Multimodal Learning using Optimal Transport for Sarcasm and Humor Detection

Shraman Pramanick*, Aniket Roy*, Vishal Patel

Johns Hopkins University

WACV WAIKOLOA, HAWAII JANUARY 4-8, 2022

Problem Statement:

- Sarcasm is a fleeting action expressed through change of tone, overemphasis in some words, drawn-out syllable, or straight looking face.
- Sarcasm and Humor are very closely related sentiments.
- Multimodal cues (visual, textual and acoustic) are necessary to detect sarcasm in conversational videos.

- In videos, all three modalities contribute in detecting sarcasm and humor.
- As shown in this example, though the utterance of Chandler apparently seems to be an appreciation, his straight-looking face and tonal specific details make the utterance sarcastic.

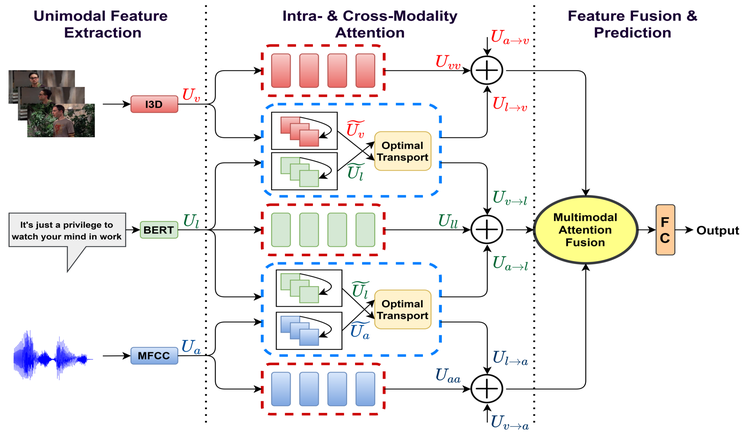
Challenges:

- For multi-modal sarcasm detection, synchronization across modalities through time is important.
- Feature dimensions across the modalities are different, therefore, feature fusion is challenging.

Our Approach:

- Feature Extraction:
  - Video: I3D features pretrained on Kinetics-400.
  - Text: Pretrained BERT features.
  - Audio: MFCC features.
  - Image: ResNet features.
- Multi-head self attention for intra modal correspondence.
- Cross-modal attention using Optimal Transport across modality.
- Finally, multi-modal attention fusion for prediction.

Our Model: MULOT:



Experimental Results:

- **Multimodal Sarcasm Datasets:**
  - For conversational videos (video + audio + text) : Multimodal Sarcasm Detection Dataset (MUStARD) , UR-FUNNY datasets.
  - For twitter posts (image + text) : Multimodal Sarcasm in Twitter Posts (MST) dataset.
  - Using MuLOT, we obtained the following performance improvements.

  On **MUStARD: 2.1%, UR-FUNNY: 1.54%, MST: 2.07%.**

Ablation Experiments:

- Visual and Language modalities are more important than Acoustic Modality.
- Both intra- & inter-modal interaction are important to detect Sarcasm Humor.

Modality	Algorithm	MUStARD Acc ↑	UR-FUNNY Acc ↑	Tiny UR-F Acc ↑	Algorithm	MST Acc ↑	MST F1 ↑	Tiny MST Acc ↑	Tiny MST F1 ↑
Trimodal	**MuLOT**	**78.57**	**73.97**	**71.46**	**MuLOT**	**90.82**	**88.52**	**88.04**	**85.93**
Unimodal	visual only	73.30	60.72	58.80	visual only	82.65	81.22	78.56	77.70
	language only	73.54	69.58	67.32	caption only	83.40	82.14	80.06	78.85
	acoustic only	64.00	64.35	55.44	OCR only	78.64	77.39	76.22	75.31
Bimodal	visual + language	77.18	70.40	69.40	visual + caption	87.35	85.93	83.94	82.45
	visual + acoustic	75.54	69.23	69.82	visual + OCR	85.66	84.37	82.30	81.38
	language + acoustic	75.72	72.10	69.12	caption + OCR	85.10	83.79	81.87	81.00
Trimodal	MuLOT w/o self-att^n	71.60	64.46	63.22	MuLOT w/o self-att^n	86.24	84.80	83.69	82.84
	MuLOT w/o cross-att^n	63.88	60.08	57.15	MuLOT w/o cross-att^n	82.32	80.20	80.07	79.28
	MuLOT w/o MAF	75.23	71.22	68.84	MuLOT w/o MAF	87.94	86.73	85.55	84.38

Quantitative Results:

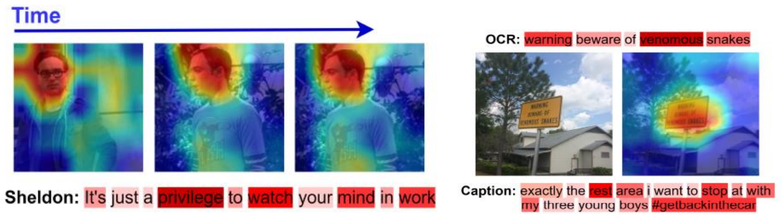
- MuLOT performs even better in limited-resource setup.
- The number of trainable parameters in MuLOT is 11 times lower than MMBT.

Algorithm	Context	Target	MUStARD Acc ↑	UR-FUNNY Acc ↑	Tiny UR-F Acc ↑
SVM	✗	✓	73.55	-	-
DFF-ATMF	✗	✓	64.45	62.55	56.35
CIM-MTL	✗	✓	67.14	63.20	56.71
TFN	✗	✓	68.57	64.71	57.23
CMFN (GloVe)	✗	✓	67.14	64.47	57.10
CMFN (GloVe)	✓	✓	70.00	65.23	59.25
MISA (BERT)	✗	✓	66.18	70.61	62.66
BBFN	✓	✓	71.42	71.68	63.20
MAG-XLNet	✓	✓	76.47	72.43	67.22
MuLOT	✗	✓	74.52	73.22	70.74
MuLOT	✓	✓	**76.82**†	**73.97**†	**71.46**†
$\Delta_{MuLOT-baseline}$			↑ 2.10	↑ 1.54	↑ 4.24

Algorithm	OCR	MST Acc ↑	MST F1 ↑	Tiny MST Acc ↑	Tiny MST F1 ↑
Concat BERT	✗	81.08	79.56	76.21	73.48
HFM	✗	83.44	80.18	77.80	74.07
D&R Net	✗	84.02	80.60	79.43	76.72
MMBT	✗	83.46	80.74	79.48	76.09
MMBT	✓	84.87	82.66	80.57	77.20
ViLBERT	✗	84.21	82.49	79.42	75.95
ViLBERT	✓	86.90	84.22	80.68	77.24
MsdBERT	✗	86.05	82.92	80.14	77.53
MsdBERT	✓	88.75	86.18	82.30	79.90
MuLOT	✗	87.41	86.33	84.46	82.62
MuLOT	✓	**90.82**†	**88.52**†	**88.04**†	**85.93**†
$\Delta_{MuLOT-baseline}$		↑ 2.07	↑ 2.34	↑ 5.74	↑ 6.03

Qualitative Results:

- We used GradCAM, to visualize the model inference.
- The model is focusing on facial expression of speakers.



Visual explanations and textual attention map for sarcastic utterances from the MUStARD dataset and sarcastic tweets from MST dataset.