Towards Increased Accessibility of Meme Images with the Help of Rich Face Emotion Captions

K R Prajwal
prajwal.k@research.iiit.ac.in
IIIT Hyderabad

C V Jawahar
jawahar@iiit.ac.in
IIIT Hyderabad

Ponnurangam Kumaraguru
pk@iiitd.ac.in
IIIT Delhi

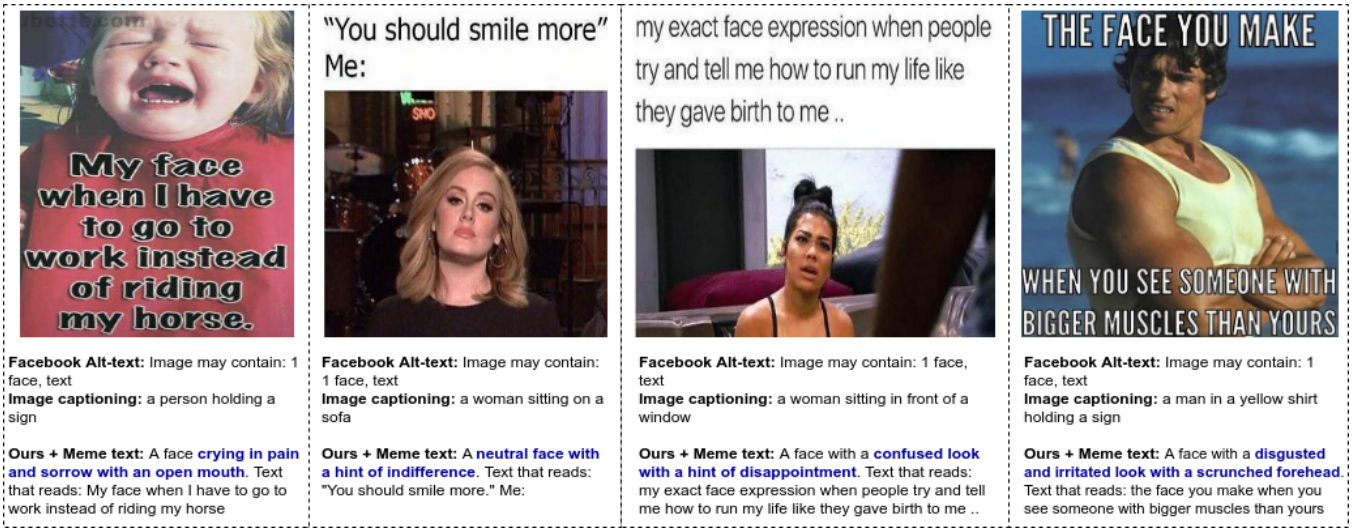

Figure 1: We address the problem of improving the accessibility of meme images, one of the most popular kinds of multimodal content in social media. We focus on face emotion, a key visual concept in meme images and describe it with a rich, fine-grained caption. In contrast, contemporary assistive technologies like Facebook Automatic Alt-Text and image captioning models miss out on the salient visual details that are necessary for a visually impaired user to understand a meme.

ABSTRACT

In recent years, there has been an explosion in the number of memes being created and circulated in online social networks. Despite their rapidly increasing impact on how we communicate online, meme images are virtually inaccessible to the visually impaired users. Existing automated assistive systems that were primarily devised for natural photos in social media, overlook the specific fine-grained visual details in meme images. In this paper, we concentrate on describing one such prominent visual detail: the meme face emotion. We propose a novel automated method that enables visually impaired social media users to understand and appreciate meme face emotions with the help of rich textual captions. We first collect a challenging dataset of meme face emotion captions to support future research in face emotion understanding. We design a two-stage approach that significantly outperforms baseline approaches across all the standard captioning metrics and also generates richer discriminative captions. By validating our solution with the help of visually impaired social media users, we show that our emotion captions enable them to understand and appreciate one of the most popular classes of meme images encountered on the Internet for the first time. Code, data, and models are publicly available[1].

