

Truthformers at Factify 2022 : Evidence aware Transformer based Model for Multimodal Fact Checking

B S N V Chaitanya^{1,2}, Potluri Prathyush^{1,2} and Vijjali Rutvik¹

¹Indian Institute of Information Technology, Sri City

²These authors contributed equally

Abstract

There has been a dramatic increase in the spread of misinformation and fake news since the growing sophistication in online communications. Despite the development of numerous fact-checking frameworks and models in research, fake news continues to be the primary concern, especially on social media platforms. Being able to identify false claims based on document references can mitigate the spreading of unverified news. We present a multi-modal fact-verification model which processes a claim data pair and corresponding verified documents to determine the claim's legitimacy. The proposed model uses both text and image content of claims and the documents to determine the level of support provided by the document for the given claim. We make use of Transformer based models for effective processing of the different modalities and use a fusion block for identifying cross-modal representations that embed combined information of the considered modalities from both claim and document to get the final predictions. Our solution ranked number three with a weighted F1-score of 0.7486. The code is available at <https://github.com/pryus/truthformers>.

Keywords

Multi-modal Fact verification, Multi-modal fusion, Self Attention, Transformers, Conv fusion

1. Introduction

In today's world, fake news is becoming a serious societal issue. People understand what fake news is, but they can't tell the difference between it and real news. Fake news may be found on nearly every social media platform. It's meant to catch people's attention and lead them astray. As a large number of people in today's world have easy access to the Internet, online and social media has become a platform for a large number of people to keep up with current events across the world. Since its inception, news coverage has been created with the goal of serving individuals as an apparatus for learning about facts and truth, and so bringing value to their lives. It has, however, evolved to now give people what they want to hear, whether it is bogus or true. Because of the amount of time people spend on social media, they are extremely

De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, co-located with AAAI 2022. 2022 Vancouver, Canada

✉ viswachaitanya.b16@iiits.in (B. S. N. V. Chaitanya); prathyush.p16@iiits.in (P. Prathyush); rutvikreddy.v16@iiits.in (V. Rutvik)

🌐 <https://github.com/chaitnayabasava> (B. S. N. V. Chaitanya); <https://github.com/PRYUS> (P. Prathyush); <https://github.com/rutvikvijjali> (V. Rutvik)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

exposed to fake news, which may quickly impact them. Fake news is difficult to rectify, and even when it is, it can have harmful consequences.

Considerable efforts have been made to combat fake news and many advances have been made in the past few years and although most existing approaches rely primarily on textual fake news detection, recently, datasets like Fakeddit [1], Politifact [2], Fever [3], LIAR [4], etc have encouraged researchers to work on multi-modal (images, memes, videos) fact-checking. Using multi-modal information to detect fake news has many advantages as different modalities capture different dimensions of the news article and they can complement each other while evaluating the genuineness of the article.

The Factify task is a multi-modal entailment task to detect multi-modal fake news. Here, each data point has a credible source of information, referred to as the "document," and its accompanying image, as well as another source whose authenticity must be determined, referred to as the "claim," which also has an associated image. The purpose is to determine if the claim includes the document. Entailment contains two verticals, textual entailment and visual entailment, and their corresponding combinations, because we're interested in a multi-modal situation with both image and text. The uniqueness of the task is that there is a single credible source of news and we need to distinguish fake/real claims from a vast number of multi-modal assertions. As a result, the task essentially is for the system to categorise a textual claim, claim image, document, and document image into one of five categories: Support_Text, Support_Multimodal, Insufficient_Text, Insufficient_Multimodal and Refute. Figure 1 shows an example from the Factify dataset in the Support_Multimodal category where both the claim text and image are similar to that of the document.

In this paper, we present our system for the task of multi-modal fake news detection. We make use of transformer based models with pseudo labelling and other strategies to achieve third position in the Factify 2022 task.

2. Related works

Fake news detection and fact checking methods mainly rely on using textual content and linguistics [5]. In [6], the authors propose a fact checking model built by using knowledge graphs of textual content to improve fact analysis in news content. In [7], a novel two stage transformer based fact checking algorithm is proposed that retrieves the most relevant facts concerning user claims about particular facts for COVID-19. In [8], the authors built a pipeline to find documents and sentences to fact-check mutated claims generated from Wikipedia pages. In [9], the authors aimed to find web pages related to given fact checking articles and predict their stances on claims in the fact checking articles. Other methods also include using temporal spreading patterns [10].

