Multimodal Fake News Detection

Nicola Fiorentino
n.fiorentino3@studenti.uniba.it

Abstract

Detecting fake news is a challenging task. This work tries to address it using a multimodal approach. We test innovative methods based on the use of vision language models and traditional methods based on supervised learning. The proposed techniques are evaluated on the Fakeddit dataset.

1. Introduction

Fake news poses a major risk to our digitized societies. The spread of false information has serious consequences on the political, economic and social life. For this reason, developing models to automatically detect fake news can be seen as a primary research goal. This work proposes a multimodal approach for detecting fake news. Indeed, today's information does not have a pure textual nature. News is most often delivered via multimedia content, including images, audio and video. Our goal is to rely on this composite nature to determine if a news item is real or not.

In this regard, we first exploit the generalization capabilities of pretrained visual language models. Thanks to deep neural networks with billions of parameters, large language models (LLMs) have had a huge impact on natural language processing. Not only do they capture the syntax and semantics of human language, but they also demonstrate general knowledge of the world. As a result, LLMs are employed in a wide range of tasks without the need for additional training. By ingesting multimodal data, visual language models achieve an even more accurate understanding of the task at hand, thus generating better results.

Second, we test a traditional approach where a task-specific classifier is trained from scratch in a supervised setup. The effectiveness of the proposed methods is evaluated on a collection of news extracted from the social news and discussion website Reddit.

2. Data

Data to be classified are sampled from Fakeddit [4], a multimodal dataset consisting of 1,000,000 Reddit posts from the decade 2008-2019. Reddit organizes posts by theme into boards called subreddits. For each post, Fakeddit provides multiple information, including the submission title, image and subreddit source.

Collected posts have an upvote/downvote score not lower than one. Each post is labeled according to its subreddit theme. Among the labels provided, we only consider the two-way label to distinguish real news from fake news. Also, in our experiments we adopt the cleaned up variant of the submission title. This is lowercased and stripped of punctuation, numbers and words revealing the subreddit source.

For the purpose of the current work, we select a stratified sample of 20,000 instances from those subreddits deemed most suitable for a fake news detection task. Selected subbreddits for the two classes are shown in Table 1. At the sampling stage, we make sure that the image of each selected instance is available and the submission title is no less than five words.

In this context, news is considered fake if its content is misleading. This may be due to a satirical intent, the presence of false information or the use of a title which does not relate to the true meaning of the image.

Real News	*10000*
nottheonion	2203
neutralnews	529
usanews	381
upliftingnews	1575
mildlyinteresting	4902
usnews	410
Fake News	*10000*
satire	680
theonion	1456
misleadingthumbnails	1894
fakehistoryporn	5970

Table 1: Subreddits considered in the sampling stage.