CCS CONCEPTS

• **Human-centered computing** → **Accessibility**; **Accessibility technologies**;

KEYWORDS

Face emotion understanding; Accessibility; Memes; Image captioning; Transfer learning

ACM Reference Format:
K R Prajwal, C V Jawahar, and Ponnurangam Kumaraguru. 2019. Towards Increased Accessibility of Meme Images with the Help of Rich Face Emotion Captions. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350939

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350939

[1] http://precog.iiitd.edu.in/research/social-media-4-all

1 INTRODUCTION

Internet Memes have risen to tremendous popularity in the past decade[2] as a universal media to rapidly propagate ideas, emotions, and cultural phenomena. Memes shared on social networking services (SNS) are almost always graphical, in the form of static images, or sometimes even as animated GIFs. They constitute a large chunk of the increasing amount of visual content in SNS [8, 33], and are virtually inaccessible to the visually impaired users. Further, the fact that a large number of visually impaired users actively participate in social networks just like everyone else [33] with a desire to engage with the growing visual content [8] is a compelling reason to study the accessibility of meme images in social media.

A typical meme image can be disentangled into two key components: i) a visual component that portrays a central character imitating a style of human behavior or an emotional reaction and ii) an overlayed language component containing a witty catchphrase or a punchline. Dancygier and Vandelanotte [11] describe memes as "multi-modal constructions" due to the intricate relationships between the visual and textual information. This aspect is also illustrated by the memes in Figure 1, where it is evident that neither the face images nor the meme texts produce the same effect when conveyed separately in the absence of the other modality of information. Consequently, a full-fledged automated assistive system for meme images must deliver both these modalities to the visually impaired user. While efforts have been taken to automatically extract the meme text [7], the more challenging problem of automatically describing the visual content of memes to the visually impaired social media users has not been explored before this work.

On the other hand, social media platforms have undertaken significant efforts to deploy automated assistive systems for natural photographs. One of the most widely used assistive features is Facebook's Automatic Alt-Text (AAT) [34] system, which automatically recognizes a set of 97 objects present in Facebook photos. However, the generated alt-text has been identified by the visually-impaired users to be grossly inadequate in terms of descriptive detail [34]. In the research community, recent image captioning models [2, 3] are capable of generating significantly more descriptive captions for natural photographs, but they focus on visual details that are, most of the time, not very useful to understand the meme. For instance, in Figure 1 most of the visual information ("woman sitting on a sofa", "man in a yellow shirt") portrayed is irrelevant to understand the meme images. As a matter of fact, along with the meme texts, these memes can be well understood just by inferring the emotion conveyed by the characters' faces. This is in stark contrast to the amount of various visual concepts (multiple salient objects, actions, relationships) present in other natural photographs. Memes, on the other hand, contain specific, fine-grained, abstract concepts such as face emotion, character identity and sometimes even a significant amount of external context [24, 25]. Thus, while conveying the visual content in meme images to the visually impaired, the automated assistive system must focus on such specific visual concepts and describe them in a vivid manner. Existing automated methods are primarily conceived for natural photographs where the aforementioned visual details are of much lesser importance. As a result, these methods fail to meet the unique requirements of meme images.

In our work, we propose an automated approach specifically for meme images. We concentrate on one of the most important visual details present in a wide range of memes: the "facial emotion". The prominence of emotion in the meme ecosystem has been repeatedly identified in several prior works. The viral nature of memes has been attributed to their ability to resonate emotions [25] among online users of SNS. Memes, as stated by Miltner [21], serve as a form of emotional expression on social media platforms, allowing the participants to take a collective stand and strongly connect with other individuals. KnowYourMeme[3], the most comprehensive encyclopedia for online memes, states that the "Internet feeds on reaction" and that the immense popularity of emotion-rich memes such as "Reaction memes" (Figure 1) is due to the fact that emotion is universally understood across all cultures. Considering these points, in this work, we propose a novel automated method that enables visually impaired social media users to understand and appreciate meme face emotions.

To this end, we studied the existing techniques in face emotion understanding and found that the current emotion models are inadequate to describe the vivid, expressive meme faces. As illustrated in Figure 2, approaches that use universal emotion labels [13], facial action units [14] or valence-arousal scores [27] do not offer the necessary detail or the ease of a natural language interface to effectively describe these emotions to a visually impaired user. Thus, we propose a novel task of captioning meme face emotions with rich natural language. We focus on Reaction memes, as they contain facial emotion as their primary visual feature[4] and are also one of the most popular[5] classes of meme images on the Internet. The emotion captions generated by our model complements existing works on meme text recognition [7] to yield accurate, meaningful descriptions for Reaction memes, thus making one of the most popular classes of memes readily accessible to the visually impaired users for the first time. Our key contributions in this paper are as follows:

- We propose a novel problem of making one of the popular types of multimodal content on the Internet, namely, Reaction memes, accessible to the visually impaired users. We identify the importance of face emotion in this space, and hence, propose the task of describing them with natural language for the first time.
- We build an end-to-end trainable model for the above task. In the process, we also create a novel face emotion captioning dataset of about 2,000 meme faces with 6,000 rich emotion captions.
- Human evaluation by visually impaired Internet users shows that the emotion captions generated by our model can enable them to understand and appreciate Reaction memes on the Internet for the first time.