Using textual only uni-modal techniques were able to produce promising findings but the brief and informal character of social media data always poses difficulty in information extraction. To get over this constraint, researchers began experimenting with characteristics taken from several modalities (such as text and image) and fusing them together to create a richer data representation and therefore using multi-modal information for fact checking and fake news detection has gained more traction in recent years. They are generally categorized into two



Candidates at Democratic debate open up about biggest setbacks they've faced. <https://t.co/HAzpFytDYo>

Claim



By Eric Bradner, Dan Merica and Gregory Krieg, CNN Updated 1457 GMT (2257 HKT) September 13, 2019 Houston, Texas (CNN)Pete Buttigieg called it "unwatchable." Amy Klobuchar warned that "a house divided cannot stand." Julian Castro said what was unfolding on stage was "called an election."

Document

Figure 1: Example of a sample from the Support_Multimodal category in the Factify dataset where both the claim text and image are similar to that of the document

different categories, one focuses on taking text and image inputs [9] to check the claim's veracity and the other one focuses on evidence-aware fact-checking where inputs are pairs of a multi-modal claim and a fact-checking-article [11]. Focusing on evidence-aware fact-checking helps to increase users' awareness of verified news. In [11], the authors use multi-modal data in social media posts to search for verified information. In [9], the authors built an end-to-end model where the extracted image and text representations are fed in two fully connected neural network classifiers, one for event discriminator and another for fake news classification.

3. Methodology

In this section, we discuss the proposed model used for extracting useful semantic information for textual and image data for carrying out the task of multi-modal fact verification. We first discuss the backbone models used for extracting embeddings from the text and image information, respectively. We then explain the fusion techniques employed for generating an effective semantic representation that combines the text and image embeddings. Each input example contains a pair of textual and image data for both the claim and document.

3.1. Truthformer Model

Figure 2 shows the high-level flow of the proposed model. Both the claim and document consist of a pair of text and image data, respectively. We used separate embedding blocks for text and image to extract useful semantic representations from the data individually for both claim and document. We also fine-tuned these blocks on the factify data, which helped us capture better inter-modal representations. We fuse these inter-modal representations using a fusion block, which carries out the multi-modal fusion of the text and image data for claims and documents individually. The output of the fusion block embeds the information from the considered modalities and generates a final embedding vector representation of (768,) size. We pass this final vector through a fully connected layer that outputs the class probabilities for the 5 considered class types.

We trained all the involved blocks simultaneously with a batch size of 8 and initialized the models involved in the embedding blocks from their respective pre-trained weights. We used a shared embedding block between claim and document for each of the 2 modalities.

In addition to the general training step, we applied another step called pseudo labelling, which resulted in a significant improvement in the performance of the proposed model. Pseudo labelling is a semi-supervised technique in which we first train the model using the train set and then use this trained model to predict the labels for the test set. We then use this newly labelled set along with the original train set for further model training.

3.2. Textual Embedding block

For extracting meaningful representations for the input text sequence, we made use of the multilingual BERT (mBERT) model [12]. The mBERT is the original BERT base model pre-trained on the top 102 languages with the same MLM objective as BERT using the Wikipedia corpus. mBERT develops complex cross-lingual representations that enable language transfer for various languages along with Hindi. This model uses multi-head self-attention to combine information from different parts of the text sequence and gives out an effective embedding representation for each token in the input sequence.

We used the base tokenizer of mBERT provided by HuggingFace [13] for tokenizing the input text sequences. We padded/truncated the input token sequence to a fixed max length of 256. The model processes the input tokens using self-attention and produces a 768 sized embedding for representing each token. We use these (256, 768) embeddings as the representation of the text data. The claim and document text are processed using the model individually to get a text representation for each. We used the weights provided by Google and fine-tuned the model on the factify data.