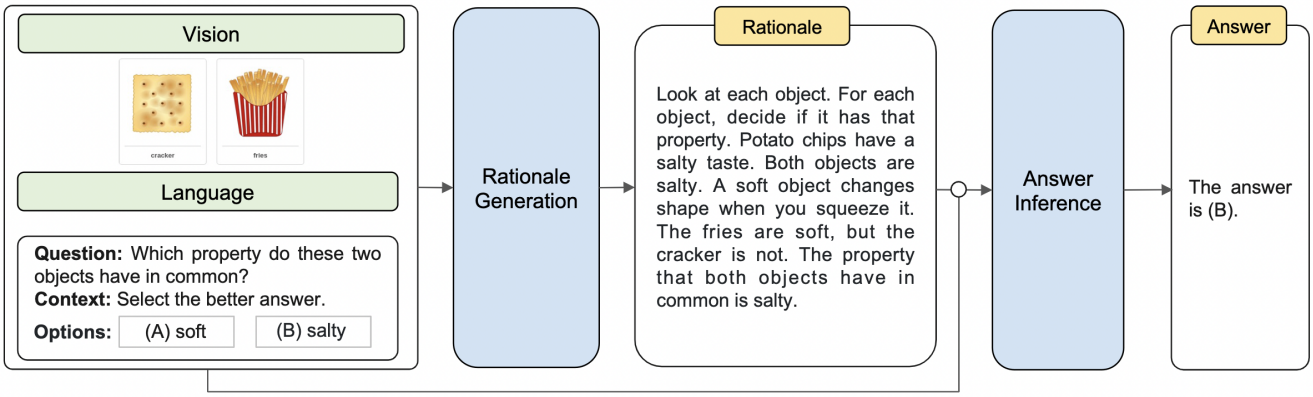


Figure 1: Overview of Multimodal-CoT framework.

3. Visual Language Models

One aspect of intelligence is the ability to quickly learn to perform a new task. When dealing with complex problems, the usual approach is to pretrain a model on a large amount of data, before fine-tuning it on the task of interest. A drawback of this approach is that fine-tuning is computationally expensive and requires large amounts of annotated examples.

More recently, powerful vision language models have been employed to solve complex tasks without additional training. This requires formulating the problem at hand via a text prompt. In this work, we test this technique with two publicly available models: Multimodal-CoT [5] and OpenFlamingo [2].

3.1. Multimodal-CoT

Language models can solve a problem by generating intermediate reasoning steps before inferring the answer. This technique, called chain-of-thought (CoT) reasoning, is investigated in [5]. The authors propose a Multimodal-CoT paradigm which incorporates text and images into a two-stage framework. Both stages share the same model architecture but differ in input and output. In the first stage, the model ingests a multiple choice question and generates a rationale describing the reasoning steps. In the second stage, the model infers an answer using the generated rationale. The overall procedure is illustrated in Figure 1.

In the current work, we test the pretrained framework released by the authors on a new task. The framework adopts *unifiedqa-t5-base* (223M parameters) as the backbone language model and is fine-tuned on the ScienceQA dataset [3]. Also, it relies on *detr-resnet-101-dc5* to extract visual features from images.

3.1.1 Experiments

In the first stage the framework is fed with a concatenation of question, submission title and possible answers to generate a rationale. The prompt is formulated as follows:

```
Question: Is this news real? {title}
Options: (A) No (B) Yes
```

In the second stage, the generated rationale is appended to the previous input to infer the answer.

The obtained results are not satisfactory. Generated rationales are redundant and often unrelated to the news. Furthermore, the accuracy achieved by the framework is comparable to a random classifier.

A close look at the data and model behavior reveals the causes of this result. First, small size language models can generate hallucinated rationales that mislead the answer inference. Second, the dataset used for training is markedly different from the one considered. In particular, ScienceQA consists of multiple choice scientific questions, annotated with detailed lectures. Each question is carefully worded and provided with contextual information. Moreover, images are schematic and references to possible answers are often included in the context.

Conversely, Fakeddit does not provide any additional information about the collected news. The only available text consists of short and often sarcastic titles written in web slang. The images provided are detailed or edited photos. Finally, to detect fake news a single general question with fixed options is asked for all the instances. Such features contribute to making Multimodal-CoT unsuitable for the task at hand.

3.2. OpenFlamingo

The second framework considered is OpenFlamingo [2]. It is an open-source reproduction of DeepMind's Flamingo [1], a family of visual language models capable of processing and reasoning about images, videos and text.

