Pro-Cap: Leveraging a Frozen Vision-Language Model for Hateful Meme Detection

Hubert Bujakowski
Jan Kruszewski
Łukasz Tomaszewski

27.11.2024

1. Introduction
2. Pro-Cap
3. Data
4. Experiments
5. Summary

Introduction

Pro-Cap: Leveraging a Frozen Vision-Language Model for Hateful Meme Detection

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

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

Adriel Kuek
adrielkuek@gmail.com
DSO National Laboratories
Singapore, Singapore

Wen-Haw Chong
whchong.2013@phdis.smu.edu.sg
Singapore Management University
Singapore, Singapore

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

Jing Jiang
jingjiang@smu.edu.sg
Singapore Management University
Singapore, Singapore

Introduction

- Memes have become a ubiquitous and influential form of daily communication
- Memes are a powerful means of expression, combining humor, visual elements, and context to convery messages

Introduction

- Memes have become a ubiquitous and influential form of daily communication
- Memes are a powerful means of expression, combining humor, visual elements, and context to convery messages

BUT

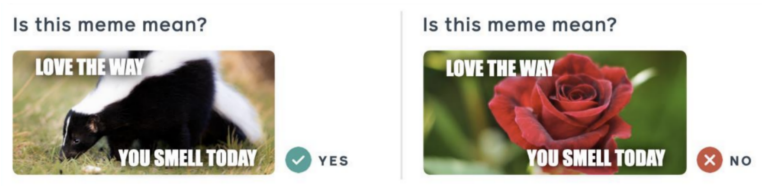
The dark side: Harmful and Offensive Content

- Popularity of memes also brings the risk of content that can be harmful or offensive

https://ai.meta.com/blog/hateful-memes-challenge-and-data-set/

Why Is Hateful Meme Detection Difficult?

- **Multimodal Nature**: Requires understanding images, text, and their interaction.



https://ai.meta.com/blog/hateful-memes-challenge-and-data-set/

The combination of text and image is neutral separately, but offensive when combined.

Methods for Hateful Meme Detection

- There are several ways to tackle hateful meme detection, each with its strengths and challenges:
- **Fine-tuning Pretrained Vision-Language Models (PVLMs)**:
  - Models like VisualBERT and ViLBERT are fine-tuned on hateful meme detection datasets.
  - Challenges: Computationally expensive.
- **Unimodal Approach (e.g., PromptHate)**:
  - Converts the multimodal task to a text-based task by generating image captions.
  - Challenges: Generic captions may miss crucial details such as race, gender, and other vulnerable identities.

Overview

The key idea is to obtain image details that are critical for hateful content detection. PVLMs can provide good descriptions, however these usually do not include critical information. To resolve this issue, additional prompts were made to the PVLM with questions important for meme classification.
Answers to these questions are called **Pro-Cap**.

VQA Questions

Previous studies have found that demographic information of people in the image or their gender significantly aids hateful meme detection.
For Pro-Cap questions about **Religion**, **Race**, **Gender**, **Nationality** and **Disability** have been introduced, as answers to these questions may be crucial for the meme to be considered hateful.

VQA Questions

Focus	Question
Content	What is shown in the image?
Race	What is the race of the person in the image?
Gender	What is the gender of the person in the image?
Religion	What is the religion of the person in the image?
Nationality	Which country does the person in the image come from?
Disability	Are there disabled people in the image?
Animal	What animal is in the image?
Val Person	Is there a person in the image?
Val Animal	Is there an animal in the image?

VQA Questions used for Pro-Cap

BERT-based Detection Model

BERT-based Detection Model is the first hateful meme classification model introduced in this paper. In this approach BERT model is fed with concatenation of meme text T and the Pro-Cap C.

$$r = BERT([T, C])$$

Next, the sentence representation $r \in \mathbb{R}^d$ is fed into a linear layer for hateful meme classification:

$$s = Sigmoid(W^T r + b)$$

where W and b are trainable parameters.

