Can Large Language Model Interpret Memes?

KeXin Xu / kx2139
PangLi Yang / py2236
YanFeng Du / yd2727
YouJun Chen / yc7093
Courant Institute of Mathematical Sciences, New York University

Abstract

Memes are a popular way for people to share their ideas and emotions online using pictures and text. Understanding memes involves decoding both visual and textual cues, which presents a unique challenge in the field of automated image understanding. This study builds upton the research introduced in the "MEMECAP: A Dataset for Captioning and Interpreting Memes" paper, by applying newer versions of the LLaMA model to interpret meme content. We replicate the original framework and test its effictiveness with the updated model on the MEMECAP dataset, which contains a diverse collection of meme images paired with human-annotated captions. To evaluate the effectiveness of the updated model, we utilized the BLEU (Bilingual Evaluation Understudy), BERT-F1 metrics, and ROUGE-L matrices to assess the linguistic and semantic alignment of the generated captions with reference captions. This report presents our experiments, quantitative results, and a qualitative analysis of the model's performance in understanding and captioning memes. Although the biases inherited in meme culture limit us from ideally evaluating the correctness of the model performance, we showed that LLMs have the strong ability to interpret the meaning of memes, and we also find the performance remains even if we make image captioning model to replace human labeling.

1 Introduction

As a pervasive form of online communication, memes encapsulate complex cultural sentiments and are often infused with metaphorical meanings that challenge straightforward interpretation.(1) These digital artifacts typically consist of an image paired with a contextual text, which conveys messages ranging from humorous to poignant. Given their ubiquity and richness, memes offer a fertile ground for exploring the capabilities of advanced computational models in understanding and generating nuanced human-like responses.

This report introduces an experimental application of Large Language Models (LLMs) to the task of meme captioning—an endeavor that goes beyond traditional image captioning by requiring a nuanced understanding of both visual and textual elements as well as their cultural contexts. We utilize the MEMECAP dataset, a resource outlined in the "MEMECAP: A Dataset for Captioning and Interpreting Memes"(1) paper, which comprises a diverse collection of memes from Reddit, annotated with expert-generated captions. Our approach harnesses the power of Zero-shot learning, leveraging the LLaMA series of models developed by Meta AI. These transformer-based models are particularly suited for tasks that integrate large volumes of text and, although not explicitly designed for Vision and Language (VL) tasks, are adapted in our framework to interpret memes through their textual components alone.

The specific challenge tackled in this project is the interpretation of the metaphorical interplay between the images and the texts of memes. Traditional VL models typically excel in either language or vision tasks but often falter when required to integrate both in culturally and contextually rich ways—as is necessary with memes. By employing LLaMA models in a zero-shot learning setting, our program aims to demonstrate that even language-focused models can effectively generate accurate and contextually appropriate captions for memes, thus extending the boundaries of what AI can comprehend and produce in the realm of digital cultural artifacts.

Our evaluation methodology employs the BLEU score, ROUGE, and BERT-F1 metric to quantitatively assess the linguistic and semantic accuracy of the captions generated by our model. This introduction sets the stage for a detailed exploration of our methods, the challenges encountered, the

solutions implemented, and the implications of our findings on the broader field of AI and cultural computation.

2 Dataset Description

The MEMECAP dataset forms the cornerstone of our project, providing a rich collection of meme images paired with annotated captions that reveal the underlying humor and cultural references. Developed to facilitate the training and evaluation of models on the task of meme captioning, this dataset uniquely combines visual content with textual annotations that interpret the meme's intended message and humor.

The dataset comprises thousands of meme images, each annotated with multiple captions and identified metaphors. These memes are sourced from Reddit. Each image in the dataset is accompanied by the following features:

- **Title and Image**: Each entry includes the title of the meme along with its corresponding visual content, providing context and enhancing understanding.
- **Human-annotated Image Captions**: These are objective descriptions of the visual content of the images, providing an unbiased perspective on what is depicted without inferring the underlying meme intent.
- **Human-annotated Meme Captions**: These annotations convey the intended message or joke of the meme as interpreted by multiple human annotators.
- **Metaphors**: These are detailed identification of entities in the image captions and their corresponding metaphorical meanings, offering insights into the interplay of text and imagery in memes.

For a detailed explanation of the annotation process, we refer the reader to the MEMECAP paper.

3 Literature Review

The study of memes intersects with the broader field of computational humor, which has been an area of interest in artificial intelligence research for decades. Memes, as a form of digital media, combine visual and textual elements to create culturally relevant jokes or commentaries. Understanding and generating meme content involves recognizing not just the explicit content but also the implicit cues and cultural contexts that inform the humor or message. Mihalcea and Strapparava(2) were among the first to explore computational models for humor recognition, setting a foundational methodology for later works focused on memes specifically.