The purpose of the framework is to develop models

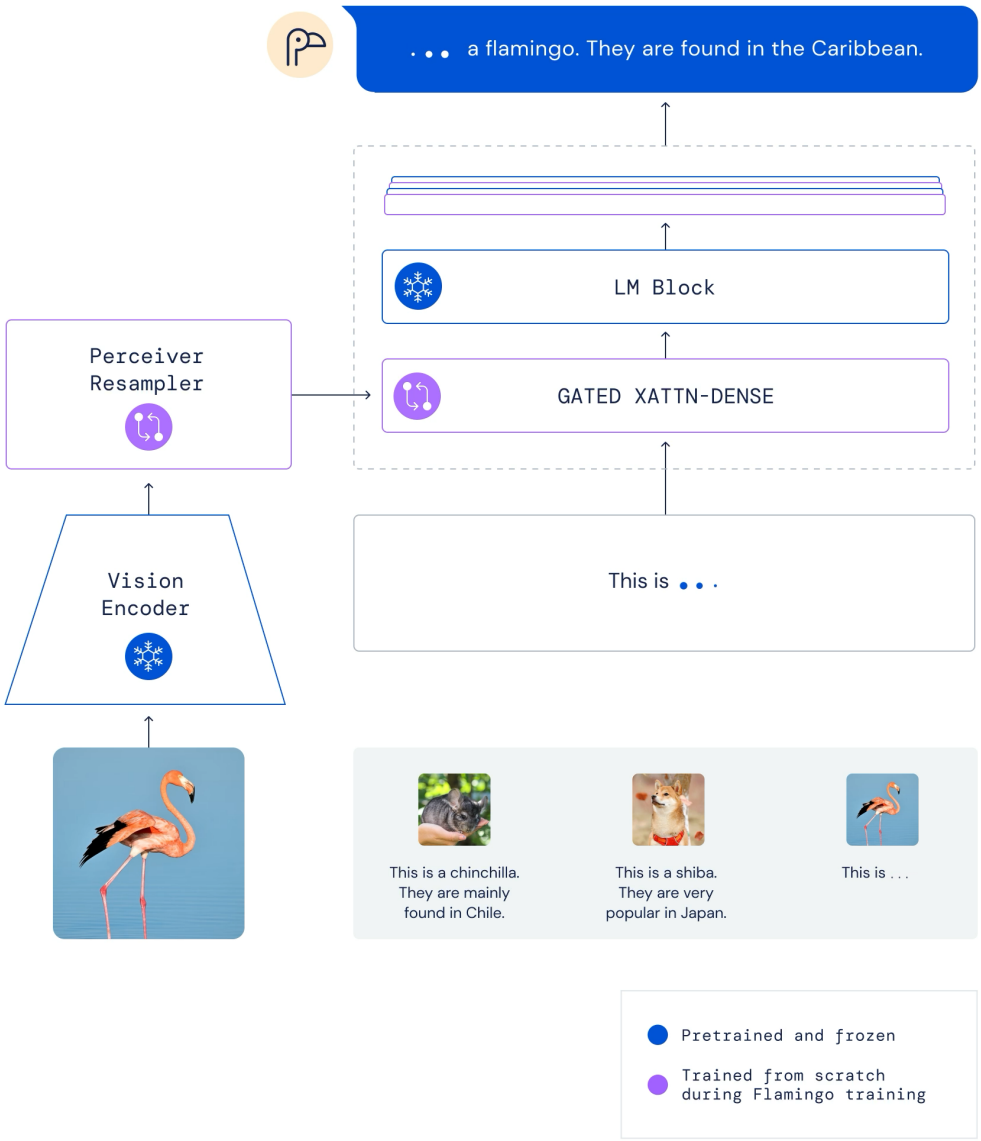


Figure 2: Architecture of Flamingo.

which can rapidly adapt to a variety of vision-language tasks. These include visual question-answering, captioning and image classification.

One single large model can achieve strong performance on a wide range of tasks by being prompted with a few examples of the task to be performed, along with a query. Differently from a pure language model, Flamingo models ingest a prompt containing images interleaved with text.

The architectural components of Flamingo are shown in Figure 2. The framework leverages pretrained vision and language models and bridges them effectively. To this end, a Perceiver Resampler receives features from a frozen Vision Encoder and generates a fixed number of visual tokens. These tokens are used to condition a frozen LLM with some cross-attention layers interleaved between the language model layers.

