Meme Detection

Problem

The dataset and task for multimodal meme detection is clearly different from other multimodal tasks such as VCR, VQA and in a way like multimodal rumor detection, rightfully identified by [1]. Most of the times, the image and text are related in a subtle way. The pretrained VL models are trained on captioning datasets such as MSCOCO or conceptual captions. These captions describe objects and their relations in the image and rarely describe abstract or metaphorical meanings. For example - in the image below, the image and text are related by loneliness of the desert.



The caption for the same using [2] , gives “desert landscape in the desert” as caption. It would be hard for language model to relate desert with an attribute - loneliness.

Word play is another kind of meme class which is hard for VL models to establish a relation. For example- in [3] , it is hard for VL model to relation between “lack toes” and “lactose”.

Another kind is metaphorical memes. For example – in [4], it is hard for VL model to relate iron and yatch.

In other kinds of memes, the context is important and it might not be feasible to derive from image or text. For example in [5], it would be hard for VL model to identify that image represents a bombing site, which is Boston marathon and then relate it to running and living longer.

Another set of memes requires knowledge of individuals in the image. For example- [9]

Related Work

In [1], authors identify that visual hint is hard to recognize for image classifier or object detector and incorporate additional information from google web entity detection API. Another additional information is race and gender of identified people in the image, as often there are racial or gender memes. Also, the work uses ensemble of different VL models and achieves top AUC ROC score.

In [6],which is the runner up, authors use a similar ensemble of models as in [1] but they miss on additional information, which seems to add a few points to AUC score. [7] proposes similar ideas.[8] uses similar ensemble of visual linguistic transformer models but does and EA optimization of the ensemble weights. This work also introduces a ranking objective of meme and it’s text confounder. Similar to [1], authors use YOLO9000 to detect race and other tags.

From the competing works, it becomes clear that adding additional information such as race,gender tags helps in detection.

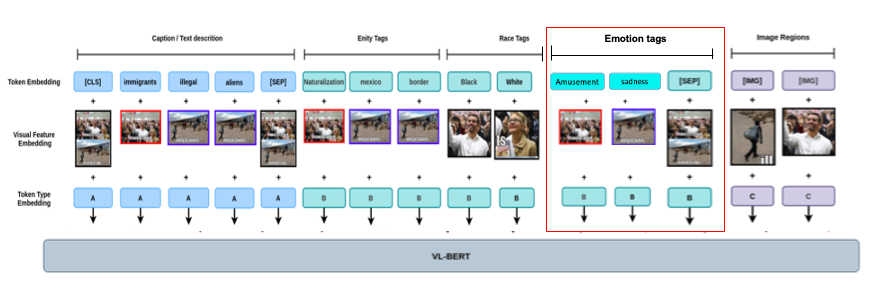
Approach

Quick approximation of how many memes are metaphorical and how many use abstract content. I use caption model from [2] and caption text from hateful meme detection dataset and compare the two for token matches. I added synonyms from wordnet for identified entities in the meme text as well. For the whole sequence, even 1 token match for a meme text and caption is counted as 1. I calculate the average for all memes in the dataset and get a value of .06. This is not an accurate measure but gives a general idea about the dataset.

I try to use emotion as a proxy for capturing overall abstract concepts of the meme image, using smiley labels predicted from [9]. I map the smiley labels to emotion tags, using the mapping provided by the authors and append the emotion tag with entity and race tags in extended VL-BERT model from [1].

A group of people holding flags

Description automatically generated with medium confidence



Highlighted part above shows the additional input on top of inputs used by [1].

For capturing emotions from image, we use smileynet[12]. The authors don’t provide the pretrained model. I take a subset (~300k out of 3.5M images) and train a model. The model generates logits for each of the 91 smiley classes which I map to emotions using the highest positive or negative value of the predicted smiley, using mapping provided by authors.

Results

Trained for 50 epochs:

|  |  |
| --- | --- |
|  | AUC-ROC |
| VL-BERT[1] | 0.76 |
| VL-BERT[1] with emotion tags | 0.75 |

Future work

I could not find any dataset or study on capturing abstract concepts from images. I think we can explore having a classification model which would classify a meme into different classes. Then, separate models for each type. Finally, we can have a weighted ensemble.

For capturing abstract concepts from images, I propose to collect a new dataset, using “representation only” images from news websites. For example- in [10], an object detection model or caption model would focus on people and stairs, but the headline is about decrease in population of young people. This needs further exploration. In some of the memes the relation between entities would not normally exist in a formal construct. For example- in [11], ISIS and goat are related but hard to recognize for a VL model. A google search for the entities ISIS and goat, gives this first result –“I am a radicalised goat hell-bent on jihad – the FBI’s new anti-Isis video game”. This could be explored.

Conclusion

Clearly, adding emotion tags don’t work. We could try with a different model to capture emotions from the images but we must look for a way to add context or learn more meaning from the images.

References:

[1] https://arxiv.org/pdf/2012.08290.pdf- top scorer in the challenge

[2] https://github.com/rmokady/CLIP\_prefix\_caption

[3] <https://www.pinterest.co.uk/pin/445012006911561614/>

[4] https://www.pinterest.com/pin/350436414750252535/

[5] <https://awwmemes.com/i/who-says-running-makes-you-live-longer-memecenter-com-memecentera-boston-986eeb87a5a44bf2ae4f66e1c65967b5>

[6] https://arxiv.org/pdf/2012.07788.pdf

[7] <https://arxiv.org/pdf/2012.12975.pdf>

[8] https://arxiv.org/pdf/2012.12871.pdf

[9] https://imgflip.com/meme/130221676/Hitler-with-dog

[10] <https://www.hindustantimes.com/world-news/amid-pandemic-germany-s-young-population-reduces-to-all-time-low-levels-101658854766383.html>

[11] <https://memegenerator.net/instance/63401128/goat-baa-hide-the-goats-isis-is-coming-to-town>

[12] https://arxiv.org/pdf/1907.06160.pdf

Text Summarization

Problem

The biggest challenge I see is to evaluate quality of generated summary. Bleu and Rouge don’t consider meaning of sentences. Also, to enforce grammatical coherence during generation is a challenge as rightly pointed out by [1].

Related work

[2] uses a 2-stage summarization, where base model generates candidates, and a ranking model predicts the best out of the candidates. The gold label is based on various metrics such as Rouge. In [3] authors use a similar 2 step process but with a ranking loss instead of binary prediction, as in [2]. [4] uses RL setting and instead of rouge to score candidates, it uses semantic encoding (BERT) to rank candidates.

Approach

I decide to introduce RL based candidate generation with [3], instead of fixed candidates. Base model (pretrained sequence transformer such as BART) is trained with an adversarial objective to defeat ranking by the other model, which has twin objectives – generate sequences and rank candidates. After training the candidate model can be thrown away.

Diagram

Description automatically generated

Results

Future work

* Study exposure bias problem more. Has there any other way to generate sequence been tried, so that error doesn’t accumulate (feedback provided and corrects itself, even if preceding tokens are incorrect)?
* Study diverse beam search
* Study PPO RL algorithm
* Explore Pegasus pretrained model
* Understanding the distribution bias in 2nd stage learning

References

[1] https://aclanthology.org/D19-1623.pdf

[2] https://arxiv.org/pdf/2203.06569v1.pdf

[3] https://arxiv.org/pdf/2203.16804v1.pdf

[4] https://aclanthology.org/D19-1623.pdf