Information Technologies Institute

Improving Multimodal Hateful Meme Detection Exploiting LMM-Generated Knowledge

Maria Tzelepi and Vasileios Mezaris
Information Technologies Institute (ITI), Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece
{mtzelepi, bmezaris}@iti.gr

CVPR Nashville JUNE 11-15, 2025
8th Multimodal Learning and Applications Workshop

Background

- Detecting hateful content in memes has emerged as a task of critical importance
- The nature of memes (images in combination with embedded text) renders hateful meme detection a **challenging** task
  - comprehension of both the involved modalities as well as their interaction is required
- Limitations of current approaches: effectively capture the multimodal semantic content, computational efficiency

Contributions

- We leverage **LMM-encoded knowledge** in a fully multimodal fashion in order to build **strong meme representations**, that include generic semantic descriptions and elicited emotions, capable of revealing the underlying meanings of the combined modalities
- We additionally use an LMM to identify hard training memes, and propose an **LMM-based hard-mining** approach that enhances the discrimination ability of the meme embeddings through the LMM-generated hard example information, achieving in turn improved classification performance
- We perform extensive experiments on two challenging datasets achieving **state-of-the-art** performance

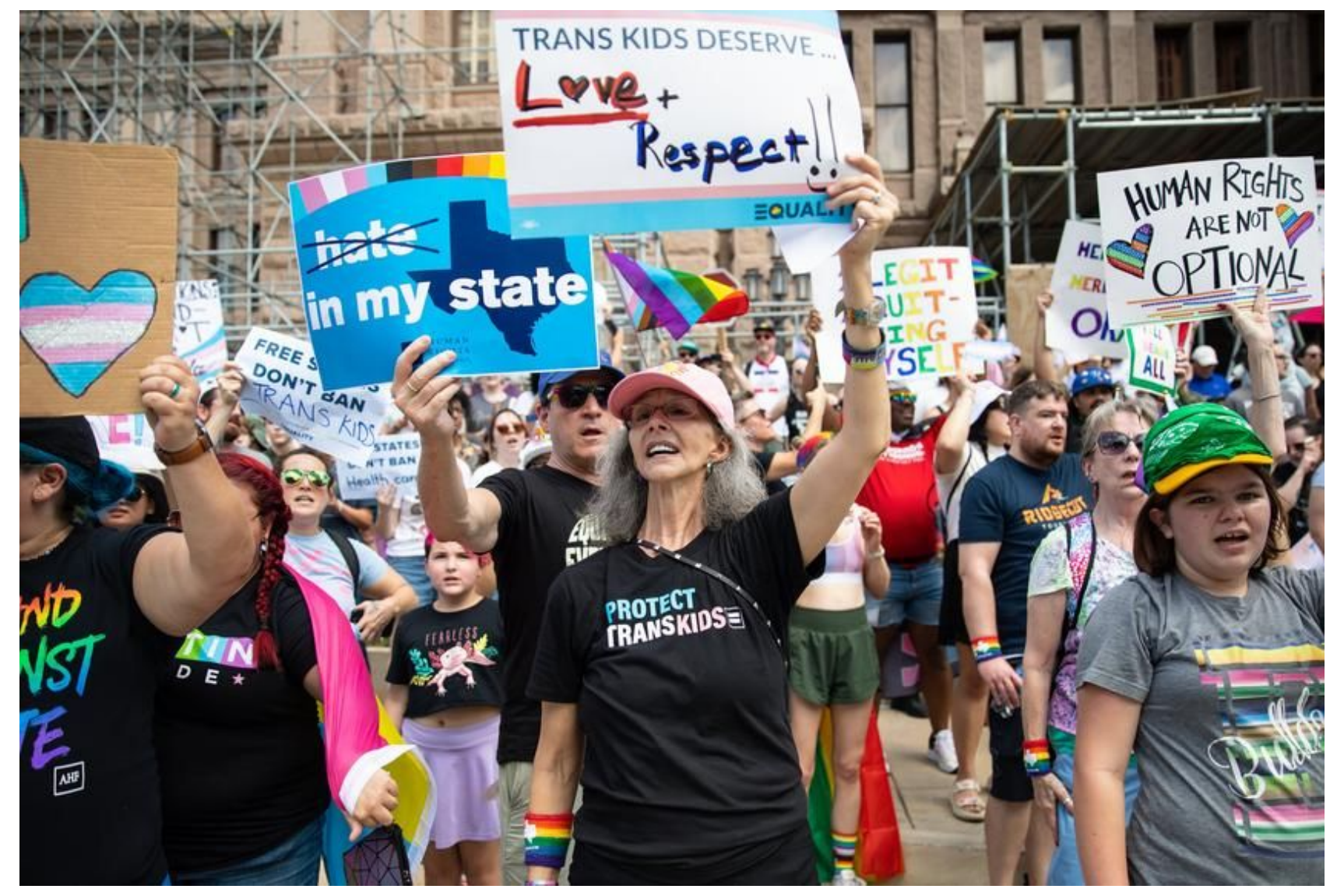
Experimental Results

Method	Harm-C	PrideMM
MOMENTA	82.44 ± 0.65	72.23 ± 0.58
DisMultiHate	81.24 ± 1.04	-
Hate-CLIPper	83.68 ± 0.62	75.53 ± 0.58
PromptHate	84.47 ± 1.75	-
ISSUES	81.31 ± 1.05	74.68 ± 1.62
Pro-Cap	85.03 ± 1.51	-
MemeCLIP	84.72 ± 0.45	76.06 ± 0.23
ExplainHM	87.00	-
LMM-CLIP (Proposed)	86.33 ± 0.42	**76.31 ± 0.39**
LMM-LongCLIP (Proposed)	**87.23 ± 0.33**	75.89 ± 0.54

Table 1: Comparisons with state of the art in terms of accuracy (%).

Embeddings					Harm-C		PrideMM	
Image	Embedded Text	Semantic Descriptions	Elicited Emotions	Hard Mining	CLIP	LongCLIP	CLIP	LongCLIP
✓	✓				84.63 ± 0.29	85.59 ± 0.25	75.21 ± 0.27	76.06 ± 0.27
✓	✓	✓			85.25 ± 0.49	86.05 ± 0.38	75.51 ± 0.29	75.38 ± 0.71
✓	✓		✓		85.31 ± 0.36	85.99 ± 0.23	75.59 ± 0.25	76.02 ± 0.41
✓	✓			✓	85.20 ± 0.46	86.16 ± 0.74	75.76 ± 0.39	**76.48 ± 0.35**
✓	✓	✓	✓		85.65 ± 0.42	86.21 ± 0.21	75.89 ± 0.53	75.47 ± 0.47
✓	✓	✓	✓	✓	**86.33 ± 0.42**	**87.23 ± 0.33**	**76.31 ± 0.39**	75.89 ± 0.54

Table 2: Ablations in terms of accuracy (%).

