Multimodal Humor Dataset: Predicting Laughter tracks for Sitcoms

Badri N. Patro *† IIT Kanpur badri@iitk.ac.in

Mayank Lunayach* IIT Kanpur mayankl@iitk.ac.in

Deepankar Srivastava IIT Kanpur deepanks@iitk.ac.in

Sarvesh IIT Kanpur sarvesh@iitk.ac.in

Hunar Singh IIT Kanpur hunar@iitk.ac.in

Vinay P. Namboodiri University of Bath vpn22@bath.ac.uk

Abstract

A great number of situational comedies (sitcoms) are being regularly made and the task of adding laughter tracks to these is a critical task. Providing an ability to be able to predict whether something will be humorous to the audience is also crucial. In this project, we aim to automate this task. Towards doing so, we annotate an existing sitcom ('Big Bang Theory') and use the laughter cues present to obtain a manual annotation for this show. We provide detailed analysis for the dataset design and further evaluate various state of the art baselines for solving this task. We observe that existing LSTM and BERT based networks on the text alone do not perform as well as joint text and video or only video-based networks. Moreover, it is challenging to ascertain that the words attended to while predicting laughter are indeed humorous. Our dataset and analysis provided through this paper is a valuable resource towards solving this interesting semantic and practical task. As an additional contribution, we have developed a novel model for solving this task that is a multi-modal self-attention based model that outperforms currently prevalent models for solving this task. The project page for our paper is https://delta-lab-iitk.github.io/Multimodal-Humor-Dataset/.

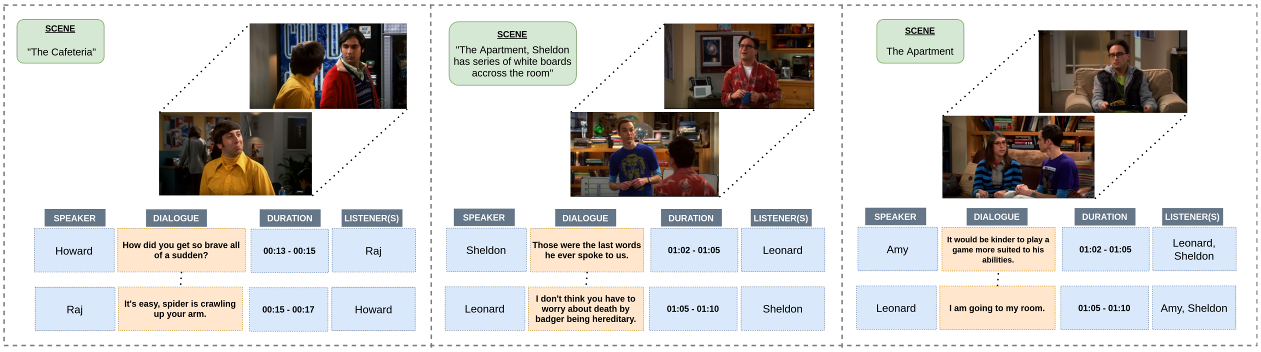
1. Introduction

Understanding humor is a quintessential human task that is not so well understood using currently prevalent AI (Artificial intelligence) systems. In this paper, we aim to solve for the following: Observing a conversation between different individuals, we aim to conclude whether the interaction is humorous or not. Solving for this task includes understanding the nuances from various cues like textual, visual etc. from the conversations. We specifically are concerned with a practical application of this task in terms of being able to predict the laughter track for a situational comedy (sitcom) show. This would also be useful for the designers of a sitcom as they will be able to use it to predict whether the dialogues and scene would be humorous or not. Until now, however, a large-scale multimodal conversational dataset containing multiple speakers in humorous dialogues was missing. Thus, we propose the Multimodal Humor Dataset (MHD). Though this task has been attempted previously (on a smaller scale) by the interesting work of Bertero *et al.* [6] and very early work by Petridis and Pantic [31], we believe that through this paper we provide a larger dataset with analysis on videos, conversations with multiple people and a more thorough analysis of the task using state-of-the-art deep learning methods.

With the rapid progress in AI, understanding emotions especially, in a multimodal context, has still not been sufficiently addressed. There has been work done in our community towards understanding sarcasm from text [34, 3, 35, 7, 10] as well as some preliminary work that aims to understand humor [30, 32, 8]. However, these efforts are evidently not as extensive as that pursued in understanding dialogue or question answering. We believe the lacuna is due to the absence of sufficiently challenging real-world datasets for solving this task. In addition to the dataset, we also provide a specific task of laughter track prediction that we address through our work. We additionally analyse data-set generalization, analysis using attention-based explanation and demonstrate the practical use of our method for generating laughter tracks in a demo video.

Given a dialogue, we aim to understand if it is humorous or not. Towards solving this challenging task, we choose to extract the dialogues spoken in a sitcom. Sitcom which is short for Situational Comedy is a comedy genre centered on fixed characters set carried over episodes. One of the distinguishing characteristics of sitcoms, as opposed to other tele-

[*] Equal contribution
[†] Currently working at Google


SCENE
"The Cafeteria"
SCENE
"The Apartment, Sheldon has series of white boards accross the room"
SCENE
The Apartment

SPEAKER	DIALOGUE	DURATION	LISTENER(S)
Howard	How did you get so brave all of a sudden?	00:13 - 00:15	Raj
Raj	It's easy, spider is crawling up your arm.	00:15 - 00:17	Howard

SPEAKER	DIALOGUE	DURATION	LISTENER(S)
Sheldon	Those were the last words he ever spoke to us.	01:02 - 01:05	Leonard
Leonard	I don't think you have to worry about death by badger being hereditary.	01:05 - 01:10	Sheldon

SPEAKER	DIALOGUE	DURATION	LISTENER(S)
Amy	It would be kinder to play a game more suited to his abilities.	01:02 - 01:05	Leonard, Sheldon
Leonard	I am going to my room.	01:05 - 01:10	Amy, Sheldon

Figure 1. Overview of the Multimodal Humor Dataset (MHD)

vision forms, is that the main character(s) barely changes from one episode to the other, and seasons for that matter. Thus whatever happens in an episode, the situation is more or less, where the episode began. This way, the dataset only scales with more and more episodes and seasons without losing much consistency.

