Learning to Evaluate Humor in Memes Based on the Incongruity Theory

Kohtaro Tanaka[1], **Hiroaki Yamane**[2,1], **Yusuke Mori**[1], **Yusuke Mukuta**[1,2], **Tatsuya Harada**[1,2]
[1]The University of Tokyo [2]RIKEN
{k-tanaka, yamane, mori, mukuta, harada}@mi.t.u-tokyo.ac.jp

Abstract

Memes are a widely used means of communication on social media platforms, and are known for their ability to "go viral". In prior works, researchers have aimed to develop an AI system to understand humor in memes. However, existing methods are limited by the reliability and consistency of the annotations in the dataset used to train the underlying models. Moreover, they do not explicitly take advantage of the incongruity between images and their captions, which is known to be an important element of humor in memes. In this study, we first gathered real-valued humor annotations of 7,500 memes through a crowdwork platform. Based on this data, we propose a refinement process to extract memes that are not influenced by interpersonal differences in the perception of humor and a method designed to extract and utilize incongruities between images and captions. The results of an experimental comparison with models using vision and language pretraining models show that our proposed approach outperformed other models in a binary classification task of evaluating whether a given meme was humorous.

1 Introduction

Humor is an essential element of human communication. Studies have shown that humor helps to build relationships in work environments (Plester, 2009), facilitates smooth discussions on controversial topics (McGhee, 1989), and helps motivate people to recognize and challenge misinformation (Yeo and McKasy, 2021).

Memes are a type of humor that has been prevalent in recent years, especially on social media. These images express multi-modal humor and often comprise a template image with superimposed upper and lower captions. In the example of a meme shown in Figure 1, the upper caption reads "JOIN THE MARINES THEY SAID" and the lower caption reads "YOU'LL BE A HERO THEY SAID". Both upper and lower captions are superimposed onto the template image of a marine in dress uniform holding an umbrella for the former president Barack Obama. The humor of memes can be explained by the incongruity theory (Raskin, 1985; Buijzen and Valkenburg, 2004). It is a well-established humor theory which states that a surprising contradiction or opposition to an expected situation or interpretation is a key element of any humor. For example, the meme shown in Figure 1 has an incongruity between the caption and the template image; the caption explains the commonsense of viewers that a marine would be a hero in the battleground, but the image is showing the contradicting reality of a marine doing a boring job of holding an umbrella for the president. A study on incongruities in memes has shown that a large number of memes have image-caption incongruity (Yus, 2021).

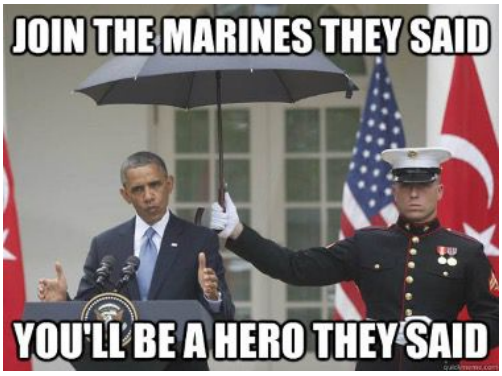


Figure 1: An example of a meme from a meme-sharing website "Best of funny memes"[1].

Given the substantial impact of memes on online communication, such as their effectiveness in correcting misinformation (Vraga et al., 2019; Kim et al., 2021a; Garrett and Poulsen, 2019), researchers have aimed to develop AI systems capable of understanding humor in memes. However, the evaluation of memes has proven a difficult task.

[1]https://www.funny-memes.org/2013/05/join-marines-they-said-youll-be-hero.html

Proceedings of the 2nd Workshop on When Creative AI Meets Conversational AI (CAI2), COLING 2022, pages 81–93
October 12–17, 2022.

Humor is subjective, and interpersonal differences may affect the perception of humor for some memes based on viewers' cultural background and personality characteristics (Ruch and Hehl, 2010). Hence, human-annotated datasets of humor in memes tend to exhibit inconsistent annotations. In addition, existing methods use vision-language pretraining models that do not explicitly extract semantic relationships between images and captions. As a result of this structure, utilizing image-caption incongruity is relatively difficult with such models, although incongruity was shown to be a significant element in humor of memes (Yus, 2021).

In the present work, we addressed the inconsistency of annotations by first creating a meme dataset with a humor annotation of memes that is not influenced by interpersonal differences in the perception of humor. These reliable humor annotations were obtained via an annotation method called best-worst scaling (BWS) (Louviere, 1992). Then, the annotations were refined by our proposed process to eliminate inconsistent examples. The consistency of each annotation was measured by quantifying the agreement of annotations between different annotators.

Based on this data, we propose a method that explicitly extracts and utilizes image-caption incongruities. Our proposed method combines a vision-language transformer with a module designed to extract the features of image-caption similarity. We validated the performance of our proposed method by conducting experiments in which we used several models for comparison to classify whether a given meme (template image + caption) was humorous.

The contributions of this study are summarized as follows.

