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

Shraman Pramanick[1,2], Shivam Sharma[2,4], Dimitar Dimitrov[3], Md. Shad Akhtar[2], Preslav Nakov[5] , Tanmoy Chakraborty[2]
[1]Johns Hopkins University
[2]Indraprastha Institute of Information Technology, Delhi, India
[3] Sofia University, Bulgaria
[4] Wipro AI Labs, India
[5]Qatar Computing Research Institute, HBKU, Doha, Qatar

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

Introduction:

- **Internet memes**
  - typically, an **image** and a **short piece of overlaid text**
  - popular medium of expression
  - empowerment through associated virality
  - funny

[1]The Hateful Memes Challenge, Kiela et al., NeurIPS’20
[2]Multimodal meme dataset for identifying offensive content, Suryawanshi et al., , LREC-TRAC ’20

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

Introduction:

- **Internet memes**
  - typically, an **image** and a **short piece of overlaid text**
  - popular medium of expression
  - empowerment through associated virality
  - funny

- Challenging for analysis
  - **multimodality**
  - **context**-dependency
  - **morphed** image
  - **noisy/manipulated** text

[1]The Hateful Memes Challenge, Kiela et al., NeurIPS'20
[2]Multimodal meme dataset for identifying offensive content, Suryawanshi et al., , LREC-TRAC '20

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

Introduction:

- **Internet memes** can be **harmful** and even **weaponized**
  - hateful memes[1]
  - offensive memes[2]

- **Harm** is a more general concept than hate and offense

[1]The Hateful Memes Challenge, Kiela et al., NeurIPS'20
[2]Multimodal meme dataset for identifying offensive content, Suryawanshi et al., , LREC-TRAC '20

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

Our Contributions:

- We extend our recently released dataset HarMeme[1], which covered **COVID-19**, with a new topic **US Politics** and thus ending up with two datasets: **Harm-C** and **Harm-P**

[1] Pramanick et al., Detecting Harmful Memes and Their Targets, ACL'21

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

Our Contributions:

- We extend our recently released dataset HarMeme[1], which covered **COVID-19**, with a new topic **US Politics** and thus ending up with two datasets: **Harm-C** and **Harm-P**

- We benchmark the two datasets against several state-of-the-art unimodal and multimodal models, and we discuss the limitations of these models.

[1]Detecting Harmful Memes and Their Targets, Pramanick et al., ACL'21

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

Our Contributions:

- We extend our recently released dataset HarMeme[1], which covered **COVID-19**, with a new topic **US Politics** and thus ending up with two datasets: **Harm-C** and **Harm-P**
- We benchmark the two datasets against several state-of-the-art unimodal and multimodal models, and we discuss the limitations of these models.
- We propose MOMENTA, a novel multimodal framework that systematically analyzes the local and the global perspective of the input meme and relates it to the background context.

[1]Detecting Harmful Memes and Their Targets, Pramanick et al., ACL'21

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

Our Contributions:

- We extend our recently released dataset HarMeme[1], which covered **COVID-19**, with a new topic **US Politics** and thus ending up with two datasets: **Harm-C** and **Harm-P**
- We benchmark the two datasets against several state-of-the-art unimodal and multimodal models, and we discuss the limitations of these models.
- We propose MOMENTA, a novel multimodal framework that systematically analyzes the local and the global perspective of the input meme and relates it to the background context.
- We perform extensive experiments on both datasets, and we show that MOMENTA outperforms the baselines.

[1]Detecting Harmful Memes and Their Targets, Pramanick et al., ACL’21

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

Data Collection & Annotation:

- Collection: Google Image, Instagram, Facebook

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

Data Collection & Annotation:

- Collection: Google Image, Instagram, Facebook
- Removal of duplicates

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

Data Collection & Annotation:

- Collection: Google Image, Instagram, Facebook
- Removal of duplicates
- Annotation guidelines

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

Data Collection & Annotation:

- Collection: Google Image, Instagram, Facebook
- Removal of duplicates
- Annotation guidelines
- Annotation process
    - ➔ Dry run
    - ➔ Final annotation
    - ➔ Consolidation

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

Dataset Summary:

Dataset	Split	#Memes	Harmfulness			#Memes	Target			
			Very Harmful	Partially Harmful	Harmless		Individual	Organization	Community	Society
Harm-C	Train	3,013	182	882	1,949	1,064	493	66	279	226
	Validation	177	10	51	116	61	29	3	16	13
	Test	354	21	103	230	124	59	7	32	26
	Total	3,544	213	1,036	2,295	1,249	582	75	327	265
Harm-P	Train	3,020	216	1,270	1,534	1,451	797	470	111	73
	Validation	177	17	69	91	85	70	12	2	1
	Test	355	25	148	182	170	96	54	12	8
	Total	3,552	258	1487	1,807	1,706	963	536	125	82

Statistics about the Harm-P and Harm-C datasets. *Very harmful* and *partially harmful* memes are annotated with one of the following four targets: *individual, organization, community,* or *society*.

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

Lexical Summary:

Dataset	Harmfulness			Target			
	Very harmful	*Partially harmful*	*Harmless*	*Individual*	*Organization*	*Community*	*Society*
Harm-C	mask (0.0512) trump (0.0404) wear (0.0385) thinks (0.0308 killed (0.0269)	trump (0.0642) president (0.0273) obama (0.0262) donald (0.0241) virus (0.0213)	you (0.0264) home (0.0263 corona (0.0251) work (0.0222) day (0.0188)	trump (0.0541) president (0.0263) donald (0.0231) obama (0.0217) covid (0.0203)	deadline (0.0709) associated (0.0709) extra (0.0645) ensure (0.0645) qanon (0.0600)	china (0.0665) chinese (0.0417) virus (0.0361) wuhan (0.0359) cases (0.0319)	mask (0.0441) vaccine (0.0430) alcohol (0.0309) temperatures (0.0309) killed (0.0271)
Harm-P	photoshopped (0.0589) married (0.0343) joe (0.0309) trump (0.0249) nazis (0.0241)	democratic (0.0164) obama (0.0158) libertarian (0.0156) republican (0.0140) vote (0.0096)	party (0.02514) debate (0.0151) president (0.0139) democratic (0.0111) green (0.0086)	biden (0.0331) joe (0.0323) obama (0.0316) trump (0.0286) putin (0.0080)	libertarian (0.0358) republican (0.0319) democratic (0.0293) green (0.0146) government (0.0097)	liberals (0.0328) radical (0.0325) islam (0.0323) black (0.0237) mexicans (0.0168)	crime (0.0201) rights (0.0195) gun (0.0181) taxes (0.0138) law (0.0135)

Top-5 most frequent words per (class/dataset). The tf-idf score per word is given within parenthesis.

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

Baselines:

- **Unimodal Models**
  - Text Only
    - BERT
  - Image Only
    - VGG19
    - DenseNet-161
    - ResNet-152
    - ResNeXt-101

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

Baselines:

- **Unimodal Models**
  - Text Only
    - BERT
  - Image Only
    - VGG19
    - DenseNet-161
    - ResNet-152
    - ResNeXt-101
- **Multimodal Models (Image + Text)**
  - Unimodal Pre-training (Text)
    - Late Fusion (Avg.)
    - Concat BERT
    - MMBT
  - Multimodal Pre-training
    - ViLBERT CC
    - VisualBERT COCO

