Humor Knowledge Enriched Transformer for Understanding Multimodal Humor

Md Kamrul Hasan[1], **Sangwu Lee** [1], **Wasifur Rahman** [1], **Amir Zadeh** [2], **Rada Mihalcea** [3], **Louis-Philippe Morency** [2], **Ehsan Hoque** [1]

[1] Department of Computer Science, University of Rochester, USA,
[2] Language Technologies Institute, CMU, USA,
[3] Computer Science & Engineering, University of Michigan, USA
mhasan8@cs.rochester.edu, slee232@u.rochester.edu, echowdh2@ur.rochester.edu, abagherz@cs.cmu.edu, mihalcea@umich.edu, morency@cs.cmu.edu, mehoque@cs.rochester.edu

Abstract

Recognizing humor from a video utterance requires understanding the verbal and non-verbal components as well as incorporating the appropriate context and external knowledge. In this paper, we propose Humor Knowledge enriched Transformer (HKT) that can capture the gist of a multimodal humorous expression by integrating the preceding context and external knowledge. We incorporate humor centric external knowledge into the model by capturing the ambiguity and sentiment present in the language. We encode all the language, acoustic, vision, and humor centric features separately using Transformer based encoders, followed by a cross attention layer to exchange information among them. Our model achieves 77.36% and 79.41% accuracy in humorous punchline detection on UR-FUNNY and MUStaRD datasets – achieving a new state-of-the-art on both datasets with the margin of 4.93% and 2.94% respectively. Furthermore, we demonstrate that our model can capture interpretable, humor-inducing patterns from all modalities.

Introduction

Ever wondered the difficulty associated for a computer algorithm to recognize the punchline of a joke? People who are funny tend to be creative as well. They experiment with words (language), gestures (vision), prosody (acoustic) and their (mis)alignments to build up a story while cross-referencing multiple sources and appropriately delivering the punchline. It is an important communication skill that removes barriers in conversations, builds trust (Vartabedian and Vartabedian 1993) and creates positive impact on mental health (Lefcourt and Martin 2012). Humor has various styles like sarcasm, exaggeration, irony, satire etc. Understanding humor in everyday communication can enable machines to adapt its behavior seamlessly while interacting with the humans, leading to a smooth and enriched user experience.

A humorous punchline is often built around background context and external commonsense knowledge. Speakers deliberately use ambiguous and sentiment evoking words to prime the audience to elicit a delightful laughter. They build up expectations in the minds of their audiences and at the opportune moment, introduce a sudden twist, funny gesture or sarcastic tone to deviate from the expectation of the story (Ramachandran 1998). Humans can naturally process all these information subconsciously. However, building an algorithm that can potentially do the same requires appropriate integration of all these disparate sources of information.

We propose to model humor centric features by capturing the diversity of meanings and the sentiment expressed in each word using ConceptNet (Speer, Chin, and Havasi 2017) and the NRC_VAD (Mohammad 2018) lexicon respectively. We also capture modality specific nuances for language, acoustic, vision, and humor centric features separately. To represent the language modality, we fine-tune a pre-trained Albert model (Lan et al. 2019). For other modalities, we train Transformer (Vaswani et al. 2017) based encoders from scratch. We enrich the language modality with humor-centric features, and then capture the cross-modality interactions using a *Bimodal Cross Attention Layer* to build a singular representation of a data instance. Since modalities often carry both complementary and supplementary information, it is crucial to model them jointly to capture their underlying interactions. During all these phases, we model the punchline in light of the context so that we can accurately capture the inter-dependency between them. We call our model **H**umor **K**nowledge enriched **T**ransformer (HKT). The main contributions of our paper are:

- We derive humor centric features (HCF) at word level by incorporating several humor theories and common sense knowledge.
- We propose HKT – a transformer based model that learns to represent a punchline in light of the background context. The model has modality specific encoders for attending to each modality and a *Bimodal Cross Attention Layer* to jointly represent pairs of modality groups effectively.
- The HKT model outperforms the state-of-the-art baselines on two different multimodal datasets of humor (Hasan et al. 2019) and sarcasm (Castro et al. 2019) detection. We perform extensive experiments to demonstrate that humor centric features, background context and cues from all three modalities are important to understand the humor.

Background

In this section, we focus on the recent research on both text-based and multimodal humor understanding. As our model is largely based on Transformer (Vaswani et al. 2017) architectures, we discuss some examples on how they have been expanded to model multiple modalities.

Text-based Humor Analysis: Early works have focused on extracting and analyzing interpretable features based on several humor theories of incongruity, ambiguity, superiority, and phonetic style (Yang et al. 2015; Miller and Gurevych 2015; Mihalcea and Strapparava 2005; Liu, Zhang, and Song 2018; Zhang and Liu 2014). The discourse of sentiments in text also plays important role in recognizing humor (Liu, Zhang, and Song 2018). One noteworthy effort has been taken by Yang et al. (Yang et al. 2015) to identify humor anchors – the most pivotal text segments in generating humor. Recent works have started to focus on deep learning based models for humor detection. Convolutional Neural Network (CNN) based model (Chen and Lee 2017) is used to detect laughter in TedTalk transcript. As a follow up, Chen et al. (Chen and Soo 2018) designs a CNN based model with highway network that achieves state-of-the-art performance on four text based humor datasets. Language model fine-tuning is applied to classify humor in a large dataset containing 300k Russian short texts (Blinov, Bolotova-Baranova, and Braslavski 2019). Transformer and BERT based architectures are also used to study humor in text (Weller and Seppi 2019; Annamoradnejad 2020).

