Multimodal Attention is all you need

Marco Saioni[1,*], Cristina Giannone[1,2]

[1]University G. Marconi, Rome, IT
[2]Almawave S.p.A., Via di Casal Boccone, 188-190 00137, Rome, IT

Abstract

In this paper, we present a multimodal model for classifying fake news. The main peculiarity of the proposed model is the *cross attention* mechanism. Cross-attention is an evolution of the attention mechanism that allows the model to examine intermodal relationships to better understand information from different modalities, enabling it to simultaneously focus on the relevant parts of the data extracted from each. We tested the model using textitMULTI-Fake-DetectiVE data from Evalita 2023. The presented model is particularly effective in both the tasks of classifying fake news and evaluating the intermodal relationship.

Keywords

Transformer, fake news classification, multimodal classification, cross attention

1. Introduction

Internet has facilitated communication by enabling rapid, immersive information exchanges. However, it is also increasingly used to convey falsehoods, so today, more than ever, the rapid spread of fake news can have severe consequences, from inciting hatred to influencing financial markets or the progress of political elections to endangering world security. For this reason, mitigating the growing spread of fake news on the web has become a significant challenge.

Fake news manifests itself on the internet through text, images, video, audio, or, in general, a combination of these modalities, which is a multimodal way. In this article, we took the two, text and image, components of news as it proposed, for instance, in a social network. In this work we proposed an approach to automatically and promptly identify fake news. We use the dataset *MULTI-Fake-DetectiVE*[1] competition, proposed in EVALITA 2023[2]. The competition aims to evaluate the truthfulness of news that combines text and images, an aim expressed through two tasks: the first, which carries out the identification of fake news (*Multimodal Fake News Detection*); the second, which seeks relationships between the two modalities text and image by observing the presence or absence of correlation or mutual implication (*Cross-modal relations in Fake and Real News*).

Our approach proposes a Transformer-based model that focuses on relating the textual and visual embeddings of the input samples (i.e., the vector representations of the text and images it receives as input).

The aim was to find a way to reconcile the two different representation embeddings because they are learned separately from two different corpora, such as text and images, trying to capture their mutual relationships through some interaction between the respective semantic spaces.

The remainder of the paper is structured as follows: section 2 presents a brief overview of related work, and section 3 describes the architecture of the proposed model. Section 4 discusses an overview of our experiments. Sections 5 and 6 present the final results and our conclusions, respectively.

2. Related Works

The Italian MULTI-Fake-DetectiVE competition [2] adds to the various datasets and challenges on multimodal fake news recently developed, for instance, Factify [3] and Fakeddint [4]. The creation of these competitions shows the interest in this task. The first task of the Italian challenge saw three completely different systems placed on the podium. While the first system POLITO[5] with a system based on the FND-CLIP multimodal architecture [6] proposing some ad hoc extensions of CLIP [7] including sentiment-based text encoding, image transformation in the frequency domain, and data augmentation via back-translation. The Extremita system [8], second classified, exploited the LLM capabilities, focusing only on the textual component of each news. They fine-tuned the open-source LMM Camoscio [9] with the textual part of the dataset. The impressive results show how the textual component plays a primary role in identifying fake news. Despite the significant contribution of the textual component to the task, more and more multimodal approaches are taking hold. In [10] proposed CNN architecture combining texts and images to classify fake news. In that direction, approaches such as CB-FAKE[11]

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy
*Corresponding author.
marco.saioni@gmail.com (M. Saioni); c.giannone@unimarconi.it (C. Giannone)

[1]https://sites.google.com/unipi.it/multi-fake-detective
[2]https://www.evalita.it[1]

incorporate the encoder representations from the BERT model to extract the textual features and comb them with a model to extract the image features. These features are combined to obtain a richer data representation that helps to determine whether the news is fake or real. Vision-language models, in general, have gained a lot of interest also in the last years, in the "large models era". Language Vision Models have been proposed during the previous year, with surprising results in many visual language interaction tasks [12],[13].

