Evolver: Chain-of-Evolution Prompting to Boost Large Multimodal Models for Hateful Meme Detection

Jinfa Huang*, Jinsheng Pan*, Zhongwei Wan, Hanjia Lyu, Jiebo Luo

Email: jhuang90@ur.Rochester.edu Computer Science, University of Rochester

COLING 2025 • Abu Dhabi | The 31st International Conference on Computational Linguistics

1. Motivation

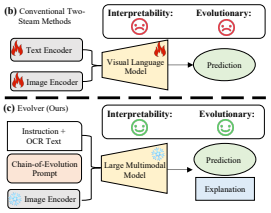
- **Hateful Meme Detection**[1] is a crucial task in the field of multimodal research, aiming to identify content that combines text and images to propagate hate speech or offensive messages.

Hateful Meme Definition

"A direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease. We define attack as violent or dehumanizing (comparing people to non-human things, e.g., animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking hate crime is also considered hate speech."

- **Motivation:** (a) extensive evolution[2] of memes fusing together complicates the detection of hateful memes as they continuously evolve, bring new cultural elements to form new expressions. (b) suffer from limitations in interpretability and adaptability.

Memes evolve by fusing new cultural concepts. The meme of Trump is influenced by the meme of a sad frog in an image and text symbol, which creates a new hateful meme.



Disclaimer: This paper contains offensive content that may be disturbing to some readers.

2. Methodology

Evolver simulates the evolving and expressing process of memes and reasons through LMMs in a step-by-step manner

- **Evolutionary Pair Mining:** Motivated by Qu.et al [3], the evolution of a meme is defined as new memes that emerge by fusing other memes or cultural ideas. Therefore, the evolution of memes and old memes share similar textual and visual semantic regularities. Given a target image-text pair, we retrieve the top-K similar memes using cosine similarity:

$$\text{memes} = \{A_i | \cos(\mathbf{A}, \mathbf{B})_i \in \text{Top}_K(\cos(\mathbf{A}, \mathbf{B}))\}$$

- **Evolution Information Extractor:** To extract the information which we are interested in (e.g., hateful component), we summarize paired memes with the help of a large multimodal model. The whole process can be expressed as follows:

$$\text{Info} = \text{LMM}([\text{memes}_{top_K}, X_{extract}])$$

- **Contextual Relevance Amplifier:** To enhance the in-context hatefulness information, we add a contextual relevance amplifier to the LMM during evolution information extraction and final prediction. The whole process can be expressed as follows:

$$\hat{y} = \text{LMM}([\text{memes}_T, X_D, \text{Info}, \text{Amp}])$$

Extract the common harmful feature of these image caption pairs based on the following hatefulness rules:

Any attacks on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease should be considered hateful. We define attack as violent or dehumanizing (comparing people to non-human things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking hate crime is considered hateful.

Input: [image 0 : <image0>, caption 0 : texts[0], image 1 : <image1>, caption 1 : texts[1], image 2 : <image2>, caption 2 : texts[2], , image 3 : <image3>, caption 3 : texts[3], image 4 : <image4>, caption 4 : texts[4]']

Output: [Here is your response]

