# This is Going Swimmingly!

Catherine Yu (cyu83), Isabelle Shapiro (iashapir), Aaron Martin (amart172), Jake Regenwetter (jregenwe)
Brown University, Department of Computer Science

## Introduction

Our project is based off of a paper implementing a relational context learning and multiplex fusion network to address the task of **multimodal sarcasm detection** - classifying several modes of inputs as sarcastic or not. This classification task is important because it can help filter misinformation online, and is also helpful for other machine learning tasks such as sentiment analysis. The paper focuses on implementing sarcasm detection *without* the use of graph structures, which are used by many existing sarcasm detection models. Using graph structures presents several limitations - most importantly, it is extremely computationally expensive to create graph networks modeling images and text. Instead, the model aims to understand dynamic relationships between the image and text pairs to capture the context needed to classify something as sarcastic or not. We chose this paper because it addresses a problem that we are all interested in and it builds off of concepts we've discussed previously in class - such as various types of attention.

## Data

In the original RCLMuFN paper, the model was trained on the MMSD and MMSD 2.0 datasets. The MMSD dataset contains pairs of images and related text sourced from social media platforms, some of which are sarcastic in nature, while others are not. The paper's original model was also trained on the MMSD dataset, which contains emojis and hashtags, in addition to MMSD 2.0. That model mainly focused on the presence of tags like "#sarcasm" to make predictions instead of the actual substance of the posts. We chose to instead use MMSD 2.0, which removes those cues.

**text:** got a nice cold for the rest of winter #lovebingill #foff
**label:** 1

We trained and tested our model on the MMSD 2.0 dataset, which contains a total of 24,635 samples, divided into 19,557 training, 2,387 validation, and 2,373 test samples. The text is broken down using the BERT tokenizer., which splits sentences into subword units. All tokenized sequences were padded or clipped to a maximum length of 77 tokens. Text features were encoded to 768-dimensional vectors using BERT. All images were resized to 224x224 pixels, and encoded using ResNet-50 to produce 768-dimensional embeddings for CLIP compatibility.

We also tested the model on a few other datasets. We combined a dataset called MORE with the Flickr8k captioned image dataset. MORE, created by researchers at Cornell for their own Multimodal Sarcasm Explanation model, contains images with only sarcastic descriptions, while none of the Flickr8k dataset is sarcastic. We took the training split from MORE (2983 samples), 2993 random samples from Flickr8k, and shuffled them together to create our custom testing set. Additionally, we tested the model on SarcNet, which contains both sarcastic and non-sarcastic image-text pairs with both English and Chinese captions. For fun, we took some of our own photos (pictured in the "Results" section) and gave them sarcastic captions to see whether our model would correctly classify them.

**text:** so sick of this view ! #besthoneymoonever

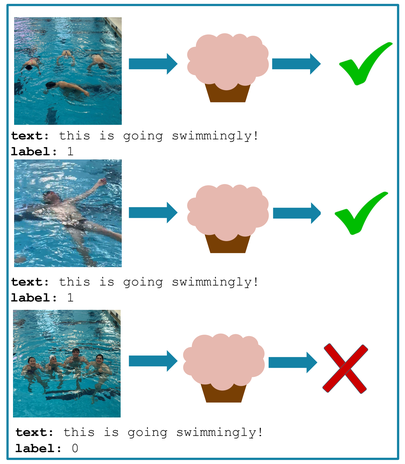
**label:** 1

## Model Architecture

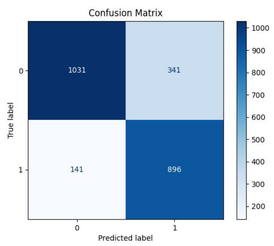
The model contains six main components:

1. **Feature Extraction**
Pretrained models CLIP Image and Text encoder, ResNet-50 and BERT are used to extract initial features from the text and images.
2. **Shallow Feature Interaction Module**
This module uses the features from ResNet-50 (images) and BERT (text), standardizes them using pooling and scaling, and performs cross attention between the two sets of features to learn shallow relational context between the text and images.
3. **Relational Context Learning Module**
This module first learns the context of the image and text features separately, then fuses these features, learning a deeper dynamic relationship between the two. The outputs from step (2) are concatenated with the CLIP features from step (1). Self attention is performed on the text and image features separately. Finally, we perform cross-attention on the self-attention outputs and take a weighted sum.
4. **CLIP-View Feature Fusion Module**
This module fuses the original CLIP features to capture relationships between the original text and image features. We perform cross attention to fuse the features together, then take a weighted sum of those outputs.
5. **Multiplex Feature Fusion Module**
The final step of the model before predicting the class of the text/image pair is to fuse the two streams of cross-attention created from the previous two modules. We first concatenate the weighted sums of steps (3) and (4) and pass these features through a small MLP consisting of a linear layer, ReLU, and Sigmoid. The same concatenated features are passed through a similar MLP that uses Layer Norm instead of Sigmoid. We take a weighted sum of these two MLPs to combine the multimodal features from both the interactions and CLIP.
6. **Prediction Module**
Finally, we pass the output of step (5) into a simple linear layer to get the logits for each class. Softmax is applied within the loss function (cross entropy) to get the output probabilities of each class (sarcastic or not).