- We created a reliable dataset of memes with annotations that quantified their degree of humor and extracted humor anchors.
- We proposed and implemented a dataset refinement process to separate memes influenced by interpersonal differences in the perception of humor. The consistency of the annotations was thoroughly examined.
- We implemented models to explicitly extract image-caption incongruities and compared their output with the baselines implemented based on pretrained vision-language models. We showed that our proposed method outperformed baselines in evaluating the humor of memes.

2 Related Work

2.1 Computational Humor Models

Due to the importance of humor in human communication, several previous studies have aimed to recognize or generate humor of a single modality. These include research on fixed forms of language-based humor such as "*I like my X like I like my Y, Z*" jokes (Petrović and Matthews, 2013), Knock-Knock Jokes (Rayz, 2004), miscellaneous short-text humor (Annamoradnejad and Zoghi, 2020), humor in dialogues (Ziser et al., 2020; Yoshikawa and Iwakura, 2020), and visual humor (Chandrasekaran et al., 2016).

However, few studies have considered multimodal humor, such as humor in memes. One study focused on the task of meme evaluation is a competition called "Memotion Analysis" (Sharma et al., 2020). This competition included the task of predicting the degree of humor of a given meme. The best-performing model adopted several pretrained feature extractors and ensemble techniques (Guo et al., 2020). However, the feature extractors were all unimodal and trained independently. Therefore, these methods do not explicitly utilize semantic relationships between images and captions.

2.2 Humor Dataset

Several methods have been developed to record human annotations of humorous content, including rating scales and BWS (Louviere, 1992).

A rating scale presents annotators with a scale and choices of integers or characters that represent a place within the scale. For example, the dataset used for the competition "Memotion Analysis" (Sharma et al., 2020) was annotated using a rating scale. Annotators were provided with four choices to choose from: not funny, funny, very funny, and hilarious.

Although this method is widely used in various disciplines, rating scales are said to have limitations, including the following (Schuman and Presser, 1996; Baumgartner and Steenkamp, 2001).

- Annotation inconsistencies between different annotators.
- Annotation inconsistencies by the same annotator.
- Bias in selection within the scale.

BWS was proposed to resolve these limitations and reduce the number of tasks required. BWS usu-

ally asks annotators to choose the best- and worst-fitting items from among four-tuples of items for the characteristics of interest. Real-valued annotations (BWS scores) can be acquired using maximum difference scaling (MaxDiff) (Finn and Louviere, 1992), a method to conduct and process BWS. To obtain real-valued scores of N items, $[1.5N, 2.0N]$ four-tuples of items were annotated so that each item was evaluated more than about five times. Then, a BWS score was calculated for an item *A* by subtracting the number of times *A* was selected as best-fitting by the number of times it was selected as worst-fitting, and dividing the result by the number of times *A* was evaluated.

The score obtained using this equation is a real value ranging from -1 (worst) to 1 (best).

This method has been proven to produce more reliable annotations compared to rating scales (Kiritchenko and Mohammad, 2017), and has also been used to evaluate the humor of jokes of the form "*I like my X like I like my Y, Z*" (Yamane et al., 2021).

3 Dataset Construction

To construct our reliable dataset, we first obtained a collection of memes from a meme-sharing website. Then, the memes were annotated by crowdworkers. Finally, the annotations were filtered and refined to eliminate inconsistent annotations. As a result, we compiled 1,450 memes with reliable and consistent annotations.

3.1 Data Collection and Preprocess

To create the dataset, we first scraped 693,465 memes (3,000 template images, 143 - 300 captions per template) from Meme Generator[2]. The scraped memes were selected in order of the number of likes they had received to ensure that the dataset included sufficient high-quality memes.

Before asking crowdworkers to annotate the humor in these memes, we conducted preprocessing to reduce the number of memes containing words that were not in English or that were profane.

First, to minimize the number of captions that were not in English, we checked whether each caption could be encoded only using ASCII characters. This filtering process eliminated captions written in languages that do not use ASCII characters and also removed emojis. However, it was not possible to eliminate captions written in languages that use the same alphabet as English, such as Spanish. Although it would be possible to strictly filter captions by checking whether all the included words were present in an English dictionary, we chose not to adopt this approach as meme captions often contain slang or deliberately misspelled words that do not appear in any English dictionary.

As we asked crowdworkers to annotate the humor in memes, we needed to minimize their exposure to profanity. Therefore, we used an open-source library called "profanity-filter" to detect and filter profanity[3]. Although this library enabled the filtration of major profanities, it was not possible to remove inappropriate words that were misspelled or partially concealed.

Finally, image templates that contained more than 150 captions after the two filtering processes were selected and compiled. The resulting preprocessed data contained 296,850 memes (1,979 image templates with 150 captions per template).

3.2 Human Annotation Task Using BWS

To obtain real-valued reliable humor annotations for these memes, we asked crowdworkers on Amazon Mechanical Turk (AMT)[4] to complete three tasks, including answering whether a meme was in English, choosing up to three words that were essential to understanding the humor in the meme, and choosing the most and least humorous meme from among four memes. In our research, we annotated 7,500 memes (100 template images with 75 captions per template) by creating 11,250 four-tuples (1.5N).

