Semantic Role Labelling for Multimodal Setup

Kumar Arjun
Department of Mathematics
2021MT10232
Course Project ELL884
mt1210232@iitd.ac.in

Harshit Singh
Department of Mathematics
2021MT10257
Course Project ELL884
mt1210257@iitd.ac.in

Shreya Gupta
Department of Mathematics
2021MT10906
Course Project ELL884
mt1210906@iitd.ac.in

Abstract

Memes wield significant influence on social media, leveraging a blend of visuals and text to shape opinions swiftly. Given their potential virality, it becomes imperative to discern their underlying intent and assess any potential harm they may cause, prompting timely intervention. One common challenge in meme analysis revolves around identifying the entities mentioned and determining the role assigned to each. Our objective is to ascertain whether a meme portrays its referenced entities positively, negatively, or as victims. Thus, our focus is on identifying the roles of entities within harmful memes, discerning the 'hero,' 'villain,' or 'victim,' if applicable.

We propose a series of techniques to decode the hidden meaning behind these memes. In contrast to conventional methods that rely on predefined entities, we also explore models that autonomously identify entities for us. This becomes particularly significant when dealing with cryptic memes aiming to conceal the identities of the parties involved.

1 Introduction

Visual Semantic Role Labeling (V-SRL) in memes is a novel task that aims to unravel the layers of meaning embedded within visual humor. Memes, as cultural artifacts, often rely on a complex interplay between images and text to convey nuanced messages and elicit specific emotions. V-SRL involves analyzing memes at a semantic level, identifying the visual elements (such as characters, objects, and actions) and their corresponding roles and relationships within the meme's narrative or joke structure. This process requires understanding not only the literal content of the image, but also the implied context, cultural references, and humor conventions. By applying V-SRL techniques, researchers can uncover the underlying semantics of memes, enabling a deeper understanding of their communicative power and cultural significance. We propose a framework for V-SRL in memes, outlining the challenges, methodologies, and potential applications in fields such as computational humor, social media analysis, and cultural studies.

Figure 1: Memes

2 Related Work

The paper "Characterizing the Entities in Harmful Memes: Who is the Hero, the Villain, the Victim?" introduces a visual-semantic role detector focused on categorizing entities into one of three classes ('HERO', 'VILLAIN', 'VICTIM', or 'OTHERS'). This model utilizes a multi-modal setup with OTK fusion, integrating both visual and textual information. While effective, this model relies on pre-labeled entities from the memes to classify them.

In our approach, we employ In-Context Learning (ICL) to achieve entity recognition, thereby eliminating the need for annotated data. This approach enables one to scale to larger datasets with-

out annotations, significantly expanding the scope of the model's applicability and capabilities. Additionally, we address the issue of poor quality OCR (Optical Character Recognition) text commonly found in memes, which often contains links to websites and lacks alignment with natural language, thus hindering the performance of language models pre-trained on typical text data.

To overcome this challenge, we introduce image captioning, allowing for better recognition of image cues. Simultaneously, by providing consistent textual input through image captions, we enhance the performance of language models, enabling them to better understand and process the textual content of memes. This combined approach lays the groundwork for a more effective analysis of multi-modal data in various applications.

3 Methodology

As Large Language Models (LLMs) continue to expand their capacity to capture vast swathes of world knowledge, Semantic Role Labelling (SRL) has evolved significantly and continues to be used in a variety of tasks. However, alongside these advancements come inherent limitations, particularly in the realm of integrating visual data. Traditional SRL techniques primarily rely on textual cues, struggling to effectively contextualize visual information. To address this gap, we propose innovative methods aimed at bridging the semantic divide between textual and visual data within the SRL framework. By incorporating visual data into the contextual understanding of language, we aim to enhance the accuracy and depth of semantic role assignments. By integrating In-Context Learning and Caption-based Semantic Role Labeling (SRL) Models into our approach, we enhance our capability to perform this task. Leveraging the more natural-looking text generated by the captioning model, we move beyond mere OCR information extraction, gaining a deeper understanding of the images at hand. This combined methodology allows us to extract richer contextual information from the captions, enabling us to comprehend the images in greater detail. By incorporating the textual descriptions provided by the captioning model, we supplement the visual cues extracted from the images, leading to a more holistic interpretation of the data.