For this dataset, we choose a very popular American sitcom, "The Big Bang Theory," and manually annotated the extracted dialogues as humor/non-humor. We first collected raw data which had Scenes, Speakers, dialogues, Start-Times, End-Times, Listeners, etc. for each of the dialogues. The episodes in sitcoms like these have a very useful entity of audience's laughter track whenever something was found humorous by the audience. Also, the show was recorded live in the front of the live audience. Thus, we are able to find annotations where people naturally laughed. We tracked all these instances of laughter tracks which helped us in labeling chunks of dialogues as humorous or non-humorous. Humorous if the chunk was followed by a laughter track and non-humorous, otherwise. In the dataset, we also manually annotate the exact duration of humor which gives a measure of how some set of dialogues were more humorous than others.

The problem we deal with in this paper is of detecting humor that we specify precisely as follows,

- *Given a dialogue consisting of various dialogue turns we leverage video and text features of each dialogue turn to classify whether the complete dialogue is humorous or not. Note that these conversations may be among more that two characters.*

The problem is comprehensively studied in this paper. We specifically consider various architectures for predicting laughter track, given a sequence of dialogue turns. Each dialogue comprises of a set of utterances spoken involving multiple people. We separately study this using text only, video only and jointly using both video and text-based features. Most existing works for humor detection use textual or visual information alone, this we believe is the first work, using both of them together. The additional use of video makes the problem more complete. In many dialogues, people involved in conversations themselves do not laugh; however, it is their interaction which is humorous. There are subtle cues such as conversation pause or surprise expressions that make the conversations humorous. The evidence that such cues are present (and are helpful) is there in our analysis as video-only methods performed competitively with the best text-only methods. Further, we observed that the best performing methods jointly used both visual and textual features.

Through this work, we propose the Multimodal Humor Dataset (MHD), having textual dialogues with the corresponding video counterparts. The contributions of this work are manifold:

- We present a large scale manually annotated dataset for a comprehensive multimodal understanding of visual humor. We present a thorough analysis of the dataset.

- We experimented with a number of state of the art methods using text-only, video-only and jointly both text and video based approaches that have been advocated for other similar tasks in the community. We also propose a novel multi-modal self attention based model for the same.

- We verify the generalization of our method by comparing the results obtained by training on our dataset and checking it on other sitcoms. We observe a small drop in accuracy (2-3%) that is improved by fine-tuning further on those datasets.

2. Related Work

With a number of challenging AI based tasks been solved, some work has also been done in understanding human sentiments such as humor. Among these are the work by [14] that provided a semi-supervised method to detect sarcasm in Twitter and Amazon data. Few works have been tried for language based detection of humor. For instance, [37] has tried a method to detect jokes in a sentence. [28] provides computational approaches to detect humor using

Field	Value
No. of dialogues	13,633
No. of speakers	210
Vocabulary size	17,336
No. of Scenes	562
Avg. no. of words in 1 dialogue turn	13.10
Avg. no. of words in scenes	7.09
Avg. dialogue length (time in sec.)	22.19
Avg. dialogue turn length (time in sec.)	3.83
No. of humorous dialogues	11,282
No. of non-humorous dialogues	2,381

Table 1. MHD Dataset Statistics (For no. of dialogue turns = 5)

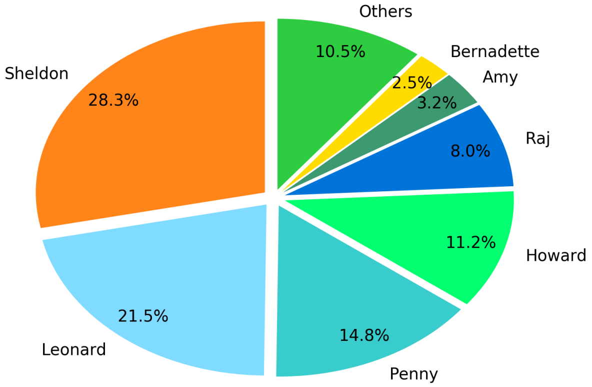
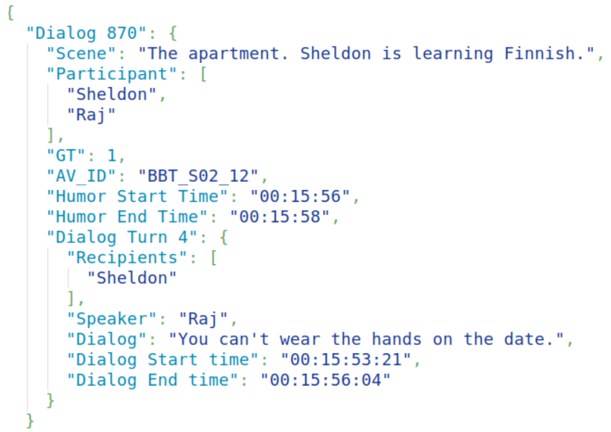


Table 2. A plot showing the distribution of various speakers' utterances across the dataset.