Annotation tasks were published on AMT and included two sections with a total of 27 questions.

Section 1 (questions 1 - 24) first asked annotators about their understanding of the meme provided. This question aimed to filter memes that were not in English and not filtered in the preprocessing phase.

Then, annotators were asked to write up to three words in the caption that were necessary to understand the meme (we refer to this data as a humor anchor). If the meme presented was not in English, they were instructed to write "NIE" (not in English) in the first box and leave the other boxes blank. This question aimed to extract humor anchors for each meme and also to evaluate the quality of the annotation (For example, if an annotator chose words that were obviously not important, such as "the", we

[2]https://memegenerator.net/

[3]https://pypi.org/project/profanity-filter/

[4]https://www.mturk.com/

concluded that annotations provided by that worker might be of low quality).

Finally, in Section 2 (questions 25 - 27), annotators were asked to choose the most and least humorous meme from among the four presented. The examples of questions that were presented to the annotators are listed in the appendix.

To ensure the quality of the annotations, the task required workers to be located in the U.S., to have a Human Intelligence Task (HIT) Approval Rate greater than or equal to 98%, and to have at least 500 HITs previously approved. In addition, workers were warned beforehand that the task may contain adult content, as the "profanity filter" did not suffice to eliminate all memes with explicit words.

3.3 Post-process

The resulting annotations were first processed to calculate and compile their BWS scores to obtain the raw dataset.

To ensure the quality of the annotations, the following additional post-processing was performed to produce the post-processed dataset.

- Only the annotations on which workers spent more than ten seconds per question were used.
- Memes which annotators identified as "Not in English" were not used.
- Only memes annotated by more than three people annotated after the other two post-processes were conducted, were used.

This filtering process has reduced the number of human-annotated memes to 6,900 for the post-processed dataset.

3.4 Refining to Filter-out Interpersonal Differences in Perception of Humor

As previous studies have shown that perceptions of humor may be influenced by individual personality characteristics, we analyzed the correlation between differences in BWS score (d) and human agreement (a) to explore how this influence affected our dataset. To do so, we first derived a total of 39,045 hierarchical pairs from the 7,809 annotated four-tuples which were used to create the post-processed dataset. (For example, when an annotator chose A as most humorous and D as least humorous from among four choices A, B, C, D, we derived five hierarchical pairs $A > B$, $A > C$, $A > D$, $B > D$, and $C > D$.) Then, for the 39,045 pairs that were retrieved, d was defined as follows, given the calculated BWS scores of meme A (S_a) and B (S_b).

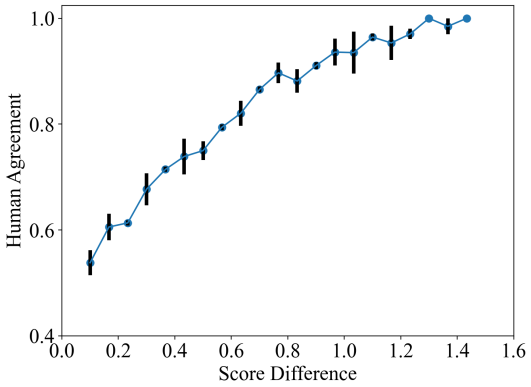


Figure 2: The figure shows the relationship between BWS score difference and human agreement. Blue dots with a blue connecting line represent the average human agreement a, and black bars represent their standard errors.

$$d = |S_a - S_b| \tag{1}$$

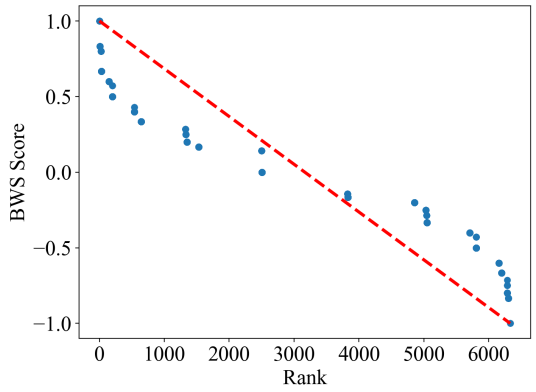
Then, let us consider N_c as the number of hierarchical pairs matching the hierarchical relationship derived from the BWS score, and N_w as the number of hierarchical pairs which contradict the hierarchical relationship derived from the BWS score (e.g., given a pair of memes A and B with BWS score of $b_A = 1$ and $b_B = -1$, and if we derived the three following hierarchical pairs $(A > B)$, $(A > B)$, $(A < B)$, then N_c and N_w would be $N_c = 2$ and $N_w = 1$). Human agreement a is defined and calculated as follows.

$$a = \frac{N_c}{N_c + N_w} \tag{2}$$

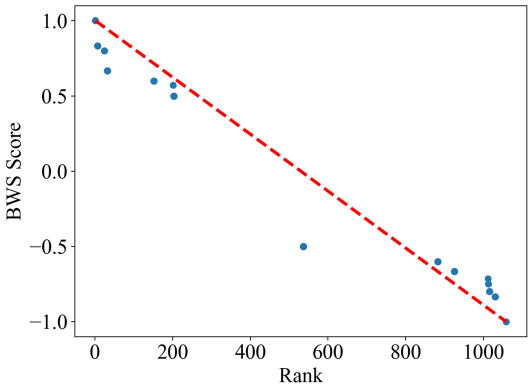
Finally, BWS score differences were binned with a unit of $\Delta d = 0.07$, and the average human agreement scores were calculated. The result is shown in Figure 2.