3.3. Image Embedding block

We used Vision Transformer (ViT) [14], a BERT based image model, to extract the semantic representations of the image data. ViT is the first transformer encoder model that achieved good results compared to the convolutional architecture models when trained on the Image-Net data. ViT in contrary to the CNN models, does not involve the usage of any convolutions for capturing the local information of images. It instead uses 16x16 non-overlapping patches of

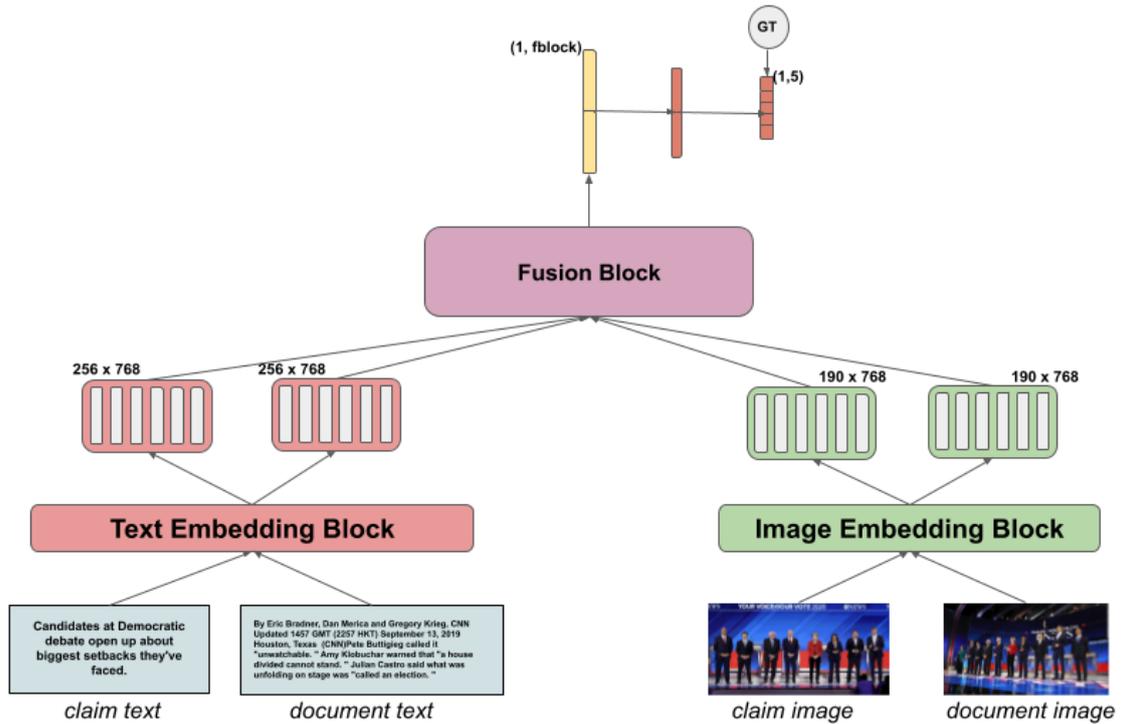


Figure 2: Block diagram of the proposed truthformer model

the image resized to 224x224 as the input sequence to the transformer model. We used the pre-trained weights provided by the authors of ViT to initialize the model and fine-tuned it on the factify data.

The model takes an image of size 224x224 as input and processes it as a patch sequence. The model splits the image into a patch sequence of 16x16 patches, where each patch acts as an input token. The splitting generates a sequence of length 190, which includes a [CLS] token added to the beginning of the sequence. ViT processes this 190 lengthed input sequence and generates an output of similar length, with each vector having a size 768. We use this (190, 768) embedding to represent the image processed by ViT. The claim and document images are processed individually using ViT to get the image representations for each.

3.4. Fusion block

For fusing the textual and image embedding information, we have tried 2 different approaches and trained separate models using either of them. We will discuss the approaches in detail.

