Network-aware Fake News mitigation on social media

Scientific Report #1

Mitrica Octavian

S.L. Dr. Ing. Ciprian-Octavian Truică
Conf. Dr. Ing. Elena-Simona Apostol

Faculty of Automation and
Computers Politehnica
University of Bucharest
Romania

Introduction

Task: We want to detect fake news in social media posts containing some text and an image.

Classification: **1 – Fake / 0 – Not fake**

Introduction

Dataset sample [1]

Introduction

The main *objective* of this scientific report is to find out, through experiments, if the similarity between image and text corresponding to a post is relevant to take into consideration for the task of fake news detection. We are going to have three types of experiments:

- Text Only detection
- Image Only detection
- Multimodal detection

Literature

Text oriented models are the most popular ones, with papers like:

- DEAP-FAKED [2]: using NLP techniques, Graph Neural Networks and Knowledge graphs
- MCDWST [3]: using word embeddings like Word2Vec and BiLSTMs networks
- FNC-1 [4]: using transfer learning from transformers to improve word embeddings

Literature

There isn’t much recent interest including fake news detection using only the image data, since the text is more descriptive, but these research papers provided insightful information:

- Jin et al. [5] presented how different patterns are common in images associated with fake news.
- ViT paper [6] which provides a large variety of pre-trained transformer models for many image-based tasks; image classification is the one that interests us and we will experiment.

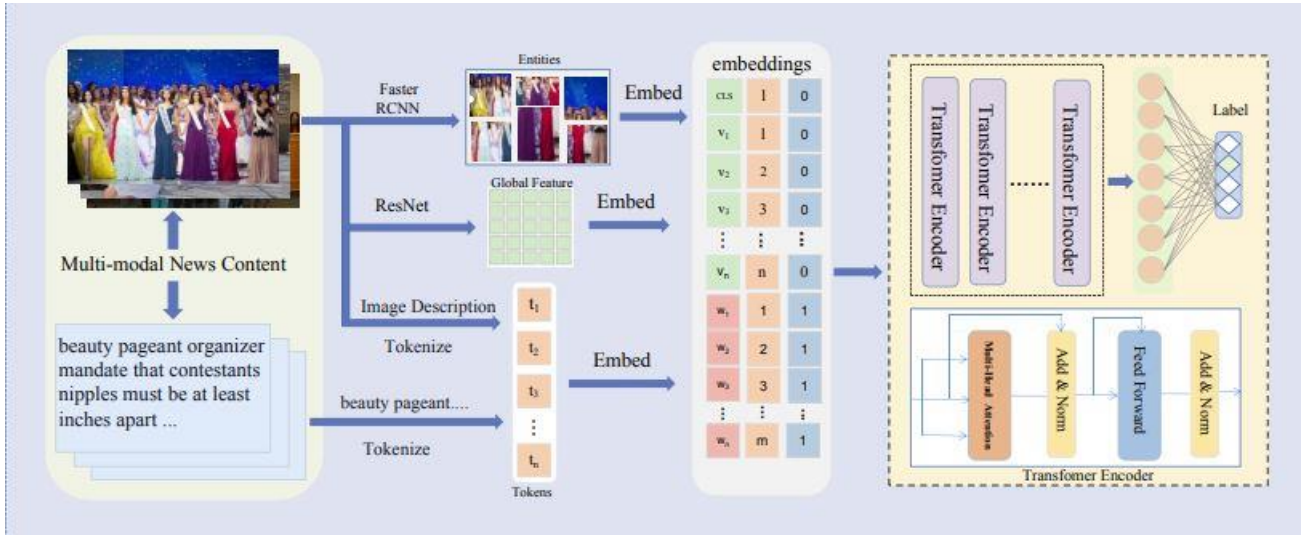
Model	Layers	Hidden size D	MLP size	Heads	Params
ViT-Base	12	768	3072	12	86M
ViT-Large	24	1024	4096	16	307M
ViT-Huge	32	1280	5120	16	632M

ViT models

Literature

Multimodal architectures have been documented in the recent years, showing relevant results and outcomes. Some of the papers that provide valuable information:

- Nakamura et al. [1] introduced the Fakeddit dataset and also experimented with detecting fake news using the title and image of a reddit post; their best performing model was using BERT for text data and ResNet50 for visual data
- Liu et al. [7] takes on the challenge of transformer fusion between text transformers and image transformers, expanding on the caption-based enhancement tactic. They achieved state-of-the-art results on Fakeddit.