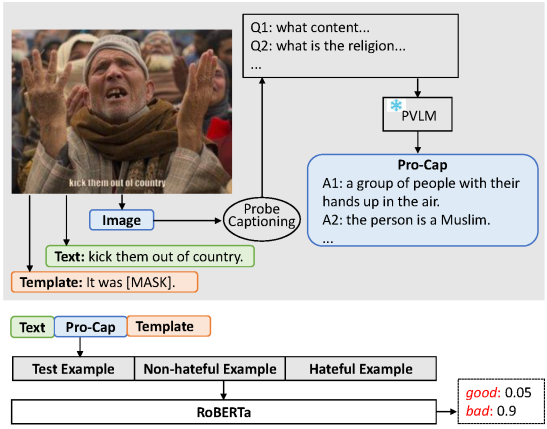
ProCapPromptHate for Hateful Meme Detection

The second approach is called ProCapPromptHate. In this approach the meme text is concatenated with Pro-Cap and also a prompt template S: "It was [MASK].". This concatenation is then forwarded to a language model to predict whether the meme is harmful. The prediction is based on the word that the model predicts for [MASK]. The probabilities for a positive and negative word are compared and based on this the meme is classified. PromptHate also includes one positive and one negative example in the context.

$$p = Sigmoid(LM([O_{test}, O_{non-hate}, O_{hate}]))$$

where $O = [T, C, S]$.

ProCapPromptHate for Hateful Meme Detection



ProCapPromptHate Diagram Source:[[1]]

Experiment Settings: FHM Dataset

Facebook Hateful Meme (FHM) Dataset:

- Synthetic memes with confounders requiring multimodal reasoning.
- Categories: Religion, Race, Gender, Nationality, Disability.
- Evaluation on *dev-seen* split (test split labels unavailable).

Hateful Meme

Non-Hateful Meme

Experiment Settings: MAMI Dataset

Multimedia Automatic Misogyny Identification (MAMI) Dataset:

- Focused on misogynistic memes targeting women.
- Reflects detection capability for female-targeted hate.

Hateful Meme

Non-Hateful Meme

Experiment Settings: HarM Dataset

Hateful Meme Detection (HarM) Dataset:

- COVID-19-related memes classified as harmless, partially harmful, and harmful.
- Partially harmful and harmful categories merged.
- Tests generalization from hateful to harmful meme detection.

Hateful Meme

Non-Hateful Meme

Experiment Settings: Dataset Statistics

Dataset	Train (Hate/Non-Hate)	Test (Hate/Non-Hate)
FHM	3,050 / 5,450	250 / 250
HarM	1,064 / 1,949	124 / 230
MAMI	5,000 / 5,000	500 / 500

Experiment Settings: Evaluation Metrics

Task: Binary Classification
Metrics:

- **Accuracy.**
- **AUC-ROC:** Area Under the Receiver Operating Characteristics curve.

Performance Reporting:

- Averaged over 10 random seeds.
- The same random seeds are used for all models.

Baselines: Overview

Objective: Compare Pro-Cap against both unimodal and multimodal models to demonstrate its effectiveness.

Key Definitions:

- **Unimodal Models:** Use one modality (text or image).
- **Multimodal Models:** Combine information from text and image modalities.

Note: Pro-Cap incorporates image information. When used with a text-based model (e.g., BERT), it is not considered unimodal.

Baselines: Unimodal Models

Text-Only Model:

- **Text-BERT:** Fine-tunes a pre-trained BERT model on meme text only.

Image-Only Model:

- **Image-Region:**
  - Extracts object-level image features using Faster-RCNN (pre-trained for object detection).
  - Applies average pooling to object features.
  - Feeds the resulting vector into a classification layer.

Baselines: Multimodal Models (Generic)

Category 1: Generic Multimodal Models

- **MMBT-Region:** Not pre-trained on multimodal data.
- **VisualBERT COCO:** Pre-trained on MS-COCO.
- **ViLBERT CC:** Pre-trained on Conceptual Captions.
- **ALBEF:** Align Before Fusion model.
- **BLIP:** Bootstrapping Language-Image Pre-training model.

Baselines: Multimodal Models (Hateful Meme Specific)

Category 2: Models Designed for Hateful Meme Detection

- **MOMENTA:**
  - Leverages CLIP for noisy meme images.
  - Uses pre-trained BERT for meme text.
  - Combines modalities via concatenation.
  - Explores multimodal interactions using local and global fusion mechanisms.
  - Utilizes augmented image tags (detected image entities).
- **DisMultiHate:**
  - Disentangles target information, essential for identifying hateful content.
- **PromptHate:**
  - Unimodal model with generic captions.