3. The proposed Model

The objective was to "engage" specialist models for natural language processing and artificial vision, making them discover and learn bimodal features from text and images collaboratively and harmoniously by applying the teachings of Vaswani et al. [14]: we decided to follow the path indicated by "Attention is all you need" Vaswani et al. very famous paper, following up on the intuition that the Attention mechanism could provide an important added value to the multimodal model of identification of fake news, becoming a Multimodal Attention (hence the title of this article), i.e. an attention mechanism applied between the two textual and visual modes of news. In fact, while Attention or Self Attention (as described in Vaswani et al. paper) takes as input the embeddings of a single modality and transforms them into more informative embeddings (contextualized embeddings), Multimodal Attention takes as input the embeddings of the two different modalities by combining them and then transforming them into a single embedding capable of capturing any existing relationships between the two input modes.

3.1. Architecture

Multimodal Attention is the heart that supports the proposed model, making it capable of exploring the hidden aspects of multimodal communication. As shown at a high level in Figure 1, the architecture of the proposed model consists of a hierarchical structure with three layers preceded by a pre-processing step. In order, there are: a pre-processing step, an input layer, a cross-modal layer and a fusion layer. It was decided to propose a network that models the consistent information between the two modalities textual and visual starting from State Of The Art pre-trained neural networks. In particular, we use a BERT [15] pre-trained model to learn the word embeddings by the textual component of news and a ResNet [16] pre-trained model to learn visual embeddings by the visual component. The two embeddings, belonging to two spaces with different dimensions, are first projected into a uniform, reduced-dimensional space, then related

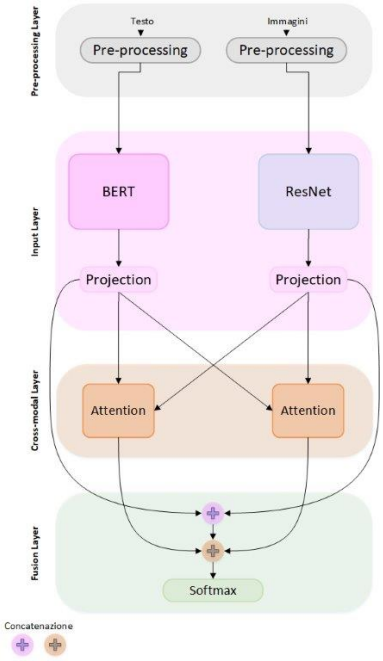

Figure 1: Proposed model architecture.

to each other with the strategy of mutual cross-attention to obtain two embeddings subsequently concatenated to provide the input of the last dense classification layer.

3.1.1. Pre-processing step

As a first step it is necessary to process the data made available by the organizers of the *MULTI-Fake-DetectiVE* competition to produce inputs that are compatible and compliant with those expected from the pre-trained models. The choices made for this preparation or for the pre-processing of the dataset and the data "personalization" strategy will then be described in the following three points:

- resolution/explosion of $1 : N$ relationships between text and images into N times $1 : 1$ relationships;
- *data augmentation* with the creation of an additional image to support the original one already present in each example;
- management of the textual component, truncated by BERT or rather by the relevant tokenizer to a fixed maximum length of tokens.

As decided for the visual and textual components, therefore following processing, for each single sample we move from the original pairs $< t, v^+ >$, where v^+ indicates the ratio $1 : N$ between text in natural language

and images in JPEG format, to the triples appropriately translated into numbers

$$< t_{trunc}, v, v_{aug} >$$

where t_{trunc} indicates, for each sample, a first-order tensor with 128 values (token), while v and v_{aug} denote third-order tensors with $(224 \times 224 \times 3)$ values (pixels). In fact, the first order tensor is the representation of the text in numerical form according to the default strategy of the BERT tokenizer, while the third order tensor is the representation of the images in numerical form according to the RGB coding for ResNet.

