MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets

Shraman Pramanick*, Shivam Sharma*, Dimitar Dimitrov, Shad Akhtar, Preslav Nakov, Tanmoy Chakraborty
Johns Hopkins University, IIIT-Delhi, Sofia University, QCRI-Doha
spraman3@jhu.edu, {shivams, shad.akhtar, tanmoy}@iiitd.ac.in, pnakov@hbku.edu.qa, mitko.bg.ss@gmail.com

Motivation

- Memes are **context dependent**.
- Notion of '**harm**' is broader than '**hate**' and '**offense**'.
- **Identifying** the **targets of harmful memes** is an important but less-studied problem.

DEPOSIT BIDEN'S VOTES HERE
I just saw Joe Biden campaigning at Macy's
DO YOU EVER LOOK AT STUFF AND WONDER HOW IT GOT THERE?

Multimodal cues are necessary to detect harm

Contributions

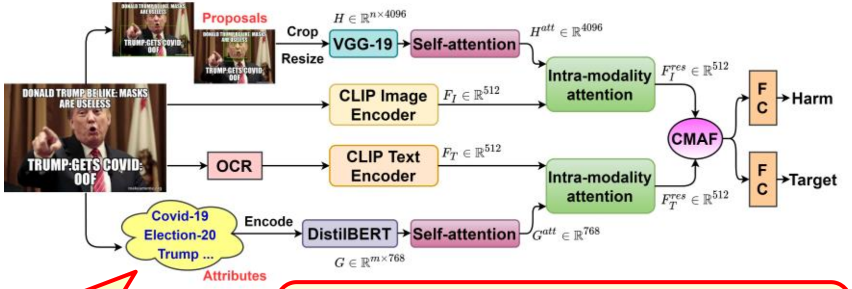
- **Harm-C** and **Harm-P** ← two large-scale datasets for harmful meme detection and target identification.

AFTER 2 WEEKS OF WFH
1 HOUR IN HOME IS EQUAL TO 7 HOURS IN OFFICE
AFTER 2 WEEKS OF WFH
1 HOUR IN HOME IS EQUAL TO 7 HOURS IN OFFICE

Attributes: {Christopher Nolan, Interstellar, work from home, humor}

- **MOMENTA** ← analyses local and global perspective of the input meme and relates to background context.

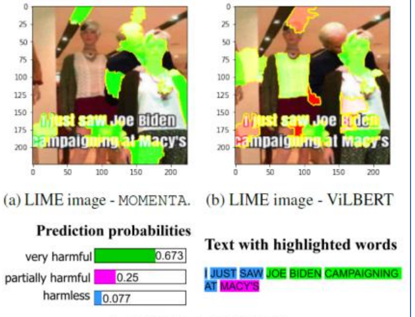
MOMENTA Architecture


Proposals
Crop
Resize
$H \in \mathbb{R}^{n\times 4096}$
VGG-19
Self-attention
$H^{att} \in \mathbb{R}^{4096}$
Intra-modality attention
$F_I^{res} \in \mathbb{R}^{512}$
CLIP Image Encoder
$F_I \in \mathbb{R}^{512}$
CMAF
F C
Harm
F C
Target
OCR
CLIP Text Encoder
$F_T \in \mathbb{R}^{512}$
Intra-modality attention
$F_T^{res} \in \mathbb{R}^{512}$
Covid-19
Election-20
Trump ...
Attributes
Encode
DistilBERT
Self-attention
$G^{att} \in \mathbb{R}^{768}$
$G \in \mathbb{R}^{m\times 768}$
DONALD TRUMP BE LIKE: MASKS ARE USELESS
TRUMP:GETS COVID: OOF

Detected attributes and proposals model the context

- ✓ CLIP Feature Representation
- ✓ Object Proposal and Image Attribute Extraction
- ✓ Inter- modality Attention
- ✓ Cross-modality Attention Fusion (CMAF)

Explainability of MOMENTA



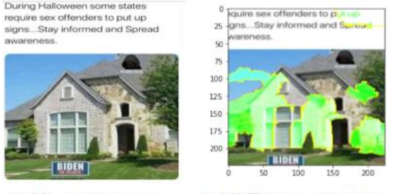
(a) LIME image - MOMENTA. (b) LIME image - ViLBERT

(c) LIME text - MOMENTA.