Overview of a multimodal architecture [7]

Datasets

The dataset used for testing was Fakeddit [1], along with Twitter15 and Twitter16 [8] for baselines. Fakeddit is a multimodal dataset containing over 1 million samples and 2-way, 3-way or 6-way classification labels.

More examples:

Look at what I found in Washington D.C!

“Meowster electric”

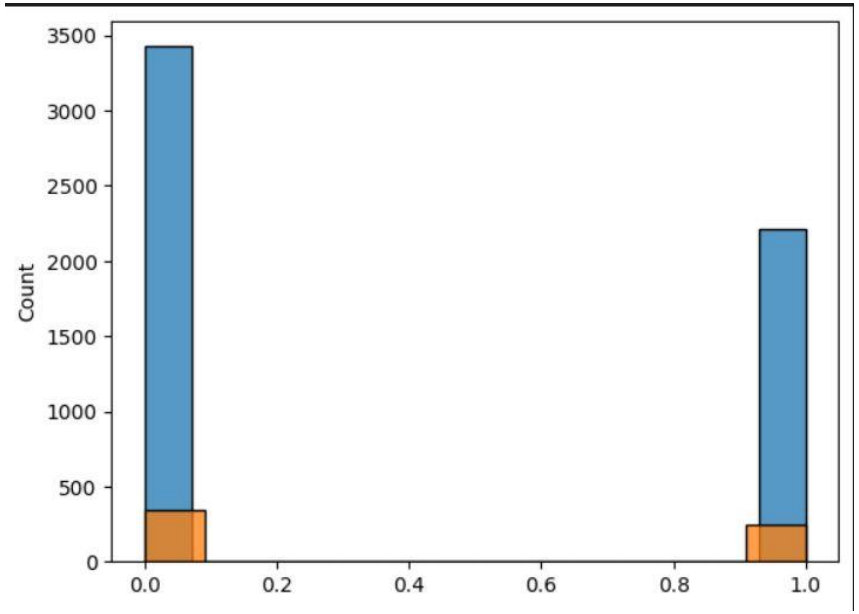
Datasets

Dataset Statistics	
Total samples	1,063,106
Fake samples	628,501
True samples	527,049
Multimodal samples	682,996
Subreddits	22
Unique users	358,504
Unique domains	24,203
Timespan	3/19/2008 - 10/24/2019
Mean words per submission	8.27
Mean comments per submission	17.94
Vocabulary size	175,566
Training set size	878,218
Validation set size	92,444
Released test set size	92,444
Unreleased set size	92,444

Fakeddit statistics [1]

Dataset	Articles	Fake news	Real news	Unverified
Fakeddit	5633	2214	3419	0
Twitter15	1340	332	670	335
Twitter16	740	187	370	181

Used datasets statistics



Label distribution (2-way labels)