It was observed that for meme pairs with a BWS score of more than 1.0, the average of human agreement reached around 0.9, and for those with a BWS score of more than 1.4, the average of human agreement was close to 1.0.

Therefore, we conducted an additional refining process to eliminate examples with BWS scores between -0.5 and 0.5. This refined dataset can be considered to contain memes that are mostly not influenced by interpersonal differences in humor perception. The refined dataset includes 1,450 annotated memes.



(a) Relationships between BWS score and ranks for the dataset before refinement.



(b) Relationship between BWS score and rank for the refined dataset.

Figure 3: The figures show the relationship between BWS score and rank for the refined dataset. Blue dots represent examples of memes, and the red line indicates the hypothetical case of a uniform distribution of BWS scores.

3.5 Dataset Statistics

Figure 3 shows the relationship between BWS score and rankings of memes based on their BWS score for the dataset before and after refinement. It may be observed that before refinement, there were fewer examples with a BWS score close to 1 or -1 compared to those close to 0.

In terms of the refined dataset, there was a large gap between ranks 200 and 900. This means that examples with a BWS score of 0.5 or -0.5 constituted about half of the dataset, and examples with scores more than 0.5 or less than -0.5 were uniformly distributed.

In Figure 4, we provide two examples of memes from the refined dataset. These are examples of memes for which annotators were consistent regarding their degree of humor.

(a) An example of memes in the refined dataset with a BWS score of 1.

(b) An example of memes in the refined dataset that has a BWS score of -1.

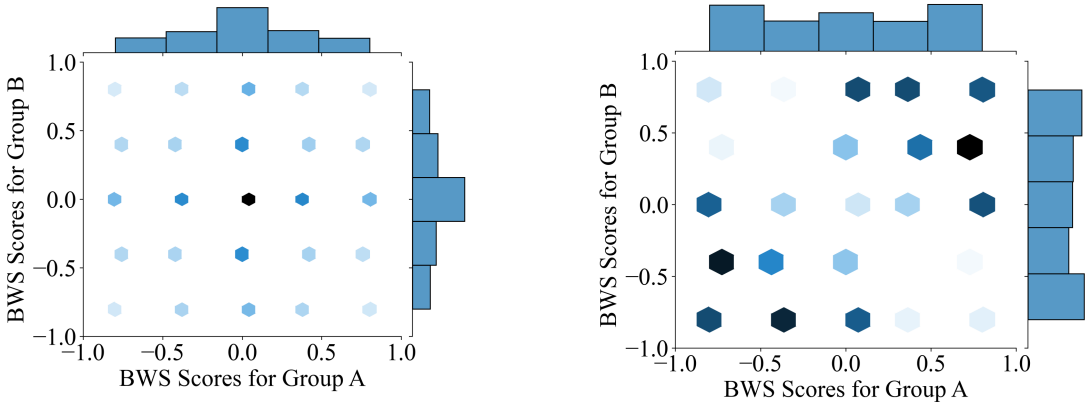
Figure 4: This figure presents two examples of memes from the refined dataset. Annotators were consistent in their evaluation of the humor of these two memes.

To evaluate the consistency of the annotations in each dataset, we examined split-half reliability (SHR). SHR was calculated by first randomly splitting the annotation tasks into two halves. Thus, of the 3,801 tasks published on AMT, 1,900 tasks were designated as group A, and the other 1,901 tasks were designated as group B. Then, memes in each group were subjected to post-processing and the BWS scores of the memes were calculated separately. Finally, we analyzed Spearman's rank correlation coefficient between two rankings of memes based on the BWS scores calculated from groups A and B. As may be observed from Figure 5, the refinement process was able to eliminate examples that involved inconsistency in the perception of humor by annotators to improve the rank correlation coefficient.

3.6 Comparison with Other Meme Datasets

A comparison between our newly created dataset and some existing meme datasets is shown in Table 1. The ImgFlip575K Meme Dataset[5] compiles memes from the meme-generating website Imgflip[6]. While this dataset is exceptionally large, it does not include humor annotations by humans. The Memotion Dataset 7k was created for the sentiment analysis competition task in (Sharma et al., 2020). Although this dataset includes humor annotations created by human annotators, they were created using a rating scale, which is known to create biases in annotations (Kiritchenko and Mohammad, 2016). In comparison, our post-processed dataset is comparable to the Memotion dataset in size, while ensuring the reliability of annotations via the comparative annotation method and filtering post-

[5]https://github.com/schesa/ImgFlip575K_Dataset
[6]https://imgflip.com/



(a) Correlation between binned BWS scores for memes in group A and group B for the dataset before refinement.