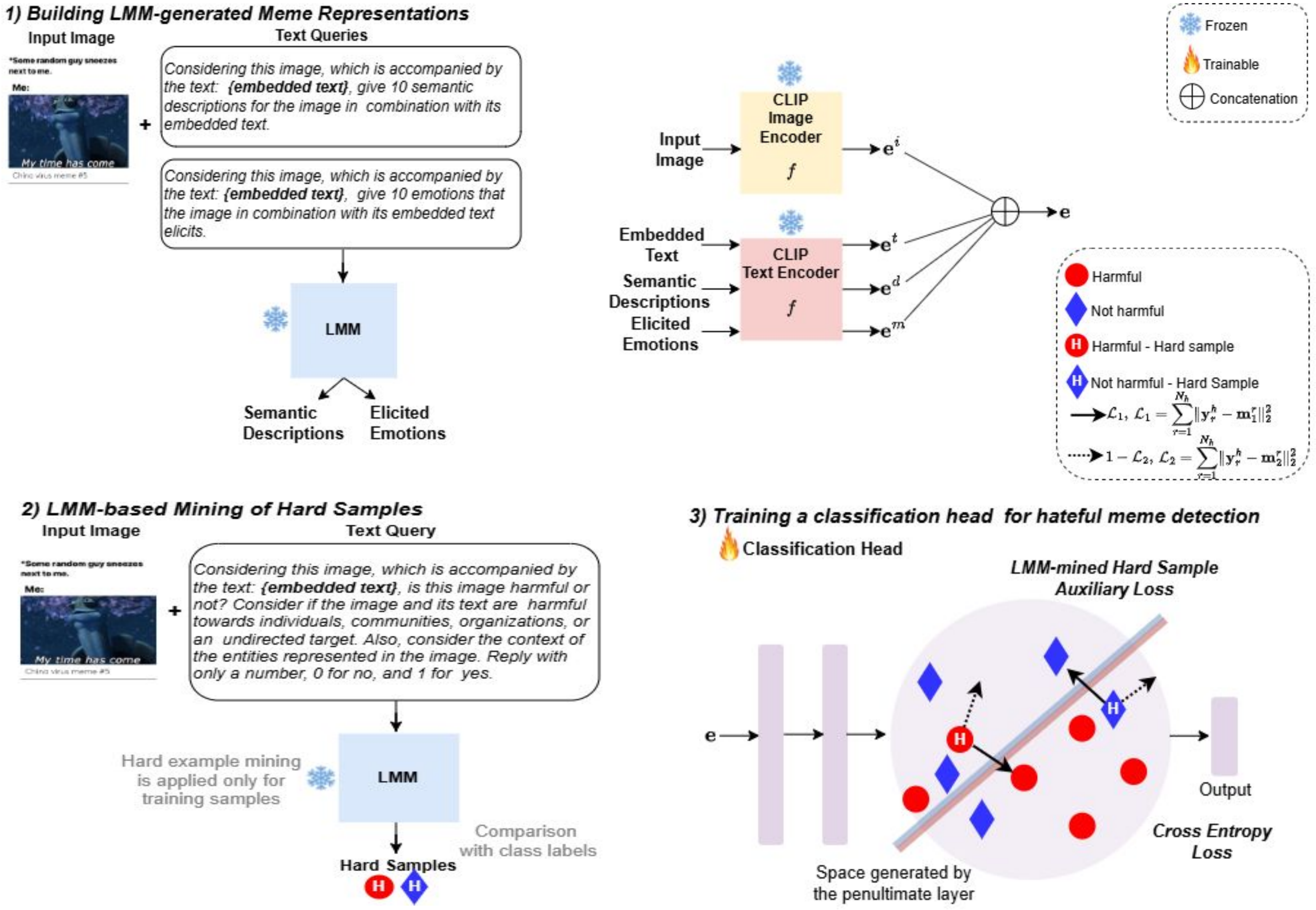


Semantic Descriptions:
1. A group of people holding up signs in support of transgender rights. 2. A crowd gathered in front of a building holding up signs and banners. 3. A demonstration in support of transgender equality with people holding up signs and banners. 4. A group of people holding up signs and banners in support of transgender rights. 5. A crowd gathered in front of a building holding up signs and banners in support of transgender equality. 6. A demonstration in support of transgender rights with people holding up signs and banners. 7. A group of people holding up signs and banners in support of transgender equality. 8. A crowd gathered in front of a building holding up signs and banners in support of transgender rights.

Elicited Emotions:
1. Joy
2. Pride
3. Hope
4. Empathy
5. Understanding
6. Acceptance
7. Love
8. Respect
9. Protection
10. Inclusion

Qualitative results on the LMM-generated semantic descriptions and elicited emotions.

Our Solution


1) Building LMM-generated Meme Representations

Text Queries

*Considering this image, which is accompanied by the text: **{embedded text}**, give 10 semantic descriptions for the image in combination with its embedded text.*

*Considering this image, which is accompanied by the text: **{embedded text}**, give 10 emotions that the image in combination with its embedded text elicits.*

2) LMM-based Mining of Hard Samples

Text Query

*Considering this image, which is accompanied by the text: **{embedded text}**, is this image harmful or not? Consider if the image and its text are harmful towards individuals, communities, organizations, or an undirected target. Also, consider the context of the entities represented in the image. Reply with only a number, 0 for no, and 1 for yes.*