Text preprocessing

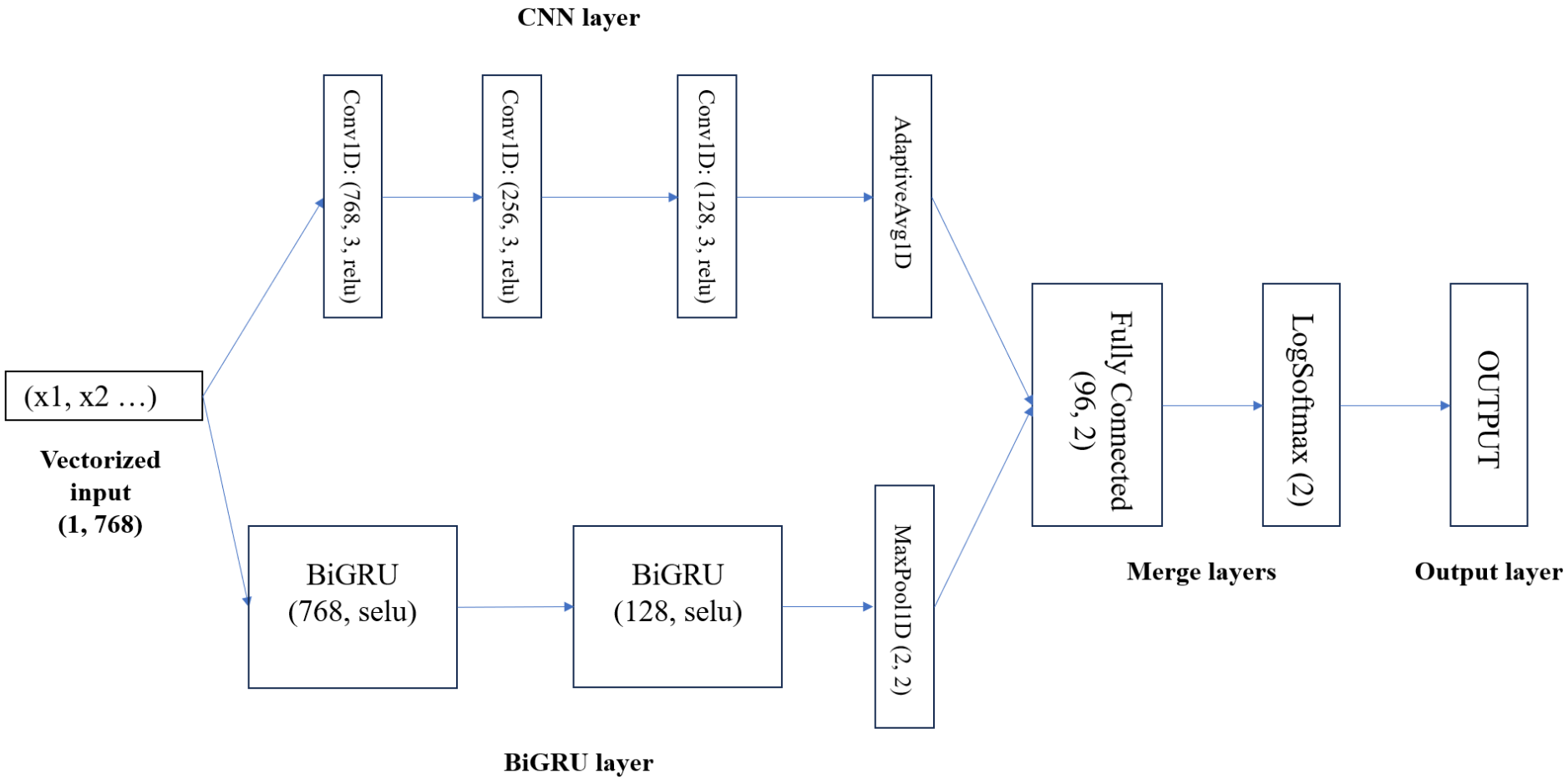
Steps before feeding the text to a model:

- Clean stopwords, punctuation and words with length < 2
- Lemmatization
- Vectorize

Example:

- “!The cats are running around...” ---------- “cat running around” (not fake)
- “join R.A.A.F” ---------- “join raaf” (fake - propaganda)

Text Only model – DistilBERT + BiGruCNN



BiGRU-CNN model overview – Inspiration taken from papers [9] and [10]

Text Only model – DistilBERT + BiGruCNN

Text input is vectorized with DistilBERT pretrained and then fed to the BiGRU-CNN model for actual training. This model uses a 2-layer BiGru in parallel with a 3-layer CNN and then concatenates the outputs from both to get a prediction.