Multimodal Humor Analysis: Due to the availability of large number of video content, researchers have started to study humor in a multimodal manner. Bertero et al. (Bertero and Fung 2016) introduce a dataset containing acoustic and text from the TV-show "Big Bang Theory" and apply Recurrent Neural Network to detect humor. UR-FUNNY (Hasan et al. 2019) is a publicly available multimodal (textual, acoustic and visual) humor dataset that is collected from TED-Talks. The dataset has punchline-context setup and the authors extend Memory Fusion Network (Zadeh et al. 2018) to incorporate context information for predicting humorous punchline. A multimodal sarcasm dataset named MUStARD (Castro et al. 2019) is collected from popular sitcom TV shows. They also provide the preceding context of each sarcastic punchline and conduct extensive evaluation. MISA (Hazarika, Zimmermann, and Poria 2020) aggregates modality-invariant and modality-specific representations and has been applied to predict humor in the UR-FUNNY dataset.

Multimodal Analysis: Learning the joint representation of multimodal data has been an active research area in NLP community (Wang et al. 2019; Pham et al. 2019; Hazarika et al. 2018; Poria et al. 2017; Zadeh et al. 2017; Liang et al. 2018; Tsai et al. 2018; Liu et al. 2018; Zadeh et al. 2018; Islam and Iqbal 2021). Recently, Transformer (Vaswani et al. 2017) based models have gained success in modeling multiple modalities. Sun et al. (Sun et al. 2019) learn joint representation of video segments and their accompanying texts from a cooking video dataset. Multimodal Transformer (Tsai et al. 2019) uses a set of transformer encoders to capture both unimodal and cross modal interactions. Similarly, Tan and Bansal (Tan and Bansal 2019) learn joint representation of text and visual through a Cross-modality Encoder. Rahman et al. (Rahman et al. 2020) integrate acoustic and visual information in the pre-trained transformers like BERT (Devlin et al. 2018) and XLNet (Yang et al. 2019). Word representations of the language models are shifted conditioned on nonverbal features by fine-tuning.

Humor Centric Features Extraction

We extract humor-centric features based on the ambiguity and superiority theories.

Ambiguity

Ambiguity occurs when a sentence expresses multiple meanings simultaneously. It can be achieved through crafting a sentence with ambiguous words. Such sentences can have both serious and funny interpretations, generating humor in that process (Charina 2017). One such example is: *Did you hear about the guy whose whole left side was cut off? He's all right now.* In this example, the word 'right' can have two primary meanings: 'good' and 'direction'. Based on the meaning we perceive to be true, we will interpret the sentence very differently, and may experience ambiguity driven humor.

Yang et al. (Yang et al. 2015) extracted the count of senses (meanings) of each word from WordNet (Fellbaum 2012) and used it as a feature for capturing ambiguity. Although the count of senses is a good starting point, we believe that finding the most frequently used senses of the words and the diversity among the senses are important to capture the ambiguity in that sentence. To that end, we use ConceptNet (Liu and Singh 2004) – a large-scale semantic structure that expresses the relationship between words and phrases through graphs. The nodes in the graph denote concepts (words or phrases) and the weighted labeled edges denote how the words are related and the confidence score of the relation. For each word, we extract neighbouring concepts after filtering out the edges with confidence score less than 1. For example, top weighted neighbours of the word 'right' are: turn, direction, correct, good, proper, etc.

For each word w in our dataset, we extract N_s senses (/concepts) and their corresponding Glove embeddings (Pennington, Socher, and Manning 2014). The Glove embeddings provide a 300 dimensional vector for each word: similar words will have similar vector representations. The matrix of all the senses of a word w is $S \in R^{N_s \times 300}$; where N_s is the number of senses/concepts. We take the summation of cosine distances of all pair of senses as a metric of ambiguity. Since cosine distance between two identical distance is zero, the metric will have higher value for more ambiguous words.

Sentiment

According to the superiority theory (Gruner 2017), humorous text often contains sentiment information and the transition of sentiments can be valuable in humor recognition (Liu, Zhang, and Song 2018). We extract valence (negative-positive), arousal (calm-excited), and dominance