empirical evidences. Recently, there has also been work towards understanding humor. This includes work by [11] that proposed a method for understanding and predicting visual humor in a given image. [9] proposed a multimodal dataset for detecting humor in an image. These works however, have not considered conversations or dialogues. Typically, context matters and more than a single image is necessary for understanding humor in conversations. The work closest to our work is the work by [5, 6, 12]. In their work the authors had proposed LSTM-based methods to detect humor in a dialogue. In contrast to our work, the authors focus only on detecting punchlines which use local context. In our work, we consider more meaningful context that includes several utterances in a dialogue, use visual with textual cues (alone and jointly) while providing thorough analysis. We also have created a dataset for the same as such datasets were not available earlier. Further, there have been a large number of works in vision and language based understanding that are also pertinent and have been considered for this work. This includes work from tasks like (a) Machine Translation, i.e. machine based translation from one language to another,[19, 36, 22, 2], (b) Image Captioning, i.e. describing contents of an observed image,[42, 27, 21, 43], (c) Visual Question Answering, i.e. answering questions based on an image, [26, 33, 1, 16, 45] (d) Visual Question Generation, i.e. generating questions based on an image,[33, 29, 41, 18, 23, 4] and (e) Visual dialogue, [13, 24, 18]. The architectures therein, however, do not directly solve but can be adapted towards solving the task of multimodal humor. Very recently, [20] has proposed a method to detect humor. The proposed method is based on the TED talks dataset, which combines vision and language modality to predict humor in the presentation of a single speaker. In contrast to our proposed work, which includes the conversation of multiple speakers on a particular scene in a specific sitcom. The conversation of multiple speakers is more challenging compared to a single speaker because each speaker has a unique way of initiating and continuing their conversation. Our work is also aimed at a more practical application that we thoroughly analyze.



```
{
  "Dialog 870": {
    "Scene": "The apartment. Sheldon is learning Finnish.",
    "Participant": [
      "Sheldon",
      "Raj"
    ],
    "GT": 1,
    "AV_ID": "BBT_S02_12",
    "Humor Start Time": "00:15:56",
    "Humor End Time": "00:15:58",
    "Dialog Turn 4": {
      "Recipients": [
        "Sheldon"
      ],
      "Speaker": "Raj",
      "Dialog": "You can't wear the hands on the date.",
      "Dialog Start time": "00:15:53:21",
      "Dialog End time": "00:15:56:04"
    }
  }
```

Figure 2. JSON snippet of a typical dialogue in the MHD dataset.

3. The MHD Dataset

To solve this semantically challenging task, the dataset has both visual and linguistic aspects to it. For all the dialogues, we provide the textual dialogues and the associated video clips. Both textual and visual features are synchronised using dialogue start and end times. The subsequent sections provide detailed descriptions and analysis of the newly created dataset.

We gave the dialogue chunks, labels of humor and non-humor using the laughter tracks in episodes. Here, laughter track is the actual track of audience laughing in the live shooting, thus achieving human annotated labels indirectly. A set of utterances spoken just before the laughter track is labelled as Humorous. To define the problem concretely, we needed to fix the size of set and it was not chosen arbitrarily, and we performed various experiments on different

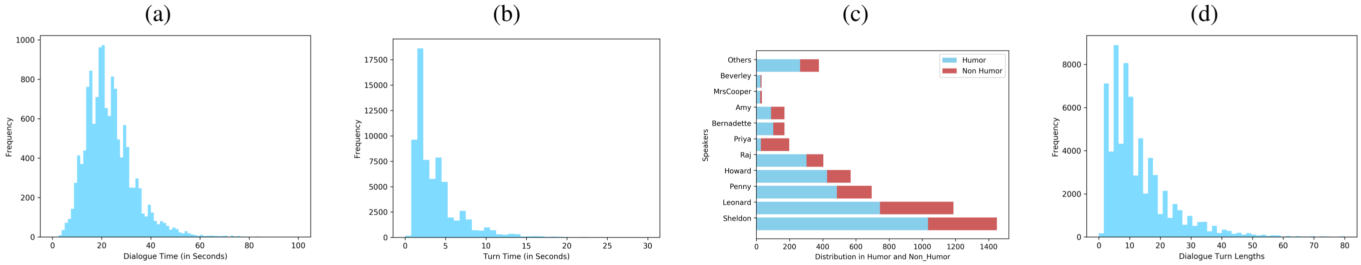


Figure 3. **(a)** *Dialogue Time distribution*: The figure indicating the average dialogue time across the dataset. **(b)** *Turn Time distribution*: The figure indicating average duration per turn in a dialogue across the dataset. **(c)** *Speaker Humor Distribution*: This plot showing contribution of each speaker for generating humor across dialogues in the dataset (figure best viewed in color). **(d)** *Turn Lengths distribution*: The plot showing distribution of number of turns in a dialogue across the dataset.

possible lengths as given in Section 5.2. We could appropriately assume that utterances which were spoken sometime before the Laughter track did not make the audience laugh. Thus, that set of utterances is labeled as Non-humorous.

We refer to the set of utterances as one **dialogue**. We label the dialogue (and not the individual utterances) as Humorous or Non-Humorous. This is done, keeping in mind that humor in these contextual utterances is a result of speaking the utterances together, and not after speaking them individually which then would otherwise sound flat. We refer individual utterances in one dialogue as one **dialogue turn**. The no. of dialogue turns is a configurable parameter in our dataset. For meaningful analysis, the number of turns in a Humorous dialogue is kept equal to that in a Non-Humorous dialogue. Thus, in the final setting, we give "dialogues" as the inputs to the models having humorous/non-humorous label as Ground Truth (GT). A dialogue also has other attributes like Scene, Participants, etc. which are described in detail in the next section. An example JSON snippet of a typical dialogue having one dialogue turn is shown in the Figure 2.