“cat running around” ---DistilBERT---> Tensor[1, 768]

Example: tensor([[1.5245e-01, -1.6061e-02, 6.2370e-02, -6.7484e-03, 4.3814e-02, …

Tensor[1,768] ---BiGRU-CNN---> Tensor[1, 2]

Example: tensor([[-3.0292e-01, -1.3419e+00]])

Final scores

Best accuracy: 76.73%

Best precision: 76.66%

Best recall: 75.11%

```
Loading data...
Training the model...
---Epoch: 0---
Saving...
Epoch Loss: 0.5157557427883148
Training Average Accuracy:  0.8023049645390071
---Epoch: 1---
Saving...
Epoch Loss: 0.47458386215670356
Training Average Accuracy:  0.8267730496453901
---Epoch: 2---
Saving...
Epoch Loss: 0.4482321009553712
Training Average Accuracy:  0.8460992907801419
---Epoch: 3---
Saving...
Epoch Loss: 0.4279957181163903
Training Average Accuracy:  0.8585106382978723
---Epoch: 4---
Saving...
Epoch Loss: 0.41153880172762375
Training Average Accuracy:  0.8716312056737588
---Epoch: 5---
Saving...
Epoch Loss: 0.39735255518863943
...
---Epoch: 14---
Saving...
Epoch Loss: 0.29500182971529576
Training Average Accuracy:  0.9085106382978724
```

