11-2023

MATK: The Meme Analytical Tool Kit

Ming Shan HEE

Aditi KUMARESAN

Nguyen Khoi HOANG

Nirmalendu PRAKASH

Rui CAO
Singapore Management University, ruicao.2020@phdcs.smu.edu.sg

See next page for additional authors

Citation

Author

Ming Shan HEE, Aditi KUMARESAN, Nguyen Khoi HOANG, Nirmalendu PRAKASH, Rui CAO, and Roy Ka-Wei LEE

MATK: The Meme Analytical Tool Kit

Ming Shan Hee
Singapore University of
Technology and Design
Singapore, Singapore
mingshan_hee@mymail.sutd.edu.sg

Aditi Kumaresan
Singapore University of
Technology and Design
Singapore, Singapore
aditi_kumaresan@mymail.sutd.edu.sg

Nguyen Khoi Hoang
VinUniversity
Hanoi, Vietnam
20nguyen.hk@vinuni.edu.vn

Nirmalendu Prakash
Singapore University of
Technology and Design
Singapore, Singapore
nirmalendu_prakash@sutd.edu.sg

Rui Cao
Singapore Management University
Singapore, Singapore
ruicao.2020@phdcs.smu.edu.sg

Roy Ka-Wei Lee
Singapore University of
Technology and Design
Singapore, Singapore
roy_lee@sutd.edu.sg

ABSTRACT

The rise of social media platforms has brought about a new digital culture called memes. Memes, which combine visuals and text, can strongly influence public opinions on social and cultural issues. As a result, people have become interested in categorizing memes, leading to the development of various datasets and multimodal models that show promising results in this field. However, there is currently a lack of a single library that allows for the reproduction, evaluation, and comparison of these models using fair benchmarks and settings. To fill this gap, we introduce the Meme Analytical Tool Kit (MATK), an open-source toolkit specifically designed to support existing memes datasets and cutting-edge multimodal models. MATK aims to assist researchers and engineers in training and reproducing these multimodal models for meme classification tasks, while also providing analysis techniques to gain insights into their strengths and weaknesses. To access MATK, please visit https://github.com/Social-AI-Studio/MATK.

CCS CONCEPTS

• **Computing methodologies → Natural language processing**; **Computer vision representations**.

KEYWORDS

meme, visual-language models, multimodal analysis

1 INTRODUCTION

Memes have become a crucial means of communication and cultural expression in recent years, influencing our digital interactions and content consumption. These images or videos, often accompanied by clever captions, spread rapidly across the internet, serving as a medium for conveying complex ideas, humor, and social commentary. As a result, this popular digital phenomenon has attracted attention from both industry and academia, leading to the analysis of its underlying topics and impact. Notably, Facebook launched "The Hateful Meme Challenge," releasing a dataset of over 10,000 memes annotated for hateful and non-hateful attributes [4]. Subsequently, further studies have emerged, exploring various applications of meme analysis, such as offensive content detection [26], sentiment analysis [22], and identification of victims and roles [3, 15].

The availability of diverse and extensive datasets has significantly contributed to the rapid development and analysis of multimodal models for memes. These datasets serve as the foundation for training and evaluating models, enabling researchers to create more accurate and robust solutions for classifying memes across different domains. Researchers have explored different approaches, including classic two-stream models that combine text and visual features for classifying hateful memes [4, 26], as well as fine-tuning large-scale pre-trained multimodal models for multimodal classification tasks [1, 13, 16, 21, 23, 29]. However, a challenge arises when models and datasets are scattered across separate GitHub repositories. With the increasing number of meme datasets and multimodal models, it becomes progressively difficult to locate and reproduce specific models on unseen datasets. This dispersion hampers research progress as substantial effort and time is required to find and adapt custom multimodal models for meme classification tasks. Therefore, establishing a centralized repository that facilitates the organization, discovery, and sharing of models and datasets can greatly enhance efficiency and foster collaborative advancements in the field of meme classification and analysis.