Visual Explanation generated by LIME on both modalities

Results

Modality	Model	Harmful Meme Detection on Harm-C						Harmful Meme Detection on Harm-P					
		2-Class Classification			3-Class Classification			2-Class Classification			3-Class Classification		
		Acc ↑	F1 ↑	MMAE ↓	Acc ↑	F1 ↑	MMAE ↓	Acc ↑	F1 ↑	MMAE ↓	Acc ↑	F1 ↑	MMAE ↓
	Human†	90.68	83.55	0.1723	86.10	65.10	0.4857	94.40	88.47	0.1028	92.12	70.35	0.6274
	Majority	64.76	39.30	0.5	64.76	26.20	1.0	51.27	33.39	0.5	51.27	22.59	1.0
Text (T) Only	TextBERT	70.17	66.25	0.2911	68.93	48.72	0.5591	80.12	78.35	0.1660	74.55	54.08	0.7742
Image (I) Only	VGG19	68.12	61.86	0.3190	66.24	41.76	0.6487	70.65	70.46	0.1887	73.65	51.89	0.8466
	DenseNet-161	68.42	62.54	0.3125	65.21	42.15	0.6326	74.05	73.68	0.1845	71.80	50.98	0.8388
	ResNet-152	68.74	62.97	0.3114	65.29	43.02	0.6264	73.14	72.77	0.1800	71.02	50.64	0.8900
	ResNeXt-101	69.79	63.68	0.3029	66.55	43.68	0.6499	73.91	73.57	0.1812	71.84	51.45	0.8422
I + T (Unimodal Pre-training)	Late Fusion	73.24	70.25	0.2927	66.67	45.06	0.6077	78.26	78.50	0.1674	76.20	55.84	0.7245
	Concat BERT	71.82	71.82	0.3156	65.54	43.37	0.5976	77.25	76.38	0.1743	76.04	55.95	0.7450
	MMBT	73.48	67.12	0.3258	68.08	50.88	0.6474	82.54	80.23	0.1413	78.14	58.03	0.7008
I + T (Multimodal Pre-training)	ViLBERT CC	78.53	78.06	0.1881	**75.71**	48.82	0.5329	**87.25**	**86.03**	**0.1276**	**84.66**	**64.70**	**0.6982**
	V-BERT COCO	**81.36**	**80.13**	**0.1857**	74.01	**53.85**	**0.5303**	86.80	86.07	0.1318	84.02	63.68	0.7020
Proposed System and Variants	CLIP	74.23	73.85	0.2955	67.04	44.25	0.6228	80.55	80.25	0.1659	77.00	56.85	0.7852
	CLIP + Proposals	77.65	76.90	0.2142	70.52	45.60	0.5955	84.16	83.80	0.1556	81.06	60.65	0.7122
	CLIP + Attributes	78.10	77.64	0.2010	71.05	45.55	0.5887	84.02	83.85	0.1508	80.75	60.23	0.7058
	MOMENTA w/o CMAF	80.75	80.17	0.1896	74.85	51.25	0.5360	86.20	85.55	0.1355	83.85	63.02	0.6990
	MOMENTA	**83.82**	**82.80**	**0.1743**	**77.10**	**54.74**	**0.5132**	**89.84**	**88.26**	**0.1314**	**87.14**	**66.66**	**0.6805**
$\Delta_{\text{MOMENTA}-baseline}$		↑2.46	↑2.67	↓0.0114	↑1.39	↑0.89	↓0.0171	↑2.59	↑2.23	↓0.0038	↑2.48	↑1.96	↓0.0177



(a) Misclassified meme. (b) LIME image - MOMENTA.

Error Analysis – MOMENTA fails here as detected attributes can't model context

Contact us – spraman3@jhu.edu
Code: https://github.com/LCS2-IIITD/MOMENTA

References:
1. The Hateful Memes Challenge, Kiela et al., NeurIPS 2020.
2. Detecting Harmful Memes and Their Targets, Pramanick et al., ACL-IJCNLP 2021.