3) Training a classification head for hateful meme detection

Input Image
*Some random guy sneezes next to me.
Me:
My time has come
China virus meme #5
+
LMM
Semantic Descriptions
Elicited Emotions
Frozen
Trainable
Concatenation
Input Image
CLIP Image Encoder
f
$\mathbf{e}^i$
Embedded Text
Semantic Descriptions
Elicited Emotions
CLIP Text Encoder
f
$\mathbf{e}^t$
$\mathbf{e}^d$
$\mathbf{e}^m$
$\mathbf{e}$
Harmful
Not harmful
Harmful - Hard sample
Not harmful - Hard Sample
$\mathcal{L}_1,\ \mathcal{L}_1 = \sum_{r=1}^{N_h} \|\mathbf{y}_r^h - \mathbf{m}_1^r\|_2^2$
$1 - \mathcal{L}_2,\ \mathcal{L}_2 = \sum_{r=1}^{N_h} \|\mathbf{y}_r^h - \mathbf{m}_2^r\|_2^2$
Input Image
Hard example mining is applied only for training samples
LMM
Comparison with class labels
Hard Samples
H
Classification Head
LMM-mined Hard Sample Auxiliary Loss
e
H
Output
Cross Entropy Loss
Space generated by the penultimate layer

- In the first step we prompt an LMM to extract semantic descriptions and elicited emotions for the memes, and we use a VLM to extract the corresponding embeddings in order to build the meme representations
- In the second step we prompt the LMM to identify hard samples
- In the third step we train a classification head for hateful meme detection using a regular supervised loss and a new LMM-mined hard sample auxiliary loss, using the identified hard samples
- The proposed objective forces the identified hard embeddings to approach their nearest non-hard embeddings inside the batch that belong to the same class (L_1) and at the same time to move away from their nearest embeddings of the opposite class (1-L_2)