Recent advances in AI have led to the development of models that process both visual and textual data. These Vision-Language Models (VLMs) are trained on diverse datasets from both domains and are typically evaluated on tasks like image captioning and visual question answering. Notable works in this area include the development of the VisualBERT(3) and VLBERT(4) models, which integrate BERT-like architectures to process multimodal inputs. These models, however, often struggle with the metaphorical and cultural layers embedded in memes, highlighting a gap in their ability to process visual metaphors and culturally specific content.

The adaptation of LLMs for meme captioning represents an innovative approach in the field. Although traditionally used for text-based applications, LLMs like LLaMA have shown promise in zero-shot learning setups where they are applied to tasks without explicit training on those tasks. This approach is particularly appealing for meme captioning, as it mimics the human ability to infer and generalize from limited data. Studies by Brown et al.(5) on GPT-3 have demonstrated the potential of LLMs to perform a wide array of tasks through such an approach, suggesting a viable pathway for meme captioning as well.

The integration of advanced LLMs in the domain of meme captioning presents a novel challenge that bridges multiple disciplines within AI. By leveraging the unique capabilities of these models in a zero-shot learning framework, researchers can explore new boundaries in the understanding and generation of culturally rich visual-textual content. As this field progresses, it will be essential to continue refining these models and datasets to capture the dynamic and intricate nature of human cultural expressions, such as memes.

4 Methodology

Figure 1 is our framework. Utilizing the MEMECAP dataset, our process begins with extracting textual content in the memes, referred to as `OCR Caption`, using the open-sourced tool EasyOCR. We then construct a detailed prompt that inte-

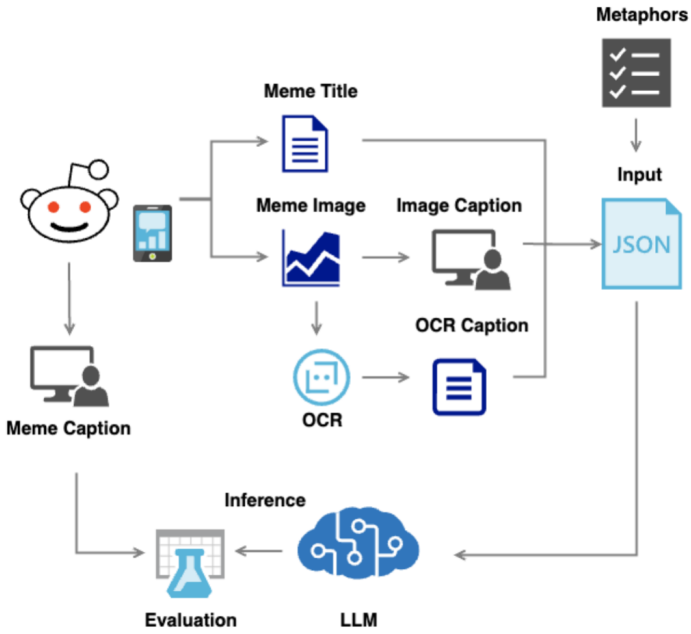

Figure 1: The framework in our study.

grates meme titles, image captions, OCR captions, and annotated metaphors of the meme. The prompt is fed into a LLaMA, a large language model developed by Meta AI, which interprets the contextual information within the memes. We adopt zero-shot learning paradigm, in which the model is not explicitly trained on meme captioning tasks but uses its general language understanding capabilities to interpret what the meme poster is trying to convey based on provided inputs. The implementation can be found at https://github.com/yDu98/MemeCaptioning.

4.1 Model Selection and Setup

We used the LLaMA model series (8) from the family of Large Language Models developed by Meta AI. This decision was guided by LLaMA's demonstrated proficiency in handling complex natural language tasks, its architectural efficiency, and its flexibility in being applied to a variety of tasks including those requiring causal language modeling.

The specific variants of the LLaMA models employed in our study include the LLaMA2, LLaMA3 8B, LLaMA3 8B with additional training on metaphors, LLaMA3 70B, and LLaMA3 70B with enhanced metaphor understanding. These models were chosen to explore the range of capabilities across different scales of model complexities.

4.2 Input Preparation

The input to the LLaMA model consists of a combination of OCR text extracted from the meme and a contextual title,and metophors. This combination is crucial as it provides the model with both the visual text component of the meme and any additional context necessary to understand the image's intent and humor. And after we have output,We can use memecaption to evaluate the output.