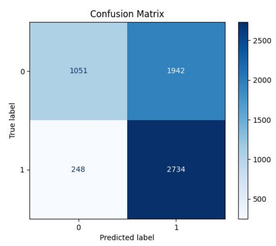
## Results



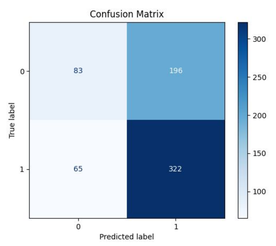

| Dataset tested | Relevant model hyperparameters | Accuracy (%) | Precision (%) | Recall (%) | F1 Score (%) |
|---|---|---|---|---|---|
| **MMSD 2.0** ***(Left figure)*** | 10% dropout before last MLP layer; BERT/Resnet/CLIP frozen; shuffling over entire training dataset; 2 epochs | 79.99 | 72.43 | 86.40 | 78.80 |
| **Muse/Flickr Custom Dataset** ***(Middle figure*)*** | BERT/Resnet/CLIP frozen; **custom frozen batch 2d norm**; shuffling over entire training dataset; 2 epochs | 72.42 | 67.83 | 85.08 | 75.48 |
| **SarcNet** ***(Right figure)*** | 10% dropout before last MLP layer; BERT/Resnet/CLIP frozen; shuffling over entire training dataset; 2 epochs | 60.81 | 62.16 | 83.20 | 71.16 |



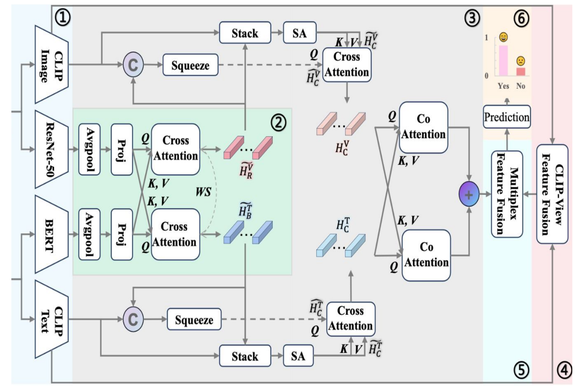

MMSD 2.0




MUSE + Flickr




SarcNet




## Conclusion

Identifying the use of sarcasm within online posts is a difficult problem, one even humans often struggle with. We were able to achieve **79.99%** accuracy on the MMSD 2.0 test split, which is in-line with many comparable models on the MMSD 2.0 dataset, but fell short of the accuracy from the original paper (91%).

## Discussion

**Limitations**
MMSD 2.0 contains only English content, limiting cross-cultural understanding of sarcasm, which varies widely in style and subtlety. Human annotation also introduces bias, as sarcasm is subjective and highly context-dependent.

**Challenges**
One problem we ran into was not being able to train the CLIP and BERT parameters of our model. When running on Oscar, we kept on running into out-of-memory errors, and were unable to request for more GPUs. Another major issue we ran into with our model was that it trained to only classify image-text pairs as sarcastic. We realized that this was due to the training data being shuffled within each batch (of 32 data points) only, instead of shuffling over the entire dataset. After fixing this, we were able to achieve higher accuracy with our model and get predicted outputs of both 0's and 1's; however, our model is still slightly skewed towards predicting things as sarcastic.

**Future Work**
Incorporating positional encoding and multi-scale deformable attention could potentially boost our model's performance. Expanding training data to multiple languages would make the model more universal, and adding cultural context from sources like X or Reddit could further improve results.

## Original Paper

Wang, Tongguan, Junkai Li, Guixin Su, Yongcheng Zhang, Dongyu Su, Yuxue Hu, and Ying Sha. 2024. "RCLMuFN: Relational Context Learning and Multiplex Fusion Network for Multimodal Sarcasm Detection." arXiv. https://arxiv.org/abs/2412.13008