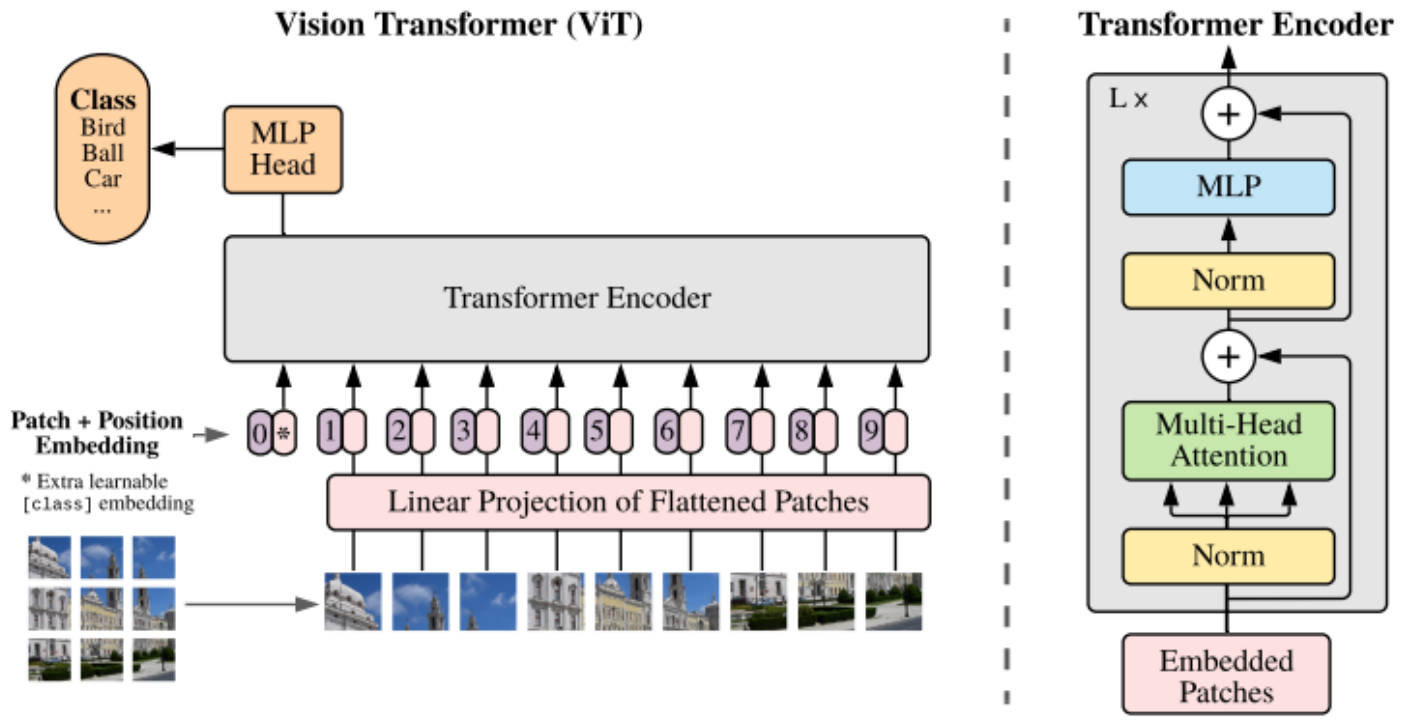
Image preprocessing

Steps before feeding the image to a model:

- Remove corrupted files
- Convert to “RGB” / “BRG”
- Resize
- Normalize
- Apply augmentations (Flip, Rotate, etc)
- Transform into tensors

Tensor[3, 224, 224]

Image Only model – ViT



ViT pipeline [6]

Image Only model – ViT

Image is preprocessed with a ViTImageProcessor from HuggingFace and then fed to a ViTForImageClassification pretrained model for fine-tuning. We modify the hyperparameters in order to get a better prediction.

Results:

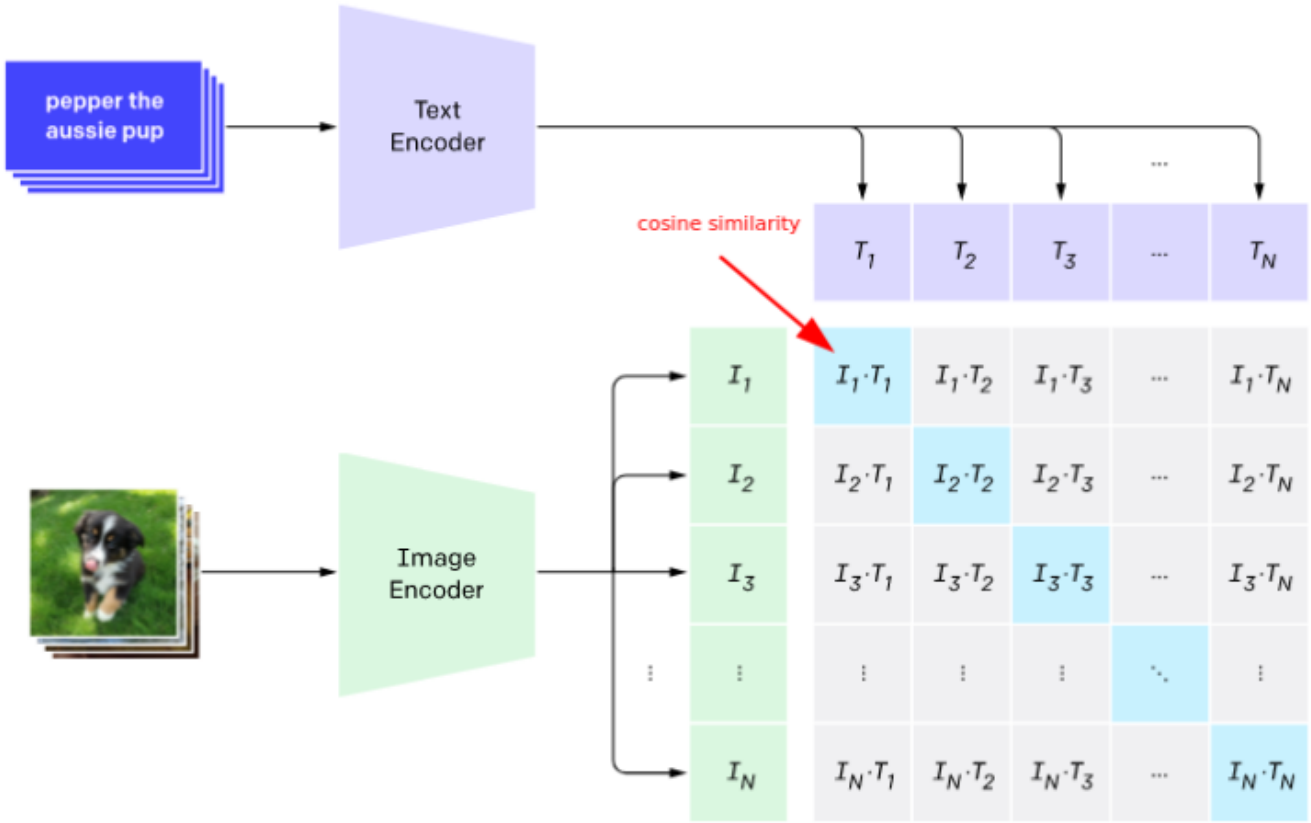
```
***** eval metrics *****
  epoch                   =        3.0
  eval_accuracy           =     0.7432
  eval_f1                 =     0.7337
  eval_loss               =     0.5341
  eval_precision          =     0.7377
  eval_recall             =     0.7315
  eval_runtime            = 0:00:17.96
  eval_samples_per_second =     32.951
  eval_steps_per_second   =      4.119
```

```
{'loss': 0.642, 'learning_rate': 0.00019858356940509917, 'epoch': 0.03}
{'loss': 0.6399, 'learning_rate': 0.0001971671388101983, 'epoch': 0.06}
{'loss': 0.7222, 'learning_rate': 0.00019575070821529746, 'epoch': 0.08}
{'loss': 0.6306, 'learning_rate': 0.00019433427762039661, 'epoch': 0.11}
{'loss': 0.6375, 'learning_rate': 0.00019291784702549575, 'epoch': 0.14}
{'loss': 0.5992, 'learning_rate': 0.0001915014164305949, 'epoch': 0.17}
{'loss': 0.6454, 'learning_rate': 0.00019008498583569406, 'epoch': 0.2}
{'loss': 0.6675, 'learning_rate': 0.00018866855524079322, 'epoch': 0.23}
{'loss': 0.6125, 'learning_rate': 0.00018725212464589238, 'epoch': 0.25}
{'loss': 0.6115, 'learning_rate': 0.00018585356940509915, 'epoch': 0.28}
```

```
c:\Users\Tavi\Desktop\AI Masters\An I\Sem I\Research I\Experiments\.venv\lib\sit
You can avoid this message in future by passing the argument `trust_remote_code=
Passing `trust_remote_code=True` will be mandatory to load this metric from the
  warnings.warn(
{'eval_loss': 0.576085090637207, 'eval_accuracy': 0.6739864864864865, 'eval_runt
{'loss': 0.603, 'learning_rate': 0.00018441926345609067, 'epoch': 0.31}
{'loss': 0.5518, 'learning_rate': 0.00018300283286118983, 'epoch': 0.34}
{'loss': 0.5103, 'learning_rate': 0.00018158640226628896, 'epoch': 0.37}
{'loss': 0.6265, 'learning_rate': 0.00018016997167138811, 'epoch': 0.4}
{'loss': 0.6039, 'learning_rate': 0.00017875354107648725, 'epoch': 0.42}
{'loss': 0.7205, 'learning_rate': 0.0001773371104815864, 'epoch': 0.45}
{'loss': 0.6277, 'learning_rate': 0.00017592067988668556, 'epoch': 0.48}
{'loss': 0.5369, 'learning_rate': 0.0001745042492917847, 'epoch': 0.51}
{'loss': 0.5864, 'learning_rate': 0.00017308781869688385, 'epoch': 0.54}
{'loss': 0.5776, 'learning_rate': 0.00017171671388101983, 'epoch': 0.57}
```

Fine-tuning the model

Multimodal model



Multimodal pipeline – contrastive-training [11]

Multimodal model

The objective of this experiment was to explore the multimodal side of fake news detection since there are just a few papers tackling the subject.