3.1.2. Input layer

This layer receives as input the previously processed dataset, i.e. the text and the images represented in numerical form, passing it to the pre-trained BERT and ResNet models to obtain the respective embeddings, subsequently projected into a space with small and common dimensions to make them comparable and to allow them to collaborate with each other in the subsequent cross-modal layers.

BERT Encoder Each sample pre-processed and represented in numerical form by the tokenizer is passed as input to the pre-trained BERT model which returns different output tensors for each of them. For the purposes of the classification task object of this study, we consider the pooled_output, a compact representation of all the token sequences given as input to the BERT model, obtained via the special token [CLS]. It is therefore a summary of the information extracted from the entire input dataset whose dimensions evidently depend on the number of hidden units of BERT. Since each text supplied as input to BERT will correspond to a tensor with 768 values real, using vector notation we have that:

$$\mathbf{e_t} = \text{BERT}(\mathbf{t_{trunc}})[pooled\_output]$$

where $\mathbf{e_t} \in \mathbb{R}^h$ is the word embeddings vector, $\mathbf{t_{trunc}} \in \mathbb{R}^{N_{max}}$ is the token input vector and $h = 768$ is the BERT hidden size. The equation shown refers to a single sample but can be extended to the entire batch of N examples processed by BERT. Indicating this batch with $\mathbf{T_{trunc}} \in \mathbb{R}^{N \times N_{max}}$, we will have:

$$\mathbf{E_t} = \text{BERT}(\mathbf{T_{trunc}})[pooled\_output]$$

where $\mathbf{E_t} \in \mathbb{R}^{N \times h}$ is the text embedding matrix learned by the BERT model.

ResNet Encoder The two images of each sample previously represented in numerical form are passed as input to the pre-trained ResNet model, which returns a visual embedding of size h_r for each example and which represents the features in a compact and semantic form extracted through convolutions and pooling within the ResNet network. In fact, to obtain visual embeddings from a pre-trained neural network like ResNet, we usually take the output of the penultimate layer, i.e. global pooling. In the proposed model, *ResNet50V2* was chosen which in global pooling reduces the spatial dimensions of the output tensor to 2048 values and therefore each input image will correspond in output to a vector with $h_r = 2048$ values, which represents the visual embeddings extracted from the network for that specific image. After obtaining the embeddings for each of the two images, they are concatenated together to obtain a single output tensor which will therefore have size $2 \times h_r = 4096$. Using the same formalism as the previous text encoder, we have:

$$\mathbf{e_v} = \text{ResNet}(\mathbf{v})[global\_pooling]$$

where $\mathbf{e_v} \in \mathbb{R}^{h_r}$ is the visual embedding vector and $\mathbf{v} \in \mathbb{R}^{L \times H \times C}$ the input third-order tensor. The indicated equation refers to a single sample but can be extended to the entire batch of N examples, therefore indicating the batch with $\mathbf{V} \in \mathbb{R}^{N \times L \times H \times C}$, we will have:

$$\mathbf{E_v} = \text{ResNet}(\mathbf{V})[global\_pooling]$$

where $\mathbf{E_v} \in \mathbb{R}^{N \times h_r}$ is the visual embedding matrix learned by the ResNet model. Similar discussion for the second image, for which it will be valid at batch level:

$$\mathbf{E_{v_{aug}}} = \text{ResNet}(\mathbf{V_{aug}})[global\_pooling]$$

where $\mathbf{E_{v_{aug}}} \in \mathbb{R}^{N \times h_r}$. By concatenating the two embeddings, we will obtain:

$$\mathbf{E_v} \oplus \mathbf{E_{v_{aug}}} = \mathbf{E_{concat(v, v_{aug})}} \in \mathbb{R}^{N \times 2h_r}.$$

From this moment and for simplicity of notation, $\mathbf{E_v}$ will refer to $\mathbf{E_{concat(v, v_{aug})}}$, knowing that this embedding is actually the concatenation of embeddings of an image and the one obtained through random transformations.