Experiment Results: Without Augmented Image Tags

Dataset / Model	FHM AUC.	FHM Acc.	MAMI AUC.	MAMI Acc.	HarM AUC.	HarM Acc.
Text BERT	$66.10_{\pm 0.55}$	$57.12_{\pm 0.49}$	$74.48_{\pm 0.60}$	$67.37_{\pm 0.57}$	$81.39_{\pm 0.91}$	$75.68_{\pm 1.59}$
Image-Region	$56.69_{\pm 1.05}$	$52.34_{\pm 1.39}$	$70.20_{\pm 0.63}$	$64.18_{\pm 0.81}$	$76.46_{\pm 0.47}$	$73.05_{\pm 1.80}$
VisualBERT COCO	$68.71_{\pm 1.02}$	$61.48_{\pm 1.19}$	$78.71_{\pm 0.59}$	$71.06_{\pm 0.94}$	$80.46_{\pm 1.04}$	$75.31_{\pm 1.44}$
ViLBERT CC	$73.05_{\pm 0.62}$	$64.70_{\pm 1.12}$	$77.71_{\pm 1.20}$	$69.48_{\pm 1.00}$	$84.11_{\pm 0.88}$	$78.70_{\pm 1.17}$
MMBT-Region	$72.86_{\pm 0.64}$	$65.06_{\pm 1.76}$	$79.17_{\pm 0.91}$	$70.46_{\pm 0.76}$	$85.48_{\pm 0.75}$	$79.83_{\pm 2.00}$
CLIP-BERT	$66.97_{\pm 0.34}$	$58.28_{\pm 0.63}$	$77.66_{\pm 0.64}$	$68.44_{\pm 1.07}$	$82.63_{\pm 3.83}$	$80.48_{\pm 1.95}$
DisMultiHate	$69.11_{\pm 0.84}$	$62.42_{\pm 0.72}$	$78.21_{\pm 0.61}$	$70.58_{\pm 1.13}$	$83.69_{\pm 1.33}$	$78.05_{\pm 0.73}$
PromptHate	$76.76_{\pm 0.95}$	$67.82_{\pm 1.23}$	$76.21_{\pm 1.05}$	$68.08_{\pm 0.58}$	$87.51_{\pm 0.74}$	$79.38_{\pm 1.72}$
BLIP	$76.80_{\pm 2.37}$	$69.20_{\pm 1.84}$	$80.59_{\pm 0.87}$	$71.84_{\pm 1.11}$	$87.09_{\pm 1.46}$	$81.81_{\pm 1.74}$
ALBEF	$79.40_{\pm 0.53}$	$70.58_{\pm 0.50}$	$\mathbf{83.24}_{\pm 0.93}$	$72.77_{\pm 1.00}$	$85.49_{\pm 1.23}$	$80.99_{\pm 0.80}$
Pro-Cap$_{\text{BERT}}$	$77.50_{\pm 0.58}$	$68.14_{\pm 0.64}$	$79.62_{\pm 0.91}$	$71.06_{\pm 0.88}$	$89.04_{\pm 1.00}$	$82.06_{\pm 1.92}$
Pro-Cap$_{\text{PromptHate}}$	$\mathbf{80.87}_{\pm 0.66}$	$\mathbf{72.28}_{\pm 0.90}$	$82.53_{\pm 0.49}$	$\mathbf{73.06}_{\pm 0.82}$	$\mathbf{90.25}_{\pm 0.54}$	$\mathbf{83.25}_{\pm 1.00}$

Experiment Results: Without Augmented Image Tags

Performance Comparison:

- Pro-CapPromptHate performs best in most of the tasks.
- Pro-CapBERT outperforms Text-BERT (text-only model), highlighting the importance of visual signals for hateful meme detection.
- Pro-CapBERT performs better than multimodal models like ViLBERT and MMBT-Region, despite having fewer parameters.
- Results show Pro-CapBERT's competitive performance against models specifically designed for hateful meme detection.

