Multimodal Laughter Reasoning with Language Models

Lee Hyun*
POSTECH
hyunlee@postech.ac.kr

Kim Sung-Bin*
POSTECH
sungbin@postech.ac.kr

Seungju Han
Seoul National University
wade3han@snu.ac.kr

Youngjae Yu
Yonsei University
yjy@yonsei.ac.kr

Tae-Hyun Oh
POSTECH
taehyun@postech.ac.kr

Abstract

Laughter is a substantial expression that occurs during social interactions between people. While it is essential to build social intelligence in machines, it is challenging for machines to understand the rationale behind the laughter. In this work, we introduce **Laugh Reasoning***, a new task that ascertains why a particular video induces laughter, accompanied by a new dataset and benchmark designed for this task. Our proposed dataset comprises video clips, their multimodal attributes including visual, semantic and acoustic features from the video, and language descriptions of why people laugh. We build our dataset by utilizing large language models' general knowledge and incorporating it into human consensus. Our benchmark provides a baseline for the laugh reasoning task with language models, and by investigating the effect of multimodal information, we substantiate the significance of our dataset. Code and dataset are available on https://github.com/SMILE-data/SMILE.*

1. Introduction

We, human beings, are immersed in laughter. Laughter is a distinctive non-verbal social signal, associated with bonding, agreement, affection, and emotional regulation [8]. It is often purposedly elicited to establish intimacy, grab attention, or build faith; *i.e.*, serving as a powerful medium to express a wide range of social and emotional implications beyond the capacity of mere words. Thus, understanding laughter is a crucial problem with huge potential in artificial social intelligence [2] to build empathetic machines with human-machine interaction. However, understanding and modeling laughter reactions is challenging.

Therefore, in this work, we take a stepping stone to tackle the challenge of understanding laughter by introducing a task, *Laugh Reasoning*, that aims to interpret the reasons behind laughter in a video. We probe through the question "why do the audiences laugh?" and reason through the answer in an unconstrained language form; thus, we define the task as a free-form text generation in which the model generates an explanation for the audience laughter with a given video clip, and we assess models according to the validity of the explanation. This task requires a paired dataset of video clips and textual reasons for laughter. Accordingly, we introduce a novel dataset, consisting of video clips and corresponding text annotations explaining laughter in each clip.

While reasoning laughter by answering the question is an effective way of probing the level of understanding, laughter itself has an inherently complex nature which can be influenced by diverse factors, *e.g.*, the subjectivity, context knowledge, and multimodality. To build a clearer resource of understanding laughter and its social norm behind it, we design dataset to focus on *audience laughter*, a cohesive form from social influence in distinct contexts, and thereby alleviate the subjectivity associated with individual laughter. Also, for our *Laugh Reasoning* task, we build baselines based on recent large language models (LLMs) and use text representation as a unified multimodal input representation by converting multimodal attributes into a textual format. These baselines deal with the requirement of sufficient context knowledge and multimodal capability, so that we can focus on the reasoning of laughter.

Given our dataset and baselines, we show the distinct characteristics of laughter across video types, highlighting the importance of our dataset, which includes a variety of laugh types. Also, our baselines show strong performance by incorporating multimodal cues in our dataset, highlighting the importance of the multimodal nature of laughters.

2. Datasets and Task Setups

We introduce a new dataset comprising 887 video clips and multimodal attributes, including visual, semantic, and acoustic features from them and the language description for