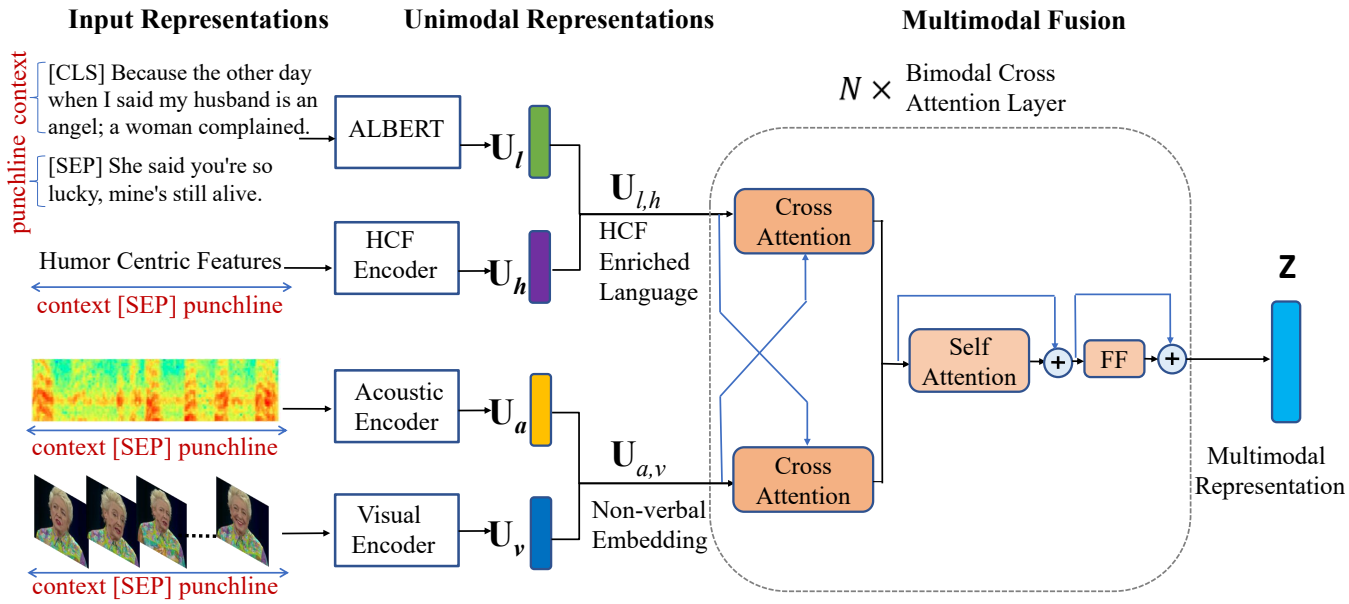

Figure 1: Humor Knowledge enriched Transformer (HKT)

(submissive-dominant) scores of each word from the NRC VAD dictionary (Mohammad 2018). This dictionary provides the above mentioned scores (in the range [0,1]) for 20k English words.

We denote the ambiguity, valence, arousal and dominance scores extracted for each word as **HCF** - Humor Centric Features (h) to be used as an additional information.

Humor Knowledge Enriched Transformer (HKT) Model

In this section, we outline the **H**umor **K**nowledge enriched **T**ransformer (HKT) model (Figure 1). First, a set of encoders create unimodal representations of punchline conditioned on context. Then, humor centric feature enriched language and non-verbal embedding go through Bimodal Cross Attention layers (details in Figure 2) to create multimodal fusion.

Task Definition

If our dataset has N data-points, we can represent the i-th data as $\{X^i = (C^i, P^i), Y^i\}$ where $i \in [1, ...N]$. Here, C^i, P^i and Y^i are the context, punchline and the label (Humor/ not Humor) associated with i-th datapoint respectively. Both the punchline and context sequences consist of four modalities: language (l), acoustic (a), vision (v) and humor centric features (h). We align the acoustic and visual features with their corresponding tokens in the language modality; therefore, the language, acoustic and vision sequences have the same length.

The total length of the i-th datapoint is $\tau^i = \tau^i_p + \tau^i_c$, where τ^i_p= punchline sequence length and τ^i_c= context sequence length. Since τ^i will have different values for different i values, we can truncate context or pad them with zero (to the right) to make sure that all datapoints have a fixed length τ. We represent the punchline of i-th datapoint as $P^i = (P^i_l, P^i_a, P^i_v, P^i_h)$; where $P^i_l \in R^{\tau^i_p \times d_l}$, $P^i_a \in R^{\tau^i_p \times d_a}$, $P^i_v \in R^{\tau^i_p \times d_v}$ and $P^i_h \in R^{\tau^i_p \times d_h}$. Here, d_l, d_a, d_v and d_h are the dimensions of the language, acoustic, visual and HCF features respectively. Similarly, the context of i-th datapoint can be represented as $C^i = (C^i_l, C^i_a, C^i_v, C^i_h)$; where $C^i_l \in R^{\tau^i_c \times d_l}$, $C^i_a \in R^{\tau^i_c \times d_a}$, $C_v \in R^{\tau^i_c \times d_v}$ and $C_h \in R^{\tau^i_c \times d_h}$.

Given a context (C^i) and punchline (P^i), our task is to predict whether the label (Y^i) is humorous or not. For achieving that goal, we will maximize the following function ϕ:

$$\phi = \prod_{i=1}^{N} p(Y^i | P^i, C^i; \theta) \quad (1)$$

In Eq. 1, ϕ represents the product of the conditional probabilities of determining the correct label given the punchline and context; θ denotes the model parameters that we want to train.

Unimodal Representation Learning

We fine-tune a *pre-trained* Albert (Lan et al. 2019) encoder for representing the language (l) only. For the other three modalities, we train a modified transformer encoder (Vaswani et al. 2017).

Language Representation: To convert the text token into vectors, we feed both the context and punchline together to Albert. We represent our language modality as: $X_l = [CLS]C_l[SEP]P_l$; where $X_l \in R^{\tau \times d_l}$, C_l=tokens of context, P_l=tokens of punchline and τ= total length of the token sequence. In essence, the $[CLS]$ token appended at the beginning will be used by the Albert encoder to create

a vector representing the whole input X_l, and the [*SEP*] token separates out the two sources of information – context and punchline – so that the punchline is modelled in light of the context. This representation is same as the one used in (Devlin et al. 2018) to model question-answering task. Albert encoder output the unimodal language representation U_l = ALBERT(X_l); where $U_l \in R^{\tau \times d_l^u}$ and d_l^u = output dimension of the Albert encoder.