The rest of the paper is organized as follows: In section 2, we survey the recent works in meme understanding, web accessibility,

[2] google.com/trends/explore?date=all&q=memes

[3] https://knowyourmeme.com/blog/meme-review/kym-review-reaction-images-of-2017

[4] https://knowyourmeme.com/memes/reaction-images/

[5] google.com/trends/explore?date=all&q=reaction meme

Method	Description
Emotion Labels	Happy
Action Units	Lips stretched
Valence-Arousal	Valence = 0.9 ; Arousal = 0.6
Ours	**a broad smile of mischievous delight and excitement**

Figure 2: A comparison of the existing emotion recognition algorithms with our proposed method. Our approach provides the ease of a natural language interface and the flexibility to add important details about the meme face emotion.

face emotion understanding and image captioning. Following this, we describe our data collection process for the Meme Face Emotion Captions dataset in section 3. Section 4 explains our emotion captioning model and presents our results. We present the results from our user study with the visually-impaired social media users in Section 5 and conclude our findings in Section 6.

2 BACKGROUND

Our work is at the intersection of four bodies of literature - Computational analysis of memes in social networks, Automated methods for improving the accessibility of web imagery, Face emotion understanding and Image Captioning.

2.1 Analysis of memes in social networks

Recent computational methods study the "why" and "how" of meme virality by studying the network structure [31] and conducting a temporal analysis of cascades [10] respectively. It was only in a very recent work [12] that meme images were subjected to analysis based on the content, by constructing a semantic space that supports further analysis like meme topic and virality prediction, clustering and tracking meme evolutionary trends. However, the off-the-shelf visual representations used in their work were learned from natural photographs. These representations lack crucial fine-grained, high-level features such as face identity and emotion. Our work is the first effort to deeply understand an abstract visual concept present in a meme. We believe that the learned emotion representations from our model can be used for other computational analysis of the meme ecosystem.

2.2 Automated accessibility of web images

Prior to the advent of automated techniques, several applications employed a "human-in-the-loop" [6, 35] to help users with vision impairment to obtain detailed, reliable information about social media images. But these applications are not scalable, degrade the user experience due to having a non-trivial latency and also come with a compromise on privacy of the visually impaired user [34]. Recently developed automated methods such as Facebook Automatic Alt-Text (AAT) [34] exploit the advances in computer vision, specifically in tasks such as object detection and recognition to tag a limited list of 97 objects present in the image. While this system requires no manual effort, it is very limited in terms of details [34]. Recent state-of-the-art models [2] in image captioning can generate descriptive captions for common everyday scenes, but as we discussed earlier, these captions are ineffective for meme images. In this work, we focus on making meme face emotions accessible to the visually impaired, as they are widely prevalent in the domain of memes [24, 25]. We verify that the face emotion captions generated by our model enable visually impaired users to understand Reaction meme images.

2.3 Face emotion understanding

The current face emotion algorithms follow one of the three major emotion models: the categorical model [13], FACS model [14] and the dimensional model of affect [27]. There has been extensive work [5, 22] with the first model of the six basic universal emotions. However, as they are inadequate for face emotions in the wild, recent efforts have been towards the latter two models. The EmotioNet database [15] presents the first large-scale dataset of 1 million face images automatically labeled with facial action units. For the dimensional model of emotion, the AffectNet database [23] contains $500K$ faces manually labeled with valence and arousal scores. In Figure 2, we compare the above models and our approach to use natural language to describe emotion. The FACS and valence-arousal models are significantly more detailed than the dimensional model. However, our approach is much more interpretable and offers the ease of a flexible natural language interface. This being said, the scale of the EmotioNet and AffectNet databases enables recent advances in deep learning to learn powerful face emotion representations. We use the AffectNet dataset in our pretraining stage and transfer the learned representations to generate accurate emotion captions without the need to collect a large emotion captioning dataset.