3.1 CaptionSRL

3.1.1 Vision Module

Although Optical Character Recognition (OCR) can provide a basic description of the content within an image, it often falls short in conveying the complete context. These descriptions are typically vague and may not capture the full picture accurately. To overcome this limitation and better understand visual cues within images, we integrate a bounding box approach alongside image captioning techniques. This combined approach allows us to identify specific elements within the image and generate descriptive captions that provide a more comprehensive understanding of its contents. By leveraging both methods, we enhance our ability to interpret and extract meaningful information from visual data. Further it provides our language model with a more consistent and natural language compared to OCR.

3.1.2 Bounding Boxes

We leverage the power of YOLOv8 to construct bounding boxes around objects within our memes. This advanced object detection algorithm allows us to precisely identify the locations of various elements within the meme images. YOLOv8, or You Only Look Once version 8, is a state-of-the-art object detection algorithm that employs bounding boxes to identify and localize objects within images or videos. Unlike traditional object detection methods that slide a window across the image and classify each patch separately, YOLOv8 divides the image into a grid and predicts bounding boxes and class probabilities for each grid cell simultaneously. The bounding boxes predicted by YOLOv8 are represented by four coordinates: the x and y coordinates of the box's center, its width, and its height. These bounding boxes are then refined and adjusted to tightly fit around the detected objects. By utilizing YOLOv8's bounding box approach, we can precisely identify the location and extent of objects within images, enabling further captioning to be more descriptive.

3.1.3 Image Captioning

The Salesforce BLIP (Bottom-Up and Top-Down Long-Short Term Memory Intersection) model is a cutting-edge approach for image captioning. Unlike traditional methods that generate captions solely based on the visual features of the entire image, BLIP takes a dual approach by integrating both bottom-up and top-down processing. In

the bottom-up phase, BLIP extracts a diverse set of visual features from the image using a pre-trained object detection model, such as Faster R-CNN. These features represent various objects, regions, and their relationships within the image. This bottom-up approach ensures that a wide range of visual information is captured and considered during caption generation. In the top-down phase, BLIP employs a Long Short-Term Memory (LSTM) network, a type of recurrent neural network (RNN), to generate captions based on the extracted visual features. The LSTM network processes the visual features along with previously generated words in a sequential manner, allowing it to incorporate both visual context and linguistic context into the captioning process. Through the integration of the Salesforce BLIP model and Bounding Boxes into our captioning workflow, we are able to provide captions that offer a holistic understanding of the image, capturing not only its visual content but also the nuances and interactions between various elements.

3.1.4 Semantic Role Modelling

We use BERT (Bidirectional Encoder Representations from Transformers) for Semantic Role Modeling (SRM). BERT, being a powerful language representation model, learns contextual embeddings of words or phrases by considering their surrounding context bidirectionally. This means that BERT can capture intricate semantic relationships between words in a sentence. We extend our BERT input by adding our labeling tokens at the conclusion of the sequence. Through this modification, we enhance the capability of our model to understand and incorporate specific labeling information relevant to our task. Subsequently, we employ a classifier to further fine-tune our model, focusing its training towards this particular objective.

3.2 In Context Entity Generation

In the context of Large Language Models (LLMs), "in context" refers to the ability of these models to understand and generate text based on the surrounding context of a given input. LLMs, such as GPT (Generative Pre-trained Transformer) models, are trained on vast amounts of text data and are designed to generate coherent and contextually relevant text based on the input provided to them. When a prompt is given to an LLM, it processes the input text along with any preceding context to generate the next sequence of words or sentences. This allows the model to produce text that is consistent with the context provided, enabling it to continue a conversation, complete a story, answer questions, or perform other language tasks in a coherent manner. The "in context" capability of LLMs is crucial for their effectiveness in natural language understanding and generation tasks. By considering the surrounding context, these models can produce more accurate and contextually relevant responses, making them highly versatile and adaptable to a wide range of applications, including text completion, summarization, translation, and dialogue generation.

In practical terms, the "in context" capability of LLMs means that they can effectively capture and leverage the nuances of language and context to generate human-like text responses that are contextually appropriate and linguistically coherent. This is achieved through the sophisticated architecture and training procedures of these models, which enable them to capture and encode complex patterns and dependencies in language data.