OpenFlamingo implements this architecture. It is trained on the open-source datasets Multimodal C4 and LAION-2B. The version used for experiments has 9B parameters and is based on the pretrained models LLaMA 7B and CLIP ViT/L-14.

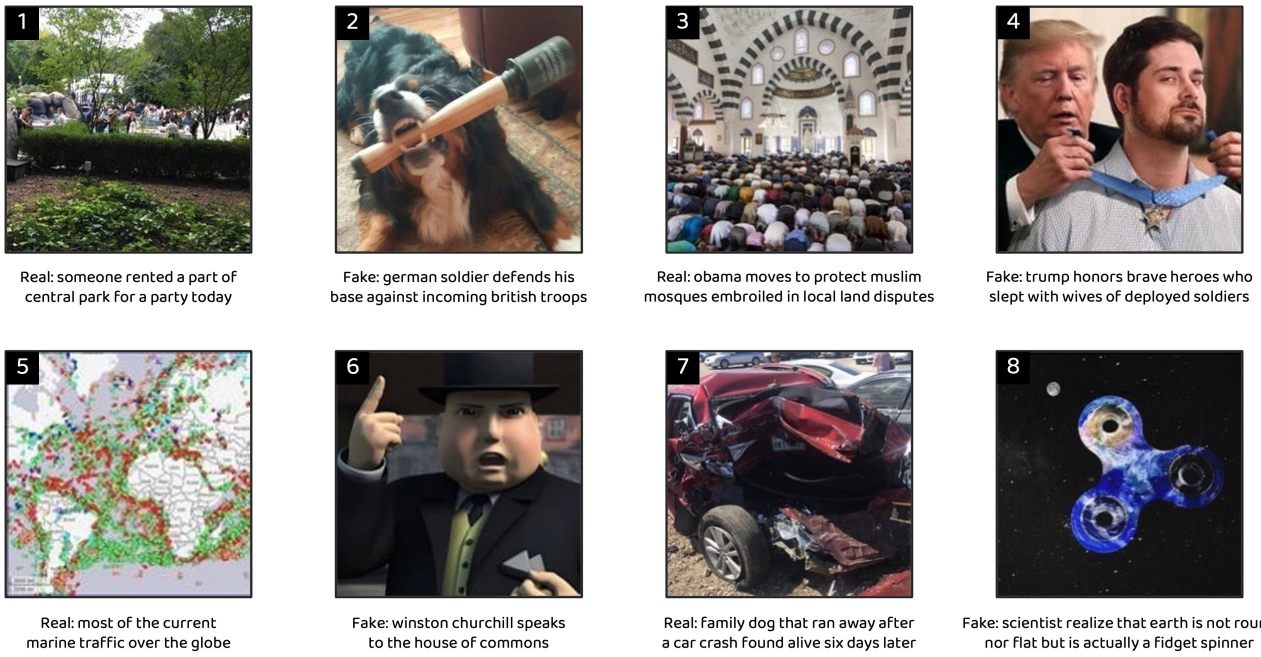


Figure 3: Support examples used by OpenFlamingo for few-shot learning. When building the prompt, an `<image>` special token indicates where an image is, while an `<|endofchunk|>` token indicates the end of the text associated with the image. The complete prompt has the following format:

```
<image>"{title1}". Question: Is this news real? Answer: Yes<|endofchunk|> ...
<image>"{title8}". Question: Is this news real? Answer: No<|endofchunk|>
<image>"{titleQ}". Question: Is this news real? Answer:
```

3.2.1 Experiments

First, we evaluate the ability of OpenFlamingo to adapt to a fake news detection task using in-context (or few-shot) learning. Specifically, we hand-pick a set of eight support examples in the form of $(image, text)$, where $text$ consists of a title, a question and an expected answer. Then, we build a multimodal prompt, by concatenating the support examples followed by the query for which the framework should make the prediction. Selected examples and prompt format are shown in Figure 3.