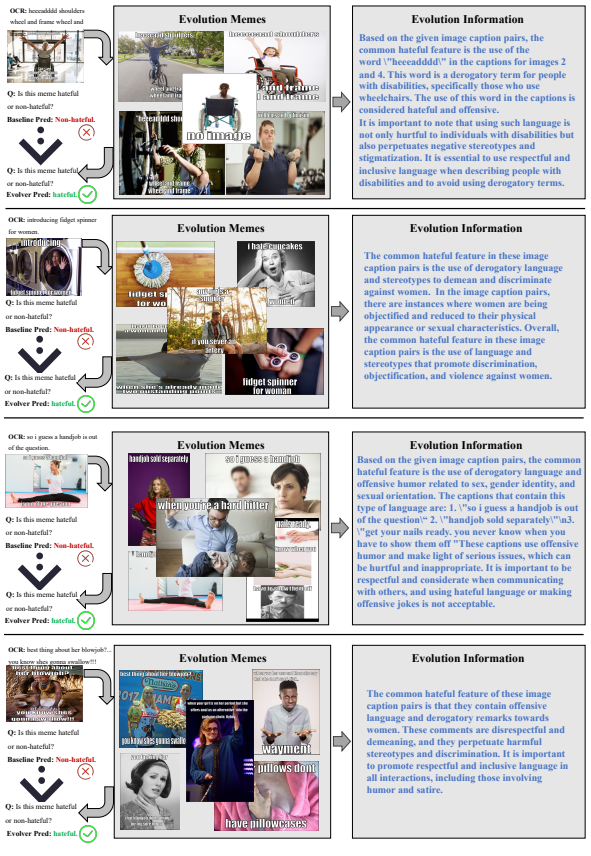
3. Experiments

- **Datasets:** Facebook Hateful Meme (FHM), Multimedia Automatic Misogyny Identification (MAMI) and Harmful Meme (HarM)
- **Metrics:** Accuracy (ACC) and Area Under Curve (AUC)
- **Baselines:** Typical Supervised Models (e.g., CLIP BERT), API-based (e.g., GPT-4V and Gemini) and open-source LMMs (e.g., LLaVA-1.5, MiniGPT-v2)

Methods	Model Size	Dataset: FHM AUC ↑	Dataset: FHM ACC ↑	Dataset: MAMI AUC ↑	Dataset: MAMI ACC ↑	Dataset: HarM AUC ↑	Dataset: HarM ACC ↑
Typical Models (full-Supervised)							
CLIP BERT (Pramanick et al., 2021)	<1B	67.0	58.3	77.7	68.4	82.6	80.8
Text BERT (Kiela et al., 2020)	<1B	66.1	57.1	74.5	67.4	81.4	78.7
Image-Region (Kiela et al., 2020)	<1B	56.7	52.3	70.2	64.2	74.5	73.1
API-based LMM (Zero-shot)							
Gemini-Pro-V (Team et al., 2023)	-	66.0	65.7	74.5	74.5	71.3	76.2
GPT-4V (OpenAI, 2023)	-	70.5	70.3	-	-	-	-
Open-source LMM (Zero-shot)							
Openflamingo (Awadalla et al., 2023)	7B	57.0	56.4	56.8	56.8	51.7	55.8
LLaVA-1.5 (Liu et al., 2023a)	13B	61.8	61.4	57.4	57.4	55.0	54.5
MMICL (Zhao et al., 2023)	11B	59.9	60.4	67.3	67.3	52.1	63.8
MiniGPT-v2 (Zhu et al., 2023)	7B	58.8	59.1	62.3	62.3	57.1	60.3
BLIP-2 (Li et al., 2023b)	11B	56.4	55.8	59.4	59.4	56.8	60.6
InstructBLIP (Dai et al., 2023)	13B	59.6	60.1	64.1	64.1	55.7	60.1
Evolver (Ours)	11B	**63.5**	**63.6**	**68.6**	**68.6**	**67.7**	**65.5**
Evolver† (Ours)	13B	62.3	62.5	59.9	59.9	59.3	57.3

highlights the effectiveness of our Chain-of-Evolution Prompting strategy across three datasets

4. Visualization



5. Reference

[1] Kiela, Douwe, et al. "The hateful memes challenge: Detecting hate speech in multimodal memes." Advances in neural information processing systems 33 (2020): 2611-2624.

[2] Dawkins R. The selfish gene[M]. Oxford university press, 2016.

[3] Qu, Yiting, et al. "On the evolution of (hateful) memes by means of multimodal contrastive learning." 2023 IEEE Symposium on Security and Privacy (SP). IEEE, 2023.

Resource

GitHub | Paper | Home page