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

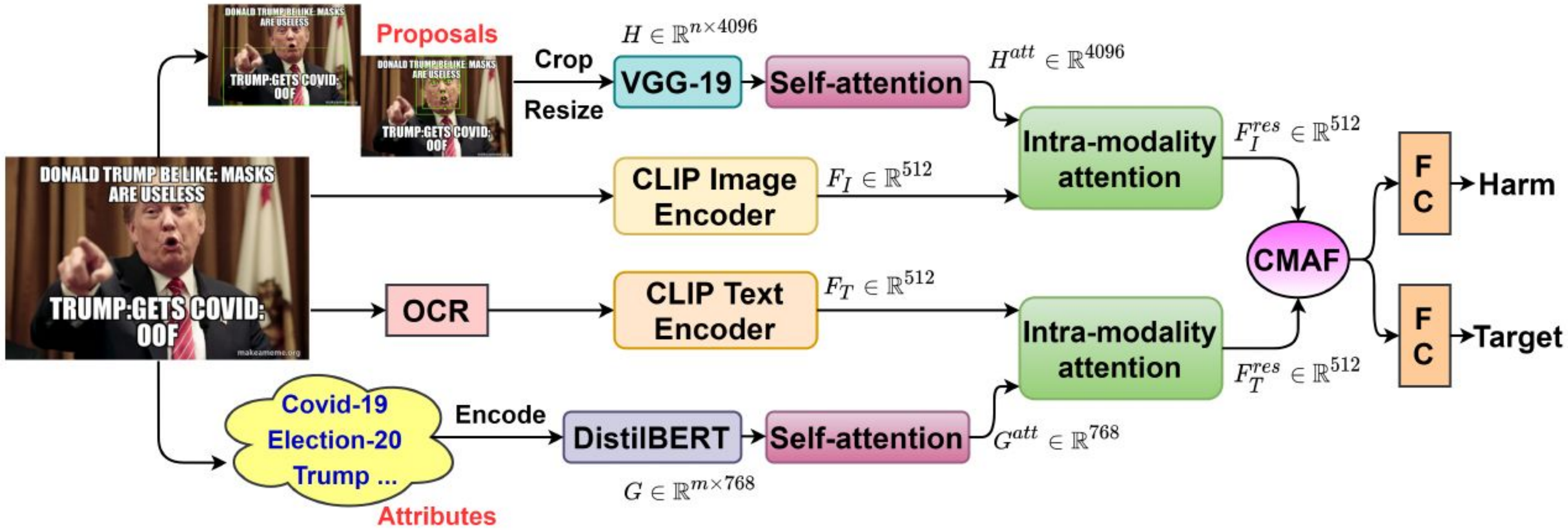
Baselines:

- **Unimodal Models**
  - Text Only
    - BERT
  - Image Only
    - VGG19
    - DenseNet-161
    - ResNet-152
    - ResNeXt-101
- **Multimodal Models (Image + Text)**
  - Unimodal Pre-training (Text)
    - Late Fusion (Avg.)
    - Concat BERT
    - MMBT
  - Multimodal Pre-training
    - ViLBERT CC
    - VisualBERT COCO

Access our dataset and implementation using this QR Code

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

MOMENTA:

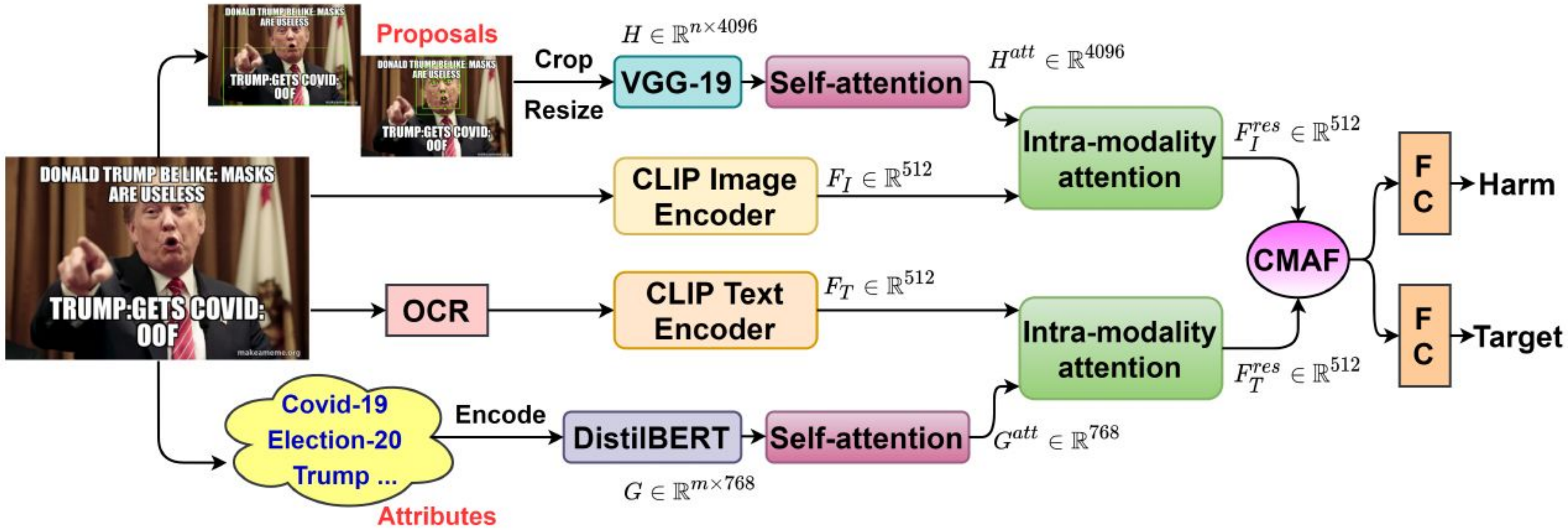


- We encode each image-text pair using CLIP[1], a pre-trained visual-linguistic model.

[1]Learning Transferable Visual Models From Natural Language Supervision, Radford et al., ICML ’21

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

MOMENTA:



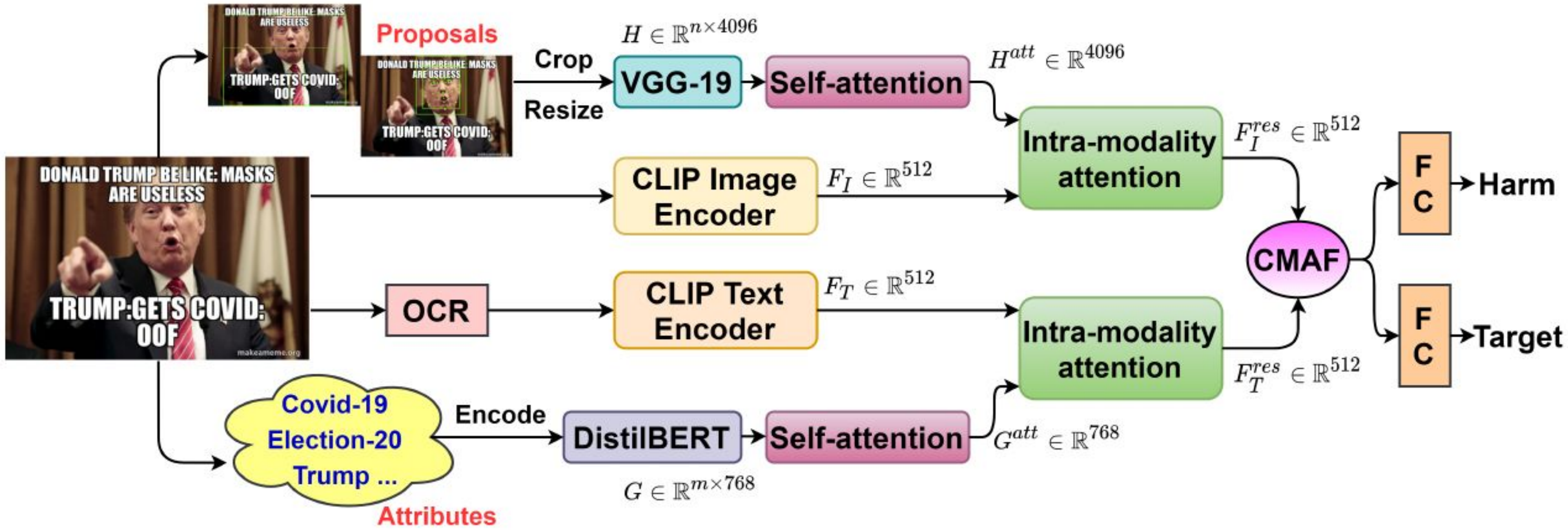
- We encode each image-text pair using CLIP[1], a pre-trained visual-linguistic model.

- In addition, we include meme object proposals (faces and foreground objects) and web attributes/entities.

[1]Learning Transferable Visual Models From Natural Language Supervision, Radford et al., ICML ’21

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

MOMENTA:



- We encode each image-text pair using CLIP[1], a pre-trained visual-linguistic model.

