Multimodal Learning using Optimal Transport for Sarcasm and Humor Detection (Supplementary)

In this supplementary document, we provide additional details on the hyperparameter values, baselines and illustrate additional qualitative results.

A. Implementation Details & Hyperparameter Values

In Table A.1, we furnish the details of hyperparameters used during training. Grid search is performed on batch size, learning rate, and the number of attention-heads to find the best hyperparameter configuration. The model is evaluated after each 10 steps on the dev set and the best model was taken to be evaluated on the test set. We use the weighted Adam optimizer [32] with step learning rate scheduler and fixed 2000 warmup steps for optimization without gradient clipping.

Hyper-parameter	**Notation**	**Value**
#dim for Dense layers in MAF	-	$[64, 8, 4, 1]$
#dim for Classification FC	-	$[192, 32, 8, 2]$
Intra-modality Attention		
#heads	k	80
#dim in Equation 1	r	30
Cross-modality Attention		
Shared feature sequence length	L_{uni}	100
Shared feature dimension	d_{uni}	64
Training		
Batch-size	-	64
Epochs	N	300
Optimizer	-	AdamW
Loss	-	BCE
Learning-rate	α	0.005
Learning-rate-decay (/10K$iter$)	-	1e-4
Betas	[β_1, β_2]	[0.9, 0.999]
Weight Decay	-	0.01
Class weights for imbalanced training data		
Training on MST $[w_{sar}, w_{non-sar}]$		$[1.2, 1]$

Table A.1: Hyper-parameters of MuLOT.

B. Baselines

Numerous methods have been proposed in the literature for multimodal sentiment analysis, but none of them has generalized their systems both on videos and images. Therefore, we chose different baselines for videos and images. Furthermore, we remove various modalities and modules at a time from our proposed MuLOT system to observe the effect in performance. Baselines on MUStARD and UR-FUNNY include:

- **Support Vector Machines (SVM)** [11] was used as the baseline model for MUStARD dataset [7]. However, they extracted individual frames from the videos and used ResNet [22] for extracting framewise visual features which does not take account the temporal dynamics of a video. We used I3D [6] for visual features, and retrain an SVM model with same configuration.
- **DFF-ATMF** [9] is a simple multimodal system which concatenates the unimodal features using an attention based modality fusion.
- **CIM-MTL** [1] leverage the interdependence of two related tasks (sentiment and emotion) in improving each others performance using an effective multi-modal attention framework.
- **Tensor Fusion Network (TFN)** [64] models intra-modality and inter-modality dynamics concurrently with local feature extraction network and 3-fold Cartesian product.
- **Contextual Memory Fusion Network (CMFN)** [20] proposes uni- and multimodal context networks that consider preceding utterances and performs fusion using the MFN model as its backbone. Originally, MFN [65] is a multi-view gated memory network that stores intra- and cross-modal utterance interactions in its memories.
- **MISA** [21] is flexible multimodal learning framework that emphasizes on multimodal representation learning as a pre-cursor to multimodal fusion.
- **Bi-Bimodal Fusion Network (BBFN)** [19] is an end-to-end network that performs fusion (relevance increment) and separation (difference increment) on pairwise modality representations, where the two parts are trained simultaneously such that the competition between them is simulated.
- **MAG-Transformer** [47] introduced Multimodal Adaption Gate (MAG) to fuse acoustic and visual information in pre-trained language transformers. Due to its superior performance on various multimodal sentiment analysis datasets, we ran MAG-XLNet on MUStARD and UR-FUNNY and it yields best baseline result on both datasets.

Baselines on MST include:

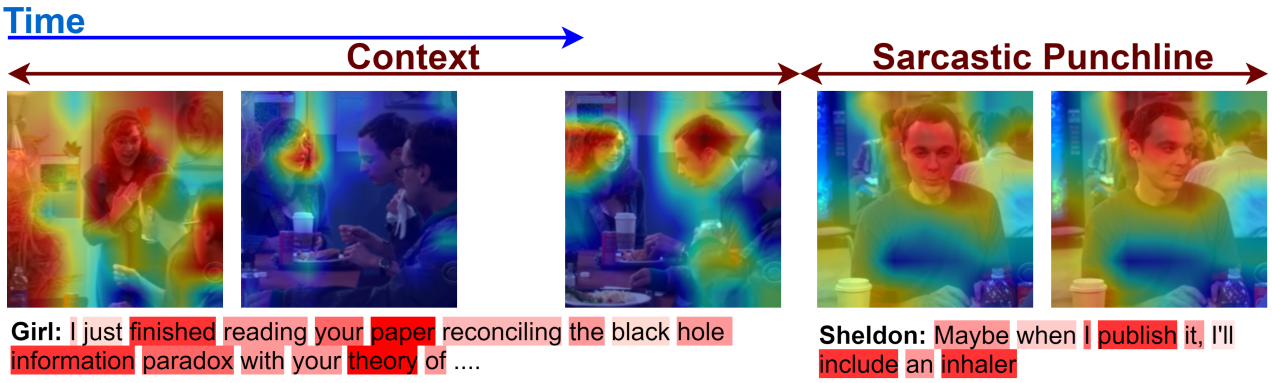

Figure A.1: Visual explanations and textual attention maps of the context and punchline of a sarcastic utterance from the MUStARD dataset.

Figure A.2: Visual explanations and textual attention maps of two sarcastic tweets and one non-sarcastic tweet from the MST dataset.

- **Concat BERT** concatenates the features extracted by pre-trained unimodal ResNet-152 [22] and Text BERT [14] and uses a simple perceptron as the classifier.
- **Supervised Multimodal Bitransformer (MMBT)** [27] captures the intra-modal and inter-modal dynamics within various input modalities.
- **Vision and Language BERT (ViLBERT)** [33], trained on an intermediate multimodal objective (Conceptual Captions) [50], is a strong model with task-agnostic joint representation of image + text.
- **Hierarchical Fusion Model (HFM)** [3] is similar to Concat BERT, but uses an hierarchical attention fusion mechanism for obtaining the multimodal feature representation.
- **D&R Net** [61] models cross-modality contrast and semantic association using Decomposition and Relation, where the decomposition network represents the commonality and discrepancy between image and text, and the relation network models the semantic association in cross-modality context.
- **MsdBERT** [37] is the SOTA on MST dataset. This system utilizes a co-attention network to exploit intra- and inter-modality incongruity between text and image.

C. Additional Qualitative Results

In this section, we illustrate a few more visual explanations over videos and images by using Grad-CAM [49] and plot the corresponding textual attention maps as generated by the multi-head attention layer of `MuLOT`. Figure A.1 shows five different frames from the context and punchline of a sarcastic utterance (video id: 1_1466) from MUStARD dataset. `MuLOT` focuses on the facial expression of the speaker in every frame. In the language modality, incongruous words to the speaker's expression are attended, and thus the system detects sarcasm. Figure A.2 shows two sarcastic tweets and one non-sarcastic tweet from MST

dataset. In the first tweet, the model attends the words "this is a stop sign" from OCR text and "I never knew" from the caption, and recognizes the writer's joking tone. Again, in the second example, MuLOT attends to the regions in image indicating lots of empty seats, which forms contradiction with the text "packed game". In the non-sarcastic tweet, no such incongruity is present.