3.1. Dataset Attributes and Creation

For each dialogue, we have the following attributes:

Scene: It describes about how the next few dialogues/visuals are going to turn out. Scenes weakly label the next few dialogues, about what they contain and may be leveraged to make the context more richer with information. We obtained the scene information from the episode scripts.

Speaker: This field stores the speaker for each dialogue turn of the dialogue. The speaker plays a central role in causing humor/non-humor. This attribute can be used to identify the most humorous speakers study to what extent their presence makes something humorous. We also obtained this information using the episode scripts.

Recipients: This is a list of all the listeners of a dialogue turn. This is manually annotated after seeing the video clips.

Participants: This gives us a list of all the participating characters that are present in the scene. This includes both the recipients and the speakers. This attribute is important as a speaker when delivering the same dialogues to different no. of people could also affect it being humorous or not.

# DT	# Train	# Test	# Val
2	14,826	3,662	1,028
3	11,873	3,039	699
4	7,998	1,806	291
5	4,956	1.206	363
6	3,312	612	411
7	2,613	285	141
8	1,545	468	183

Table 3. Train, Test and Val splits when humor:non-humor is 2:1 in each (i.e. Train, Test and Val) (*DT* stands for no. of turns in a dialogue)

Dialogue Turns: These are the utterances delivered by the speaker. We feed a set of dialogue turns as one dialogue to the model. For the textual turns, we use context vectors made of word vectors using the dataset vocabulary, and for video turns, we feed the model with C3D [38] features of the video turn clips. This is obtained using the episode scripts.

Dialogue Start/End time: These are the starting/ending times of the specific dialogue turn obtained using the *.srt* files. They were used in extracting video clips corresponding to the textual dialogue turns.

Humor Start/End time: For humorous dialogues, these are the starting/ending times of the laughter track. In case of multiple laughter tracks within one dialogue, the last one is considered. This is used in separating the humorous instants from the non-humorous pinpoints to the exact instances of humor. This could have been done using the available audio pattern detection techniques, which could have detected the Laughter track. However, we observed that the duration and pattern of the Laughter track is not uniform in an episode. Automatic methods made errors in this very important aspect of the dataset. Thus, we chose to annotate this manually for all the episodes as this forms the ground-truth to evaluate the system.

We believe that all the above attributes affect the humorous nature of a dialogue in some way or the other. Figure 3 shows distribution of dialogue time, turn time, speakers' contribution towards humor and average turn length per dialogue.

Ratio	# Train	# Test	# Val
1	3,382	942	260
1.5	4,265	905	267
2	4,956	1,206	363

Table 4. Train, Test and Val splits for different humor:non-humor ratios (No. of dialogue Turns = 5)

3.2. Dataset Analysis

As Hu *et al.*[17] suggests, humor is very context dependent. A particular sentence may be humorous, but the overall context in which it was said, also plays a major role. However, having very long conversations would dilute the ability to localize humor. In an effort to balance this, we analysed the various number of turn lengths for the dialogues and studied how it affected the model's performance. With a lower number of turns (like 2 ,3, 4) and thus with lesser context, the model didn't perform as well as it did when it used a higher number of turns (5 & 6). When we increased the turn lengths further, the performance started to decrease again. This behaviour may be attributed to the lesser capacity of the models of modeling longer contexts. Detailed results are shown in Table 6. Table 1 shows some statistics of our dataset when the turn length was 5. Here, we have a total of 13,633 dialogues available for training, testing and validating.

Episodes Distribution We are currently releasing the dataset of the first five seasons. These seasons have 110 episodes in total (i.e., 17, 23, 23, 24, and 23 episodes, respectively). Leveraging the already existing dataset divisions in the form of episodes, we pick our train, test, and val sets from entirely different episodes. That is, among the total 110 episodes, we randomly selected six episodes (5%) for Validation, 22 episodes (20%) for Test and remaining (82) episodes (75%) for Training. This is done keeping in mind that sometimes the model may recognize similar dialogues from the same episodes and could thereby "cheat." Segregating these sets in this way helps in more robust evaluation. Also, doing this would not increase the task's difficulty by very much as episodes across the seasons in sitcoms have more or less, the same characters and thus the same humor patterns.

Words Distribution Distribution of the speakers' utterances (dialogue turns) across the dataset is shown in the Table 2. It shows the percentage of dialogue turns spoken by the top 9 characters. Sheldon and Leonard together, seem to cover almost half the no. of spoken dialogues. A bubble plot to visualize the vocabulary of humorous and non-humorous dialogues is given in the supplementary.

As evident in the Figure 2 of the supplementary, the word distribution of humorous and non-humorous words came out to be similar, suggesting Bag of Words based approaches may not be very effective. An additional plot showing the individual word distribution of the dialogues spoken by each of the Top 6 characters is attached in the supplementary. We observed that the distribution for different speakers is more or less same.

Further, a t-SNE plot of the last video frames of the last dialogue turns for both humorous and non-humorous visual dialogue turns is attached in the supplementary.