In our Semantic Role Labeling (SRL) process, we implement in-context learning to generate entities. This approach allows us to leverage the contextual information present within the text to better understand and identify the entities involved in the semantic roles. Instead of relying solely on predefined lists or dictionaries of entities, in-context learning enables our model to dynamically generate entities based on the surrounding context of the text. By considering the broader linguistic context, including nearby words and phrases, our model can make more informed decisions about which words represent entities within the sentence.

3.3 OpenFlamingo

OpenFlamingo provides a framework for implementing multimodal in-context learning, where the model learns to integrate information from different modalities within the context of a given task or environment. This approach allows the model to capture rich contextual dependencies between different modalities, leading to more accurate and meaningful representations of the data.

Data Fusion in OpenFlamingo facilitates the fusion of textual and visual data by injecting images alongside text inputs during the training process. This integration allows the model to learn from both modalities simultaneously, capturing

the rich interactions and dependencies between them.

Link to code here

4 Experiments

4.1 Entity Generation

In our research endeavors, we explored innovative techniques for entity generation tasks by employing a zero/few-shot prompting strategy with the OpenFlamingo Model. This approach is particularly noteworthy as it enables us to initiate the entity generation process with minimal input data, thereby alleviating the need for extensive training datasets. By leveraging the zero/few-shot prompting mechanism, we provided the OpenFlamingo Model with a small set of annotated examples extracted from our dataset. These examples were sampled randomly and served as prompts or guidelines for the model, guiding its understanding of the task at hand and facilitating the generation of entities based on the provided context.
Given below is one of the examples of In-Context Learning for our task. Here, the model generated both entities and corresponding labels. However, the focus of the model in this case was identification of entities from the multi-modal data which could then be re-labelled using our model. The illustration below also showcases the inadequacy of the OCR text provided and how that leads to a cascading of errors in the entire task. This also shows how a Zero-Shot setup causes the model biases to propagate, making it harder for the model to adapt to the task at hand.

> *OCR: WHO WOULD WIN? Thanos One hantavirus boy Sorry for the spoiler*
> **hero: romeo vs romeo vs romeo vs romeo vs romeo vs**

However in a few shot setting we are able to much better grasp of knowledge by our model.

> *Example 1: OCR: HOW THE MEDIA CREATES PANIC BEHIND CAMERA IN FRONT OF CAMERA NOTICE ANYTHING DIFFERENT?*
> **hero: [] villain: ['media'] victim: [] other: []**

> *Example 2: OCR: China spent the crucial first days of the Wuhan coronavirus outbreak arresting people who posted about it online and threatening journalists Start tacking the problem Immedately OR draw25*
> **hero: [] villain: ['xi jinping', 'china'] victim: ['journalists'] other: ['jounralists', 'coronavirus', 'wuhan coronavirus', 'wuhan', 'people', 'journallists']**

> *Example 3: OCR: YEAH.UM HUM. IM GOING TO HAVE TO HAVE YOU JUST DO ALL THE WORK WHILE I WORK FROM HOME*
> **hero: [] villain: [] victim: [] other: ['work from home']**

> *Example 4: OCR: Manish eanh Apr 15 RCB dont lose even a pngle manch mas year. What a team kyal Chaleng argioe ta u*
> hero: ['royal challengers bangalore (rcb)'] villain: [] victim: [] other: ['royal challangers bangalore']

> *Test Case:*
> *OCR: WHO WOULD WIN? Thanos One hantavirus boySorry for the spoiler guys*
> **hero: [] villain: ['thanos'] victim: [] other: ['hantavirus']**

Despite being provided with just four examples, the OpenFlamingo model demonstrates adaptability in adjusting to the data, enabling it to identify entities. However, it still struggles to accurately classify entities such as 'Thanos' and 'Hantavirus' into specific classes.

4.2 Captioning Model

The bad text from OCR doesn't allow us to use the full power of large laguage models. OCR often returns illformed words or misses end-tken symbols and concatenates these texts. For Example the text used above best illustrates it.