To address this issue, we have developed the Meme Analysis Tool Kit (MATK), an open-source tool kit explicitly created to support the expanding collection of meme datasets and cutting-edge models for meme classification tasks. MATK serves as a valuable resource for machine learning practitioners and socio-linguistic researchers, simplifying the training process of state-of-the-art models. Additionally, it incorporates post-hoc model interpretation techniques
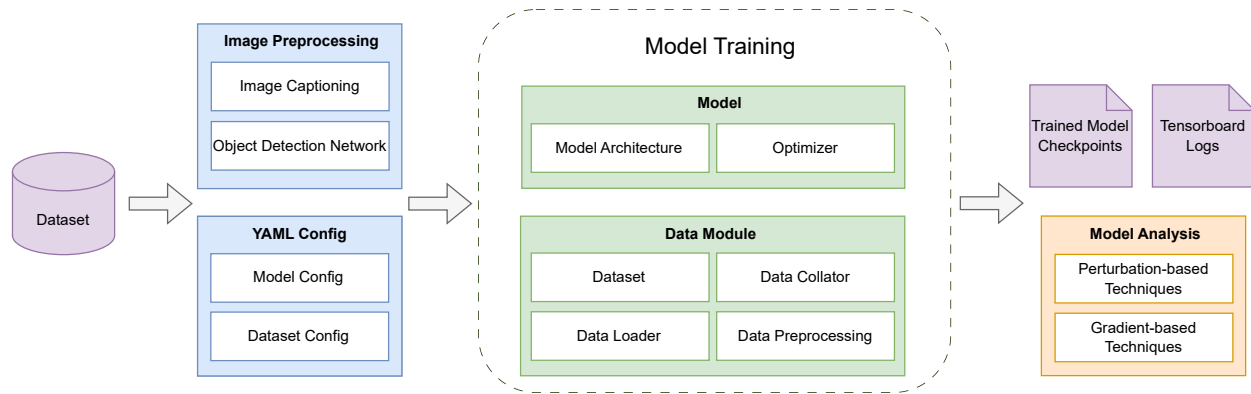
Figure 1: Overall framework of MATK repository

such as LIME [20] and SHAP [9], facilitating the analysis of model strengths and weaknesses. Notably, MATK encourages researchers to contribute new techniques and datasets, promoting reproducibility and fostering collaboration. The contributions of MATK can be summarized as follows:

- MATK is a centralized repository that organizes, discovers, and shares meme datasets and multimodal models. It currently supports the training and analysis of seven models and eight meme datasets, with future additions planned.
- MATK is designed as a user-friendly tool for researchers and practitioners, especially those new to the field, to efficiently train state-of-the-art models on multiple meme datasets.
- MATK encourages collaborations and contributions by offering a straightforward training framework with modular components and a flexible configuration system.

2 MATK: MEME ANALYSIS TOOL KIT

In this section, we present a comprehensive overview of MATK. MATK makes use of the PyTorch Lightning framework [1] to simplify the training and development process for multimodal memes. It adopts an object-oriented programming approach, utilizing composition and inheritance among classes to avoid redundant implementation. Furthermore, MATK incorporates a flexible configuration system, enabling researchers to seamlessly switch between datasets and model modules. The overall framework architecture of MATK is depicted in Figure 1.

2.1 Design Principles

MATK embodies user-friendliness and simplicity through the following design principles:

Modularity: Leveraging PyTorch Lightning, MATK allows for effortless integration of components using LightningModel and LightningDataModule. Independent modules can be developed for different tasks and input modalities, enabling seamless interfacing between new LightningDataModules and LightningModels.

Composable Configurations: MATK provides a user-friendly configuration system based on YAML files. These YAML files offer flexibility in specifying components and training parameters. For

example, researchers can define the training parameters under the "trainer" configuration, swap the LightningModel using the "model" parameter, or swap the meme dataset using "data" parameter.

Reproducibility: MATK ensures reproducible implementations of multimodal models for diverse meme classification tasks. By specifying a random seed using "seed_everything" in the YAML file, researchers and practitioners can easily reproduce results.

2.2 Image Preprocessing

MATK provides three sets of useful tools designed to extract meaningful visual features from memes, specifically focusing on eliminating text overlays, converting images into embedding features, and translating images into text-based features.

Image Cleaning: MATK offers image cleaning tools to enhance the extraction of visual features from memes. These tools effectively remove superimposed text from images without compromising the visual elements. This is achieved by using MMedit [11] and KerasOCR [2] to perform inpainting on the removed text areas.

Embedding Features: Vision-language models (VLMs) often utilize object detection networks to extract prominent visual features from images. However, running the object detection network concurrently with the VLMs can slow down model training and increase GPU RAM usage. Hence, it is beneficial to preprocess images into embedding features before running the model. MATK provides helpful scripts for converting image modalities into embedding features. These scripts include visual feature extraction scripts that employ Faster-RCNN [19] and CLIP [17] models.

Text-Based Features: Researchers have achieved success in multimodal meme classification tasks by leveraging text-based models. One common approach to extracting and representing the semantics of an image as textual descriptions is through image captioning techniques. MATK includes image captioning scripts that utilize ClipCap [12] and BLIP-2 [7] models, simplifying the process of generating image captions for text-based feature extraction.