(b) Correlation between binned BWS scores for memes in groups A and B for the refined dataset.

Figure 5: The figures show the correlation between BWS scores for memes in groups A and B. To obtain this figure, BWS scores were binned in to 5 bins (-1 to -0.6, -0.6 to -0.2, -0.2 to 0.2, 0.2 to 0.6, 0.6 to 1). The darkness of each hexagon represents the number of memes plotted in each spot, with darker shades representing more memes. The Spearman's rank correlation coefficient for the post-processed dataset was 0.01, whereas that of the refined dataset was 0.52.

processing. Finally, to the best of our knowledge, our refined dataset is the only available dataset that considers interpersonal differences in the perception of humor and includes examples with consistent annotations.

4 Meme Evaluation Model Based on the Incongruity Theory

Studies have shown that many memes exhibit incongruities between images and their captions, which express humor (Yus, 2021). Therefore, we hypothesized that a module designed explicitly to extract incongruities between an image and its caption would improve a model's ability to classify whether a given meme is humorous.

To extract incongruities between image and text, we propose an incongruity extraction module consisting of CLIP image and text encoder (Radford et al., 2021), which is highlighted in orange in Figure 6. In this proposed method, a template image and caption are each fed into the corresponding pretrained CLIP encoder to obtain feature vectors of both the image ($\boldsymbol{v}_{\mathrm{I}} \in \mathbb{R}^{512}$) and the caption ($\boldsymbol{v}_{\mathrm{T}} \in \mathbb{R}^{512}$). Since pretrained CLIP encoders are trained such that the encoded feature vectors of a similar image and caption are located close to each other in the same latent space, we hypothesized that a feature vector $\boldsymbol{v}_{\mathrm{CLIP}} \in \mathbb{R}^{512}$ calculated by equation 3 would include encoded semantic information on the relativity of the input image to the input caption.

$$\boldsymbol{v}_{\mathrm{CLIP}} = \boldsymbol{v}_{\mathrm{I}} - \boldsymbol{v}_{\mathrm{T}} \tag{3}$$

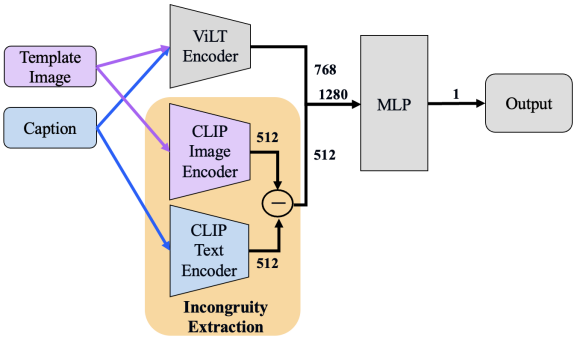


Figure 6: The figure shows an overview of the proposed method, which combines a module designed to extract incongruity between image and text with a ViLT encoder. In the incongruity extraction module, highlighted in orange, feature vectors of a template image and caption are extracted using the CLIP image and text encoders. The two resulting feature vectors are then subtracted and concatenated with the output of the ViLT encoder to be fed into an MLP. The numbers written next to the arrows represent the dimension of each vector.

We considered this information encoded in $\boldsymbol{v}_{\mathrm{CLIP}}$ to be useful in determining the level of incongruity between an image and its caption because image-text incongruity can be considered as a type of semantic relationship between image and text.

After obtaining $\boldsymbol{v}_{\mathrm{CLIP}}$ as a feature representing incongruity between an image and a caption, $\boldsymbol{v}_{\mathrm{CLIP}}$ is concatenated with the output features of a ViLT encoder ($\boldsymbol{v}_{\mathrm{ViLT}} \in \mathbb{R}^{768}$) and fed into a multilayer perceptron (MLP) model to predict whether a given meme is humorous, as shown in Figure 6.

Dataset	Instances	Humor Annotation by Human	Annotation Method	Ensured Reliability of Annotation	Ensured Consistency of Annotation
ImgFlip575K	575,948				
Memotion	6,991	✓	Rating-scale		
Raw (Ours)	7,500	✓	BWS		
Post-processed (Ours)	6,900	✓	BWS	✓	
Refined (Ours)	1,450	✓	BWS	✓	✓

Table 1: Comparison of our datasets to other meme datasets. Our post-processed dataset is comparable to the Memotion Dataset 7k in size, with the advantages of guaranteed reliability via BWS and the additional filtering processes. Our refined dataset ensures the consistency of the included humor annotations.

5 Experiments

5.1 Experimental Setting

To compare and evaluate the performance of the proposed method with the other models compared on the task of classifying humor in memes, we used 1,411 memes in the refined dataset with both upper and lower captions.