OCR: WHO WOULD WIN? Thanos One hantavirus boySorry for the spoiler guys

In such scenarios, we leverage Bounded Boxes in conjunction with the Salesforce-BLIP model to conduct captioning of images. This approach enables us to generate more natural and contextually relevant texts that not only describe the contents of the image but also provide a comprehensive understanding of the depicted scene. By utilizing

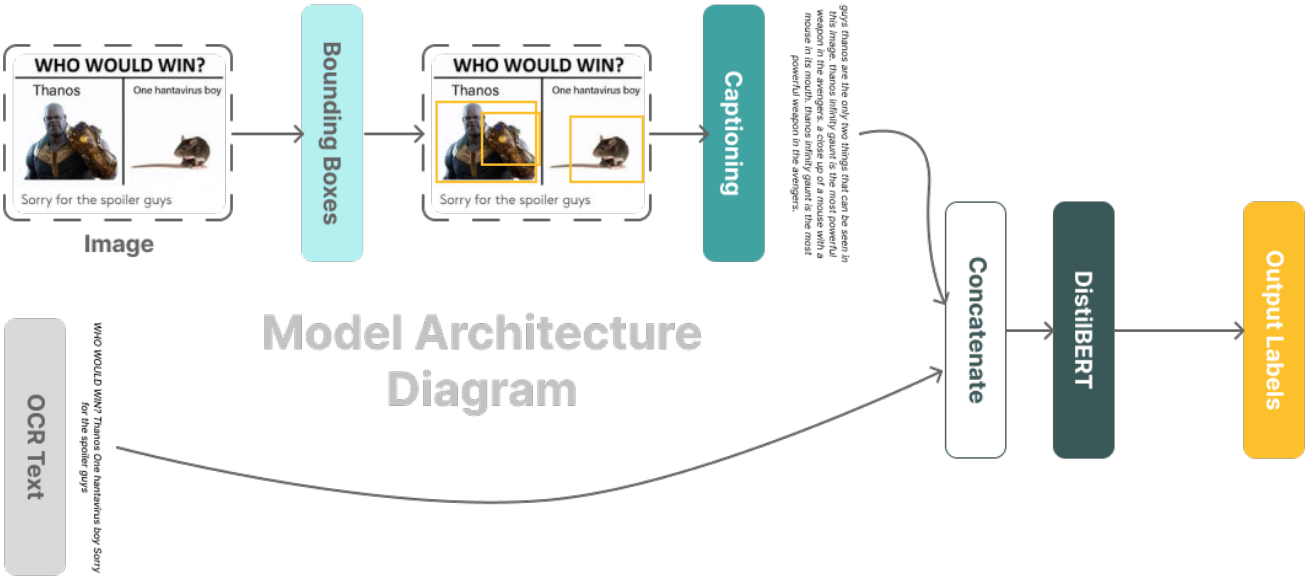

Figure 2: Model Architecture Diagram

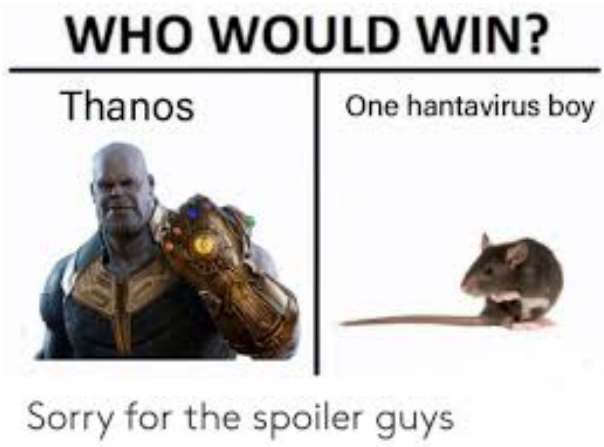

Figure 3: Query Image

Bounded Boxes, we precisely delineate the regions of interest within the image, ensuring that the captioning process focuses on the relevant visual elements. The Salesforce-BLIP model, renowned for its proficiency in understanding and interpreting complex visual data, is then employed to generate captions that encapsulate the essence of the image in a natural and fluent manner. Through this combined methodology, we are able to produce captions that go beyond simple descriptions, offering insights into the actions, interactions, and contextual nuances depicted within the image.

OCR: WHO WOULD WIN? Thanos One hantavirus boy Sorry for the spoiler guys thanos are the only two things that can be seen in this image. thanos infinity gaunt is the most powerful weapon in the avengers. a close up of a mouse with a mouse in its mouth. thanos infinity gaunt is the most powerful weapon in the avengers.