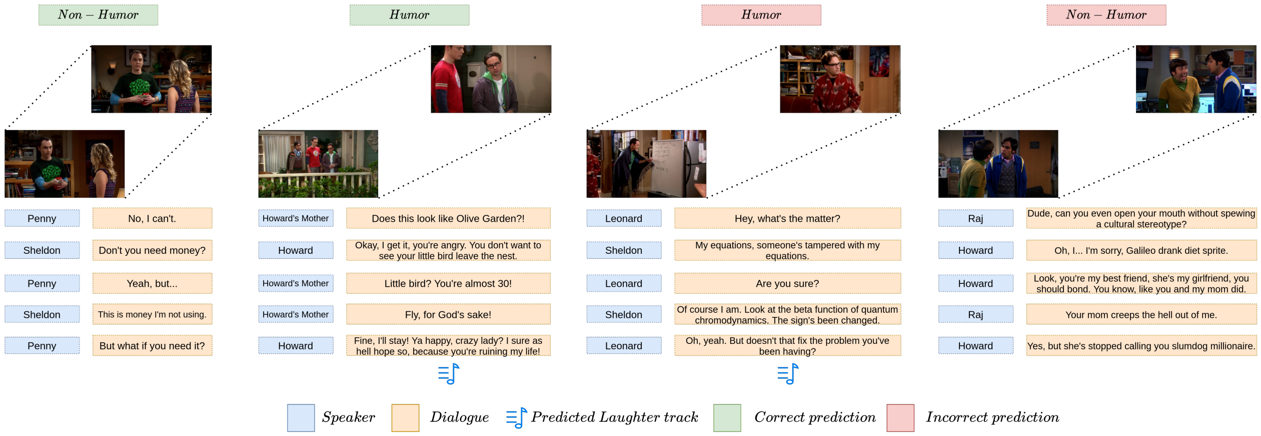
3.2.1 Dataset Distribution

As the Table 1 suggests, our dataset is not quite balanced in terms of the number of humor and non-humor labels. Humor labels are about 82.5% of the total samples. We expect this kind of distribution from a humorous TV series like Big Bang Theory. However, this also makes the dataset unusable in its current form. We tackle this problem by doing sampling. That is, we take all the non-humorous samples (say their no. is "*NH*") and then we randomly sample *k* times the above chosen no. of non-humorous samples (*from all the humorous samples, i.e. H = k*NH*). We tried with 1, 1.5 and 2 as the values of *k*. A lesser value of *k* would mean a more balanced but lesser data at the same time. Thus, we found 2 to be a reasonable choice. Data split statistics for *k* = 2 for all dialogue turns is shown in the Table 3. For no. of dialogue Turns = 5, we vary *K* to 1, 1.5 and 2 and the dataset split stats are given in the Table 4. For tackling the bias, we could also have augmented the data by reordering the turns in non-humorous dialogues. It would have increased the non-humor numbers, but we would have lost the context of the dialogues, which plays a crucial role in a semantically challenging task like humor detection.

4. Baseline Humor Models

In this section, we propose and evaluate various models for the proposed classification task. The model is given a input, dialogue $D = \{(d_1, d_2, \dots d_t), C\}$, where $\{d_i\}$ is a dialogue turn, C is the ground-truth label (humorous/non-humorous) and 't' is the turn length. The model creates a context feature vector using all the dialogue turns and solves a binary classification problem. We experimented with three types of models: Attention based, Fusion based and Sequence based models. Existing works uses only text or visual features, we used both text and visual features while we also experimented using them separately. Below, we elaborate about the attention based model using both text and video based features which were performing well for this task.

- **Text Fusion Model (TFM)**: After obtaining encoding feature of each dialogue turn, we fuse them and feed the output feature vector to binary classifier.
- **Text Sequential Model (TSM)**: Individual dialogue turn features are fed to sequential model (LSTM) for obtaining the final feature vector.
- **Text Attention Model (TAM)**: Here, we used a self attention module to attend to features. We used recent



Penny	No, I can't.
Sheldon	Don't you need money?
Penny	Yeah, but...
Sheldon	This is money I'm not using.
Penny	But what if you need it?

Howard's Mother	Does this look like Olive Garden?!
Howard	Okay, I get it, you're angry. You don't want to see your little bird leave the nest.
Howard's Mother	Little bird? You're almost 30!
Howard's Mother	Fly, for God's sake!
Howard	Fine, I'll stay! Ya happy, crazy lady? I sure as hell hope so, because you're ruining my life!

Leonard	Hey, what's the matter?
Sheldon	My equations, someone's tampered with my equations.
Leonard	Are you sure?
Sheldon	Of course I am. Look at the beta function of quantum chromodynamics. The sign's been changed.
Leonard	Oh, yeah. But doesn't that fix the problem you've been having?

Raj	Dude, can you even open your mouth without spewing a cultural stereotype?
Howard	Oh, I... I'm sorry, Galileo drank diet sprite.
Howard	Look, you're my best friend, she's my girlfriend, you should bond. You know, like you and my mom did.
Raj	Your mom creeps the hell out of me.
Howard	Yes, but she's stopped calling you slumdog millionaire.

Figure 4. Randomly sampled results (MSAM model) of each prediction category, $(correct/incorrect) \times (humor/non-humor)$. Eg. Humor label in a red box means ground truth label was non-humor but predicted label was humor (figure best viewed in color).

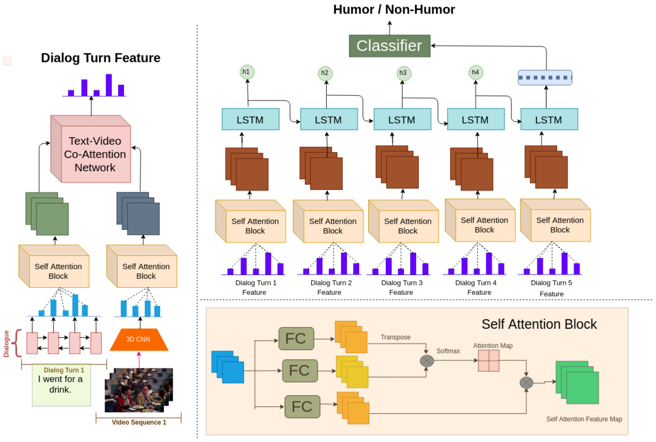


Figure 5. The figure describing the proposed Multimodel Self Attention Model (MSAM) for the laughter detection task. We obtain features of each joint dialogue turn using Multimodel Self attention network. We then obtain the final feature vector using a sequential network before feeding the resultant vector to the binary classifier.

attention techniques like Transformer [40] and BERT [15] to get encoding features and denote the model variants as TAM_Tran and TAM_BERT.