To train and evaluate the models, we conducted a ten-fold cross-validation, in which 1,411 memes were randomly divided into ten subsamples such that all memes with the same template image belonged to the same subsample. The memes were distributed such that all subsamples had approximately the same amount of memes. The subsample with a minimum number of memes had 131 memes, and that with a maximum number had 153.

To evaluate the models, each model was trained and evaluated ten times with the data divided into training, validation, and testing sets with a ratio of 8:1:1. For each evaluation step, the weight of the model that achieved the highest classification accuracy on the validation set was chosen to be evaluated on the testing set. We recorded classification accuracy scores calculated on ten different testing sets.

In addition, to minimize the effect of random initialization of the MLP model on the results of the evaluation, we conducted ten-fold cross validation with eight runs over different random seeds. Therefore, a total of 80 accuracy scores were obtained from each model, and the average accuracy score and standard error for each model were used for quantitative comparisons.

5.2 Models for Comparison

To validate the performance of the proposed model, we experimented with two additional models.

The first model encoded meme template images and captions into visual and textual features using a pretrained ViLT model (Kim et al., 2021b). As meme captions can be divided into upper and lower captions, a [SEP] token was inserted between these two parts before they were transformed into word embeddings. The output features of the ViLT encoder were then fed into an MLP model designed to output the probability with which a given meme could be classified as humorous.

In our proposed model, we supposed that the subtracted features of CLIP represented incongruity between an image and its caption, and considered this useful to improve performance on the humor classification task. To validate this statement, we implemented another model which did not subtract features provided by CLIP, but instead concatenated both encoded features of images and their captions to the ViLT output.

5.3 Parameters and Optimization Settings

All models in the experiment used three-layer MLPs with two hidden layers with a dimension of 768. A dropout layer with a dropout probability of 0.5 was added to all models to prevent overfitting. The models were trained with the objective of minimizing the binary cross-entropy loss using the Adam optimizer (Kingma and Ba, 2015). The weight decay parameter was set as 0.01, and the learning rate as 0.0001 for all models.

6 Results and Discussion

6.1 Quantitative Analysis

The result of the experiment is shown in Table 2. From the Table, it was observed that the proposed model (ViLT + CLIP incongruity) outperformed all other models for comparison. First, the proposed model was able to achieve around 5% better results compared to the model using only ViLT. This shows that the module designed to extract incongruities between image and text improved performance. Furthermore, the proposed model also outperformed the model that used ViLT and full

Model	Accuracy
ViLT	53.0 ± 0.2
ViLT+CLIP full feature	56.7 ± 0.4
ViLT+CLIP incongruity	**57.7 ± 0.4**

Table 2: The table show results of humor classification performance of the proposed model (ViLT + CLIP incongruity) and the other models compared.

CLIP features. This further strengthens our proposition that a model able to extract incongruities between image and text performs well in evaluating humor in memes, as the subtraction process of our proposed model extracted incongruities more explicitly compared to the baseline model using full CLIP features.

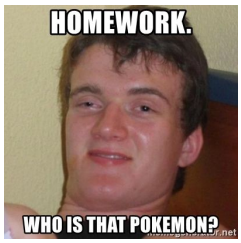
6.2 Qualitative Analysis

We also performed a qualitative analysis of the results to explore the characteristics of the proposed model. To conduct the analysis, we analyzed the classification results of the same testing set for all models used in the experiment.

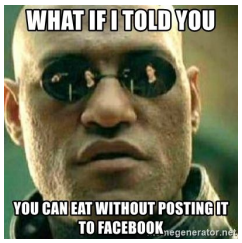
We first recognized that both ViLT and the proposed model were able to identify memes with BWS scores greater than 0.7 as humorous with high accuracy. Out of ten memes with a BWS score greater than 0.7, ViLT was able to correctly identify nine as humorous, and the proposed model was able to correctly identify all ten. This signifies that memes that almost all annotators agreed were humorous were evaluated accurately by both ViLT and the proposed model.

In addition, the proposed model also outperformed other models in evaluating memes with BWS scores less than 0.7. Figure 7 shows two examples of memes that only the proposed model was able to correctly identify as humorous. The two memes involve incongruity between the image and the caption. The meme on the left is humorous because of the incongruity of an adult man making a childish statement about not wanting to do homework. In addition, the meme on the right also involves an incongruity between the image and the caption; it shows an intimidating man with a serious face, but the caption is pointing out a trivial notion, mocking people who post their every meal on social media.

In contrast, some examples of the analyzed memes showed a limitation of the proposed model; for some meme image templates, the proposed model seems to have output the classification based only on the image. For example, for all memes created from a template image called "sad-trooper", the proposed model predicted the memes as not being humorous regardless of their captions. While we could not identify the cause of this limitation, it is possible that for some template images, the image feature vector obtained by CLIP was embedded in a space far from the embedded vectors of other meme image templates and captions. This would produce subtracted feature vectors that are almost the same for all memes with a given image template regardless of their captions.



(a) This meme is humorous because of the incongruity of an adult man making a childish statement about not wanting to do homework.



(b) This meme is humorous because of the incongruity between an intimidating man with a serious face and the trivial notion of mocking people who post their every meal on social media.