In the chase for a conclusive research and a nice comparison between models, time was running short and the limitations of my computing power were reached.

My goal was to fine-tune a pretrained DistilBERT + ViT encoder / decoder but the local machine was at it’s limit already and time was ticking.

Finally I combined the two outputs from the last model and applied a softmax over them resulting in better metrics but leaving ample room for improvement

Results:

Overall accuracy: 77.7 %

Overall precision: 77.27 %

Overall recall: 76.72 %

```
OutOfMemoryError: CUDA out of memory.
```

Limitations

Results

Model	Accuracy	Precision	Recall	F1
DistilBERT BiGruCNN	76.73	76.66	75.11	
Fine-tuned ViT	74.32	73.77	73.15	
Multimodal	**77.70**	**77.27**	**76.72**	

Comparison of the experiments

Model	Accuracy	Precision	Recall	F1
MCWDST [4]	76.90	77.27	76.89	76.82
Fakeddit [8]	89.09	-	-	-
DEAP-FAKED [3]	89.55	-	-	-
(BERT+Dense)+ Xception [9]	91.87	93.39	93.29	93.25
Multimodal transformers [10]	92.51	93.83	93.74	93.79
Ours	77.70	77.27	76.72	76.99

Comparison with state-of-the-art

Conclusion

From the results we can see that combining text and image proves to be relevant. There are a lot of ways to improve the model forward and do some more sophisticated layering at the end. With more computing power we can train on the whole dataset and see how the results improve.

A big limiting factor was the local environment. Moving forward, switching to cloud GPUs will be more time effective.

References

1. Nakamura, K., Levy, S., & Wang, W.Y. (2019). r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. ArXiv, abs/1911.03854.
2. M. Mayank, S. Sharma, and R. Sharma, “DEAP-FAKED: knowledge graph based approach for fake news detection,” CoRR, vol. abs/2107.10648, 2021.
3. C. -O. Truică, E. -S. Apostol, R. -C. Nicolescu and P. Karras, "MCWDST: A Minimum-Cost Weighted Directed Spanning Tree Algorithm for Real-Time Fake News Mitigation in Social Media," in IEEE Access, vol. 11, pp. 125861-125873, 2023, doi: 10.1109/ACCESS.2023.3331220.
4. V. Slovikovskaya, “Transfer learning from transformers to fake news challenge stance detection (fnc-1) task,” 2019.
5. Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, “Novel visual and statistical image features for microblogs news verification,” IEEE Transactions on Multimedia, vol. 19, no. 3, pp. 598–608, 2017.
6. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in International Conference on Learning Representations, 2021.
7. P. Liu, W. Qian, D. Xu, B. Ren, and J. Cao, “Multi-modal fake news detection via bridging the gap between modals,” Entropy, vol. 25, no. 4, 2023.
8. X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, “Real-time rumor debunking on twitter,” in Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM ’15, (New York, NY, USA), p. 1867–1870, Association for Computing Machinery, 2015.

References

9. J. Duan, H. Zhao, W. Qin, M. Qiu and M. Liu, "News Text Classification Based on MLCNN and BiGRU Hybrid Neural Network," 2020 3rd International Conference on Smart BlockChain (SmartBlock), Zhengzhou, China, 2020, pp. 1-6, doi: 10.1109/SmartBlock52591.2020.00032.
10. Ma, Yuqun, Chen, Hailong, Wang, Qing, and Zheng, Xin, “Text classification model based on cnn and bigru fusion attention mechanism,” ITM Web Conf., vol. 47, p. 02040, 2022.
11. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever “Learning Transferable Visual Models From Natural Language Supervision” https://arxiv.org/abs/2103.00020