- **Video-based models**: Similar to text-based models, we obtain context features using fusion, sequential and attention models, denoting them as Video Fusion Model (**VFM**), Video Sequential Model (**VSM**) and Video Attention Model (**VAM_Self**) respectively.

- **Video-Text Fusion Model(VTFM)**: Text and Video features are fused together before feeding to the classifier. We are also proposing a novel mulimodal self-attention modal which we believe would be more suitable for this task.

- **Multimodal Self Attention Model (MSAM)** Given a sequence of dialogue turns $(d_1, d_2, \ldots d_t)$, We obtain text encoding representation for each dialogue turn using Bidirectional Encoder Representations from Transformers (BERT) [15] and video encoding representation using C3D [39]. As laughter track prediction depends on the context and the position information, we propose a multimodal self attention model to capture both of them. We obtain text and video features with the help of self attention [44] module as shown in the Figure 5. Then attention features are fed to a co-attention [25] based multimodal network to obtain a joint feature. This was done for each dialogue turn separately. This joint feature was passed through a self-attention block for capturing positional information along each multimodal turn. Finally, the entire dialogue's feature is obtained by modelling the final context vector by feeding all the dialogue turns sequentially to an LSTM. We train the proposed model with binary cross entropy loss predicting humor/non-humor labels.

5. Experiments and Results

We evaluated our dataset on various models while also considering different dataset parameters and modalities. For the number of dialogues turns = 5 and humor:non-humor ratio = 2, we have shown experiments on Attention, Fusion and Sequential models using only Text, only Video and both of them together, while reporting their Accuracy, ROC and F_1 score.

5.1. Baseline Results

The results of our various baseline models are depicted in Table 5. For text-only, *TAM_BERT* achieves the highest F_1 score of 79.96% for the humor class. For video-only, VAM_Self[44] performed best and achievd a F_1 score of 79.30% for the humor class. When the model was fed with both visual and textual dialogues, we saw a clear increase in performance. MSAM (Multimodal Self Attention Model)

Method	Accuracy	F_1 Score	ROC
TFM	65.92	74.29	0.52
TSM	66.06	76.59	0.57
TAM	66.95	77.83	0.59
TAM_Tras	67.34	79.73	0.59
TAM_BERT	67.98	79.96	0.60
VFM	61.02	70.50	0.47
VSM	65.12	77.83	0.52
VAM_Self	67.72	79.30	0.60
VTFM	68.48	79.12	0.60
MSAM	**72.37**	**81.32**	**0.68**

Table 5. Accuracy, F_1 score (Humor class) and ROC (macro-avg) for various models evaluated on the dataset with dialogue Turn Length = 5. Where, *trans* stands for Transformer and *self* stands for Self Attention. F_1 score for the non-humor class for MSAM model was 51.82%

Method	Accuracy	F_1 score	ROC
Dialogue-2	56.55	62.20	0.54
Dialogue-3	60.02	70.66	0.54
Dialogue-4	61.35	72.67	0.56
Dialogue-5	66.06	76.59	0.57
Dialogue-6	68.95	79.48	0.58
Dialogue-7	65.61	77.10	0.55
Dialogue-8	62.82	75.35	0.52

Table 6. Analysing the effect of number of dialogue turns across all the available metrics. F_1 is for Humor class. ROC is in macro-average.

achieved the highest scores in all the three measures (Accuracy, F_1 score, and ROC). F_1 score for the non-humor class for MSAM model was 51.82%. The ROC plots can be found in the supplementary.

From the results, we observe that text based and video based models do not perform as well as the video-text joint models. One possible reason is that a lot of the humor is based on sarcasm or satire, thus the text alone would not reflect the true sentiments. Thus, sometimes the text-based methods are misled in predicting humor. On the other hand, the video provides visual cues such as the facial expressions of people, their movements etc. that helps in the classification of humor. The given provisions of this dataset shall enable further research in this area, and we would develop models that will really understand humor.

5.2. Number of Dialogue Turns

To see the effect of the number of dialogues turns, we performed the experiments using sequential model TSM (on Textual dialogues) while varying no. of dialogue turns from 2 to 8. These results are listed in Table 6. We found a peak near 5-6 dialogue turns which further validates our choice for the no. of dialogue turns used for other experiments. We found the minimum at 2 despite a large training data (see Table 3), suggesting the importance of context in this semantically difficult problem. As we increase the no. of dialogue turns, the performance increases until the peak at 6 as the context becomes richer. Slight inferior performance on dialogue turns 7 and 8 could be attributed to lesser training data.

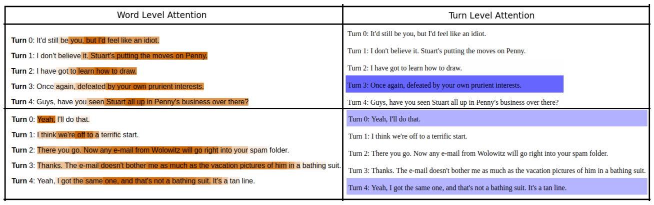
5.3. Analysing the Speaker Effect

In sitcoms like these, sometimes some characters are more humorous than others. They being one of the speakers in the dialogue could potentially affect the dialogue's chance of becoming humorous/non-humorous. Thus, to study the effect of this modality in the problem other than the dialogues alone, we chose to experiment while providing the TSM with both the textual dialogues and the speakers. Doing this increased the accuracy to 67.26% as shown in Table 7. Considering both the dialogue information and speaker information the model accuracy, F_1 score and ROC increases as mentioned in the Table 7.