2.4 Image Captioning

The release of large-scale captioning datasets such as MS-COCO captions [9] have facilitated rapid advances in the image captioning task. In the second stage of our approach, we leverage a recent state-of-the-art captioning architecture [3] to generate emotion captions. We propose a novel method to pretrain the image encoder of this captioning model with face emotion features.

3 MEME FACE EMOTION CAPTIONS DATASET

As there are no publicly available datasets for face emotion captioning, we construct a dataset of $2,000$ meme faces annotated with 3 captions each.

3.1 Challenges in face emotion captioning

Prior efforts in creating large face emotion datasets [15, 23] emphasize the immense difficulty of recognizing emotions from static

face images in the wild. Our goal to annotate these emotions using fine-grained natural language is thus, far from being a trivial task.

3.1.1 Annotator constraints. Although crowd-sourcing platforms like Mechanical Turk[6] are fast, efficient approaches to annotate large datasets, there is a significant variation in quality as well [5]. This point was also noted during the creation of the AffectNet dataset [23]. Therefore, we hire in-house annotators with English proficiency to caption our meme faces. During our pilot study, we found that, on an average, an annotator requires close to one and a half minutes to write a caption of high quality. This is about eight times longer [15] than it takes to annotate other complex emotion attributes like facial action units. The above constraints significantly impact the scale of our dataset. In the first stage of our approach, we illustrate the immense utility of our fine-grained dataset when used effectively in conjunction with existing large coarse-grained datasets.

3.1.2 Subjective nature of emotion. Emotion recognition from unconstrained real-world face images is not only difficult but is also prone to significant differences in the perception of the displayed emotion [5, 23]. Further, the lack of context makes the task inherently ambiguous [16]. During our pilot study, we found that imitating the face expression shown [32] eases the annotation task by enabling the annotators to decipher the emotion in the face better. Further, after collecting the dataset, we quantitatively analyze the inter-annotator agreement of the captions and draw conclusions regarding this topic.

3.2 Data collection

To collect meme faces, we first collect a database of Reaction memes from a popular meme website called me.me[7] by searching for keywords “my reaction when", “my face when" and “how I feel when". Using the dlib library [18], we detect and deduplicate faces extracted from these memes. False-positive face detections are removed by a human annotator. Out of over 36,000 memes collected, we only choose and annotate about 2,000 meme faces because, as mentioned earlier, rich emotion caption annotation is time-consuming and expensive. However, we show in a later section, that by transferring learning from existing large face emotion label datasets like AffectNet, our dataset is very useful to train existing captioning models to generate rich, meaningful face emotion captions.

Each face was independently described by 3 annotators with a rich caption and with one of the six universal emotion labels. To ensure annotation quality, the annotators are initially tutored using 20 example faces that are not part of our final dataset. We only show the face crop to the annotator and instruct them to write a rich emotion caption and also categorize them to one or more of the basic emotion classes. In Table 1 and 2, we present descriptive information about our dataset and the frequency of the basic emotion classes.

3.3 Measuring inter-annotator agreement

We use standard captioning metrics to measure agreement across our emotion captions. For each image in the dataset, we treat one of its captions as a “test caption" and the remaining ones as ground-truth reference captions and compute BLEU [26], CIDEr-D [30], METEOR [4], ROUGE$_L$ [19] and SPICE [1] scores. The mean inter-annotator agreement score A for a face image across its set of emotion captions C using a metric M is the mean of the scores obtained by considering each of the image's captions as a “test caption". Concretely:

$$A = \frac{1}{|C|} \sum_{i=1}^{|C|} M(c_i, C \setminus c_i)$$

To put these scores in perspective and for comparison, we also compute the agreement scores of the MS-COCO training split. The agreement scores are shown in Table 3.

[6] https://www.mturk.com/
[7] https://me.me

Attribute	Value
Number of face images	2,000
Number of captions	6,000
Median face image resolution	128 x 128
Median length of description	8
Median term frequency	3
Total vocabulary size	1,782

Table 1: Descriptive statistics of our dataset

Emotion label	Number of images
Happy	724
Surprise	444
Fear	153
Disgust	94
Angry	484
Sad	566

Table 2: Distribution of basic emotion classes in our captioning dataset. Emotion datasets are typically skewed and classes like "Disgust" and "Fear" contain far fewer samples than the others.

