# Meme Analysis using Image Captioning Model and GPT-4

[1]Marvin John Ignacio, [1]Thanh Tin Nguyen, [1]Jia Wang, [1]Yong-Guk Kim
[1]Dept. of Computer Engineering, Sejong University, Seoul, Korea

**Abstract**

We present a new approach to evaluate the generated texts by Large Language Models (LLMs) for meme classification. Analyzing an image with embedded texts, i.e. meme, is challenging, even for existing state-of-the-art computer vision models. By leveraging large image-to-text models, we can extract image descriptions that can be used in other tasks, such as classification. In our methodology, we first generate image captions using BLIP-2 models. Using these captions, we use GPT-4 to evaluate the relationship between the caption and the meme text. The results show that $OPT_{6.7B}$ provides a better rating than other LLMs, suggesting that the proposed method has a potential for meme classification.

## 1. Introduction

Meme images are pictures with embedded texts that are meant to provide amusement for the audience. However, some memes convey harmful messages and are unsafe to view by certain people. For this reason, a challenge called Memotion [1] is introduced to classify the sentiment, emotion, and intensity of the meaning of meme images using machine learning or deep learning techniques. Most solutions combine the image and textual features in a multimodal setup, such as in the work of Nguyen et al. [2]. However, the results show that encoding images is the main challenge for the task. This is because meme images contain embedded images and texts. Models tend to read the text rather than see the objects inside the image. It is important to recognize the picture minus the text to fully understand the meaning behind the meme. Figure 1 shows an example of meme images.


(a) (b)

Figure 1. The image above shows two examples of meme images. The first image (a) is a funny meme, while the second image (b) is a meme with racist text.

Many researchers have utilized Large Language Models (LLMs) in their work, especially in understanding contexts in input modalities. This leads to Visual Question Answering (VQA) tasks, where a human prompt is used as a question related to an input image. Proper understanding of the image leads to better results in computer vision tasks such as classification. Contextual information also provides explainability. This information allows humans to assess the reason behind the model's output.

This paper presents our approach to evaluating LLMs using GPT-4. Figure 2 describes the overall diagram of our work. In the next section, we discuss the methodology of our study. We then write how our experiment was conducted in Section 3. We provide the results and give our insights in Section 4. Lastly, we summarize and give future directions written in Section 5.

## 2. Methodology

### 2.1 Image Captioning using BLIP-2

Combining and training state-of-the-art computer vision models and LLMs for an image-to-text task is challenging due to being resource-intensive. One simple approach is to train a separate model to align these two models. The work of Li et al. [3] named BLIP (Bootstrapping Language-Image Pre-training)-2 combines Vision Transformer (ViT) [4] with two LLMs: OPT [5] and Flan T5 [6].

To bootstrap a frozen image encoder model with a frozen LLM, an intermediate module called Q-Former is introduced. It performs vision-language representation learning using three objectives. First, it uses Image-Text Matching (ITM) to align the image features with text queries. Images and texts are correlated using self-attention, cross-attention, and a feedforward layer. The second part of the Q-Former is Image-grounded Text Generation (ITG). It comprises self-attention and a feedforward layer, and its purpose is to generate texts using images as the condition. The third objective is the Image-Text Contrastive Learning (ITC). It aligns the image and text representations by utilizing ITM and ITG. Multiple self-attention masking strategies are performed depending on the objective.