*equally contributed

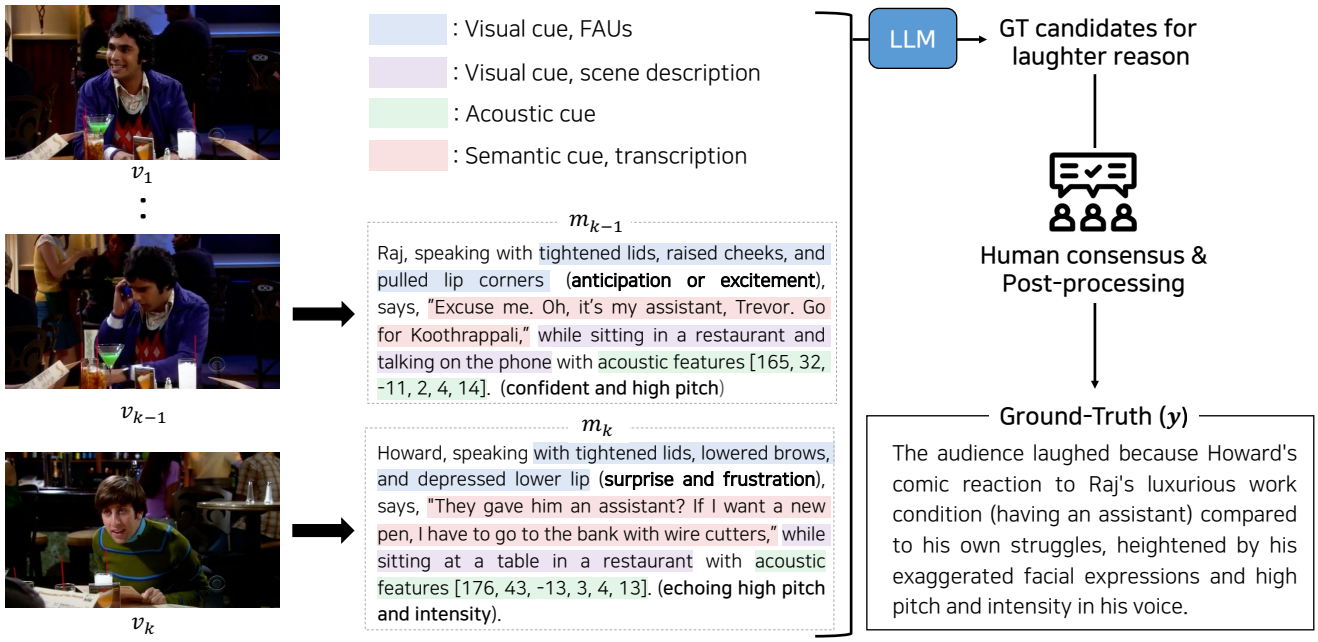

Figure 1. **Dataset curation pipleline.** Each video clip (d) is trimmed into list of video segments (v_i), and each video segment is encoded into multimodal textual representation (m_i). Then, we use LLM for creating a GT (ground-truth) reason candidates for laughter, and human annotators verify these candidates. Finally, we run a manual post-processing, and obtain the final GT reason (y) for laughter. The bold text in parentheses on the m shows that LLM is semantically aware of the multimodal textual representation.

laughter reasons. The dataset focuses on audience laughter among many types of laughter since audience laughter usually has a clearer signal than other laughter and represents a general and cohesive form of laughter.

To encompass a wider range of videos that contain the situation where audiences laugh, we construct our dataset using two different sources: TED talks and sitcoms.[1] TED talks employ humor more sparingly, often to enhance audience engagement. In contrast, sitcoms are intentionally scripted to evoke laughter, *e.g.*, frequently using punchlines and comedic scenarios. This heterogeneity allows our data to cover diverse situations that induce laughter.

2.1. Data Collection

We curate video clips that span between 10 and 90 seconds for TED talks and 7 and 60 seconds for sitcoms. If a video is too short, it might fail to provide necessary contexts for laughter. In contrast, if a video is too long, it may dilute specific laughter-inducing contexts with uncorrelated information. The average duration for TED talk clips is longer than sitcoms, given the protracted nature of talks.

Given that a single video clip often contains multiple instances of laughter, we focus on the last laugh in a clip for easier annotation. We only use video clips that meet the following filtering criteria, using a laugh detector [5] to identify audience laughter instances. Our filtering criteria are: laughter should last at least 0.5 secs., and be no more than 1 second interval between the video clip's last utterance and the onset of laughter. The latter criterion filters out the laughter events that are not related to the punchlines but are induced by something else.

[1]We source the video clips from MUStARD [4] and UR-Funny dataset [6].

2.2. Multimodal Video Encoding

Videos are multimodal, which include visual, acoustic, and semantic cues (*i.e.*, transcription). We encode video clips into textual representation, embracing their multimodal information, so that we can leverage the pre-trained knowledge of LLMs while exploiting multimodal inputs in our baselines. First, starting from a video clip, we build a list of video segments by trimming the clip based on the utterances. The definition of the utterance varies upon to the source of the video: for TED talks, each sentence is defined as an utterance, since TED talk usually has a single speaker. If the utterance is too short (2 seconds or less), we concatenate adjacent utterances into one. For sitcoms, we define consecutive sentences from the same speaker as an utterance.