Metric \ Dataset	Ours	MS-COCO
BLEU-1	0.35	0.51
BLEU-2	0.15	0.32
BLEU-3	0.07	0.20
BLEU-4	0.03	0.12
ROUGE$_L$	0.26	0.40
METEOR	0.12	0.20
CIDEr-D	0.18	0.84
SPICE	0.11	0.24

Table 3: Comparison of inter-annotator agreement scores between our Meme Face Emotion Captions dataset and MS-COCO Captions dataset that illustrates: i) the inherent ambiguity of our task and ii) that our set of captions for a given image captures different emotion perspectives.

The lower agreement scores in our dataset compared to that of MS-COCO highlight the inherent ambiguity in the task of describing a face emotion. The emotion captions are significantly more subjective and abstract. An image in MS-COCO typically contains concrete objects (e.g. dog, table, bus, road) which are more likely to be described using the same word by multiple annotators. However, in our dataset, very similar emotions can be described using very different words, resulting in low inter-annotator agreement scores using the above metrics. On the other hand, the lower agreement scores could also be attributed to the inefficacy of the standard automatic metrics to capture the high-level semantic similarities between the captions. In many of the recent works [1, 29], automatic metrics are found to correlate poorly with human judgement in several cases. The fine-grained, abstract nature of our problem calls for future research efforts to develop better metrics that are tailored to judge rich emotion captions. Another plausible solution that can be addressed in future work is to augment this dataset with more number of captions per image, to enhance the reliability of the automated metrics and to capture a wider range of emotion interpretations. Finally, the agreement scores also point out that our set of captions are diverse and capture multiple emotion perspectives of a given face. It would be interesting to see future approaches that can generate emotion captions by also taking into account the additional context present in the meme such as the meme text.

4 GENERATING RICH EMOTION CAPTIONS

Our approach consists of two stages. We first design a model to learn fine-grained emotion representations from face images. We then incorporate this model in our captioning framework to obtain superior results over the baseline captioner.

4.1 Stage 1: Learning robust emotion representations

The success of the current captioning methods for natural scenes can be credited to the existence of several robust image encoders that are pretrained on the ImageNet dataset. But the visual representations of these pretrained models do not contain fine-grained features such as face emotions. The absence of a good image encoder means that we would need a large face emotion captioning dataset to be able to generate high-quality emotion captions. As it is resource-intensive to collect captioning data, we instead develop an approach to learn a robust face emotion encoder from external data sources.

Transfer learning has been a powerful strategy [28] to reduce the need for large manually annotated datasets for deep neural networks. We propose to exploit the recently released AffectNet database [23] and pretrain a face emotion encoder that can produce robust, generalizable face emotion features. AffectNet contains $500K$ images manually annotated with valence and arousal values in the range $[-1, 1]$. The most straightforward method to train a network to learn emotion features is to regress for valence and arousal scores as done by Mollahosseini et al. [23]. But we argue that naive regression is not an ideal training strategy for this data. Firstly, valence and arousal values are inherently ambiguous and subjective. Secondly, it is hard to even for a human, and hence a learning algorithm, to translate the complex concept of face emotion to an absolute real-value between $[-1, 1]$. We propose a more intuitive strategy in which we train a model to compare a pair of faces and rank them according to their valence and arousal values.

We implement this using a siamese architecture consisting of two convolutional neural network branches with shared weights. Given a pair of images, a, b each branch of the siamese network learns a mapping $f : x \mapsto f(x)$ to output a set of face emotion features $f(a)$ and $f(b)$ respectively. We concatenate these emotion features and use a fully connected layer to predict two probability distributions over three rank labels 0, 1 and 2. Each of the three rank labels indicates whether a is ranked lower, equal or higher than b on the valence and arousal scales. Given the absolute valence and arousal scores for a pair of input images a, b, we determine the rank label r for each attribute based on its absolute scores s_a and s_b. Concretely:

$$r = \begin{cases} 0 & |s_a - s_b| > \lambda; s_a > s_b \\ 1 & |s_a - s_b| \leq \lambda \\ 2 & |s_a - s_b| > \lambda; s_a < s_b \end{cases}$$

Typically, attribute-based ranking methods use just two rank labels to indicate the higher or lower ranked sample among the input pair. However, in our case, as the valence and arousal scores are ambiguous and noisy, we found it very beneficial to train the model to label a pair of input images as “equal rank" if the absolute difference in their scores is less than a threshold λ. In our experiments, we found a threshold of $\lambda = 0.25$ to give the best results. As we have formulated our ranking task as 3-way classification of rank labels, we can train our face emotion encoder network by minimizing the standard cross-entropy loss.