Fusion using Conv1D layer (FusConv1D): In this, we used a Conv1D layer with a kernel size of 3 to process the sequence embeddings and fuse them into a single vector. We initialized a separate Conv1D layer for both modalities. The text conv1D layer would process the (256, 768) embedding from textual embedding block and generate a (768,) text vector. Similarly,

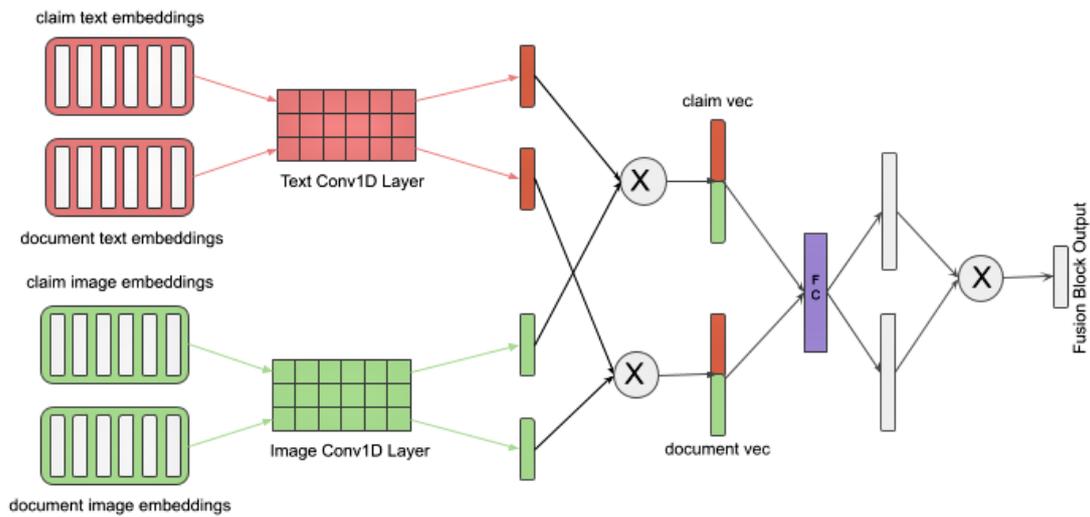


Figure 3: Fusion block using Conv1D and fully connected layers for fusing the textual and image embeddings of claim and document data.

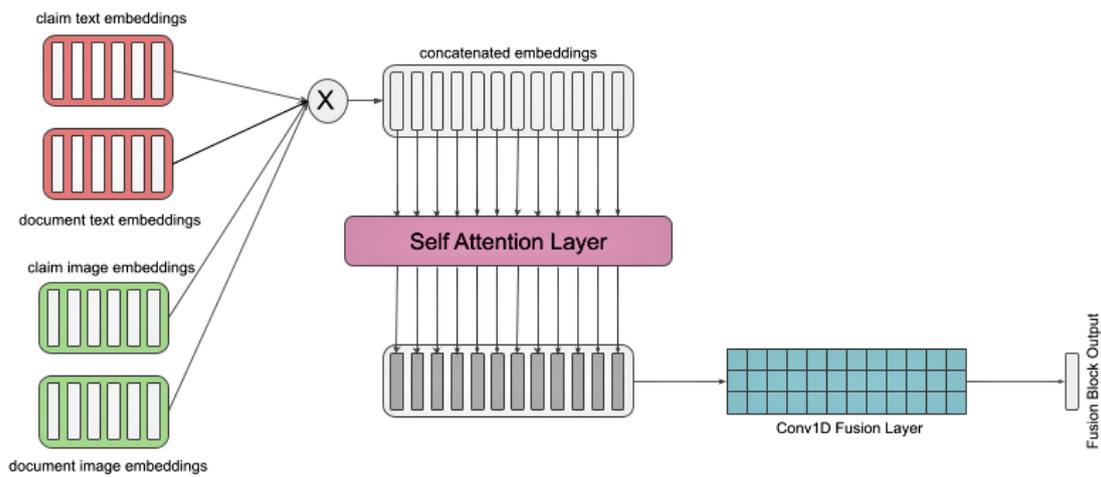


Figure 4: Fusion block using Self-Attention and combined Conv1D layer for fusing the textual and image embeddings of claim and document data.

the image conv1D layer would process the (190, 768) embedding from the image embedding block to generate a (768,) image vector. We then concatenate the claim pair vectors and the document pair vectors to get the final representations for the claim and document pairs. We further process these vectors individually using a shared fully connected layer and concatenate the respective outputs to get the output of the fusion block. Figure 3 shows the workings of this approach.

Fusion using a common self-attention layer (FusAttn): In this, we concatenate the embeddings of the text claim, text document, image claim and image document data and pass them through a self-attention layer that encourages the model to learn efficient cross-modal representations. We then pass the output of the self-attention layer through a Conv1D layer to fuse the sequence embeddings into a single vector of (768,) size, which would be the output of the fusion block. Figure 4 shows the workings of this approach.

