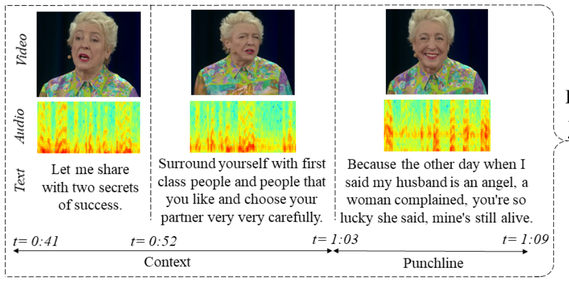
Humor Knowledge Enriched Transformer for Understanding Multimodal Humor

Md Kamrul Hasan (mhasan8@cs.rochester.edu), Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency and Ehsan Hoque

University of Rochester, Carnegie Mellon University & University of Michigan, USA

Abstract



1) Can a computer recognize the punchline of a joke using different modalities (text, acoustic & vision) and context ?
2) Will Humor Centric external Knowledge help?

Contributions

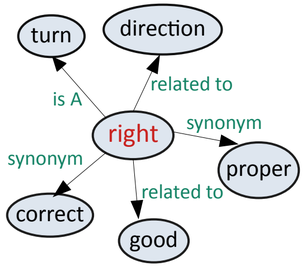
1. Extract Humor Centric Features (HCF) on word level
2. Design Humor Knowledge Enriched Transformer (HKT) model
3. Achieve state-of-art performance in multimodal humor & sarcasm detection
4. Identify humor inducing multimodal patterns

Humor Centric Features (HCF)

Ambiguity

Did you hear about the guy whose whole left side was cutoff? He's all right now.

- Different meanings of right: 'good' & 'direction' creates ambiguity
- ConceptNet (Liu 2004) used to extract different senses of each word
- Metric: summation of cosine distances of all pair of senses



Different senses of word 'right' in ConceptNet

Superiority

Don't you hate it when someone answers their own questions? I do.

- Humorous text often contains sentiment information
- For each word extracted: Valence, Arousal & Dominance
- Used NRC VAD dictionary (Mohammad 2018).

Datasets

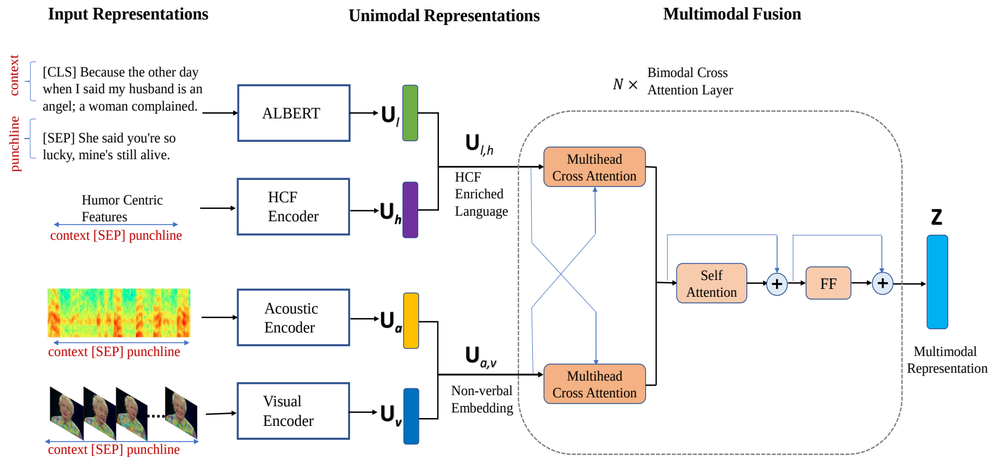
UR-FUNNY (Hasan et al. 2019)

- Multimodal dataset of humor detection
- Collected from TedTalk videos
- Video utterances with context & punchline

MUStARD (castro et al. 2019)

- Multimodal dataset of sarcasm detection
- Collected from sitcoms: Friends, The Big Bang Theory etc
- Video utterances with context & punchline

Humor Knowledge Enriched Transformer (HKT)



Model Punchline Conditioned on Context Story

Transformer Encoders: Learn Unimodal Representations

Bimodal Cross Attention Layer: Learn Joint Representation

Results

Performance of HKT Model

Binary Accuracy

Models	UR-FUNNY	MUStARD
C-MFN	65.23	-
MISA	69.82	66.18
MAG - XLNet	72.43	76.47
HKT	**77.36**	**79.41**
Δ SOTA	**4.93** ↑	**2.94** ↑

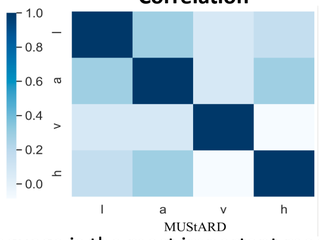
- MISA : Current SOTA for multimodal humor detection in UR-FUNNY dataset
- MAG-XLNet : Current SOTA for multimodal sentiment analysis in CMU-MOSI and CMU-MOSEI dataset
- **HKT** outperforms all the baselines

Role of Modalities

Binary Accuracy

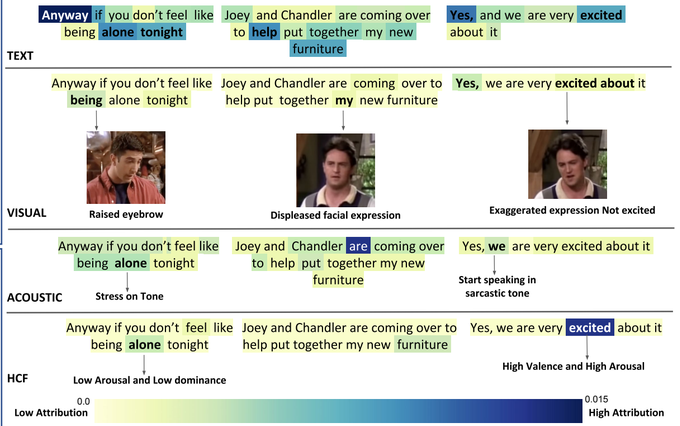
Models	UR-FUNNY	MUStARD
HKT	**77.36**	**79.41**
Language Only (l)	73.54	73.53
Remove acoustic (a)	74.14	76.47
Remove visual (v)	76.06	76.47
Remove HCF (h)	76.36	75.00

Correlation



- Language is the most important modality
- Output of unimodal encoders have low correlations → each component learning complementary information

Multimodal Humor Anchors



- Integrated gradients (Sundararajan 2017) used to measure the contribution of features to model's final decision.
- Darker color → higher importance

Project: https://roc-hci.com/current-projects/multimodal-humor-understanding/

UR-FUNNY Dataset : https://github.com/ROC-HCI/UR-FUNNY