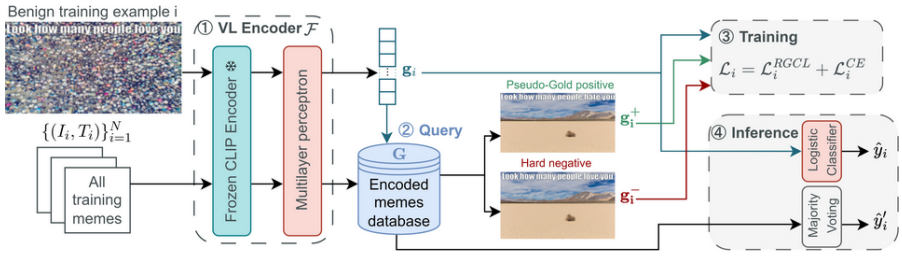
1 Proposed Approach - Retrieval Guided Contrastive Learning (RGCL)

One model we analyzed for Hateful Meme Classification was the RGCL model.



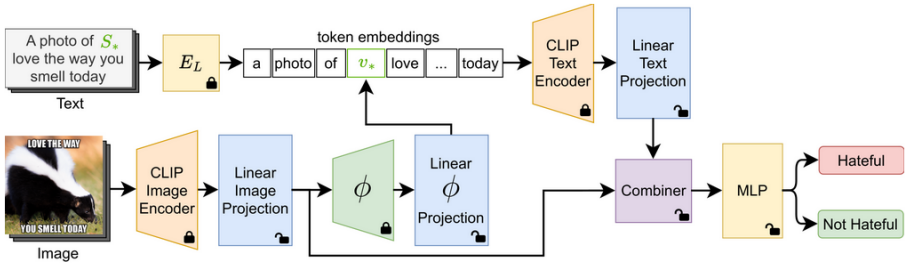
1.1 Custom Input Integration

A method was successfully implemented to allow users to provide custom memes as input. The system processes these memes and determines whether they contain hateful content.
The RGCL model uses as input the meme and a JSON file which contains the image ID and text in the meme. For the custom input integration,

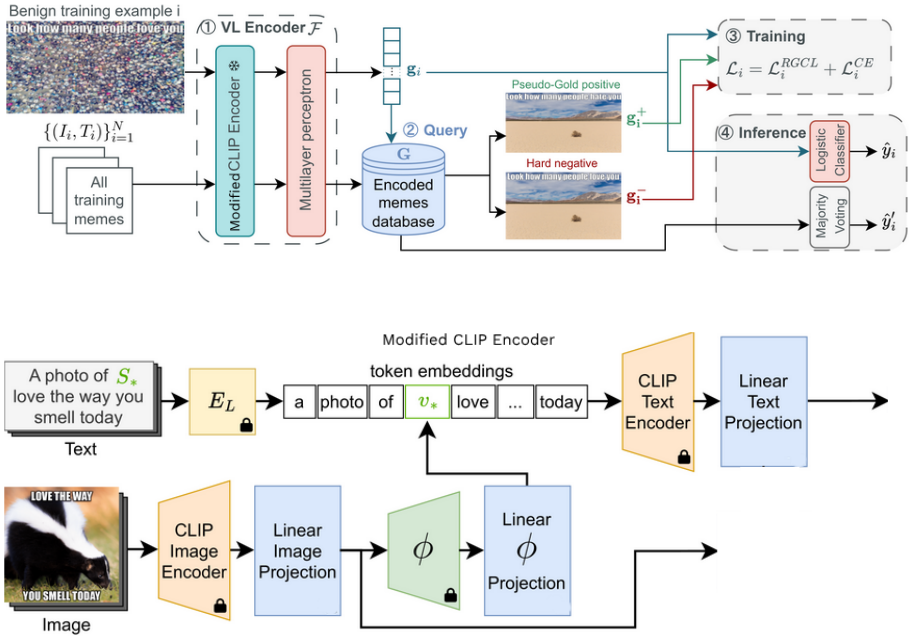
- First, text was extracted from the meme using OCR packages from Jaided AI team.
- Using the text extracted above, the JSON file was created for the memes to be tested.
- The RGCL model was changed to return the classification results for each input
- Finally, the altered RGCL model was run using the above created JSON file and the input images.

1.2 Pre-processing Input Approach

Similarities were noted between RGCL and another model - ISSUES.



The RGCL architecture uses CLIP embeddings in the front and then uses pseudo gold positives and hard negative examples to correctly classify memes. ISSUES on the other hand altered the CLIP architecture by introducing SEARLE textual inversion architecture. Using the pre-trained weights for ϕ the SEARLE architecture was implemented in the CLIP embeddings computation step in RGCL.



1.3 Points of note

- In the custom input integration, the accuracy is lower than the original model as the original model had the correct text in the JSON files, but the new model uses OCR to extract text from the memes and the accuracy of the OCR negatively affects the overall accuracy of the model.
- In the pre-processing input approach, unlike the original implementation of the ISSUES model, no method was added to change the linear projection weights for text and image embeddings during training, the pre-processing was static and was done before training of the RGCL model
- Also, in the pre-processing input approach, unlike the original implementation of the ISSUES model, no combiner was used. The modified CLIP embeddings were directly used by the RGCL model

2 Results and Conclusion

The model performed poorly after the pre-processing. The accuracy dipped from ~ 0.8 to ~ 0.6.
Conclusion:

- Weights from another model should not be used, despite similarities in model structure.
- Training method should be implemented which changes the weights for the projection layers using the loss of the RGCL model.

3 References

- **RGCL** - Mei J, Chen J, Lin W, Byrne B, Tomalin M. Improving hateful meme detection through retrieval-guided contrastive learning. arXiv. 2023 Nov 14. Available from: https://doi.org/10.48550/arXiv.2311.08110
- **ISSUES** - Burbi G, Baldrati A, Agnolucci L, Bertini M, Del Bimbo A. Mapping memes to words for multimodal hateful meme classification. arXiv. 2023 Oct 12. Available from: https://doi.org/10.48550/arXiv.2310.08368
- **HarMeme Dataset** - Singh A, Goswami V, Natarajan V, et al. MMF: A multimodal framework for vision and language research. GitHub. 2020. Available from: https://github.com/facebookresearch/mmf