Projection The pre-trained models provide embeddings with different sizes. It is, therefore, necessary to transform them into a space with the same dimensionality to obtain comparable representations. The *projection* function carries out this task, introduced both to reduce the dimensions of the two embeddings and reduce the computational load, improving the performance of the multimodal model and allowing it to learn more complex patterns. The projection of embeddings is particularly useful in cases where you want to compare the semantic representations of two objects, ensuring that both are aligned in the same reduced semantic space, making

them comparable in terms of similarity or distance or facilitating the comparison and analysis of relationships. For this model, we selected $d_{prj} = 128$ as the projection size, reducing both embeddings sizes of the input components.

3.1.3. Cross-modal layer

This layer is the heart of the model, which is developed taking inspiration from the behavior of human beings when faced with news made up of text and images. Intuitively, we try to read in the image what is written in the text and to represent in the text what is shown by the image. It can be said that cross-modal attention relations exist between image and text. This is why, to simulate the human process described in a neural model, we relied on the cross attention between the two modalities, a variant of the standard component of *multi-head attention* capable of capturing global dependencies between text and images.

In the proposed model, two blocks of crossed attention are activated in the two text-image and image-text perspectives. In the first case, we consider the textual embeddings as queries for the *multi-head attention* block, while the visual ones as key and value. This should allow the characteristics of the text to guide the model to focus on regions of the image semantically coherent with the text: in fact, if the textual embeddings are considered as queries and the visual ones as key and value, then the attention will be applied to the images in based on compatibility with the text, which is therefore considered the context on which to evaluate the relevance of an image. In this way, attention is focused on the images with respect to how relevant they are to the text, i.e. we try to give importance to the visual features in relation to the context provided by the text. Conversely, in the second case the visual embeddings are the queries, while the keys and values are the textual embeddings, and this should allow the visual features to make the model pay attention to those parts of text consistent with the images. That is, the same thing as in the previous case applies, but the roles between text and image are reversed.

Wanting to formalize the bidirectional cross-attention between the embeddings of the text $\mathbf{E_{t-projected}}$ and those of the images $\mathbf{E_{v-projected}}$, we can write:

$$\mathbf{E_{cross-tv}} = \text{Attention}(\mathbf{E_{t-projected}}, \mathbf{E_{v-projected}})$$

$$\mathbf{E_{cross-vt}} = \text{Attention}(\mathbf{E_{v-projected}}, \mathbf{E_{t-projected}})$$

where $\mathbf{E_{cross-tv}}$ represents the attention embeddings of image information with respect to the text and $\mathbf{E_{cross-vt}}$ represents attention embeddings of text information compared to images.

In this layer the dimensions of the embeddings are not modified in any way, therefore we remain in $\mathbb{R}^{N\times 128}$.

3.1.4. Fusion layer

Once you have available the embeddings (textual and visual) learned unimodally in the network, and the cross-attention embeddings learned intermodally, it is necessary to implement a fusion strategy that can best balance their respective contributions in the multimodal classification task. Although the architecture of the model would seem to suggest the implementation of the *late fusion* strategy, it is necessary to observe how the cross-attention of the *cross-modal layer* is already a fusion strategy adopted in the network during learning before the one explicitly implemented in the next *fusion layer*: this allowed the model to learn shared features during training while maintaining the suitable flexibility between the multimodal components, i.e. without excessively influencing the learning process of each modality separately.