4.1.1 Training details. We train our siamese ranking network on the AffectNet dataset which contains $500K$ faces labeled with absolute valence and arousal scores. We discard a large number of generic “neutral" faces in the dataset, as we are interested in learning representations that are useful for vivid meme face emotions. For each face image in the dataset, we sample another random face from the dataset and use this as the input pair. We use 90% of the data for training and a 10% validation split to monitor the validation loss. We use the ResNet50 architecture as our CNN encoder, and train it with an initial learning rate of $1e^{-2}$ and a batch size of 200 input pairs. We decay the learning rate when the validation loss does not improve for 2 consecutive epochs. We train until the validation stops improving for six consecutive epochs. We obtain a validation accuracy of 75.4% and 66.5% for valence and arousal ranking respectively.

In the process of ranking based on valence and arousal scores, the CNN model learns fine-grained face emotion representations that can be used for our meme face captioning task to generate rich emotion captions.

4.2 Stage 2: Face emotion captioning

In this section, we describe multiple variants of our captioning model by first describing the common framework followed in all of them. In our experiments, we modify a recently proposed captioning model [3] to generate emotion captions for our task. The key difference in these variants is the use of different image encoders to extract face emotion features from the face image. One of the

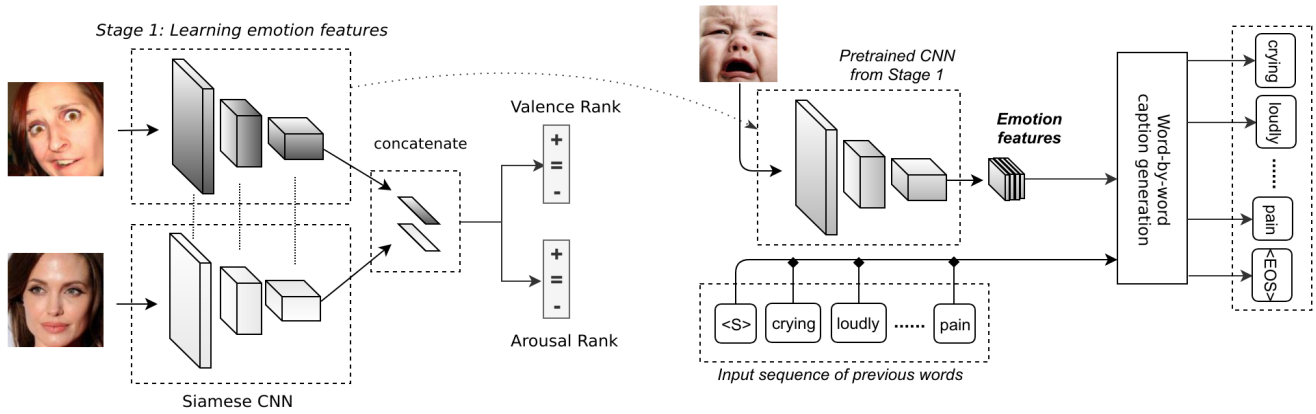

Figure 3: Our two-stage approach to generate rich captions. In Stage 1, we pretrain an image encoder to learn discriminative emotion representations from ambiguous valence-arousal scores in the AffectNet dataset. Instead of naive regression, we propose a more intuitive training strategy to learn to compare two faces and identify how they rank relative to each other in the valence-arousal scales. Using these learned emotion features in the captioning model (Stage 2) helps us generate accurate, descriptive captions with limited training data.

variants is our pretrained face emotion encoder obtained from Stage 1.

4.2.1 List of experiments. We train three different variants of the captioning model, by incorporating different image encoder networks. All the image encoders follow the ResNet50 [17] architecture but are pretrained using different methods. Our baseline uses the ResNet50 network pretrained on the ImageNet image classification dataset. The second variant is similar to the one proposed by Mollahosseini et al. [23] where we pretrain a ResNet50 on AffectNet by regressing the valence-arousal scores. The last variant is again a ResNet50, but it is pretrained using our siamese ranking approach as described in detail in the previous section.