Acoustic, Visual and Humor Centric Feature Representations: Transformer encoders are used to learn the unimodal representations of acoustic (a), vision (v) and HCF (h).

Transformer Encoder Layer contains a Multihead self-attention sub-layer and a Feed Forward (FF) sub-layer. The self-attention layers calculate the weighted summation of values; where the weights are computed from the scale dot product of query and key vector.

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_h}})V \quad (2)$$

Multiple self-attention layers operating in parallel, hence the name Multi-Head Self Attention – each potentially focusing on complementary aspects of the input. Following the convention of (Vaswani et al. 2017), we add layer normalization and residual connections after each sub-layers. N_a, N_v and N_h is the number of encoder layers are used in the Acoustic, Visual and HCF encoder respectively. To align the input representation with language, we create an input representation X_m for the modality m: $X_m = [PAD]C_m[SEP]P_m$.

$m = \{a, v, h\}$ represents the acoustic, visual and HCF respectively; where $X_m \in R^{\tau \times d_m}$ and d_m = the dimension of features in the corresponding modalities. [*PAD*] is used as a placeholder token mimicking the [*CLS*] token used for language, and [*SEP*] token is used to separate the context tokens from the punchline tokens. We send each input sequence X_m to the modality-specific encoder to obtain the unimodal representation U_m = TransformerEncoder(X_m). In the above equation, $U_m \in R^{\tau \times d_m^u}$ and d_m^u = output dimension of the transformer encoder in the corresponding modality $m \in [a, v, h]$.

Grouping together modality information: As discussed in the preceding sections, Albert and the Transformer Encoders give the unimodal representations of language (U_l), acoustic (U_a), vision (U_v) and HCF (U_h). To infuse the language information with the knowledge gained from humor centric features, we create an HCF-Enriched language representation $U_{l,h} = U_l \oplus U_h$; $\oplus$ represents concatenation and $U_{l,h} \in R^{\tau \times (d_l^u + d_h^u)}$. Similarly, we combine acoustic and visual representations to create a non-verbal embedding $U_{a,v} = U_a \oplus U_v$ where $U_{a,v} \in R^{\tau \times (d_a^u + d_v^u)}$.

Bimodal Cross Attention Layer

A Bimodal Cross Attention Layer is added to learn the joint representation of $U_{l,h}$ and $U_{a,v}$ (Figure 1). We modified the original transformer encoder layer (Vaswani et al. 2017) to fuse the information across two modalities (Detailed architecture is shown in Figure 2). For the sake of brevity, the rest of this section will assume that we instantiate this layer with two vectors (U_{m1},U_{m2}): representing

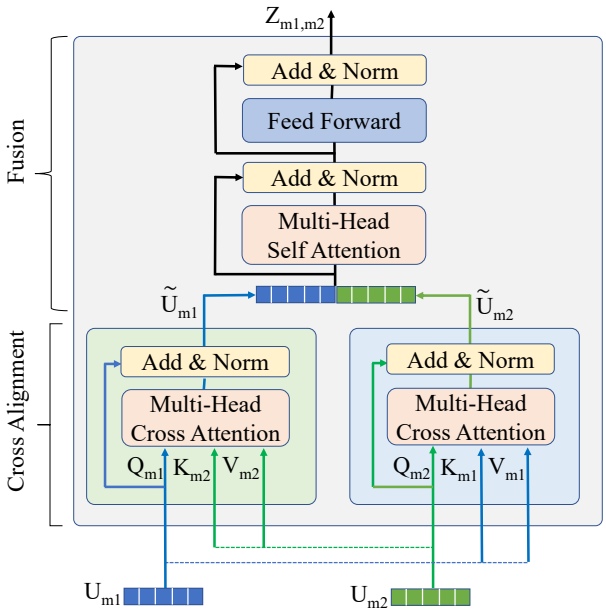

Figure 2: Bimodal Cross Attention Layer

modalities $m1$ and $m2$. However, these vectors can represent a group of modalities as well. For example, we assume that $U_{m1} = U_{l,h}$ and $U_{m2} = U_{a,v}$ in the model presented in Figure 2.

To exchange information between the two vectors (U_{m1},U_{m2}), we create two sets of queries (Q_{m1}, Q_{m2}), keys (K_{m1}, K_{m2}) and values (V_{m1}, V_{m2}) matrices following the convention of standard transformer (Vaswani et al. 2017). Each set is be attached to one of the two Multihead Cross Attention sub-layers. The sub-layers exchange the key and value matrices to compute the cross-aligned heads using the Equation 2. In essence, each sub-layer will create a vector representation while absorbing information from its pair. Similar approach of accomplishing cross-alignment among modalities has been studied in other language-vision tasks as well (Tan and Bansal 2019; Lu et al. 2019).

Although each of the outputs of the two Multihead Cross Attentions, ($\tilde{U}_{m1}, \tilde{U}_{m2}$) contain information about the other, we need to facilitate further exchange of information to build a more contextual and unified vector representation. For achieving that goal, we concatenated and passed them to a Multihead Self Attention followed by a Feed Forward sub-layer. Multihead Self Attention updates the representation of each element of input in light of the information gained from all the other elements and help us create the final joint representation $Z_{m1,m2}$. We add layer normalization and residual connections after each sub-layers as well.