Second, we evaluate OpenFlamingo using zero-shot prompting, i.e. without providing any support examples. In this setup, the question is slightly modified as follows: *Is this news real or fake? Explain your reasoning step by step by comparing text and image*. This is useful for highlighting possible options and requesting a rationale which explains the chosen answer. The maximum number of new tokens the model can generate is set to 50. Moreover, in order to obtain a non-redundant output, repeated sentences are removed. To extract a class from the generated text we define the following rule: if the rationale contains the string *"is real"* the news is classified as real. Otherwise it is considered fake.

Due to GPU memory limitations, inference is performed using half numerical precision (16 bits instead of 32). The framework relies on beam search to reduce the risk of missing high probability word sequences. Specifically, it keeps the most likely num_beams hypotheses at each step and eventually chooses the hypothesis with the overall highest probability. To ensure a good trade-off between computational efficiency and quality of results, the value of num_beams is set to 3.

Since the two classes are perfectly balanced, we use accuracy for the evaluation. The achieved values are shown in Table 2. While the zero-shot setup results in limited performance, providing some support examples allows to better leverage the generalization capabilities of the framework. On the one hand, this confirms that OpenFlamingo can adapt to a completely new task. On the other hand, analyzing the rationales reveals that generated text is not always reliable. Examples of rationales generated by Multimodal-Cot and OpenFlamingo are shown in Figure 4.

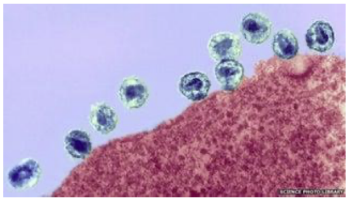
News	Multimodal-CoT Rationale Generation	OpenFlamingo Zero-shot Prompting
Real: this store i was at displays the calories you burn walking up the stairs 	A fake news story is made up of lies. Many people believe the stories they're telling are true. The number of people paying attention to your health is called your metabolism. The number of people paying attention to your health is called your metabolism. You burn calories when you walk. So, you burn the calories you burn walking up the stairs	I think this is real because it is on a staircase and it says calories you burn walking up the stairs
Fake: apple calculating the price of new iphone 	A company is making sure to tell the consumers the price of their products. The price of a new iphone is going up. So, the price of new iphone is going up. Apple is calculating the price of new iphone	This is a fake news because it is not possible to calculate the price of the new iPhone by using an apple. The apple is a fruit and not a calculator. It is not possible to calculate the price of the new iPhone by using an apple
Real: us man ordered to stop spreading hiv 	A news is a fact. It tells you how hiv is spread. The united states has declared the disease noncommunicable. People can get sick when they eat other people. But people are still living with hiv. The united states has declared the disease noncommunicable. People don't think this news is real. This is a fake story. People are afraid that the news will change	This is a fake news. The image is taken from a video game called "The Last of Us" which is a post-apocalyptic action-adventure survival horror video game developed by Naughty Dog and published by Sony
Fake: donald trump to sponsor bernie sanders independent run for presidency 	This news is fake. Donald trump to sponsor bernie sanders independent run for presidency. He is an independent candidate. This is a fake news. The news is a lie	This news is fake because it is a joke. It is a joke because Donald Trump is a republican and Bernie Sanders is a democrat. Also, Donald Trump is a billionaire and Bernie Sanders is a socialist

Figure 4: Examples of rationales generated by Multimodal-CoT and OpenFlamingo.

Prompting	Accuracy
Few-shot	**0.6897**
Zero-shot	0.5795

Table 2: Accuracy achieved by OpenFlamingo.

4. Supervised Learning

The second line of research involves training a task-specific classifier using a supervised setup. To this end, news items are first processed to extract relevant features. Then a machine learning model is trained on the extracted features. This approach is evaluated by combining different types of inputs, including the submission title, the rationale generated by OpenFlamingo and the image associated with the news.