4.2.2 Training details and Evaluation. We split our dataset of 2,000 images and use 80% for training and 10% each for validation and test splits. To ensure a fair comparison, all our captioning variants follow the same hyper-parameters and the only difference is the usage of the image encoder pretrained using different methods. The image encoder is also frozen during training and the word embedding dimension is set to 64. We apply a dropout of 0.1 after every layer in the convolutional decoder. We use weight decay of $5e^{-4}$ and apply gradient clipping of 0.1. We use a batch size of 32 and train until the validation CIDEr score does not improve for 20 epochs. All the other hyper-parameters are set to the same as in the setup by Aneja et al..

We choose the best-performing model on the validation split and evaluate it on our held-out test split. In Table 4, we report the test scores using standard captioning metrics: BLEU, CIDEr, ROGUE$_L$, METEOR, and SPICE. In Table 5, we report individual captioning scores obtained by our approach for each of the basic emotion classes, as the imbalance of emotion labels causes significant variation. In Figure 4, we also show some qualitative results of our best model against the ground-truth and naive regression-based model.

4.3 Analysis of the captioning results

In this section, we draw inferences from our model's predictions and present the benefits of our pretraining methodology. We also investigate the common failure cases of our model and discuss possible remedies.

4.3.1 Advantage of pretraining. The key variation across our experiments is the different pretraining strategies used to learn face emotion features. From Table 4, it is evident that the current captioning models that use a CNN pretrained on ImageNet, lack face emotion information to generate accurate emotion captions. Further, Table 4 also highlights the impact of choosing an effective pretraining strategy on the final captioning results. Naive regression on the ambiguous, subjective valence and arousal scores leads to noisier gradients and does not encourage the model to learn high-level features that can discriminate subtle face emotions. On the other hand, our siamese ranking model is much more discriminative and obtains more accurate richer captions. As our image encoder was trained on AffectNet, our model is specifically good at discriminating between different levels of valence and arousal. In Figure 4, phrases such as *soft smile, laughing loudly* (top row samples) is very indicative of capturing different extremes of the arousal scale. Similarly, phrases like *warm gaze, disgusted look* show that the model also captures different values of valence. Notice that the faces in the top left corner and the bottom row all comprise a smiling face but are captioned very differently as they reflect different emotions. The ability of our ranking model to describe emotions at such a granular level will be very helpful for visually impaired users to accurately imagine the face emotion depicted in the meme image.

4.3.2 Inherent ambiguity and imbalanced classes. In Table 5, we report individual captioning scores for each basic emotion label. The results are in line with recent findings [23] that negative emotions are harder to decipher. A typical case of ambiguity in describing face

Image encoder \ Metric	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE$_L$	METEOR	CIDEr	SPICE
ResNet50 (ImageNet pretrained)	0.43	0.20	0.10	0.06	0.33	0.12	0.14	0.10
ResNet50 (regression)	0.44	0.24	0.14	0.09	0.33	0.13	0.18	0.12
Ours (siamese ranking)	**0.48**	**0.28**	**0.17**	**0.12**	**0.36**	**0.15**	**0.28**	**0.14**

Table 4: Captioning scores on the held out test split of our dataset. Our approach (siamese ranking) of pretraining significantly outperforms other existing baselines across all standard metrics.

Emotion \ Metric	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE$_L$	METEOR	CIDEr	SPICE
Happy	0.53	0.33	0.23	0.16	0.41	0.18	0.39	0.17
Surprise	0.47	0.27	0.18	0.12	0.34	0.16	0.28	0.14
Fear	0.51	0.32	0.23	0.18	0.31	0.16	0.30	0.09
Disgust	0.33	0.19	0.13	0.10	0.26	0.10	0.37	0.06
Angry	0.45	0.25	0.15	0.10	0.34	0.14	0.18	0.12
Sad	0.45	0.24	0.14	0.08	0.34	0.14	0.19	0.12

Table 5: Captioning scores by our siamese ranking model for individual emotion labels. The results illustrate that negative emotions are more ambiguous and are significantly harder to understand.