The concatenation preserves each modality's distinctive features, allowing the model to exploit them during learning, unlike the sum which could lead to the loss of information due to values that can cancel each other out, taking away the model's descriptive capacity. For these reasons, the fusion occurs taking into consideration all four embeddings learned by the model $\mathbf{E_{t-projected}}$, $\mathbf{E_{v-projected}}$, $\mathbf{E_{cross-tv}}$, $\mathbf{E_{cross-vt}}$, where the first two provide distinctive unimodal features, while the other two provide correlated and mutually "attentioned" cross-modal features. The hybrid fusion strategy then completes the recipe, providing that pinch of flexibility necessary to give balance to the multimodal classifier. Formally we have the following equation, which aims to make the most of both the information provided by the individual modalities as such, and that provided jointly:

$$\mathbf{E_{global}} = (\mathbf{E_{t-projected}} \oplus \mathbf{E_{v-projected}}) \oplus$$

$$\mathbf{E_{cross-tv}} \oplus \mathbf{E_{cross-vt}}$$

where $\mathbf{E_{global}}\ in \mathbb{R}^{N\times 4d_{prj}}$, where N is the size of the batch of examples given as input to the network and $d_{prj} = 128$.

The final output of the multimodal model is obtained by applying a densely connected layer with $C = 4$ units and a softmax activation function that returns the probabilities of the four classes. Formally:

$$\mathbf{Y} = (\mathbf{E_{global}}\mathbf{W} + \mathbf{b})$$

$$\mathbf{O} = \text{softmax}(\mathbf{Y})$$

with $\mathbf{W} \in \mathbb{R}^{4d_{prj}\times C}$, $\mathbf{b} \in \mathbb{R}^{1\times C}$ and therefore $\mathbf{O} \in \mathbb{R}^{N\times C}$ is a matrix in which each row is a vector with $C = 4$ values representing the conditional (estimated) probability of each class for the relevant sample.

4. Experimental Setup

4.1. Split dataset into *training* and *validation*

To guarantee that the proportions relating to the classes and sources are maintained uniformly in the two sets, the 1034 samples of the dataset are randomly divided following the 80%-20% proportion between training and validation in a stratified way both with respect to the labels, as also happens in the baseline model of the competition *MULTI-Fake-DetectiVE* and, with respect to the type of source of the news.

4.2. Training and validationn

For our experiment, the model was trained up to 80 epochs with early stopping on using the *focal loss* [17] function. It is a dynamically scaled loss *cross entropy* function, where the scaling factor decays to zero as confidence in the correct class increases. Intuitively, this scaling factor can automatically scale the contribution of easy examples during training and quickly focus the model on difficult examples. For the optimizer we chosed *AdamW*, given that the models used to analyze text and images were originally pre-trained using this algorithm, which applies weight regularization directly to the model parameters during weight updating, helping to improve the stability and generalization of the model.

5. Results

5.1. Official *baseline* models

In the notebook provided by the *MULTI-Fake-DetectiVE* organizers there is an evaluation strategy on the official dataset which is developed by comparing the performance of the unimodal pre-trained models with a multimodal model:

- *Text-only model*: model trained only on textual features, extracted with a pre-trained BERT network.
- *Image-only model*: model trained only on the visual features of images, extracted with a pre-trained ResNet18 network.
- *Multi-modal model*: model trained on the concatenation of text and image features, extracted separately with the previous two *only-model*.

The F1-weighted score values of the three baseline models are shown in Table 1. The textual model is therefore the most effective among the three baseline models in classifying fake news and the visual one has lower performance than the textual model. The multimodal model obtained an F1-weighted score lower than that obtained by the unimodal textual model, but higher than the score of the unimodal visual model, indicating that the integration of visual and textual information led to an improvement in performance compared to the model visual, but not enough to outperform the text model. This suggests that there may be potential to perform additional optimizations or modality integration strategies to achieve better performance from the multimodal model.

Model	Accuracy	F1-weighted
Text-only	0.498	**0.462**
Multi-modal	0.480	0.442
Image-only	0.438	0.371

Table 1
Summary and comparison of the main metrics for the three baseline models on the official dataset.