4. Experiments

In this section, we discuss the different experiments and compare them with the baseline model performance provided by the Defactify team. We trained all the variations of the proposed models on Google colab pro machines.

4.1. Dataset

For training and validating the performance of the proposed fact verification model, we used the dataset provided by the Defactify team [15] [16]. The 5 categories involved in the classification task are Support Multi-modal, Support Text, Refute, Insufficient Multi-modal and Insufficient Text. The dataset consists of 35,000 training, 7,500 validation and 7,500 test samples. There is an equal distribution in the number of data points for each category. Each data point consists of a text and image pair of claim and document, respectively. We carried out data analysis and estimated the average length of each claim text to be 184 and that of document text to be 2779. Due to the resource limitations, we had to truncate the document text to a max length of 256 even though we would lose considerable information, especially in document text.

We pre-processed the text data by expanding the contractions and removing the URLs, user mentions, stop words and hashtags. We replaced the emojis with respective identification word tags. We observed multilingual data and hence used the BERT version trained on multilingual text data. All the images were standardized and resized to a resolution of (224, 224) before being passed to the ViT model.

4.2. Experimental setup

In all the models, the text has been either padded or truncated to a max length of 256 as the BERT model is compute-intensive and, we were also using the ViT model, which is also a BERT based model. We were limited to a smaller batch size of 8 due to these resource limitations. These limitations in resources led us to choose between accurate batch statistics or processing more text information by BERT, resulting in a trade-off situation between batch size and max length of textual token data.

Adam optimizer with 0.9 beta1 and 0.999 beta2 was used for optimizing the cross-entropy loss function. All the models were trained for 15 epochs each, with a learning rate of 0.0002. We validate and compare the efficacy of the trained models using accuracy and weighted F1-score.

Table 1

Weighted F1-scores of all the methods on the validation and test sets. Following are the abbreviations **PL**: Pseudo Labelling, **FusConv1D**: Fusion using Conv1D layer, **FusAttn**: Fusion using a common self attention layer, **NSP-singleBERT**: Model using the Next Sentence Prediction setting to process claim and document textual data.

| Model | val | test |
|-------------------------------------|---------------|---------------|
| BERT + ViT + FusConv1D + PL | 0.7723 | 0.7486 |
| BERT + ViT + FusConv1D | 0.7414 | 0.7179 |
| BERT + ViT + FusAttn | 0.7385 | 0.7140 |
| BERT + ResNet50 + FusConv1D | 0.7190 | 0.7047 |
| NSP-singleBERT + ViT + FusAttn | 0.7219 | 0.7103 |
| NSP-singleBERT + ResNet50 + FusAttn | 0.7180 | 0.6912 |
| Baseline | 0.5411 | 0.5309 |

4.3. Results and Comparison

Table 1 contains a detailed comparison of the proposed model with either of the 2 fusion block techniques and the benchmark provided by Mishra et al. [15]. We reported the weighted F1-score on train, val and test set data. From the table, we can see that the model with the Conv1D fusion block could outperform the rest of the models. The Conv1D layer captured inter-modal relations efficiently compared to the other fusion technique used. We also compare the model performances with and without using the pseudo label fine-tuning step and observed a significant boost when using the additional pseudo labelling step.

We also experimented with using a Next Sentence Prediction setting, where we passed the claim and document text data as a whole to the BERT model for extracting cross-document embeddings. These text embeddings would now represent the text data of claim and document as a single sequence, unlike in the proposed model, where claim and document text data are processed individually using the same BERT model. We then fused the unified text embeddings with the claim and document image embeddings using the self-attention fusion block. We included the performance of this experiment in Table 1 along with that of the proposed model.

We carried out experiments to compare the performance when using the ViT and ResNet [17] models for image embedding extraction. The improvement observed when using ViT shows that the transformer-based model extracted better embedding representations of the image compared to ResNet by using the self-attention mechanism, resulting in a significant boost in the overall model performance.

Overall the proposed model with the Conv1D fusion block and fine-tuned using the pseudo labelling step was the best performer with a weighted F1-score of 0.7486 on the test set, which helped us secure 3rd position in the factify task. Table 2 summaries the top 2 models performance over each class.