4.3 Zero-Shot Learning Approach

In the zero-shot learning setup, the model uses a prompt that combines the meme's OCR text and its title without prior direct training on meme captioning. This approach tests the model's ability to apply its general understanding of language and culture to a new and untrained task, mimicking how humans often interpret memes based on their cumulative knowledge and cultural understanding.

```
# Create the formatted prompt
prompt = f"Human: \nThis is a meme with
the title: '{title}'. \n"
prompt += f"The image description
is: '{image_description}'. "
prompt += f"The following text is written
inside the meme: '{ocr_captions}'. "
prompt += f"Rationale: '{keyword1}' is
a metaphor for '{meaning1}', "
prompt += f"'{keyword2}' is a metaphor
```

```
for '{meaning2}'"
prompt += "What is the meme poster
trying to convey? \n"
prompt += "Please summarize it to one
sentence."
```

4.4 Caption Generation Process

The caption generation process involves feeding the combined input (OCR text ,image caption and title) into the LLaMA model, which then generates several potential captions for each meme. These captions are intended to not only describe the meme but also interpret its humor or underlying message.

4.5 Evaluation Metrics

To evaluate the effectiveness of our captioning approach, we used two primary metrics:

- BLEU Score: The BLEU score measures the linguistic accuracy of the machine-generated captions against the reference captions provided in the MEME CAP dataset. It assesses how closely the model's captions match the human-generated captions in terms of word choice and sentence structure.
- BERT-F1 Score: This metric evaluates the semantic similarity between the generated captions and the reference captions. It uses the BERT model to encode captions into vectors and calculates the F1 score based on the overlap of these semantic vectors, providing a more nuanced measure of the model's performance beyond the surface-level lexical similarity.
- ROUGE-L measures the longest common subsequence between the generated text and the reference text, focusing on the sequence rather than individual words or semantic similarity. This metric is particularly useful for assessing the fluency and order of information in generated captions. It provides insights into how effectively the captioning model reproduces the sequence of ideas and factual content of the reference captions, which is crucial for maintaining logical coherence and informativeness.

5 Example

Metaphors Explained

- **Bot → Apple Corporation:** The term "Bot" represents the Apple Corporation, symbolizing automated or predictable behavior in product updates.
- **Cartoon character → Apple Corporation:** The surprised cartoon character represents Apple, shocked at consumer reactions.
- **Seeing → Lack of sales in Apple's latest iPhone:** The act of seeing here highlights the recognition of declining sales for the new iPhone model despite added features.

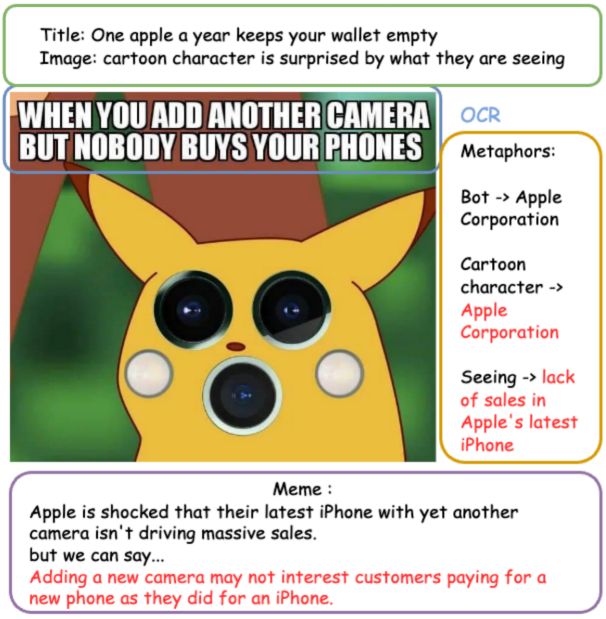

Figure 2: A example of memes

The non-uniqueness of memecaption

Note that meme does not have only one meaning, we can also say:Adding a new camera may not interest customers paying for a new phone as they did as they did for an iPhone.

Sample Output

Input with Metaphors: The meme poster is humorously conveying that Apple's pricing strategy, where they release new iPhones with minimal changes, is not convincing consumers to buy new products.

Input without Metaphors: The meme poster is humorously expressing their surprise and frustration when they add a new camera but nobody is buying their products.

6 Analysis

6.1 Performance Evaluation

We evaluated the performance of the various models with both automatic metrics.

Table 1: Evaluation Results

Evaluation	BLEU	BERT-F1	ROUGE-L
Llama2	13.55	61.58	16.23
Llama3 8B	16.41	60.85	18.18
Llama3 8B+metaphors	16.5	60.61	30.13
Llama3 70B	18.08	62.56	20.21
Llama3 70B+metaphors	19.36	63.61	26.14

Table 1 shows the performance of the various models and input setups in terms of these metrics.