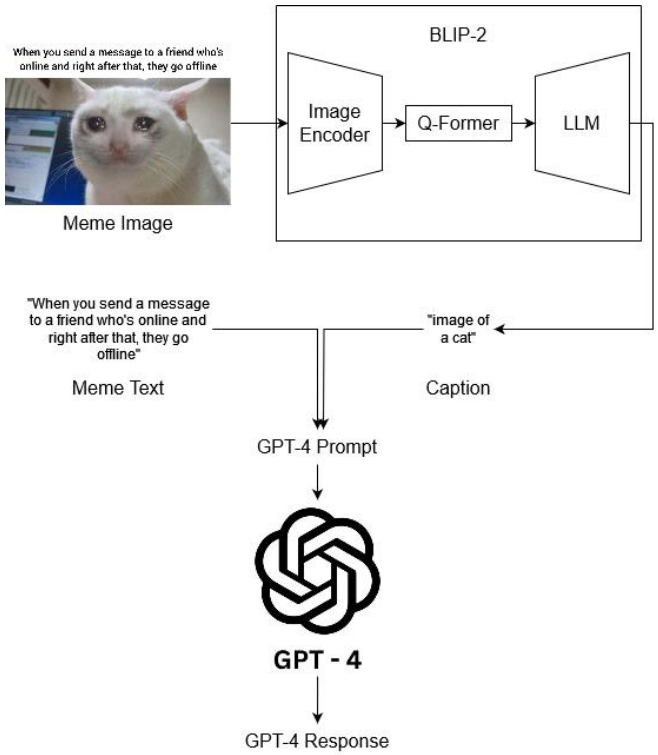


Figure 2. The overall diagram of our methodology. Using BLIP-2, we create an image caption of the meme image to describe the image. Using the meme text included in the Memotion 1 dataset, we create a prompt to GPT-4 asking for the relationship between the meme caption and text.

To use an LLM decoder, BLIP-2 encodes the image using the frozen ViT model. Q-Former then generates queries, which the frozen LLM then uses. Depending on the type of LLM, a prompt can be added to the query. OPT, which is a decoder-based model, does not require a prompt. Flan T5, however, is an encoder-decoder-based model and can accept a prompt.

Li et al. evaluated their setup with existing models for VQA and image captioning tasks. They achieved a score of 65.0 in the VQAv2 dataset in a zero-shot visual question answering task. The model also provided an outstanding CIDEr score of 145.8 in the image captioning task using the COCO dataset. Given this outcome, our methodology will use BLIP-2 to generate image captions.

### 2.2 Context Generation and Evaluation using GPT-4

GPT-4 [7] is a very large multimodal model by OpenAI, surpassing the performance of GPT-3.5 (ChatGPT). It accepts image and text input and generates text responses according to the user prompt. Many researchers use GPT-4 in their research, especially when it comes to understanding contexts and generating human-like responses. However, GPT-4 is not open-source and requires a fee via the OpenAI web interface or API. Our methodology uses GPT-4 to evaluate the relationship between the image caption and the meme text. This provides useful information in classifying meme images. We also ask GPT-4 to evaluate its analysis to determine whether the image caption generated can increase the understanding of GPT-4 regarding the meme context.

## 3. Experiment

The first stage of this experiment is to evaluate the best BLIP-2 configuration for the image captioning task. Five models were used: $OPT_{6.7B}$, $OPT_{6.7B}$-COCO, Flan T5 XL, Flan T5 XL-COCO, and Flan T5 XXL. The models with COCO suffixes are finetuned by the original authors using the COCO dataset. The image encoder for all these LLMs is ViT. We generate image captions on Memotion 1 dataset. This dataset contains 6991 training images. Figure 3 shows an example of a meme with the generated image captions from different BLIP-2 models.

We used a computer having 32 GB RAM and 10 GB dedicated GPU memory. Since we have limited resources, we used quantized models from Hugging Face [8]. The Flan T5 XL models used approximately 7 GB of dedicated GPU memory, while $OPT_{6.7B}$ used 9 GB of dedicated GPU memory. For the Flan T5 XXL, we distributed the layers to utilize both RAM and GPU memory. Approximately 16 GB of RAM and 9 GB of GPU memory were used by this model.


| | |
|---|---|
| Flan T5 XL | i'm sorry ania please don't hate me 3 |
| Flan T5 XL-COCO | two minions are standing next to each other |
| Flan T5 XXL | i'm sorry amina please don't hate me |
| OPT 6.7B | i'm sorry anna please don't hate me |
| OPT 6.7B-COCO | two minions are standing next to each other with a caption |

Figure 3. The image shows an example meme image, together with the captions generated by the BLIP-2 models.

Given that GPT-4 has a cost per API call, we sampled 100 meme images from Memotion 1 dataset, giving us a total of 500 API calls. For each caption generated by the BLIP-2 models, we create a prompt to relate the image caption and the meme text. The meme text is provided in the Memotion 1 dataset. Unfortunately, GPT-4 uses Chat Completions API. It does not provide log probabilities, unlike in Completions API. The log probabilities are helpful in evaluating the response confidence of the generative models. As an alternative solution, we asked GPT-4 to evaluate its responses and give a rating between 1 and 5, with 5 being the highest score. If GPT-4 couldn't produce a response due to its limitations, we put a score of 0 on the evaluation. Figure 4 shows an example of a meme image and the GPT-4 response.