We denote each video clip in our dataset as $d = \{(v_1, m_1), (v_2, m_2)), .., (v_k, m_k)\}$, where v stands for a video segment trimmed by utterance and m denote multimodal textual representation that includes visual, acoustic and transcription for the corresponding video segment v (See Figure 1). Note that we use transcription for semantic cues in the video clip.

Visual cues We compose visual cues with facial expression and scene description to perceive human-specific and scene-wide contextual information. Specifically, to process

human-specific information, we utilize the active speaker detection algorithm [9] and face detector [13] to crop the face of the speaking person in each video segment. This process effectively identifies the active speaker, especially for sitcoms where many people appear in a single scene, allowing to align visual features with utterances. For facial expression description, we extract 14 facial action units (FAUs) [2] from each frame in the video segment with 10 frames per second (FPS). Then, we accumulate them and take the three most dominant units. For scene-wide contextual cues, we use the video captioning [10] to extract scene description. The scene description provides high-level context for the visual cues including the surrounding objects and background that interact with the speaker.

Acoustic cues We extract the mean and the variance of pitch, intensity, jitter and shimmer as acoustic features from speech utterance using off-the-shelf speech processing models [1]. Since the extracted values are real numbers, we initially try to convert them to a linguistic format with certain criteria (*e.g.*, map to "high pitch" if the mean pitch value is greater than 200). However, it is challenging to set an objective criterion that considers various factors, including the gender, context, and identity of the speaker. Instead of putting real numbers into text, we use themselves as acoustic features by giving a description of them as a prompt to LLMs, leveraging their knowledge on understanding numerical number.

2.3. Annotation for Laughter Reason

We employ human annotators from Amazon Mechanical Turk (AMT) to label videos with reasons for laughter. Given the inherently subjective nature of humor and the extensive variability in laughter triggers, constructing a ground truth (GT) by free-form annotations posed significant challenges. To mitigate these issues, we utilize the large language model, ChatGPT (GPT-3.5 [7]), with multimodal textual representation m to generate candidates for laughter reason, these candidates are subsequently presented to annotators with the corresponding video clip. The annotators are asked to choose the most appropriate explanation among them. If none of the candidates were suitable, we instruct them to write or correct the reason in free-form.

After annotation, the candidate with the most votes is selected as the GT, and if the annotator provided the reason for laughter in free-form, we manually check the validity of the reason for laughter. Additionally, we verify all GT and manually refine it if it is not plausible for laughter reasons with video. This approach alleviates the annotation workload and facilitates evaluation by developing a more concise GT for this complex and subjective task. Finally, our dataset is represented as $d = \{(v_1, m_1), (v_2, m_2)), .., (v_k, m_k),\ y\}$, where y is a GT explanation for laughter in the video clip d.

[2]We use https://github.com/CVI-SZU/ME-GraphAU to extract FAUs.

2.4. Task Definition

We present *Laugh Reasoning*, a new task that challenges the model to understand the reason for laughter in a given video. Our task is designed to enable a model to generate an explanation that clarifies why a particular situation incited laughter in given video. We formally define this task as, $\hat{y} = f(d)$, where $\hat{y}$, f, and d stand for the estimated explanation about laughter, the model, and the input video clip.

3. Experiments

In our finalized dataset, we split the dataset into 5 cross-validation splits except for test set. We use the training set either for fine-tuning, or as a few-shot context for the language model. To assess models in the laugh reasoning task, we measure the similarity between the generated explanations and the human-annotated references

Model f We use LLMs as baselines for our benchmark. To use LLMs in our benchmark, we replace the input video clip d with our multimodal textual representation m. We can rewrite the task formula as, $\hat{y} = f(\mathcal{P},\ \{m_1,\ m_2, ...,\ m_k\})$, where $\mathcal{P}$ denotes a fixed prompt that describes input representation and instructing the generation task to language model. This approach is based on the success of prior arts [12, 11] using text as an intermediate representation to leverage language model for a wide range of tasks.