4.1. Feature extraction

To extract the features, we first define a baseline using only text input. In particular, the submission title and OpenFlamingo rationale are converted into a numerical representation using a TF-IDF Vectorizer. The scikit-learn library is adopted for this purpose.

We then extract multimodal features using more advanced tools. Specifically, we adopt the Python framework SentenceTransformers to compute the embeddings related to title, rationale and image. Text embeddings are computed relying on the pre-trained model *all-mpnet-base-v2*. Image embeddings are computed with the pre-trained model *clip-ViT-L-14*. All the embeddings are 768-dimensional vectors.

4.2. Training

For training we consider a Support Vector Classifier (SVC) and a Multi-layer Perceptron (MLP) classifier. Since our primary goal is to evaluate the contribution of each input, the models are initialized with the default hyperparameters provided by scikit-learn. When training the MLP, the maximum number of iterations is extended to 500 to ensure convergence of the training procedure.

The performance of the classifiers is evaluated for each combination of the inputs: title only, image only, rationale only, title and image, title and rationale, image and rationale, title and image and rationale. For multi-input experiments, we append the texts when using TF-IDF Vectorizer, and concatenate the embeddings when using SentenceTransformers. Given the limited computational cost required for training, all experiments are performed using 10-fold cross validation.

4.3. Evaluation

The results obtained are shown in Table 3 and Table 4. As expected, the use of embeddings guarantees a clear performance improvement compared to TF-IDF features. The news title turns out to be more informative than the rationale generated by OpenFlamingo. Furthermore, using visual features helps to increase the overall accuracy.

If SVC achieves better results with TF-IDF features, neither model can be considered better than the other when embeddings are used. Training the MLP classifier by concatenating all the embeddings ensures the best overall result. However, this is barely superior to the setup with only title and image.

Text	SVC	MLP
Title	**0.8158**	0.7756
Rationale	0.7622	0.7160
Title + Rationale	0.8131	0.7698

Table 3: Classification accuracy using TF-IDF features.

Embeddings	SVC	MLP
Title	0.8485	0.8417
Image	0.8595	0.8511
Rationale	0.7768	0.7572
Title + Image	0.8653	0.8961
Title + Rationale	0.8604	0.8503
Image + Rationale	0.8635	0.8667
Title + Image + Rationale	0.8682	**0.9000**

Table 4: Classification accuracy using embeddings.

5. Conclusion

Experiments conducted show that vision language models can prove effective in performing unexpected tasks. If Multimodal-CoT is not able to adapt to fake news detection, OpenFlamingo demonstrates significant generalization capabilities. Using a few support examples provides the model with an understanding of the task at hand. This allows for decent performance in data-starved regimes. On the downside, such models are exposed to risks, such as propagating social biases or hallucinating.

The second line of research shows that, having enough labeled data, training a task-specific classifier can ensure better performance. In this respect, using transformers to compute embeddings is more effective than a simple Bag

of Words approach. Also, leveraging multimodal features proves to be beneficial.

As the capabilities of pretrained models are closely related to their size and the amount of training data, future research concerns the adoption of more powerful models. Larger models better exploit a greater number of shots [1]. Moreover, given sufficient computational resources, such models can be adapted to a fake news detection task by fine-tuning their weights. This would lead to more reliable results.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. *arXiv e-prints*, page arXiv:2204.14198, Apr. 2022.

[2] Anas Awadalla, Irena Gao, Joshua Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo (v0.1.1), Mar. 2023. https://doi.org/10.5281/zenodo.7733589.

[3] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[4] Kai Nakamura, Sharon Levy, and William Yang Wang. r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. *arXiv e-prints*, page arXiv:1911.03854, Nov. 2019.

[5] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal Chain-of-Thought Reasoning in Language Models. *arXiv e-prints*, page arXiv:2302.00923, Feb. 2023.