Figure 7: The figures show two examples of memes with image-caption incongruity, which only the proposed model was able to correctly identify as humorous.

7 Conclusion

Constructing a computational system to evaluate humor in memes is difficult due to the lack of datasets of memes with reliable and consistent humor annotations and the complexity of searching and extracting cross-modal incongruities between images and their captions. To overcome these challenges, we first created a dataset of memes annotated using BWS and proposed a refining process which was able to eliminate examples of memes affected by interpersonal differences in the perception of humor. Then, we used the refined dataset to train and validate the effectiveness of the proposed method, which was designed to extract incongruities between images and their captions to accurately classify whether a given meme is humorous. The experimental results showed that the proposed model was able to extract and utilize incongruities between images and their associated captions to outperform other multi-modal models on the humor classification

task. This demonstrates the importance of using features that represent incongruities when evaluating humor in memes. Possible future work includes using the features representing incongruities not only to evaluate but also to generate new humorous memes from text or image input.

Acknowledgments

This work was partially supported by JST AIP Acceleration Research JPMJCR20U3, Moonshot R&D Grant Number JPMJPS2011, CREST Grant Number JPMJCR2015, JSPS KAKENHI Grant Number JP19H01115, and Basic Research Grant (Super AI) of Institute for AI and Beyond of the University of Tokyo. We would like to thank Hiromichi Kamata and Kohei Uehara for the helpful discussions. Furthermore, we would like to thank Yusuke Kurose and Miyuki Kajisa for their support in creating the dataset, and Editage (www.editage.com) for English language editing.

References

Issa Annamoradnejad and Gohar Zoghi. 2020. Colbert: Using bert sentence embedding for humor detection. *arXiv preprint arXiv:2004.12765*.

Hans Baumgartner and Jan-Benedict EM Steenkamp. 2001. Response styles in marketing research: A cross-national investigation. *Journal of marketing research*, 38(2):143–156.

Moniek Buijzen and Patti M Valkenburg. 2004. Developing a typology of humor in audiovisual media. *Media psychology*, 6(2):147–167.

Arjun Chandrasekaran, Ashwin K Vijayakumar, Stanislaw Antol, Mohit Bansal, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2016. We are humor beings: Understanding and predicting visual humor. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference*.

Adam Finn and Jordan J Louviere. 1992. Determining the appropriate response to evidence of public concern: the case of food safety. *Journal of Public Policy & Marketing*, 11(2):12–25.

R Kelly Garrett and Shannon Poulsen. 2019. Flagging facebook falsehoods: Self-identified humor warnings outperform fact checker and peer warnings. *Journal of Computer-Mediated Communication*, 24(5):240–258.

Yingmei Guo, Jinfa Huang, Yanlong Dong, and Mingxing Xu. 2020. Guoym at SemEval-2020 task 8: Ensemble-based classification of visuo-lingual metaphor in memes. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1120–1125, Barcelona (online). International Committee for Computational Linguistics.

Sojung Claire Kim, Emily K Vraga, and John Cook. 2021a. An eye tracking approach to understanding misinformation and correction strategies on social media: The mediating role of attention and credibility to reduce hpv vaccine misperceptions. *Health communication*, 36(13):1687–1696.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021b. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. The Proceedings of Machine Learning Research (PMLR).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Svetlana Kiritchenko and Saif M. Mohammad. 2016. Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. In *Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, San Diego, California.

Jordan J. Louviere. 1992. Experimental choice analysis: Introduction and overview. *Journal of Business Research*, 24(2):89–95.

Paul E. McGhee. 1989. Chapter 5: The contribution of humor to children's social development. *Journal of Children in Contemporary Society*, 20(1-2):119–134.

Saša Petrović and David Matthews. 2013. Unsupervised joke generation from big data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–232, Sofia, Bulgaria. Association for Computational Linguistics.

Barbara Plester. 2009. Healthy humour: Using humour to cope at work. *Kotuitui: New Zealand Journal of Social Sciences Online*, 4:89–102.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. The Proceedings of Machine Learning Research (PMLR).

Victor Raskin. 1985. *Semantic mechanisms of humor*. D. Reidel.

Julia Rayz. 2004. Computationally recognizing wordplay in jokes. *Cognitive Science - COGSCI*.

Willibald Ruch and Franz-Josef Hehl. 2010. A two-mode model of humor appreciation: Its relation to aesthetic appreciation and simplicity-complexity of personality. In *The sense of humor*, pages 109–142. De Gruyter Mouton.

Howard Schuman and Stanley Presser. 1996. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Chhavi Sharma, Deepesh Bhageria, William Paka, Scott, Srinivas P Y K L, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis-The Visuo-Lingual Metaphor! In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain. Association for Computational Linguistics.

Emily K Vraga, Sojung Claire Kim, and John Cook. 2019. Testing logic-based and humor-based corrections for science, health, and political misinformation on social media. *Journal of Broadcasting & Electronic Media*, 63(3):393–414.