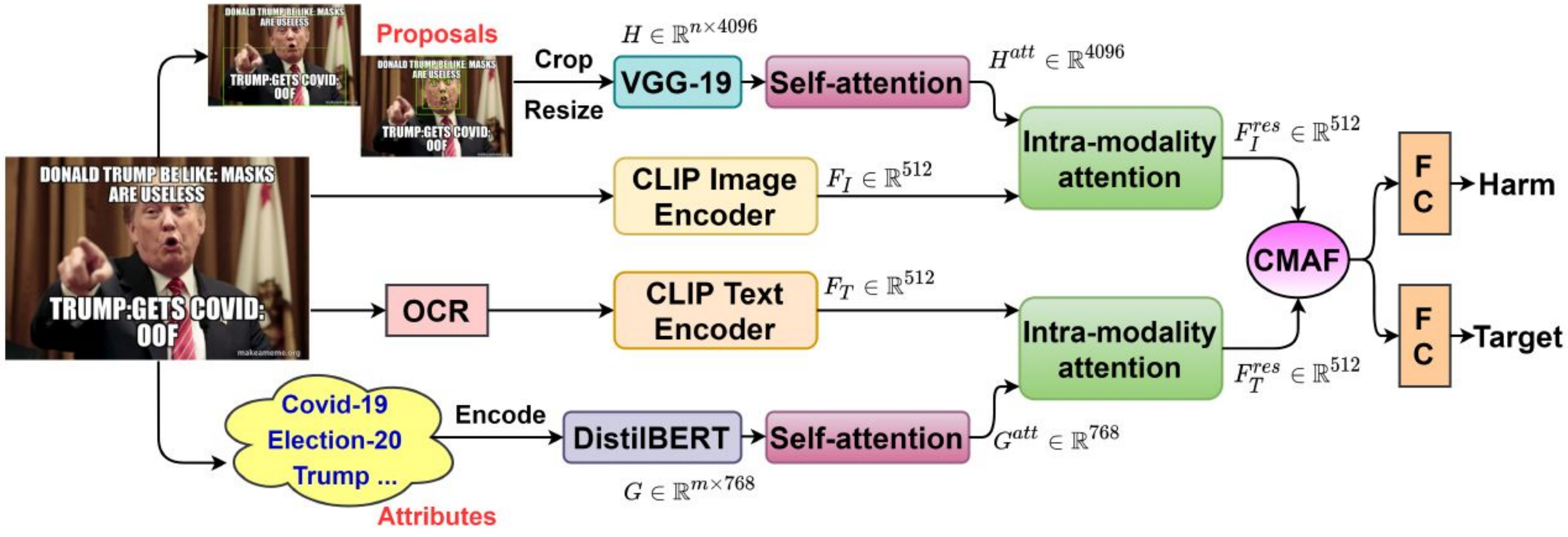
- In addition, we include meme object proposals (faces and foreground objects) and web attributes/entities.

- Intra-modality attention - object proposals + CLIP image features and web attributes/entities + CLIP text features

[1]Learning Transferable Visual Models From Natural Language Supervision, Radford et al., ICML '21

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

MOMENTA:



- We encode each image-text pair using CLIP[1], a pre-trained visual-linguistic model.

- In addition, we include meme object proposals (faces and foreground objects) and web attributes/entities.

- Intra-modality attention - object proposals + CLIP image features and web attributes/entities + CLIP text features

- Cross-modality attention fusion (CMAF) with two major parts: modality attention generation and weighted feature concatenation

[1]Learning Transferable Visual Models From Natural Language Supervision, Radford et al., ICML ’21

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

Evaluation:

- In terms of accuracy, we observe that MOMENTA achieves sizable improvements for the 2-class and 3-class tasks over the best multimodal models on both Harm-C and Harm-P datasets.

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	0.5000	64.76	26.20	1.0000	51.27	33.39	0.5000	51.27	22.59	1.0000
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**
$\Delta_{\text{MOMENTA}-best\_model}$		2.46	2.67	0.0114	1.39	0.89	0.0171	2.59	2.23	0.0038	2.48	1.96	0.0177

Performance of MOMENTA for harmful meme detection (2-class, 3-class) on both Harm-C and Harm-P datasets.