Experiment Results: With Augmented Image Tags

Dataset / Model	FHM AUC.	FHM Acc.	MAMI AUC.	MAMI Acc.	HarM AUC.	HarM Acc.
VisualBERT COCO	$72.56_{\pm 0.80}$	$64.28_{\pm 1.27}$	$80.84_{\pm 0.67}$	$72.86_{\pm 0.71}$	$82.96_{\pm 0.98}$	$78.81_{\pm 0.80}$
ViLBERT CC	$75.72_{\pm 0.91}$	$68.24_{\pm 0.44}$	$80.33_{\pm 1.01}$	$71.75_{\pm 1.14}$	$84.79_{\pm 1.23}$	$81.39_{\pm 1.62}$
MOMENTA	$69.17_{\pm 4.71}$	$61.34_{\pm 4.89}$	$81.68_{\pm 2.80}$	$72.10_{\pm 2.90}$	$86.32_{\pm 3.83}$	$80.48_{\pm 1.95}$
DisMultiHate	$79.89_{\pm 1.71}$	$71.26_{\pm 1.66}$	$80.08_{\pm 0.55}$	$71.87_{\pm 0.47}$	$86.39_{\pm 1.17}$	$81.24_{\pm 1.04}$
PromptHate	$81.45_{\pm 0.74}$	$72.98_{\pm 1.09}$	$79.95_{\pm 0.66}$	$70.31_{\pm 0.64}$	$90.96_{\pm 0.62}$	$84.47_{\pm 1.75}$
BLIP	$76.40_{\pm 1.49}$	$69.29_{\pm 1.44}$	$80.63_{\pm 1.05}$	$70.62_{\pm 1.48}$	$86.88_{\pm 1.15}$	$82.66_{\pm 1.13}$
ALBEF	$80.77_{\pm 0.81}$	$71.70_{\pm 0.98}$	$82.45_{\pm 0.85}$	$72.45_{\pm 0.96}$	$86.91_{\pm 0.72}$	$81.78_{\pm 1.20}$
Pro-Cap$_{\text{BERT}}$	$79.75_{\pm 1.15}$	$71.28_{\pm 0.91}$	$81.20_{\pm 0.69}$	$71.80_{\pm 1.42}$	$89.75_{\pm 1.49}$	$82.71_{\pm 1.60}$
Pro-Cap$_{\text{PromptHate}}$	$\mathbf{83.58}_{\pm 0.60}$	$\mathbf{75.10}_{\pm 0.97}$	$\mathbf{83.77}_{\pm 0.75}$	$\mathbf{73.63}_{\pm 0.75}$	$\mathbf{91.03}_{\pm 1.51}$	$\mathbf{85.03}_{\pm 1.51}$

Experiment Results: With Augmented Image Tags

Performance Comparison:

- Pro-CapPromptHate performs best in all of the tasks.
- Pro-CapBERT continues to outperform multimodal models like VisualBERT and ViLBERT even with augmented image tags.
- Pro-CapBERT also surpasses models designed for hateful meme detection such as MOMENTA and DisMultiHate.
- Pro-CapBERT performs consistently well across all datasets with the additional image tags.

Experiment Results: Ablation Study

Focus	Questions
Content	what is shown in the image?
Race	What is the race of the person in the image?
Gender	What is the gender of the person in the image?
Religion	What is the religion of the person in the image?
Nationality	Which country does the person in the image come from?
Disability	Are there disabled people in the image?
Animal	What animal is in the image?
Val Person	Is there a person in the image?
Val Animal	Is there an animal in the image?

VQA Questions

Ans. Length	FHM	MAMI	HarM
No Centric	$70.08_{\pm 1.57}$	$72.78_{\pm 0.63}$	$80.11_{\pm 1.14}$
Penalty = 1	$71.94_{\pm 0.97}$	$73.06_{\pm 0.82}$	$82.09_{\pm 1.21}$
Penalty = 2	$72.28_{\pm 0.90}$	$72.91_{\pm 1.16}$	$82.85_{\pm 1.51}$
Penalty = 3	$71.40_{\pm 1.06}$	$72.47_{\pm 0.74}$	$83.25_{\pm 1.00}$
Pro-Cap$_{\text{PromptHate}}$	$\mathbf{72.28_{\pm 0.90}}$	$\mathbf{73.06_{\pm 0.82}}$	$\mathbf{83.25_{\pm 1.00}}$

Ablation results

Experiment Results: Ablation Study

Key Focus Areas:

- Impact of asking different hateful-content centric questions.
- Impact of the length of answers to probing questions.