2.3 Data Modules

The MATK repository incorporates various meme datasets, each encapsulated within a LightningDataModule class. This class serves

---

[1]https://www.pytorchlightning.ai/index.html

[2]https://keras-ocr.readthedocs.io/en/latest/

the primary purpose of facilitating dataset loading, preprocessing, batch size definition, and data collation functions. The supported datasets are listed below:

- **Facebook's Hateful Memes (FHM)** [4]. This dataset consists of 10,000 memes, some of which contain discriminatory content towards protected categories/classes. Each meme is labeled as hateful or non-hateful.
- **Facebook's Fine-Grained Hateful Memes (FHM-FG)** [10]. An extension of Facebook's Hateful Memes. This dataset comprises 10,000 memes annotated with labels indicating whether they are hateful or non-hateful, along with the attack type and the targeted protected category.
- **Harmful Memes (HarMeme)** [22]. This dataset contains 3,544 memes related to COVID-19, annotated with the intensity of harmfulness and the target class.
- **Multimedia Automatic Misogyny Identification** [2]. This dataset consists of 10,000 memes focused on misogyny. The memes are annotated with multiple aspects of common hatred towards women, including misogyny, shaming, stereotypes, objectification, and violence.
- **Memotion** [22]. This dataset contains 9,871 memes annotated with sentiment, type of emotion and the corresponding emotional intensity.

2.4 Models

MATK includes a collection of state-of-the-art models specifically designed for meme classification. These models are encapsulated within a LightningModel class, which encompasses the model's architecture and associated functions for training, validation, and testing. The following models are supported:

- **BART** [6]. A transformer encoder-decoder model that utilizes a denoising autoencoder during pre-training. It learns to reconstruct the original text by corrupting it with a customizable noising function.
- **T5** [18]. A transformer encoder-decoder model trained on a diverse range of text-to-text tasks, both supervised and unsupervised. Each supervised classification tasks are transformed into a language problems for training.
- **PromptHate** [1]. An encoder model that effectively utilizes pre-trained language models by employing prompts and leveraging implicit knowledge for classifying hateful memes.
- **VisualBERT** [8]. A unified single-stream vision-language model that processes both textual and visual inputs within the same module.
- **LXMERT** [27]. A two-stream vision-language model that separately processes text and images using distinct modules.
- **FLAVA** [24]. A foundational vision-language model pre-trained on multimodal data, including image-text pairs and unimodal data with unpaired images and text.

2.5 Analysis Techniques

MATK incorporates analysis techniques that allow for the assessment of the model's strengths and weaknesses. These techniques provide practitioners and researchers with valuable insights into the model's performance on unseen datasets. Currently, MATK supports the following techniques:

Table 1: Reproducibility Experiments on FHM. A bolded score indicates successful reproduction of a similar model performance, while an <u>underlined score</u> denotes a considerable discrepancy in model performance.

Model	MATK		Reported Results	
	Acc.	AUROC	Acc.	AUROC
BART	50.60	46.60	-	-
T5	55.60	59.60	-	-
PromptHate	**72.45**	**81.24**	**72.98**	**81.45**
VisualBERT	61.00	<u>67.50</u>	-	<u>74.10</u>
LXMERT	61.20	66.10	-	-
FLAVA	69.40	**77.40**	-	**76.70**

- **LIME** [20]. This perturbation-based technique approximates any black box machine learning model by creating a local and interpretable model to explain individual predictions.
- **Integrated Gradients** [25]. This gradient-based technique evaluates feature importance by calculating the average of the model's output gradient interpolated along a straight-line path in the input data space.

3 EMPIRICAL RESULTS

In this section, we showcase the reproducibility and utilization of our library. We evaluated our models using the Facebook's Fine-Grained Hateful Memes dataset and performed LIME analysis on a fine-tuned FLAVA model.

3.1 Reproducibility Experiments

Table 1 shows the performance of implemented models in MATK. MATK reproduced the results of FLAVA and PromptHate models almost perfectly, with a difference of (+/- 1). However, the variation in performance of the VisualBERT model can be attributed to differences in implementation and the pre-trained model checkpoint used. MATK currently uses VisualBERT from Huggingface [3], while the reported results are based on VisualBERT from Facebook's MMF [4]. This highlights the importance of a unified repository, as it ensures consistency for practitioners and researchers who might make similar assumptions. Consequently, we intend to support the model variants from Facebook's MMF within MATK in the future.