5.4. Humor and Non-humor Ratio

In order to validate that our results are not tending towards the frequency baselines, we change Humor:Non-Humor ratio (i.e 1, 1.5 and 2) in the dataset. These results are shown in Table 8. Decrease in performance may be attributed to the decrease in available data (as shown in Table 4).

5.5. Explaining Humor



Word Level Attention	Turn Level Attention
Turn 0: It'd still be you, but I'd feel like an idiot. **Turn** 1: I don't believe it. Stuart's putting the moves on Penny. **Turn** 2: I have got to learn how to draw. **Turn** 3: Once again, defeated by your own prurient interests. **Turn** 4: Guys, have you seen Stuart all up in Penny's business over there?	Turn 0: It'd still be you, but I'd feel like an idiot. Turn 1: I don't believe it. Stuart's putting the moves on Penny. Turn 2: I have got to learn how to draw. Turn 3: Once again, defeated by your own prurient interests. Turn 4: Guys, have you seen Stuart all up in Penny's business over there?
Turn 0: Yeah, I'll do that. **Turn** 1: I think we're off to a terrific start. **Turn** 2: There you go. Now any e-mail from Wolowitz will go right into your spam folder. **Turn** 3: Thanks. The e-mail doesn't bother me as much as the vacation pictures of him in a bathing suit. **Turn** 4: Yeah, I got the same one, and that's not a bathing suit. It's a tan line.	Turn 0: Yeah, I'll do that. Turn 1: I think we're off to a terrific start. Turn 2: There you go. Now any e-mail from Wolowitz will go right into your spam folder. Turn 3: Thanks. The e-mail doesn't bother me as much as the vacation pictures of him in a bathing suit. Turn 4: Yeah, I got the same one, and that's not a bathing suit. It's a tan line.

Figure 6. The left column shows visualization of attention at the word level and the right column shows attention visualization at turn level.

We experimented with starting with basic models, then going for BERT based (TAM-BERT) and then finally proposed the Multi-modal Self Attention-based (MSAM) model. We observed good F_1 scores, suggesting that our model was able to learn some useful features for detecting the Laughter tracks. Another beneficial thing would be to study the reasons behind positive detections and, if possible, provide explanations for the detected laughter. To the best of our knowledge, no previous work has studied the explainability part of humor (especially in a multimodal setting). Towards temporally localizing the humor causes, first, we analyzed the contribution of each word of every dialogue turn (Figure 6 (left)). Word based attention was challenging for understanding the model. Hence, we also visualized the contribution of each turn as a whole. This turn level attention for attending to context was possible with the unique characteristic of our dataset (Figure 6 (right)). The explanation analysis is a preliminary one and we are hoping that

Type	Accuracy	F_1 score	ROC
Speaker (S)	62.15	69.34	0.52
Dialogue (D)	66.06	76.59	0.57
Joint (D+S)	67.26	78.53	0.61

Table 7. Comparing the results when feeding the model with Speaker or Dialogue or Dialogue + Speaker both. (S stands for Speaker.) F_1 is for Humor class. ROC is in macro-average.

our MHD dataset would be a valuable resource to explore and dive deep into various areas of computational humor.

5.6. Discussion on model's performance

Success cases: In Figure 4, we have shown random qualitative results. In simple situations like the first column, where neither textual nor visual clues support the presence of humor, our model does an excellent job of not detecting humor. The second column represents a very tricky situation where both speakers of the conversation are not visible to the model. The model has to decide based only on the textual information and (non-humorous) facial expressions of the listeners. The model's ability to keep track of dichotomy developed in the last couple of dialogues seemed to help. Otherwise, the last dialogue turn (Howard's part) alone has no humor present in it. Situations like these make this dataset very interesting and challenging. A good performing model will have to take care of how to weigh in both the textual and visual information correctly.

Failure cases: There's a pattern in the sitcom that when speakers use complicated scientific terms, incongruity develops, leading to humor. In the fourth dialogueue of the third column (Figure 4), we observe this pattern. Also, visual incongruity (Sheldon wearing a blanket, which is rare) further adds to the false signal. In the future, we expect that the proposed models would be more robust to such outlier signals and would capture more of the long tail of humor/non-humor examples. In the fourth column, if the model has to detect humor correctly, it should know that Raj is an Indian, and the Slumdog Millionaire movie is set in India. This is a classic case of dependence on some knowledge base. Still, in this multimodal scenario, the model might have taken clues from the visual information as Howard (man with a yellowish sweater) teasingly laughs at Raj. We hope that future models would mitigate dependence on knowledge bases in these situations like these by more cleverly leveraging the visual information.

5.7. Generalisation and Human Evaluation

To evaluate if our models are not over-fitting on our dataset and are actually learning semantically meaningful features for detecting humor, we decided to evaluate a few episodes of some similar sitcoms using our model. The Sitcoms that we tried were *Friends*, *How I Met Your Mother*, and a very recent sitcom, *2 Broke Girls*. Due to relatively different test settings, we did observe a drop in detection accuracy, but the drop was only about 3%, 2%, and 3%, respectively. We simply evaluated on the test episodes without fine-tuning as such. Then, we also evaluated after doing the fine-tuning, and the relative accuracy drop decreased in all three cases. This hints towards the generalizability of our dataset and model.