5.2. Proposed model

To evaluate the model proposed on the *Multimodal Fake News Detection* task, we chose to follow the approach used by the organizers in the notebook of the baseline models, i.e. we performed an ablation study on the proposed model: first a unimodal textual model was trained, then a unimodal visual one, then a multimodal one without *cross-bi-attention*, finally a multimodal one with *cross-bi-attention*. Table 2 reports the respective accuracy and F1-weighted values.

Model	Accuracy	F1-weighted
Proposed Multi-modal $\otimes$	0.541	**0.537**
Proposed Text-only	0.472	0.469
Proposed Multi-modal $\oplus$	0.460	0.445
Proposed Image-only	0.418	0.422

Table 2
Ablation study on the proposed model: accuracy and F1-weighted. The $\otimes$ symbol indicates *cross-bi-attention* enabled, while $\oplus$ indicates *cross-bi-attention* disabled (i.e. concatenation enabled).

The results for the unimodal and multimodal models without *cross-bi-attention* are in perfect harmony with those of the similar baseline models.
But the data that catches the eye is that of the accuracy and F1-weighted values of the multimodal model with *cross-bi-attention*. In particular, its F1-weighted score is almost seven percentage points higher than the proposed textual unimodal model, more than eleven compared to the visual unimodal model and more than nine compared to the multimodal one without *cross-bi-attention*.

Let's see the accuracy and F1-weighted values of the multimodal model proposed with *cross-bi-attention* against finalist models. Its F1-weighted score is two and a half points higher than that of the winning model of

the *MULTI-Fake-DetectiVE* competition, as evident from the Table 3. As supposed and hoped, the mechanism

Model	Accuracy	F1-weighted
Proposed Multi-modal	0.541	**0.537**
PoliTo - FND-CLIP-ITA	-	0.512
ExtremITA - Suede_LoRA	-	0.507
Baseline Multi-modal	0.480	0.442

Table 3
Final comparison between all the analyzed models and the proposed model.

of crossed attention seen from the two text-image and image-text perspectives enriched by the skip connection provided by the simple concatenation of the two different embeddings, provides the model with that extra edge that allows it to dig background in the relationships between textual and visual features. By combining bilateral cross-attention and residual connection, tasks of the *cross-modal layer* and the *fusion layer* respectively, significant semantic and semiotic interrelations are obtained in favor of the performance of the classifier which becomes more precise and sensitive.

In fact, if on the one hand the *cross-modal layer* allows the model to learn multimodal semantics between text and images, the *fusion layer* enhances it by improving its stability, capacity and performance thanks to the skip connection which provides the gradient with a useful direct path during backpropagation to flow without tending to zero, bringing significant and additional information into each layer of the network.

All the results described up to this point are obtained by measuring the model on the *Multimodal Fake News Detection* task of the competition covered by this work. As mentioned, the organizers also proposed a second task *Cross-modal relations in Fake and Real News*, aimed at verifying the robustness of the model to changing tasks without any human intervention. Table 4 shows the accuracy and F1-weighted values for the proposed model called to express itself on the *Cross-modal relations* task, together with the baseline and winner models of the *MULTI- competition Fake-DetectiVE*. The results show

Model	Accuracy	F1-weighted
Proposed Multi-modal	0.529	**0.527**
PoliTo - FND-CLIP-ITA	-	0.517
Baseline Multi-modal	-	0.442

Table 4
Result summary on Task 2.

a clear improvement in performance in solving the task even compared to the winning model of the competition. This is a very important result, because it demonstrates the network's ability to adapt to changes in tasks and changes in training data, which is not at all a given.

The data preparation strategy in the *Pre-processing step* provides the model with more information to learn from, the real strength can be identified in the *Cross-modal Layer*.

6. Conclusions

The Internet has facilitated the multimodality of communication by enabling rapid information exchanges that are increasingly immersive but increasingly used to convey falsehoods. In this study, a multimodal model for identifying fake news was proposed which is based on the mechanism of cross attention between the representations of the features learned by the network on the textual component of the news and those learned on the visual component associated with it.