Hiroaki Yamane, Yusuke Mori, and Tatsuya Harada. 2021. Humor meets morality: Joke generation based on moral judgement. *Information Processing & Management*, 58(3):102520.

Sara K. Yeo and Meaghan McKasy. 2021. Emotion and humor as misinformation antidotes. In *Proceedings of the National Academy of Sciences of the United States of America(PNAS)*.

Tomohiro Yoshikawa and Ryosuke Iwakura. 2020. Study on development of humor discriminator for dialogue system. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 24(3):422–435.

Francisco Yus. 2021. Incongruity-resolution humorous strategies in image macro memes. *Internet Pragmatics*, pages 131–149.

Yftah Ziser, Elad Kravi, and David Carmel. 2020. Humor detection in product question answering systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 519–528. Association for Computing Machinery.

A Appendix

A.1 AMT Interface for Annotating Humor in Memes

In this section, we provide examples of the interface shown to annotators of AMT to obtain humor annotations of memes.

In Section 1 (questions 1 - 24), annotators were first asked to select one of three choices on their understanding of the meme provided, as shown in Figure 8. Memes identified as not in English were eliminated in the post-process.

Then, annotators were asked to write up to three words in the caption that were necessary to understand the meme, as shown in Figure 9. If the presented meme was not in English, the annotators were asked to input “NIE” in the first box and leave the other two boxes blank. It was designed such that if an annotator entered a word that is not in the meme presented, the interface would show an error saying, "You may not input a word that is not in the caption".

The two questions shown in Figure 8 and 9 were asked for 12 separate memes within a task, constituting the first 24 questions presented to the annotators.

Finally, in Section 2 (questions 25 - 27), annotators were asked to choose the most and least humorous meme from among the four presented, as shown in Figure 10. It was ensured that annotators could not select the same meme as most and least humorous. Memes presented in questions 25 - 27 are identical to the memes that were annotated in questions 1 - 24.

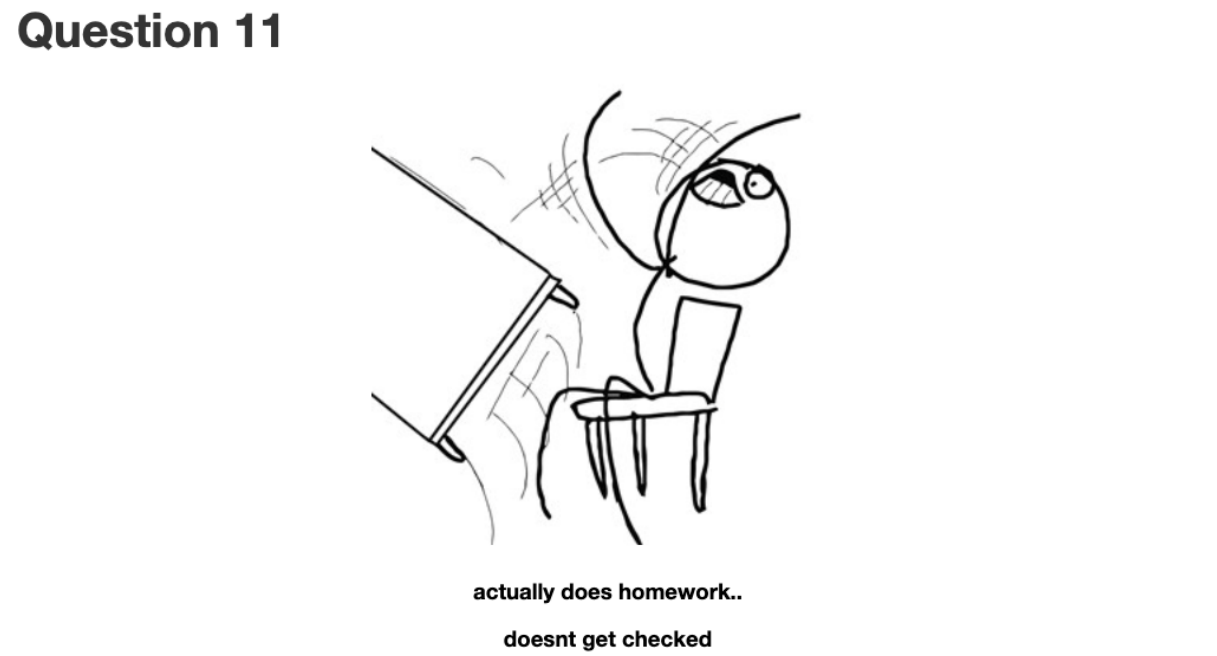


Figure 8: AMT interface asking annotators to select their understanding of the meme. Annotators were asked to choose whether they understood the humor in the meme or if the meme was not in English. This question was used to filter-out memes that were not in English. This question was asked for 12 separate memes in each task.

Figure 9: AMT interface asking annotators to write up to three words that are necessary to understand the meme. This question was used to extract important words to understand the humor in memes (humor anchor).

Figure 10: AMT interface asking annotators to choose the most and least humorous meme out of the four presented. The annotations were used to calculate the BWS score of each meme.