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

Evaluation:

- In terms of accuracy, we observe that MOMENTA achieves sizable improvements for the 2-class and 3-class tasks over the best multimodal models on both Harm-C and Harm-P datasets.

- The corresponding Macro-F1 scores also improve by a similar margin.

Modality	Model	Harm-C 2-Class Acc ↑	Harm-C 2-Class F1 ↑	Harm-C 2-Class MMAE ↓	Harm-C 3-Class Acc ↑	Harm-C 3-Class F1 ↑	Harm-C 3-Class MMAE ↓	Harm-P 2-Class Acc ↑	Harm-P 2-Class F1 ↑	Harm-P 2-Class MMAE ↓	Harm-P 3-Class Acc ↑	Harm-P 3-Class F1 ↑	Harm-P 3-Class 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	0.5000	64.76	26.20	1.0000	51.27	33.39	0.5000	51.27	22.59	1.0000
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**
$\Delta_{\text{MOMENTA}-best\_model}$		2.46	2.67	0.0114	1.39	0.89	0.0171	2.59	2.23	0.0038	2.48	1.96	0.0177

Performance of MOMENTA for harmful meme detection (2-class, 3-class) on both Harm-C and Harm-P datasets.

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

Evaluation:

- Similar trend is observed for target identification

Modality	Model	Target on Harm-C Acc ↑	F1 ↑	MMAE ↓	Target on Harm-P Acc ↑	F1 ↑	MMAE ↓
	Human†	87.55	82.01	0.3647	90.58	72.68	0.6324
	Majority	46.60	15.89	1.5000	56.47	18.05	1.5000
Text (T) only	TextBERT	69.35	55.60	0.8988	72.54	60.36	0.8895
Image (I) only	VGG19	63.48	53.60	1.0549	68.24	55.24	1.0225
	DenseNet-161	64.52	53.51	1.0065	69.40	57.95	0.9540
	ResNet-152	65.75	53.78	1.0459	68.75	57.00	0.9667
	ResNeXt-101	65.82	53.95	0.9277	70.22	59.67	0.9245
I + T (Unimodal Pretraining)	Late Fusion	72.58	58.43	0.6318	73.25	64.28	0.8541
	Concat BERT	67.74	49.77	0.8879	72.46	60.87	0.8655
	MMBT	72.58	58.35	0.6318	74.65	65.12	0.8441
I + T (Multimodal Pretraining)	ViLBERT CC	72.58	57.17	0.8035	77.25	**67.39**	**0.8410**
	V-BERT COCO	**75.81**	**65.77**	**0.5036**	**77.28**	66.90	0.8536
Proposed System and Variants	CLIP	72.47	62.14	0.6312	72.40	65.66	0.8557
	CLIP + Proposals	74.85	64.38	0.5746	75.85	66.13	0.8482
	CLIP + Attributes	74.56	61.38	0.6015	76.20	66.34	0.8491
	MOMENTA w/o CMAF	76.16	64.80	0.5422	77.54	67.25	0.8430
	MOMENTA	**77.95**	**69.65**	**0.4225**	**78.54**	**68.83**	**0.8295**
$\Delta_{\text{MOMENTA}-best\_model}$		2.14	3.88	0.0811	1.26	1.44	0.0115

Performance of MOMENTA for target identification of harmful memes on both Harm-C and Harm-P datasets.

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

Evaluation:

- Similar trend is observed for target identification
- MOMENTA outperforms the best models by 2.14 points absolute in terms of accuracy and by 3.88 points in terms of F1 score on Harm-C, and by 1.26 points of accuracy and 1.44 points of F1 on Harm-P