Table 2

Category-wise weighted F1-scores for the top-2 models and the baseline

| Category | BERT + ViT + FusConv1D + PL | BERT + ViT + FusConv1D | Baseline |
|-------------------------|-----------------------------|------------------------|----------|
| Support Text | 0.7765 | 0.7613 | 0.8267 |
| Support Multimodal | 0.8505 | 0.8986 | 0.7546 |
| Insufficient Text | 0.7942 | 0.7932 | 0.7442 |
| Insufficient Multimodal | 0.8448 | 0.7157 | 0.6967 |
| Refute | 0.9881 | 0.9711 | 0.4235 |
| Weighted F1-score | 0.7486 | 0.7179 | 0.5309 |

5. Conclusion

In this work, we propose a Transformers based fact verification model. The model makes use of inter-modal and cross-modal relations between image and textual data to identify whether a claim is supported by the provided document. Our proposed model outperformed the baseline model based on classic ML algorithms when verified on the test set of the Factify dataset. The proposed model extracts meaningful embeddings from both text and image data, the fusion of which generated better representations for the multi-modality data combining the knowledge. This effective fusion technique helped us secure 3rd position in Factify’s fact verification task.

Despite this performance, we can further boost the proposed model by using better pre-training methods and more advanced Transformer models which process longer text sequences. In future research, we would explore using the improved Transformer versions, which are more resource-intensive but can process longer text sequences, like Roberta [18], Big-Bird [19] and Longformer [20].

References

- [1] K. Nakamura, S. Levy, W. Y. Wang, r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection, arXiv preprint arXiv:1911.03854 (2019).
- [2] S. Garg, D. K. Sharma, New politifact: A dataset for counterfeit news, in: 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), IEEE, 2020, pp. 17–22.
- [3] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, Fever: a large-scale dataset for fact extraction and verification, arXiv preprint arXiv:1803.05355 (2018).
- [4] W. Y. Wang, ” liar, liar pants on fire”: A new benchmark dataset for fake news detection, arXiv preprint arXiv:1705.00648 (2017).
- [5] K. Shu, D. Mahudeswaran, H. Liu, Fakenewstracker: a tool for fake news collection, detection, and visualization, Computational and Mathematical Organization Theory 25 (2019) 60–71.
- [6] J. Z. Pan, S. Pavlova, C. Li, N. Li, Y. Li, J. Liu, Content based fake news detection using knowledge graphs, in: International semantic web conference, Springer, 2018, pp. 669–683.

- [7] R. Vijjali, P. Potluri, S. Kumar, S. Teki, Two stage transformer model for covid-19 fake news detection and fact checking, arXiv preprint arXiv:2011.13253 (2020).
- [8] Y. Nie, L. Bauer, M. Bansal, Simple compounded-label training for fact extraction and verification, in: Proceedings of the Third Workshop on Fact Extraction and VERification (FEVER), 2020, pp. 1–7.
- [9] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, J. Gao, Eann: Event adversarial neural networks for multi-modal fake news detection, in: Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining, 2018, pp. 849–857.
- [10] J. Ma, W. Gao, K.-F. Wong, Rumor detection on twitter with tree-structured recursive neural networks, Association for Computational Linguistics, 2018.
- [11] N. Vo, K. Lee, The rise of guardians: Fact-checking url recommendation to combat fake news, in: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018, pp. 275–284.
- [12] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.
- [14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, CoRR abs/2010.11929 (2020). URL: <https://arxiv.org/abs/2010.11929>. arXiv:2010.11929.
- [15] S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, P. Patwa, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Factify: A multi-modal fact verification dataset, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022.
- [16] P. Patwa, S. Mishra, S. Suryavardan, A. Bhaskar, P. Chopra, A. Reganti, A. Das, T. Chakraborty, A. Sheth, A. Ekbal, C. Ahuja, Benchmarking multi-modal entailment for fact verification, in: Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022.
- [17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. arXiv:1512.03385.
- [18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
- [19] M. Zaheer, G. Guruganesh, K. A. Dubey, J. Ainslie, C. Alberty, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, et al., Big bird: Transformers for longer sequences., in: NeurIPS, 2020.
- [20] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).