"The meme image uses a man holding a glass of wine as a visual representation of someone making a toast or sharing a profound thought. The caption ""4 things people can't choose"" is the punchline of the meme, listing four things that are generally out of an individual's control, including a humorous political jab at the lack of choice in the Russian presidential elections. I would rate this explanation a 5 as it accurately describes the meme's content and context."

Figure 4. The meme image (left) and the GPT-4 response (right). The caption in this example is: "a man holding a glass of wine and a caption that says, 4 things people can't choose". We can see in the GPT-4 response how it describes the context of the meme and the rating of its response.

Our experiment shows that $OPT_{6.7B}$ performs better than Flan T5 XL/XXL. Table 1 shows the value counts of the rating for each model. We can see in the results that the COCO finetuned models perform less than the non-finetuned models. Most GPT-4 responses are between 4 and 5, indicating that BLIP-2 models are good candidates for image captioning.

Table 1. Value counts of rating for each BLIP-2 models.

| **Rating** | **Flan T5 XL** | **Flan T5 XL-COCO** | **Flan T5 XXL** | **$OPT_{6.7B}$** | **$OPT_{6.7B}$-COCO** |
|---|---|---|---|---|---|
| 5.0 | 39 | 19 | 38 | **47** | 34 |
| 4.5 | 13 | 24 | 20 | 18 | 15 |
| 4.0 | 43 | 45 | 35 | 27 | 38 |
| 3.5 | 0 | 3 | 3 | 2 | 5 |
| 3.0 | 3 | 7 | 2 | 4 | 4 |
| 2.0 | 2 | 1 | 0 | 0 | 2 |
| 0 | 0 | 1 | 2 | 2 | 2 |

## 4. Discussion

Referring to the paper of Li et al., their study showed that $OPT_{6.7B}$ performs better in VQA tasks than the other models. Our experiment also showed that OPT6.7B is better at describing the meme image. This suggests that GPT-4 can evaluate the quality of another model's image captions or VQA responses.

Utilizing LLMs is challenging because they are resource-intensive and expensive to use. Our study shows that using quantized models can reduce a lot of computational resources. This makes future research using LLMs as part of the methodology more feasible.

## 5. Conclusion

We present a method for evaluating BLIP-2 generated image contexts in meme images using GPT-4. Using the image captions generated by the BLIP-2 models and meme texts, we ask GPT-4 to understand and evaluate the quality of the context. Our experiments show that $OPT_{6.7B}$ performs better than Flan T5 XL/XXL. The COCO finetuned models perform poorly than the non-finetuned models.

We also present a way to utilize LLMs in a machine with limited resources by using quantized models from Hugging Face. By using this technique, we can reduce the cost of using proprietary LLMs such as GPT-4. The methodology presented in this paper can be used to improve the meme classification task. In our future work, we will utilize the contexts generated by GPT-4 for the meme classification.

## References

[1] Sharma, Chhavi, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari and Björn Gambäck. "SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor!" *International Workshop on Semantic Evaluation* (2020).

[2] Nguyen, Thanh Tin, Nhat Truong Pham, Ngoc Duy Nguyen, Hai Nguyen, Long H. Nguyen and Yong-Guk Kim. "HCILab at Memotion 2.0 2022: Analysis of Sentiment, Emotion and Intensity of Emotion Classes from Meme Images using Single and Multi Modalities (short paper)." *DE-FACTIFY@AAAI* (2022).

[3] Li, Junnan, Dongxu Li, Silvio Savarese and Steven C. H. Hoi. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." *ArXiv abs/2301.12597* (2023).

[4] Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit and Neil Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." *ArXiv abs/2010.11929* (2020).

[5] Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang and Luke Zettlemoyer. "OPT: Open Pre-trained Transformer Language Models." *ArXiv abs/2205.01068* (2022).

[6] Chung, Hyung Won, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping

Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai-hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le and Jason Wei. “Scaling Instruction-Finetuned Language Models.” *ArXiv abs/2210.11416* (2022).

[7] OpenAI. “GPT-4 Technical Report.” *ArXiv abs/2303.08774* (2023).

[8] Hugging Face. https://huggingface.co/