4.3 CaptionSRL

We compared both distillBERT and BERT based models and received good results compared to the State of the Art. In our research, we utilized the

Models	Accuracy	Precision	Recall	F1
Dist BERT	0.752	0.724	0.752	0.744
BERT	0.801	0.676	0.822	0.722

Table 1: Example Table

HVV harmful memes dataset, encompassing a diverse range of content that includes both US political memes and memes pertinent to the COVID-19 pandemic. By fine-tuning our BERT models on this dataset, we aimed to refine the focus of the models specifically towards the political context embedded within the memes.

5 Discussion

The implementation of in-context generation of entities represents a significant advancement in natural language processing and Computer Vision, particularly in the realm of understanding memes and uncovering hidden meanings within them. By dynamically generating entities based on contextual cues, we broaden our horizons to a more nuanced and adaptive approach to processing textual and visual data in a multimodal setup.

6 Conclusion

Through the integration of BERT and Image Captioning with bounded boxes, SRL transcends traditional limitations by capturing intricate contextual nuances and dependencies from image and OCR, enabling a more comprehensive and nuanced understanding of visual-language semantics. Decoding important visual cues from the image captioning alongside with BERT's bidirectional processing and contextual embeddings empower SRL models to discern subtle semantic relationships and nuances, resulting in more accurate and insightful semantic role annotations.

7 Future Works

1. **Dataset Expansion** We aim to demonstrate the effectiveness of our In-Context Learning model in generating meme datasets without relying on annotators or incurring additional costs. This approach allows us to leverage the inherent capabilities of the model to learn from the data contextually, eliminating the need for manual annotation and associated expenses.

2. **ICL based approches for few shot labelling** While the In-Context Learning (ICL) model excels at identifying entities within text, it may still exhibit biases when labeling roles based on context. This issue becomes particularly salient in the context of memes, where the data is often cryptic and challenging to interpret based on limited examples.

   Despite its proficiency in entity identification, the ICL model's role labeling process may be influenced by inherent biases present in the training data or the model architecture itself. These biases can manifest in the form of skewed role assignments, potentially leading to misinterpretations or misrepresentations of the underlying message conveyed by the memes.

3. **Vision Module Improvements** Despite our ability to grasp the visual content of images through bounded box captioning, we still struggle to contextualize the fundamental elements essential for human-level interpretation of memes. To overcome this bottleneck, it is imperative to develop stronger embeddings for visual cues, which would enable us to capture the intricate nuances and implicit meanings embedded within the visual content. Enhancing the embedding representations of visual cues would allow us to encode a richer and more comprehensive understanding of the images, going beyond mere object recognition to capturing the subtleties of visual metaphors, symbolism, and cultural references inherent in memes. By leveraging advanced techniques in computer vision and multimodal learning, we can extract deeper insights from the visual content, facilitating a more nuanced interpretation of the memes.

Acknowledgments

We would like to express our sincere gratitude to the ELL884 course and Tanmoy Sir and the TAs for their invaluable support and guidance throughout our learning journey. The course has provided us with a comprehensive understanding of NLP and Language Modelling, equipping us with essential knowledge and skills that will undoubtedly benefit us in our future endeavors.

References

1. Waheeb Abu-Ulbeh, Maryam Altalhi, Laith Abualigah, Abdulwahab Almazroi, Putra Sumari, and Amir Gandomi. 2021. Cyberstalking victimization model using criminological theory: A systematic literature review, taxonomies, applications, tools, and validations. Electronics, 10:1670.

2. Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multimodal memes.

3. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

4. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models, Anas Awadalla and Irena Gao and Josh Gardner and Jack Hessel and Yusuf Hanafy and Wanrong Zhu and Kalyani Marathe and Yonatan Bitton and Samir Gadre and Shiori Sagawa and Jenia Jitsev and Simon Kornblith and Pang Wei Koh and Gabriel Ilharco and Mitchell Wortsman and Ludwig Schmidt 2023,

5. Real-Time Flying Object Detection with YOLOv8, Dillon Reis and Jordan Kupec and Jacqueline Hong and Ahmad Daoudi,2023