Modality	Model	Target on Harm-C Acc ↑	Target on Harm-C F1 ↑	Target on Harm-C MMAE ↓	Target on Harm-P Acc ↑	Target on Harm-P F1 ↑	Target on Harm-P MMAE ↓
	Human†	87.55	82.01	0.3647	90.58	72.68	0.6324
	Majority	46.60	15.89	1.5000	56.47	18.05	1.5000
Text (T) only	TextBERT	69.35	55.60	0.8988	72.54	60.36	0.8895
Image (I) only	VGG19	63.48	53.60	1.0549	68.24	55.24	1.0225
	DenseNet-161	64.52	53.51	1.0065	69.40	57.95	0.9540
	ResNet-152	65.75	53.78	1.0459	68.75	57.00	0.9667
	ResNeXt-101	65.82	53.95	0.9277	70.22	59.67	0.9245
I + T (Unimodal Pretraining)	Late Fusion	72.58	58.43	0.6318	73.25	64.28	0.8541
	Concat BERT	67.74	49.77	0.8879	72.46	60.87	0.8655
	MMBT	72.58	58.35	0.6318	74.65	65.12	0.8441
I + T (Multimodal Pretraining)	ViLBERT CC	72.58	57.17	0.8035	77.25	**67.39**	**0.8410**
	V-BERT COCO	**75.81**	**65.77**	**0.5036**	**77.28**	66.90	0.8536
Proposed System and Variants	CLIP	72.47	62.14	0.6312	72.40	65.66	0.8557
	CLIP + Proposals	74.85	64.38	0.5746	75.85	66.13	0.8482
	CLIP + Attributes	74.56	61.38	0.6015	76.20	66.34	0.8491
	MOMENTA w/o CMAF	76.16	64.80	0.5422	77.54	67.25	0.8430
	MOMENTA	**77.95**	**69.65**	**0.4225**	**78.54**	**68.83**	**0.8295**
$\Delta_{\text{MOMENTA}-best\_model}$		2.14	3.88	0.0811	1.26	1.44	0.0115

Performance of MOMENTA for target identification of harmful memes on both Harm-C and Harm-P datasets.

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

Transferability:

		Harm-C			Harm-P			Combined		
		H-2†	H-3‡	Tar*	H-2†	H-3‡	Tar*	H-2†	H-3‡	Tar*
Harm-C	ViLBERT	78.06	48.82	57.17	74.20	51.39	54.10	74.85	44.15	46.52
	V-BERT	80.13	53.85	68.77	74.56	52.87	53.46	75.04	45.20	47.66
	MOMENTA	**82.80**	**54.74**	**69.65**	80.25	61.87	58.39	81.66	49.83	50.12
Harm-P	ViLBERT	71.28	42.57	48.20	86.03	64.70	67.39	75.88	44.18	45.82
	V-BERT	72.58	45.10	54.07	86.07	63.68	66.90	76.20	45.69	47.38
	MOMENTA	76.30	50.46	58.33	**88.26**	**66.66**	**68.83**	80.75	49.70	50.28
Combined	ViLBERT	73.48	43.11	51.45	76.92	56.50	60.20	79.20	53.65	58.12
	V-BERT	74.88	46.28	60.82	76.85	56.07	58.22	80.45	53.98	58.76
	MOMENTA	79.50	51.07	62.56	81.09	62.85	61.87	**85.20**	**58.44**	**61.20**

Transferability of the two best-performing baselines and MOMENTA on Harm-C, on Harm-P, and on the combination.
The models are trained on the dataset in the row and tested on the one in the column. All scores are Macro F1.

- When training and testing on the same dataset, all models yield high F1 scores.

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

Transferability:

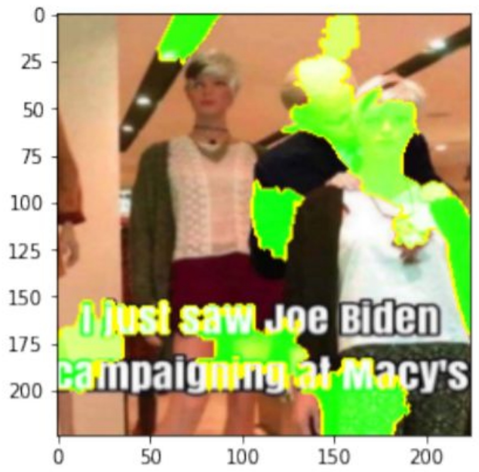
		Harm-C			Harm-P			Combined		
		H-2†	H-3‡	Tar*	H-2†	H-3‡	Tar*	H-2†	H-3‡	Tar*
Harm-C	ViLBERT	78.06	48.82	57.17	74.20	51.39	54.10	74.85	44.15	46.52
	V-BERT	80.13	53.85	68.77	74.56	52.87	53.46	75.04	45.20	47.66
	MOMENTA	**82.80**	**54.74**	**69.65**	80.25	61.87	58.39	81.66	49.83	50.12
Harm-P	ViLBERT	71.28	42.57	48.20	86.03	64.70	67.39	75.88	44.18	45.82
	V-BERT	72.58	45.10	54.07	86.07	63.68	66.90	76.20	45.69	47.38
	MOMENTA	76.30	50.46	58.33	**88.26**	**66.66**	**68.83**	80.75	49.70	50.28
Combined	ViLBERT	73.48	43.11	51.45	76.92	56.50	60.20	79.20	53.65	58.12
	V-BERT	74.88	46.28	60.82	76.85	56.07	58.22	80.45	53.98	58.76
	MOMENTA	79.50	51.07	62.56	81.09	62.85	61.87	**85.20**	**58.44**	**61.20**

Transferability of the two best-performing baselines and MOMENTA on Harm-C, on Harm-P, and on the combination.
The models are trained on the dataset in the row and tested on the one in the column. All scores are Macro F1.

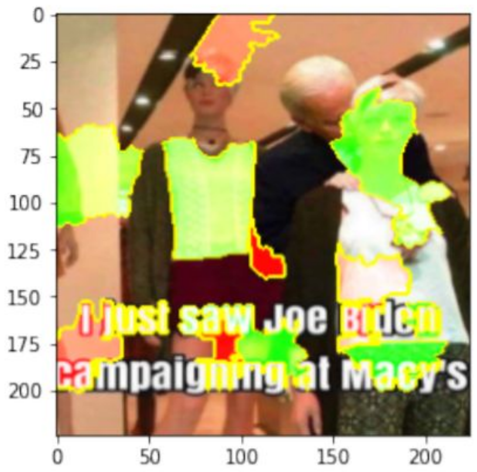
- When training and testing on the same dataset, all models yield high F1 scores.
- However, MOMENTA shows much better transferability capabilities. When trained on one dataset and tested on a different one, MOMENTA yields much better results both for harmful detection and for target identification.

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

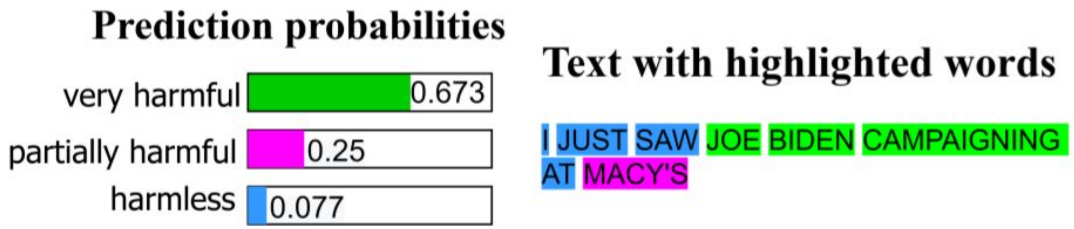
Analysis:



(a) LIME image - MOMENTA.



(b) LIME image - ViLBERT



(c) LIME text - MOMENTA.

- The fine-grained face detection and the robust CLIP encoder help MOMENTA to better identify subtle harmful elements in the image.

Example of explanation by LIME on both modalities for MOMENTA and ViLBERT.

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

Conclusion:

- We introduced two large-scale datasets, **Harm-C** and **Harm-P**, for detecting harmful memes and their targets.
- We benchmarked **Harm-C** and **Harm-P** against state-of-the-art unimodal and multimodal models.
- We proposed **MOMENTA**, a novel multimodal deep neural network that systematically analyzes the local and the global perspective of the input meme.
- Extensive experiments showed the efficacy of MOMENTA, which outperforms various state-of-the-art baselines for both tasks.
- We demonstrated model transferability and interpretability.
- In future work, we plan to extend the datasets with more domains and languages.

Thank You!