Specifically, we introduce three different baselines for our task: (1) LLM fine-tuned on our dataset, (2) LLM with zero-shot learning, and (3) LLM with few-shot in-context learning. We employ the *Davinci* model of GPT-3 [3] for all baselines. For the fine-tuning scenario, we utilize the training split of our dataset. In the cases of zero-shot and in-context learning, we give an instruction to GPT-3 to reason why the audience laughed, using a sample from the test set. Additionally, for in-context learning, we provide the model with three randomly chosen labeled examples from the training set. To generate the outputs, we use sampling with a temperature of 0.5. Note that our task has the flexibility to switch to other language models or vision-language models for f as our dataset comprises video clips and their multimodal textual representation as pair.

Multimodal information v.s. Transcript only As introduced in Sec. 2.2, our dataset provides multimodal information, which includes acoustic cues, visual cues, and transcription. This multimodal information is vital for discerning laugh-inducing reasons, as diverse multimodal factors can trigger laughter. To validate the importance of multimodal information in understanding the reason behind the laughter, we conduct an ablation study comparing the use of all multimodal information versus using only the transcription.

Model	Modality	$BLEU_4$ (↑)	METEOR (↑)	$ROUGE_L$ (↑)	BERTScore (F1) (↑)
GPT-3 (zero-shot)	T	0.126	0.155	0.313	0.389
	A+V+T	0.157	0.184	0.364	0.454
GPT-3 (in-context)	T	0.187	0.198	0.368	0.431
	A+V+T	0.232	0.230	0.413	0.476
GPT-3 (FT)	T	0.230	0.243	0.429	0.488
	A+V+T	**0.279**	**0.267**	**0.475**	**0.523**

Table 1. **Evaluation on laugh reasoning.** We evaluate whether the model can explain why the audience laughed. GPT3 [3] is used as a language model for fine-tuning, in-context (3 shots), and zero-shot experiments on our proposed dataset. Each modality cue in our dataset is denoted as Transcript (T), Audio (A), and Visual (V). FT denotes fine-tuning the model.

Results The quantitative results for the laugh reasoning task are summarized in Table 1. Across all models, from zero-shot to fine-tuning, there is an overall improvement in performance on the laughter reasoning task when utilizing all the modality cues from our dataset, compared to using the transcript alone. We believe such performance gain is due to the various modality cues embedded in the video that trigger laughter.

Furthermore, the results show that the model trained with all modalities successfully discerns the reasons for laughter by referencing multimodal information, while a transcript-only model achieves a partial understanding. Interestingly, GPT3 (in-context) provided with all modalities resulted in comparable performance to the one fine-tuned model with the transcript-only dataset. This also shows that providing multimodal cues to the model may further help the model in reasoning about the laughter.

Acknowledgment

This work was supported by IITP grant funded by Korea government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub; No.RS-2022-00164860, Development of human digital twin technology based on dynamic behavior modeling and human-object-space interaction; No.2022-0-00290, Visual Intelligence for Space-Time Understanding and Generation based on Multi-layered Visual Common Sense).

References

[1] Tomas Arias-Vergara, Juan Camilo Vásquez-Correa, and Juan Rafael Orozco-Arroyave. Parkinson's disease and aging: analysis of their effect in phonation and articulation of speech. *Cognitive Computation*, 2017.

[2] William Sims Bainbridge, Edward E Brent, Kathleen M Carley, David R Heise, Michael W Macy, Barry Markovsky, and John Skvoretz. Artificial social intelligence. *Annual review of sociology*, 1994.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. 2020.

[4] Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy, July 2019. Association for Computational Linguistics.

[5] Jon Gillick, Wesley Deng, Kimiko Ryokai, and David Bamman. Robust laughter detection in noisy environments. In *Interspeech*, 2021.

[6] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 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)*, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[7] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. 2022.

[8] Sophie K Scott, Nadine Lavan, Sinead Chen, and Carolyn McGettigan. The social life of laughter. *Trends in cognitive sciences*, 2014.

[9] Ruijie Tao, Zexu Pan, Rohan Kumar Das, Xinyuan Qian, Mike Zheng Shou, and Haizhou Li. Is someone speaking? exploring long-term temporal features for audio-visual active speaker detection. In *MM*, 2021.

[10] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

[11] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. In *NIPS*, 2022.

[12] Andy Zeng, Adrian Wong, Stefan Welker, Krzysztof Choromanski, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, et al. Socratic models: Composing zero-shot multimodal reasoning with language. In *ICLR*, 2022.

[13] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *IEEE ICCV*, 2017.