Multimodal Fusion

The Bimodal Cross Attention Layer is used to create a fusion vector of humor-centric-feature enriched language embeddings ($U_{l,h}$) and non-verbal embeddings ($U_{a,v}$) shown in Figure 1. That fusion vector is: Z =

BimodalCrossAttention$(U_{l,h}, U_{a,v})$. In order to create a single vector representation unifying all components of our model, we create five vectors:$[e_l, e_a, e_v, e_h, e_z]$; e_l is the vector corresponding to the $[CLS]$ token in Albert model, and the rest of them are created by applying Max-pooling layers on $[U_a, U_v, U_h, Z]$ respectively. Max-pooling gives us a computationally efficient method of extracting the most salient features across the time dimension and yields a fixed dimensional vector. Finally, we concatenate all these representations to get the final embedding $o = e_l \oplus e_a \oplus e_v \oplus e_h \oplus e_z$; $o \in R^{d_o}$. The output probability is computed as $p = \textit{softmax}(oW + b)$, where $W \in R^{d_o \times l}$ and $b \in R^l$ denote model parameters and l denotes the number of classes.

Experiments

In this section, we discuss our experimental methodology: datasets we use, features we extract and baselines we compare with.

Datasets

We work with datasets that have language, acoustic and vision modalities and the preceding context of the punchline with the humor(/sarcastic) label. Only UR-FUNNY and MUStaRD fulfill these criteria.

UR-FUNNY: The UR-FUNNY (Hasan et al. 2019) is collected from TED talk videos and therefore, has language, acoustic and visual modalities and the context preceding the punchline. Punchline is extracted using the 'laughter' markup – indicating when audience laughed during the talk – in the transcripts. The sentences preceding the punchline form the context. Negative samples are also extracted in similar manner where target punchline utterances are not followed by 'laughter'. In total, the dataset consists of 5K humor and 5K non-humor instances (from 1741 distinct speakers). Version 2 of the UR-FUNNY dataset is used for all experiments.

MUStARD: Multimodal Sarcasm Detection Dataset (MUStARD) (Castro et al. 2019) is compiled from popular TV shows like Friends, The Big Bang Theory, The golden Girls and Sarcasmaholics. It provides 690 video segments that are manually annotated with sarcastic/non-sarcastic labels. They provide target punchline utterance and the associated historical dialogues as context.

Feature Extraction

The following standard features are used by the baseline models and our **HKT** model.

Language: Albert (Lan et al. 2019) language model is fine tuned to learn the contextual word representations. P2FA forced alignment model (Yuan and Liberman 2008) is used to extract the timing of all the words used in punchline and context. Once we extract the acoustic and visual features for the whole video, we use each word's timing to slice off the relevant range of acoustic and visual features for that word. Those two feature arrays are averaged out across the time dimension separately and in the end, we get the acoustic and visual feature vectors for each word (Chen et al. 2017). We extract 'senses' of word from ConceptNet and use GloVe embeddings (Pennington, Socher, and Manning 2014) to measure ambiguity.

Acoustic: We use COVAREP (Degottex et al. 2014) to extract low-level acoustic features. This feature set includes Melcepstral coefficients, fundamental frequency, voiced/unvoiced segments , normalized amplitude quotient, quasi open quotient (Kane and Gobl 2013), glottal source parameters (Drugman et al. 2012), harmonic model and phase distortions, the formants etc.

Visual: OpenFace 2 (Baltrusaitis et al. 2018) is used to extract facial Action Units (AU) features and Rigid and non-rigid facial shape parameters. Facial action unit features are based on the Facial Action Coding System (FACS) (Ekman 1997) which are widely used in human affect analysis.

Baseline Models

The performance of our **HKT** model is compared with the following baselines:

Contextual Memory Fusion Network (C-MFN) (Hasan et al. 2019) was used to detect humor punchlines in UR-FUNNY dataset. They extended the Memory Fusion Network (Zadeh et al. 2018) by incorporating the information from the preceding context.

Support Vector Machines (SVM) was used as the baseline model for MUStARD dataset (Castro et al. 2019). They used ResNet (He et al. 2016) and Librosa (McFee et al. 2018) to extract visual and acoustic features respectively and did not align all three modalities. While we have attempted to extract the acoustic and visual features using the Covarep and Openface and align the modalities, we lost 14 samples. To present a fair comparison, we extract all the features (mentioned above) and retrain an SVM model on the train, dev, and test sets that we define for MUStARD.

MISA (Hazarika, Zimmermann, and Poria 2020) achieved SOTA performance on the UR-FUNNY dataset by projecting their data to a modality invariant and three modality-specific spaces and then aggregating all those projections. They used BERT language encoder and worked with punchline only. For fair comparison, we rerun MISA by concatenating both punchline and context. As our model use ALBERT language encoder, so we run a variant of MISA with ALBERT. We experiment with these variants of MISA on both UR-FUNNY and MUStARD datasets.

MAG-Transformer (Rahman et al. 2020) introduced Multimodal Adaption Gate (MAG) to fuse acoustic and visual information in pretrained language transformers. During fine tuning, the MAG shifts the internal representations of BERT and XLNet in the presence of the visual and acoustic modalities. Both the MAG-BERT and MAG-XLNet achieved SOTA performance in CMU-MOSI and CMU-MOSEI datasets of multimodal sentiment analysis. Moreover, the MAG-XLNet achieved human level performance on the CMU-MOSI dataset and outperformed all the state of the art multimodal fusion models. We apply the MAG-XLNet on both UR-FUNNY and MUStARD datasets due to it's superior performance. For fair comparison with our model, we also run a variant of MAG-Transformer where we use ALBERT pretrained transformer.