Key Findings:

- Asking target-specific questions improves performance by over 2% on FHM and over 3% on HarM, but has less impact on MAMI.
- Longer answers do not drastically change detection performance, highlighting the robustness of Pro-Cap.
- HarM benefits from longer answers, while MAMI prefers shorter answers.

Case study: Pro-CapPromptHate vs basic PromptHate

Meme			
Ground Truth	Hateful (religion)	Hateful (religion)	Hateful (race)
Basic PromptHate	Non-hateful	Non-hateful	Non-hateful
Pro-Cap$_{\text{PromptHate}}$	Hateful	Hateful	Hateful
Meme text	changing every single country it touches	no that is not his daughter that is his wife yet the world is silent	the definition of utter disgust in plain black and white
Basic caption	mughal structure is one of the largest mosques in the world.	portrait of a father hugging his daughter while smiling at camera in the living room at home.	love is in the air!.
Pro-Cap	(Content:) a black cat sitting on a blue and white tiled floor. (Race:) a black person is standing on a blue and white tiled floor in islamic. (Gender:) a man in a black shirt is standing on a blue and white tiled floor with a clock on top of his head. (Country:) islamic. (Religion:) the person is a muslim and he is wearing a black t-shirt and a black sleeveless.	(Content:) a man and a woman hugging on a couch. (Race:) a white man and a white woman hugging on a white couch. (Gender:) a man and a woman hugging on a white couch. (Country:) islamic. (Religion:) an muslim man and woman hugging on a white couch.	(Content:) a black and white photo of a man and a woman. (Race:) a black man and a white woman in a black and white photo. (Gender:) a man and a woman in a black and white photo. (Country:) afghanistan. (Religion:) he is a christian.

Case study: Pro-CapPromptHate vs Basic PromptHate

Key Observations:

- Generic captions are not always sufficient for hateful meme detection.
- Probing questions about vulnerable targets (e.g., race, religion) provide critical keywords.
- Examples highlight how target-specific questions enhance detection compared to generic captions.

Case study: Error cases of Pro-CapPromptHate

Meme		
GT	Hateful (gender)	Non-hateful
Pred	Non-hateful	Hateful
Meme text	scientist are working hard to cure them all	islam is a religion of peace stop criticizing my religion
Pro-Cap	(Content:) two women in wedding dresses kissing each other. (Race:) a white woman kissing a brunette woman in a wedding dress. (Gender:) a woman is kissing a man in a wedding dress. (Country:) the person in the image comes from a country in the philippines. (Religion:) the person in the image is a christian.	(Content:) a man with a beard laughing in the woods. (Race:) a african man with a beard and a red hat is smiling in the woods. (Gender:) a man with a beard and a red hat in front of a wooded area. (Country:) egypt is the country that the person in the image comes from. (Religion:) he is a muslim man with a beard and a red tiara on his head.

Case study: Error cases of Pro-CapPromptHate

Key Insights:

- Complex reasoning remains a challenge for current language models, despite accurate probe-captioning.
- Small-scale datasets hinder models from performing more sophisticated reasoning.
- Mis-predictions (e.g., "a woman kissing a man" for gender-related questions) highlight model limitations.
- Debiasing techniques may be necessary to address potential biases in hateful content detection.

Conclusion

- **Pro-Cap** leverages a frozen vision-language model to detect hateful memes, a critical task in mitigating online harm.
- Model demonstrated competitive performance against models specifically designed for hateful meme detection.
- The article is a great example of usage deep learning in key areas of our lives.

Bibliography


Rui Cao, Ming Shan Hee, Adriel Kuek, Wen-Haw Chong, Roy Ka-Wei Lee, and Jing Jiang.
Pro-cap: Leveraging a frozen vision-language model for hateful meme detection.
In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5244–5252, 2023.