- **Performance Increases with Model Size:** As the model size increases from Llama2 to Llama3 70B, there is a general increase in the metrics, indicating better performance with larger, more capable models.
- **Impact of Metaphors:** Models trained with metaphors (*Llama3 8B+metaphors* and *Llama3 70B+metaphors*) generally perform better in terms of ROUGE-L, significantly so in the case of *Llama3 8B+metaphors*. This suggests that the inclusion of metaphorical understanding improves the model's ability to generate more contextually and structurally coherent captions.
- **Consistent Improvement in BERT-F1 with Model Complexity:** Larger models (70B) with metaphors also show an improvement in BERT-F1 scores, suggesting better semantic understanding and alignment with human reference captions.

6.2 Image Captioning Model

Our previous work relies heavily on human labeling. An intriguing problem is whether we can take advantage of the image-optioning models now that they are mature. Therefore, we use the Microsoft GIT (short for GenerativeImage2Text) model (7) to automate the image labeling processes that used to require human effort. The result is shown in the following table.

Table 2: Evaluation Results

Evaluation	BLEU	BERT-F1	ROUGE-L
Llama3 8B	15.92	60.53	17.24
Llama3 8B+metaphors	16.12	60.46	22.13

Compared to Table 1, we find no significant difference between using human-labeled captions and using the results from the image captioning model. The worse performance when the model interprets images that contain serious inner images may lead to a slightly lower score.

One possible interpretation of this result is that the meaning of a meme mainly lies in the text of the image or other more profound metaphors. Since the OCR model can address the previous one, the remaining metaphor problems will be the last critical problem for meme captioning.

7 Limitation

First, we cannot pass images with other text information together because Llama is a language-oriented model. It will induce inconsistent behavior due to outer dependencies.

Second, memes' true meaning is subjective. Automatic matrices may not always provide the most accurate feedback on the correctness of model inferences. Human evaluation is necessary for rigorous experiments.

Third, metaphors tend to be volatile under different cultures or use cases. Human-labeled metaphors cannot adjust based on such conditions.

8 Conclusion

This study embarked on the innovative endeavor of employing Large Language Models (LLMs) to generate captions for memes, a complex task that intertwines visual elements with layered textual and cultural nuances. Our approach leveraged the zero-shot learning capabilities of the LLaMA models to process and interpret memes solely through their textual components extracted via OCR, without direct training on meme-specific data.

The evaluations conducted using metrics such as BLEU, BERT-F1, and ROUGE-L have demonstrated promising results, showcasing that LLMs can indeed grasp and articulate the subtle interplay of humor and meaning in memes. The incorporation of metaphorical content into the training further enhanced the models' ability to generate contextually and structurally coherent captions, a testament to the potential of LLMs in handling multimodal and culturally rich content.

However, the limitations of our approach, primarily the inability of language-only models to directly process visual information, suggest areas for future research. Integrating actual visual pro-

cessing capabilities, possibly through advanced Vision-Language Models (VLMs), could provide a more holistic understanding of memes. This would not only improve the accuracy of caption generation but also enhance the models' applicability to a broader range of multimodal communication forms.

In conclusion, the success of this project opens up new avenues in the field of computational humor and meme interpretation. It also sets a foundation for further exploratory work that could bridge the gap between purely textual understanding and multimodal communication, expanding the reach of AI in understanding and generating human-like responses in the digital age.

References

[1] Hwang, E., & Shwartz, V. (2023). MemeCap: A Dataset for Captioning and Interpreting Memes. arXiv preprint arXiv:2305.13703.

[2] Mihalcea, R., & Strapparava, C. (2006). Learning to Laugh (Automatically): Computational Models for Humor Recognition. Computational Intelligence, 22(2), 126-142. https://doi.org/10.1111/j.1467-8640.2006.00278.x

[3] Li, L. H., Yatskar, M., Yin, D., Hsieh, C., & Chang, K. (2019). VisualBERT

[4] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2019). VL-BERT: Pre-training of Generic Visual-Linguistic Representations. ArXiv, arXiv:1908.08530.

[5] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.

[6] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and Efficient Foundation Language Models.: A Simple and Performant Baseline for Vision and Language. ArXiv, arXiv:1908.03557.

[7] Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L. (2022). GIT: A Generative Image-to-text Transformer for Vision and Language. ArXiv, abs/2205.14100.

[8] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. ArXiv, abs/2302.13971. https://api.semanticscholar.org/CorpusID:257219404