Multimodal Models	UR-FUNNY	MUStARD
C-MFN (Glove)	65.23	-
C-MFN (Albert)	61.72	-
SVM	-	71.6
MISA (BERT) [punchline only]	70.61	-
MISA (BERT)	69.62	66.18
MISA (ALBERT)	69.82	66.18
MAG-ALBERT	67.20	69.12
MAG-XLNet	72.43	76.47
HKT	**77.36**	**79.41**
Δ SOTA	**4.93** $\uparrow$	**2.94** $\uparrow$

Table 1: Performances (binary accuracy) of multimodal models on the UR-FUNNY & MUStARD datasets.

Models	UR-FUNNY	MUStARD
HKT	**77.36**	**79.41**
- acoustic (a)	74.14	76.47
- visual (v)	76.06	76.47
- HCF (h)	76.36	75.00
language only (l)	73.54	73.53
acoustic only (a)	64.99	73.53
visual only (v)	55.84	64.71
HCF only (h)	56.54	60.29

Table 2: Role of modalities in our HKT model. Here '-' denotes removal of the corresponding feature set. Binary accuracy is reported here as performance.

Experimental Design

Adam optimizer and Linear scheduler are used to train the **HKT** model. We use different learning rates for language, acoustic, visual and HCF encoders. The search space of the learning rates is $\{0.001, 0.0001, 0.00001, 0.000001\}$. Binary cross entropy is used as loss function. We experiment with $\{1, 2, 3, 4, 5, 6, 7, 8\}$ encoder layers and $\{1, 2, 3, 4, 6\}$ cross attention heads for the language, acoustic, visual and HCF encoders. For the Bimodal Cross Attention we experiment $\{1, 2\}$ layers and $\{1, 2, 4\}$ attention heads. Dropout $[0.05 - 0.30]$ (uniform distribution) is used to regularize the model. For other baseline models, we first experiment with best configurations that were presented in their respective papers. In addition, we run experiments with extensive hyper parameter search for fair comparison. We provide details of the best model configurations and hyper-parameters search spaces in the supplementary material [1]. In our framework, it is possible to reproduce the same experiment on K80 gpu for specific hyper-parameters and seed. Both the UR-FUNNY and MUStARD have balanced test set. Hence, we use **Binary Accuracy** as our performance metric.

Results and Discussions

In this section, we compare the performance of **HKT** model with the baselines, conduct ablation studies to show the importance of including multiple modalities and HCF features, and demonstrate our model's capability of capturing multimodal humor anchors.

Comparison with Baselines

Table 1 shows that the **HKT** model outperforms the baselines significantly on the UR-FUNNY dataset (4.93% increase) and MUStARD dataset (2.94% increase). The original C-MFN model was trained with Glove embeddings on UR-FUNNY dataset. We re-train the C-MFN model with the embeddings extracted from the pre-trained Albert model to ensure a fair comparison. However, it performed poorly on humorous punchline detection in UR-FUNNY. The MISA baseline model only used punchline of UR-FUNNY to predict humor (Hazarika, Zimmermann, and Poria 2020). Therefore, we train a variation of our HKT model by removing the context and achieved 71.33% accuracy (0.72% increase compare to MISA). However, our full model achieves 77.36% accuracy on UR-FUNNY that indicates the importance of context in detecting humorous punchline.

MISA (BERT) and MISA (ALBERT) that are trained on the full sequence of context and punchline do not achieve better performance. They perform worse than the punchline only MISA (BERT) model in UR-FUNNY dataset. The possible explanation is that MISA encodes the full temporal unimodal sequence into a single latent space and reconstructs the unimodal embeddings from the latent space. So, information might get lost during encoding the long sequence. The authors also worked with punchline only in their experiments for the UR-FUNNY dataset. Similarly, MISA variants perform worse than the SVM baseline in MUStARD dataset. This generative approach does not work well on the small dataset like MUStARD. In all of theses cases, we experiment with their reported best model configurations. Then additional extensive hyper-parameter search is done for fair comparison. However, MISA variants do not achieve reasonable performance on the long sequence of punchline followed by a context.

MAG-XLNet is the current state-of-the-art model for multimodal sentiment analysis in CMU-MOSI and CMU-MOSEI datasets. In here, it also achieves competitive performance compare to other baselines in both datasets. However, MAG-ALBERT does not achieve similar performance. Our **HKT** model achieves better result than MAG-XLNet and shows the importance of this kind of architecture for modeling multimodal humorous punchline in the light of background context and humor centric features.