Ratio	Accuracy	F_1 score	ROC
2:1	66.06	76.59	0.57
1.5:1	65.09	75.79	0.56
1:1	57.00	61.53	0.57

Table 8. Studying the effect of humor vs non-humor ratio. Accuracy, F_1 score (Humor class) and ROC(macro-avg) for the TSM model

Further, we compared the performance of our laughter detection models with the humans who were asked to label the same set of test dialogues, which they found to be leading to laughter. Test episodes were randomly selected and were divided into short clips based on the annotated humor/non-humor labels. A randomized group was asked to give the clips a humor-non/humor label, and an accuracy of about 81% was achieved. For another related evaluation, we let our model predict laughter tracks in a random test episode. The episode's visual and text features were fed to our best model, and wherever the model predicted laughter, an artificial laughter track was added. Both videos (original and with predictions) are displayed side-by-side in the above-mentioned project website. The demo nicely illustrates that our approach could also be used to add laughter tracks on various other comedic shows where a live audience is absent.

6. Conclusion

In this paper, we solve for predicting laughter in sitcom videos. Towards solving this, we propose a new dataset termed as 'Multimodal Humor Dataset'. This manually curated dataset has annotations that provide information whether some dialogue (comprising of several dialogue turns) is humorous or not. It also includes other information such as the identity of people speaking and the recipients. We have analyzed the proposed dataset and evaluated it thoroughly. We propose a number of methods that are based on state of the art techniques prevalent in vision and language tasks. These solve for text-based, video-based and joint text & video-based prediction of humor. We also proposed a multimodal self attention model (MSAM) which performed relatively well for the task. With the best of these methods, we obtain an F_1 score accuracy of around 80% which is encouraging. In future, we would be interested in further progress towards solving and understanding humor, for instance, considering audio and other cues such as scenes, intensity, recipient etc.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] Francesco Barbieri and Horacio Saggion. Modelling irony in twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–64, 2014.

[4] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis*, volume 3, 2017.

[5] Dario Bertero and Pascale Fung. A long short-term memory framework for predicting humor in dialogues. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 130–135, 2016.

[6] Dario Bertero and Pascale Fung. Predicting humor response in dialogues from tv sitcoms. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5780–5784. IEEE, 2016.

[7] SK Bharti, B Vachha, RK Pradhan, KS Babu, and SK Jena. Sarcastic sentiment detection in tweets streamed in real time: a big data approach. *Digital Communications and Networks*, 2(3):108–121, 2016.

[8] Piot Bilal, Olivier Pietquin, and Matthieu Geist. Imitation learning applied to embodied conversational agents. In *Machine Learning for Interactive Systems*, pages 1–5, 2015.

[9] Giuseppe Boccignone, Donatello Conte, Vittorio Cuculo, and Raffaella Lanzarotti. Amhuse: a multimodal dataset for humour sensing. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 438–445. ACM, 2017.

[10] Mondher Bouazizi and Tomoaki Ohtsuki. Sarcasm detection in twitter:” all your products are incredibly amazing!!!”-are they really? In *2015 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2015.

[11] Arjun Chandrasekaran, Ashwin K Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. We are humor beings: Understanding and predicting visual humor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4603–4612, 2016.

[12] Peng-Yu Chen and Von-Wun Soo. 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)*, pages 113–117, 2018.

[13] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. *CoRR*, abs/1611.08669, 2016.

[14] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116. Association for Computational Linguistics, 2010.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[16] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. *CoRR*, abs/1505.05612, 2015.

[17] Shuqin Hu. A relevance theoretic analysis of verbal humor in the big bang theory. *Studies in Literature and Language*, 7(1):10–14, 2013.

[18] Unnat Jain, Ziyu Zhang, and Alexander G Schwing. Creativity: Generating diverse questions using variational autoencoders. In *CVPR*, pages 5415–5424, 2017.

[19] Aditya Joshi, Pushpak Bhattacharyya, and Sagar Ahire. Sentiment resources: Lexicons and datasets. In *A Practical Guide to Sentiment Analysis*, pages 85–106. Springer, 2017.

[20] Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*, 2019.

[21] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[22] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*, 2017.

[23] Yikang Li, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, and Xiaogang Wang. Visual question generation as dual task of visual question answering. *CoRR*, abs/1709.07192, 2017.

[24] Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. *CoRR*, abs/1706.01554, 2017.

[25] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[26] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *CoRR*, abs/1410.0210, 2014.

[27] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *CoRR*, abs/1412.6632, 2014.

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

[29] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Larry Zitnick, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. *CoRR*, abs/1603.06059, 2016.

[30] Antinus Nijholt. *Humor and embodied conversational agents*. Centre for Telematics and Information Technology, University of Twente, 2003.

[31] Stavros Petridis and Maja Pantic. Audiovisual discrimination between laughter and speech. pages 5117 – 5120, 05 2008.

[32] Bilal Piot, Olivier Pietquin, and Matthieu Geist. Predicting when to laugh with structured classification. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[33] Mengye Ren, Ryan Kiros, and Richard S. Zemel. Image question answering: A visual semantic embedding model and a new dataset. *CoRR*, abs/1505.02074, 2015.

[34] Antonio Reyes, Paolo Rosso, and Tony Veale. A multidimensional approach for detecting irony in twitter. *Language resources and evaluation*, 47(1):239–268, 2013.

[35] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, 2013.

[36] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[37] Julia M Taylor and Lawrence J Mazlack. Computationally recognizing wordplay in jokes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26, 2004.

[38] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[39] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[41] Ashwin K. Vijayakumar, Michael Cogswell, Ramprasaath R. Selvaraju, Qing Sun, Stefan Lee, David J. Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424, 2016.

[42] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015.

[44] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *stat*, 1050:21, 2018.

[45] Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. *CoRR*, abs/1511.03416, 2015.