It is Okay to Not Be Okay: Overcoming Emotional Bias in Affective Image Captioning by Contrastive Data Collection – Team 6 Joao Vasco, Lee Kyumhi, Oh sujeong April 25, 2023

1. What was the bias problem tackled and in what task does it appear?

The bias problem tackled in this paper is emotional bias in affective image captioning. Affective image captioning is a task aimed at generating captions that express emotions from images. In this paper, the ArtEmis dataset is important throughout. ArtEmis, termed [Art Emotions], is a large-scale dataset of emotional reactions to images along with language explanations (of chosen emotions). The authors identified two significant issues with previous researches and ArtEmis dataset. Firstly, the emotional distribution in the dataset was unbalanced that it is biased towards positive emotions. Secondly, the captions provided has low specificity for each painting, which means that the language explanations used to explain the images are not accurate enough.

Tackling emotional bias in affective image captioning is important because it can lead to more fair and accurate representation of emotions in images. By using methods like contrastive data collection to collect image-caption pairs that evoke opposite emotions, machine learning models can be trained to generate captions that are less biased towards specific emotions or sentiments, and thus provide a more accurate and diverse range of captions for a given image. The importance of tackling the bias is to enable more realistic and diverse emotional expressions from images. If an image captioning system is trained on a biased dataset, it may generate captions that are always positive, even when the image contains negative or ambiguous emotions. This can be harmful for applications such as mental health support, where the system needs to empathize with user's feelings and provide appropriate feedback.

2. Write a couple of paragraphs explaining what you have understood about the proposed methods to detect and/or deal with the bias.

To solve the unbalanced emotional distribution problem and low specificity problem of ArtEmis dataset, they propose a method called Contrastive Data Collection, which involves collecting and using a set of images that are similar to the training data but have different emotional labels. For details, they made a score called "emotional score" to identify emotionally biased paintings based on the number positive and negative emotions the painting has. For each of painting, they retrieved 100 most similar paintings based on high-level semantic features using layer fc7 from CNN model-vgg16. Out of these 100, they chose 24 paintings, 12 from the nearest neighbors, 12 from paintings had the highest emotional score. The authors conducted collection experiments on Amazon Mechanical Turk(AMT) using two tasks. In the first task, AMT workers were asked to select the most similar painting from a list of 24 visual nearest neighbors that evoked an opposite emotion to a given query painting. In the second task, annotators were asked to specify the primary emotion they felt by observing the selected painting. Trough collection experiment, Authors identified 52933 emotionally biased paintings, total of 260,533 instances to make a new dataset and named it as *Contrastive.* Additionally, dataset that combine it with a random subset from ArtEmis called *Combined.*

Originally, ArtEmis dataset had 62% positive emotions and only 26% of negative emotions. However, after using Contrastive Data Collection method, dataset has a more balanced distribution, with 47% of the samples being positive and 45% being negative. According to the semantic space theory (emphasize emotion labels more) Authors fine tune a RoBERTa language model and use it to predict the extended emotion set of both Combined and ArtEmis. Using histogram of the emotion responses and the Pearson correlation between all pairs of emotion types, they show that the combined dataset is more representative of the Semantic Space theory emotions. To assess the effectiveness of the Contrastive Data Collection method, the authors evaluated the performance of several models on the newly created dataset. The models included a K-nearest neighbors (NN) model, a "Show, Attend and Tell" (SAT) model, and two versions of the M^2 transformer (vanilla and modified). They used several evaluation metrics, such as CIDEr, METEOR, and ROUGE-L, to compare the performance of these models. The results showed that the Contrastive Data Collection method outperformed the baseline method in terms of CIDEr and METEOR scores. While SAT model outperformed the NN model by 28 and 29% higher on METEOR and ROUGE-L scores originally, it increases to 65, 63% respectively, when evaluated on the newly created and combined datasets. In conclusion, the Contrastive Data Collection method resulted in better performance than the baseline method in terms of CIDEr and METEOR scores. The authors used both ArtEmis dataset and new datasets with different models to evaluate the effectiveness of their new method. The results showed that the Contrastive Data Collection method outperformed the baseline method in terms of CIDEr and METEOR scores. Therefore, the authors concluded that their method could be used to create more comprehensive and balanced multi-modal datasets for emotion recognition tasks.

3. Besides the mentioned bias, state and explain 5 other new concepts that you have learned while reading the selected paper or reference papers.

CONTRASTIVE LEARNING : Contrastive learning can be used in order to find similar and dissimilar examples. A complementary dataset was collected in a contrastive manner, pairs of images were presented to amazon workers that would elicit another emotion from the given one. In result we could see improvements in reducing bias.

MESHED-MEMORY TRANSFORMERS : This was also used to improve the accuracy of image captioning. It was used in the context of the paper to generate captions that can be able to distinguish images. It is also trained to describe the accuracy of the images. It takes into account how can the images be related or unrelated to eachother and this model is able to capture more captions than just the simple ones, what will lead to better accuracy results and better captions.

EVALUATION METRICS:
CIDER: Consensus-based Image Description Evaluation & METEOR: Metric for evaluation of translation with explicit ordering. Were metrics used that helped on evaluation the performance of the existing models.
CIDER represents how close can a computer generate a result is in relation to human captions, the higher the score the better in terms of performance.
METEOR on the other hand is used to compare the semantic of the sentence between the generated sentences and the human ones.

NEAREST NEIGHBOURS : Nearest neighbours is usually used in order to present similarities between captions of images. Usually it helps on getting groups of similartities between groups. On the context of the paper it was used as part of the contrastive learning process. The NN were important to calculate examples in order to find out about things that were similar or dissimilar based on the features that each one has.

NEURAL SPEAKERS: Neural speakers generate textual descriptions and can also generate captions for a picture. They are usually used in DL(Deep Learning) techniques. CNN(Convolutional Neural Network-used mainly for image processing and computer vision things) and RNN (Recurrent Neural Network-more used to topics related with natural language processing) are examples of were this could be implemented in order to learn features from the picture provided and generate the correspondent caption.

4. What is your overall opinion about this paper? What insights did you gain? Are there things you agree or disagree with? Is there something you want to add about tackling the bias problem addressed by the paper?

This paper proposed a contrastive data collection approach to balance ArtEmis with a new complementary dataset. The dataset is a pair of similar images that have contrasting emotions such as positive or negative.

We gained some insights from this paper:
- Emotional bias is a common and critical issue in natural language generation tasks, especially dealing with subjective and affective domains.
- Contrastive data collection is an effective way to mitigate emotional bias by exposing the model to diverse opinions on same or similar inputs.

We agreed about this paper:
- Emotional bias in machine learning models can lead negative attitudes toward certain groups.
- Tackling this problem is necessary to promote more fair and inclusive representation of emotions in image captioning.

We have some disagreements or questions regarding this paper:
- This paper does not provide any ethical or social implications of affective image captioning. It would be important to discuss how this task can be used for. It would also be relevant to consider how different cultures or contexts may influence or interpret affective captions differently.

We think that contrastive data collection is a promising direction for reducing bias in natural language generation tasks, but it may not be sufficient or applicable for all cases. It would be interesting to investigate other methods or techniques for mitigating emotional bias beyond contrastive data collection such as applying fairness metrics.

Works Cited

P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. Guibas. Artemis: Affective language for visual art. In *CVPR*, 2021.