Role of Different Components

Role of Modalities: we retrain our model by removing a set of modalities (one at a time) and thus show its importance in Table 2. Language manifests itself as the dominant modality in the UR-FUNNY dataset. On the other hand, both language and acoustic shows highest importance in MUStARD dataset. Although removing acoustic and visual features drops the performance for both datasets, the drop is very negligible for visual in the UR-FUNNY dataset. Since the cameras move a lot and seldom focuses on the faces

[1] https://github.com/matalvepu/HKT

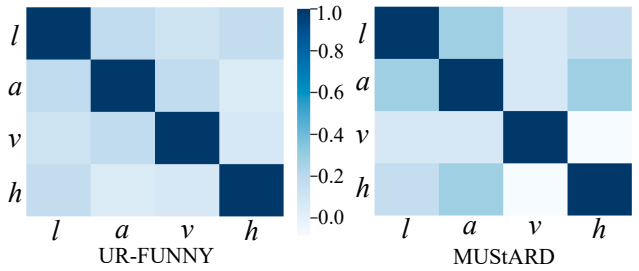

Figure 3: Correlation among the prediction outputs (dev & test set) of unimodal models.

of the speaker in TED-Talk recordings, the visual features tend to carry a lot of noises (Hasan et al. 2019). We find that 70% of the visual vectors are zero in UR-FUNNY dataset. Removing HCF is also consequential, especially for the MUStARD dataset. In our early exploration, we have experimented with transformer-based language encoder trained on Glove embeddings. ALBERT language encoder achieves better results (8% increase on UR-FUNNY), which is why we use the ALBERT encoder in our HKT model.

We try to understand how much modality-specific information our unimodal encoders can capture. Each encoders are trained separately with its modality-specific information only (while detaching it from other parts of the model). Next, we take the outputs from each of those of the unimodal encoders to predict humor/sarcasm (Table 2). These results are in accordance with the results mentioned in the preceding paragraph. The acoustic modality alone works surprisingly well for MUStARD dataset: we observe that actors exaggerate their voice to deliver sarcastic punchlines in sitcom shows, which can explain the superior performance of the acoustic modality. HCF features also capture some meaningful insight from the MUStARD dataset. Figure 3 shows how much complementary information is present in each modality with respect to the other modalities. We extract the predictions from each unimodal encoders and calculate the Pearson correlation among them. The correlation values are low, which indicates that each modality covers different aspects of information. The HKT model brings all this complementary information together, which can explain its higher accuracy when compared to previous models.

Bimodal Cross Attention Layers: We experiment with different numbers of Bimodal Cross Attention Layers. However, in both dataset we observe that increasing the number layers do not improve performance (highest 76.47% accuracy on MUStARD and highest 75.75% accuracy on UR-FUNNY). Specifically, in MUStARD dataset the model overfits very quickly due to high number of parameters compare to the small amount of data.

HCF: Integrated gradients technique (Sundararajan, Taly, and Yan 2017) is used to analyze the relative importance of the humor centric features in the model's inference. The model puts more weight (1.88 times higher) on the "Sentiment" features compared to the "Ambiguity". The test and dev split of the MUStARD dataset are used to compare.

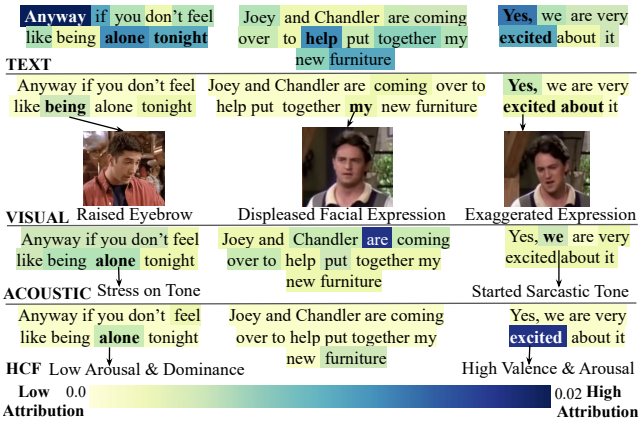

Figure 4: Multimodal Humor Anchors extracted by the HKT model. Integrated gradients (Sundararajan, Taly, and Yan 2017) method is used to decipher how each input token (across modalities) contributed to the model's final decision. Since acoustic, visual and HCF features are aligned on word level, we color coded the words to indicate the timestamps where model put attention on the corresponding modalities. For example, model puts highest attention to the visual features corresponding to the time-periods when the words 'being' and 'yes' were spoken. (Best view in color and zoomed)

Multimodal Humor Anchors

Humor anchors are the input tokens that play a pivotal role in creating humor (Yang et al. 2015). We want to see if our HKT model can identify humor anchors present in other modalities than text. Figure 4 shows the visualization of an example from MUStARD dataset (video id: 2_524). We use integrated gradients method (Sundararajan, Taly, and Yan 2017) to find the candidate tokens which have high impact in the model's decision making process. As the acoustic and visual modalities are aligned with text in word level, we know the timestamps where each feature resides. We get the attribution value (measure of impact) for each feature vector. Then we manually go through the video to understand how those high-attribution achieving features are related to humor. We have found that the model puts high attribution to meaningful patterns like eye brow raise, exaggerated facial expressions, stress on tone, high valence and arousal.

Conclusion

In this paper, we introduce HKT – a humor knowledge enriched multimodal model that can effectively learn the multimodal representation of a punchline conditioned on the context story. Our experiments show significant improvements in the task of humorous/sarcastic punchline detection on two publicly available datasets: UR-FUNNY and MUStARD. We demonstrate that context, humor centric features based on humor theories and cues from all three modalities are important for recognizing the humorous punchline. Additionally, we demonstrate that our model is able to find meaningful humor anchors across the modalities

Acknowledgments

This research was supported in part by grant W911NF-19-1-0029 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO). Authors AZ and LPM were supported in part by the National Science Foundation (Awards #1750439, #1722822) and National Institutes of Health.

References

Annamoradnejad, I. 2020. ColBERT: Using BERT Sentence Embedding for Humor Detection. *arXiv preprint arXiv:2004.12765* .