3.2 LIME Analysis

Figure 2 shows LIME analysis of the FLAVA model's explanation for a meme incorrectly predicted as hateful. Our observations indicate that the image's objects (Arab and riffle) and strong language ("radicalization" and "types") significantly influence the model's prediction. While this could suggest the presence of a potential bias in the trained model, we also note that the model demonstrates an understanding of negation through the negative contribution of the term "different." Consequently, a comprehensive and thorough analysis is warranted to investigate these findings further. Nonetheless, the utilization of LIME analysis provides researchers with an initial yet valuable tool for assessing model performance.
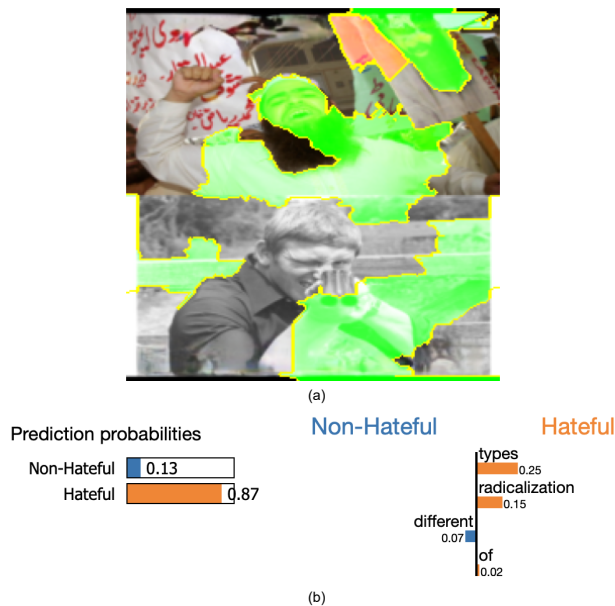
---

[3]https://huggingface.co/
[4]https://github.com/facebookresearch/mmf

Figure 2: LIME analysis of meme image.

4 CONCLUSION

We present the Meme Analytic Tool Kit (MATK), a comprehensive tool designed to facilitate meme classification tasks by incorporating advanced multimodal models and supporting various meme datasets. MATK focuses on modularity, flexibility, and user-friendliness, making model training and analysis across diverse domains effortless. Importantly, MATK integrates model analysis techniques that offer practitioners and researchers valuable insights into the strengths and weaknesses of their models. Our future plans for MATK include the following: (i) expanding dataset and model support to include the TotalDefMeme dataset [14], MET-Meme dataset [28], and the DisMultiHate model [5]; (ii) enabling multi-task training with different datasets, such as using FHM for hateful classification and HarMeme for harmful classification; (iii) enhancing the documentation to provide comprehensive and detailed information.

5 ACKNOWLEDGMENTS

REFERENCES

[1] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2023. Prompting for multimodal hateful meme classification. *arXiv preprint arXiv:2302.04156* (2023).
[2] Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 Task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022).* 533–549.
[3] Shaik Fharook, Syed Sufyan Ahmed, Gurram Rithika, Sumith Sai Budde, Sunil Saumya, and Shankar Biradar. 2022. Are you a hero or a villain? A semantic role labelling approach for detecting harmful memes.. In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations.* 19–23.
[4] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems* 33 (2020), 2611–2624.
[5] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM International*

[6] *Conference on Multimedia.* 5138–5147.
[7] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).
[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).
[8] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).
[9] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
[10] Lambert Mathias, Shaoliang Nie, Aida Mostafazadeh Davani, Douwe Kiela, Vinodkumar Prabhakaran, Bertie Vidgen, and Zeerak Waseem. 2021. Findings of the WOAH 5 shared task on fine grained hateful memes detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021).* 201–206.
[11] MMEditing Contributors. 2022. MMEditing: OpenMMLab Image and Video Editing Toolbox. https://github.com/open-mmlab/mmediting.
[12] Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734* (2021).
[13] Niklas Muennighoff. 2020. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788* (2020).
[14] Nirmalendu Prakash, Ming Shan Hee, and Roy Ka-Wei Lee. 2023. TotalDefMeme: A Multi-Attribute Meme dataset on Total Defence in Singapore. In *Proceedings of the 14th ACM Multimedia Systems Conference (MMSys '23).* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3587819.3592545
[15] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. *arXiv preprint arXiv:2110.00413* (2021).
[16] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184* (2021).
[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning.* PMLR, 8748–8763.
[18] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.
[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).
[20] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. [n.d.]. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD.*
[21] Vlad Sandulescu. 2020. Detecting hateful memes using a multimodal deep ensemble. *arXiv preprint arXiv:2012.13235* (2020).
[22] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020. SemEval-2020 Task 8: Memotion Analysis–The Visuo-Lingual Metaphor! *arXiv preprint arXiv:2008.03781* (2020).
[23] Shivam Sharma, Firoj Alam, Md Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. *arXiv preprint arXiv:2205.04274* (2022).
[24] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *CVPR.* 15638–15650.
[25] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning.* PMLR, 3319–3328.
[26] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying.* 32–41.
[27] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).
[28] Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. MET-Meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 2887–2899.
[29] Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. 2022. Multimodal Hate Speech Detection via Cross-Domain Knowledge Transfer. In *Proceedings of the 30th ACM International Conference on Multimedia.* 4505–4514.