningful results, with 1 MuRel cell, mean pooling and no pairwise. When we had pairwise enabled and multiple MuRel cells, the training speed became slower, so we had to train it on a subset of data, and the results were not significant. The performance of the MuRel models were also not as strong as we would expected, as the models were underfitting and some of the models were only trained on a subset of the data.

Table 1: Evaluation Error Rate (%) of Model Variants on Overall and each Label

	Text only	Baseline freeze all BERT layers	Baseline unfreeze 2 BERT layers	mBERT freeze all BERT layers	mBERT unfreeze 2 BERT layers
Overall	16.8	28.0	13.1	15.3	11.6
True (Not Fake News)	10.6	16.8	10.3	10.5	9.3
Satire	32.8	65.2	27.9	35.2	21.8
False Connection	28.6	52.3	21.4	23.4	16.2
Imposter Content	44.0	94.0	42.5	64.0	42.9
Manipulated Content	12.1	13.3	6.5	8.4	7.3
Misleading Content	19.4	42.9	13.8	21.5	13.4

5 Model Analysis

In this section, we would like to answer 3 main questions:

1. How do different model variants perform on each label?
2. Does multi-modality help with fake news classification task?
3. Does text used in fake news classification task have distinct attention pattern from pre-trained BERT model?

Notice that when doing this analysis, we only uses variants of baseline model and mBERT model trained on full training dataset. Detailed classification error rate is shown in Table 1.

5.1 How do different model variants perform on each label?

From Table 1, we have the following observations:

1. In general all variants perform well in True label and manipulated content, likely because these 2 labels represent the most common scenarios in fake news detection, and therefore the samples are easy to train.
2. On the other hand, imposter content label has very high error rate on all model variants. Through inspecting samples of this label, we cannot grasp a static connection or pattern that these samples have.
3. Models with 2 BERT layers unfrozen have lower error rate than models with frozen BERT layers, which indicates that text distribution of Fakeddit dataset is vastly different from general Wikipedia dataset used to train BERT.
4. Baseline model with all BERT layers frozen has much worse performance than all other variants, which is reasonable because in this model, only the last linear layers are trainable, which greatly limits its ability to finetune, comparing to any other model with complex trainable architectures (i.e. attention layer).

5.2 Does multi-modality help with fake news classification task?

We want to understand whether multi-modality, in our scenario it's image, helps with this particular dataset. To do that, we trained a model using only text from Fakeddit dataset. In order to make sure that result comparison is unbiased between text-only model and regular model from the perspective of model size and architecture complexity, we reused mBERT architecture, with a slight modification which sets output of ResNet-50 to zero. The result is demonstrated in Table 1 as well. We can see that the text-only model outperforms baseline model with all BERT layer frozen, due to the reason explained in Section 5.1. The model performance is noticeably worse than all other multi-modal models though, which indicates the significance of image signal in this dataset.

Figure 10 is an example where the text-only model classifies as "False Connection", while the true label is "True". It's corresponding text is "kick failed", which accurately describes what's happening in the image. However, without the image, the "failed" word can easily be associated to negative labels. On the other hand, this example is accurately classified in regular mBERT model.

In sum, the text-only model cannot capture the actual relationship between image and text, and therefore multi-modality is very helpful for the classification of this task.

Figure 10: Example where the text-only model fails to classify. The corresponding text is "kick failed"

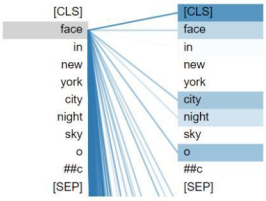
.

5.3 Does text used in fake news classification task have distinct attention pattern from pre-trained BERT model?

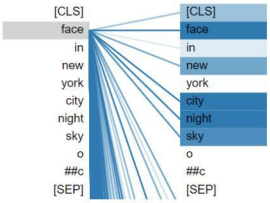
For this question, quantitatively, the short answer is yes, since we see from Table 1 that models with unfrozen BERT layers outperform models with frozen BERT layers. We will still like to closely examine what extra information does BERT captures while finetuning with Fakeddit classification task.

BertViz [11] is a tool for visualizing attention in the Transformer model including BERT. We loaded both mBERT with frozen weight model and mBERT with unfrozen weight model, and only apply the BERT submodule into BertViz. Figure 11 illustration the attention visualization of an example where mBERT with frozen BERT layers falsely classifies as True (not fake news) whereas the unfrozen mBERT model correctly classifies as False Connection. The keywords of classifying this example are "face" and "sky". We can see that "face" token has a strong attention with "sky" in unfrozen mBERT model, but there is no connection in frozen mBERT model.

(a) example image



(b) text visualization of frozen mBERT model



(c) text visualization of unfrozen mBERT model

Figure 11: Example where mBERT with frozen BERT layers fails to classify. The corresponding text is "face in new york city night sky oc".

6 Conclusion

In this project, we explored multiple multi-modality classification approaches, and applied these approaches on a new domain: fake news detection. Through the exploration, we demonstrated effectiveness in evaluation accuracy when image and text have rich interaction in model evaluation.

With more time and resource in the future, we will train ViLBERT model on full dataset to get a full measurement of model performance, and also longer time as we believe the models are underfit. Image feature extraction for ViLBERT was computationally expensive with MaskRCNN. We planned to switch to Detectron 2 and extract features on a multi-GPU machine. We will also try other fusion methods such as FiLM [5] and MAC [6] network. Furthermore, model ensemble can further improve classification accuracy.

References

[1] Kai Nakamura and Sharon Levy. Fakeddit. `https://fakeddit.netlify.ap`, 2020.

[2] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*, 2019.

[3] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering. *arXiv:1505.00468*, 2015.

[4] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv:1707.07250*, 2017.

[5] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. 2017.

[6] Drew A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. *arxiv:1803.03067*, 2018.

[7] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1989–1998, 2019.

[8] Jianfei Yu and Jing Jiang. Adapting bert for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5408–5414. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[9] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. 2019.

[10] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[11] Jesse Vig. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*, 2019.