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 59–66. IEEE.

Bertero, D.; and Fung, P. 2016. Deep learning of audio and language features for humor prediction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 496–501.

Blinov, V.; Bolotova-Baranova, V.; and Braslavski, P. 2019. Large Dataset and Language Model Fun-Tuning for Humor Recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4027–4032.

Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; and Poria, S. 2019. Towards Multimodal Sarcasm Detection (An _Obviously_ Perfect Paper). *arXiv preprint arXiv:1906.01815* .

Charina, I. N. 2017. Lexical and Syntactic Ambiguity in Humor. *International Journal of Humanity Studies (IJHS)* 1(1): 120–131.

Chen, L.; and Lee, C. 2017. Predicting Audience's Laughter During Presentations Using Convolutional Neural Network. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 86–90.

Chen, M.; Wang, S.; Liang, P. P.; Baltrušaitis, T.; Zadeh, A.; and Morency, L.-P. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 163–171.

Chen, P.-Y.; and Soo, V.-W. 2018. Humor recognition using deep learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 113–117.

Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, 960–964. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Drugman, T.; Thomas, M.; Gudnason, J.; Naylor, P.; and Dutoit, T. 2012. Detection of glottal closure instants from speech signals: A quantitative review. *IEEE Transactions on Audio, Speech, and Language Processing* 20(3): 994–1006.

Ekman, R. 1997. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA.

Fellbaum, C. 2012. WordNet. *The encyclopedia of applied linguistics* .

Gruner, C. R. 2017. *The game of humor: A comprehensive theory of why we laugh*. Routledge.

Hasan, M. K.; Rahman, W.; Bagher Zadeh, A.; Zhong, J.; Tanveer, M. I.; Morency, L.-P.; and Hoque, M. E. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2046–2056. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1211. URL https://www.aclweb.org/anthology/D19-1211.

Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; and Zimmermann, R. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, 2122–2132.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. MISA: Modality-Invariant and-Specific Representations for Multimodal Sentiment Analysis. *arXiv preprint arXiv:2005.03545* .

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Islam, M. M.; and Iqbal, T. 2021. Multi-GAT: A Graphical Attention-based Hierarchical Multimodal Representation Learning Approach for Human Activity Recognition. *IEEE Robotics and Automation Letters* 1–1. doi:10.1109/LRA.2021.3059624.

Kane, J.; and Gobl, C. 2013. Wavelet maxima dispersion for breathy to tense voice discrimination. *IEEE Transactions on Audio, Speech, and Language Processing* 21(6): 1170–1179.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* .

Lefcourt, H. M.; and Martin, R. A. 2012. *Humor and life stress: Antidote to adversity*. Springer Science & Business Media.

Liang, P. P.; Liu, Z.; Zadeh, A.; and Morency, L.-P. 2018. Multimodal language analysis with recurrent multistage fusion. *arXiv preprint arXiv:1808.03920* .

Liu, H.; and Singh, P. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal* 22(4): 211–226.

Liu, L.; Zhang, D.; and Song, W. 2018. Modeling sentiment association in discourse for humor recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 586–591.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064* .

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 13–23.

McFee, B.; McVicar, M.; Balke, S.; Thomé, C.; Lostanlen, V.; Raffel, C.; Lee, D.; Nieto, O.; Battenberg, E.; Ellis, D.; et al. 2018. WZY. *Rachel Bittner, Keunwoo Choi, Pius Friesch, Fabian-Robert Stter, Matt Vollrath, Siddhartha Kumar, nehz, Simon Waloschek, Seth, Rimvydas Naktinis, Douglas Repetto, Curtis" Fjord" Hawthorne, CJ Carr, Joo Felipe Santos, JackieWu, Erik, and Adrian Holovaty,"librosa/librosa: 0.6* 2.

Mihalcea, R.; and Strapparava, C. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 531–538. Association for Computational Linguistics.

Miller, T.; and Gurevych, I. 2015. Automatic disambiguation of English puns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 719–729.

Mohammad, S. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 174–184.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Poczos, B. 2019. Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities. *arXiv preprint arXiv:1812.07809* .

Poria, S.; Cambria, E.; Hazarika, D.; Mazumder, N.; Zadeh, A.; and Morency, L.-P. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*, 1033–1038. IEEE.

Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A. B.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2359–2369.

Ramachandran, V. S. 1998. The neurology and evolution of humor, laughter, and smiling: the false alarm theory. *Medical hypotheses* 51(4): 351–354.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; and Schmid, C. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 7464–7473.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3319–3328. JMLR. org.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* .

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. *arXiv preprint arXiv:1906.00295* .

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176* .

Vartabedian, R. A.; and Vartabedian, L. K. 1993. Humor in the Workplace: A Communication Challenge. .

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, Y.; Shen, Y.; Liu, Z.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2019. Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors. *arXiv preprint arXiv:1811.09362* .

Weller, O.; and Seppi, K. 2019. Humor Detection: A Transformer Gets the Last Laugh. *arXiv preprint arXiv:1909.00252* .

Yang, D.; Lavie, A.; Dyer, C.; and Hovy, E. 2015. Humor recognition and humor anchor extraction. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2367–2376.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, 5753–5763.

Yuan, J.; and Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123(5): 3878.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* .

Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Memory fusion network for multi-view sequential learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zhang, R.; and Liu, N. 2014. Recognizing humor on twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 889–898. ACM.