Many multimodal models are based on the concatenation of features learned from distinct modalities which, despite having good performance, however, limit the potential of the interaction between the features themselves.

From the experiments carried out, the use of cross-attention demonstrated significant improvements in the performance of the model proposed in this work compared to the first two models classified in the *MULTI-Fake-DetectiVE* competition for both tasks requested by the organizers, despite the dataset available for training is very small in size and unbalanced both with respect to the categories to be predicted and with respect to the source of the news. Despite the intrinsic complexity of the two tasks, the cross-layer of the proposed model manages to express the representations learned from the text and images of a news story in a harmonious, collaborative and synergistic way, balancing their contributions and preventing one from taking over the other.

Future developments concern the components of the model which could use a Visual Transformer [18] instead of the ResNet in order to relate textual embeddings and visuals both generated by training a Transformer network.

References

[1] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3473.

[2] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. C. Passaro, M. Sabbatini, Multi-fake-detective at evalita 2023: Overview of the multimodal fake news

detection and verification task, CEUR WORKSHOP PROCEEDINGS 3473 (2023). URL: https://ceur-ws.org/Vol-3473/paper32.pdf.

[3] S. Suryavardan, S. Mishra, P. Patwa, M. Chakraborty, A. Rani, A. N. Reganti, A. Chadha, A. Das, A. P. Sheth, M. Chinnakotla, A. Ekbal, S. Kumar, Factify 2: A multimodal fake news and satire news dataset., in: A. Das, A. P. Sheth, A. Ekbal (Eds.), DE-FACTIFY@AAAI, volume 3555 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: http://dblp.uni-trier.de/db/conf/defactify/defactify2023.html#SuryavardanMPCR23.

[4] K. Nakamura, S. Levy, W. Y. Wang, Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 6149–6157. URL: https://aclanthology.org/2020.lrec-1.755.

[5] L. D'Amico, D. Napolitano, L. Vaiani, L. Cagliero, Polito at multi-fake-detective: Improving FND-CLIP for multimodal italian fake news detection, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3473/paper35.pdf.

[6] Y. Zhou, Q. Ying, Z. Qian, S. Li, X. Zhang, Multimodal fake news detection via clip-guided learning, 2022. URL: https://arxiv.org/abs/2205.14304. arXiv:2205.14304.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. arXiv:2103.00020.

[8] C. D. Hromei, D. Croce, V. Basile, R. Basili, Extremita at EVALITA 2023: Multi-task sustainable scaling to large language models at its extreme, in: M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi (Eds.), Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023, volume 3473 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023. URL: https://ceur-ws.org/Vol-3473/paper13.pdf.

[9] A. Santilli, E. Rodolà, Camoscio: an italian instruction-tuned llama, 2023. URL: https://arxiv.org/abs/2307.16456. arXiv:2307.16456.

[10] I. Segura-Bedmar, S. Alonso-Bartolome, Multimodal fake news detection, Information 13 (2022). URL: https://www.mdpi.com/2078-2489/13/6/284.

[11] B. Palani, S. Elango, V. K, Cb-fake: A multimodal deep learning framework for automatic fake news detection using capsule neural network and bert, Multimedia Tools and Applications 81 (2022). doi:10.1007/s11042-021-11782-3.

[12] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, J. Xu, B. Xu, J. Li, Y. Dong, M. Ding, J. Tang, Cogvlm: Visual expert for pretrained language models, 2024. URL: https://arxiv.org/abs/2311.03079. arXiv:2311.03079.

[13] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, 2024. URL: https://arxiv.org/abs/2310.03744. arXiv:2310.03744.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2017. arXiv:1706.03762.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[17] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, 2018. arXiv:1708.02002.

[18] 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, 2021. arXiv:2010.11929.