Regression: *a confident gaze with a slight smile*
Ranking: *a soft smile with a warm gaze*
GT: affectionate smile with a warm gaze and loving way

Regression: *a happy and honest smile with mouth open*
Ranking: *laughing loudly with joy*
GT: laughing with pure joy and excitement

Regression: *a happy and honest smile with teeth showing*
Ranking: *a disgusted look with a scrunched forehead*
GT: an irritated and disgusted look accompanied with frustration

Regression: *a neutral face with no emotion*
Ranking: *a disappointed and grieving look with a little wrath*
GT: sorrow with mouth open in a sad way and eyes almost teary

Regression: *a wide smile of happiness and joy*
Ranking: *a wide smile with a creepy look*
GT: a very creepy smile with an unsettling gaze

Regression: *a neutral face with a hint of indifference*
Ranking: *a confident and mischievous look with a half smile*
GT: a confident look with lips in a half smirk

Figure 4: An illustration of the captions generated by our siamese ranking model on unseen images in the held-out test set. For each example, along with our best model, we show the caption generated by the baseline naive regression and the ground truth caption as well. The siamese ranking model generates more descriptive captions for the visually impaired users to visualize the meme face emotion.

emotions can be seen in Figure 4, bottom-row left example. While the captions by our ranking model and the ground truth match, one could also argue that the caption generated by the regression model could make sense in some contexts. A possible resolution would be to collect more captions per image to capture more emotion perspectives. As visually impaired users tend to place a lot of trust in the provided descriptions [20], it is important to smoothly handle such limitations in future efforts.

5 VALIDATION OF THE GENERATED CAPTIONS BY THE VISUALLY-IMPAIRED USERS

As our captions are generated specifically to assist the visually impaired users on social media, it is essential to validate whether these captions fulfill their intended goal to improve the accessibility of Reaction memes. The validation process also allows us to identify the limitations of our work and propose promising future directions.

To interact with visually impaired social media users, we visited a reputed organization[8] that strives to enable visually-challenged individuals to be self-reliant and independent. We conducted a study with 6 visually impaired individuals who regularly access the Internet. All of them possess some form of serious visual impairment and interact with their devices using screen readers. The age group is $21-36$ years, with 1 female and 5 male participants. In this section, we present our findings and feedback from our interaction with these participants.

We interviewed six visually impaired individuals with various percentage of vision loss. We found that five individuals are active on social media, i.e. they use SNS for at least an hour per day. They are most active on Facebook, as the platform provides alt-text descriptions to understand the visual content. They regularly post status updates, share and like photos and selfies. All the five active users had come across memes several times while browsing on Facebook, both in English and regional languages. Their unanimous opinion was that they can never understand meme images without asking for external visual assistance.

After this initial discussion, we described 25 randomly chosen Reaction memes from our unseen test set to obtain their feedback for our generated captions. We also compared the effectiveness of our approach against existing automated techniques including the widely used Facebook AAT [34] and a state-of-the-art image captioner [2]. We systematically compare across five techniques by instructing the visually impaired users to rate their understanding of a particular meme on a scale of $1-5$ based on the five types of descriptions that are read out. To ensure the absence of any order bias, captions $(a-e)$ are read out in a random order for a particular meme image. The results of this study are presented in Figure 5.

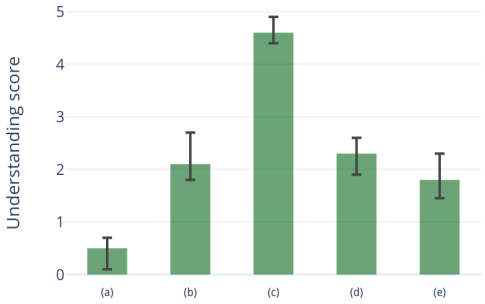

Figure 5: Ratings provided by the visually impaired users for various automated techniques: (a) Facebook AAT, (b) Meme text only, (c) Our Emotion caption + Meme text, (d) Facebook AAT + Meme text, and (e) Image caption + Meme Text.

We can make the following inferences. Firstly, Facebook AAT (a), in its current state is almost of no use to understand meme images. In the case of (b) the meme text, which can be extracted using recently proposed systems [7], provides valuable context to understand the meme image, but it is still very much incomplete without any sort of visual information. Supplementing this meme text with rudimentary visual cues from Facebook AAT helps only marginally as shown in (d). In (e), we obtain a very interesting result where despite adding more descriptive visual captions using a state-of-the-art image captioner, the visually impaired participants evaluate this case as even worse than just plain meme text. One of them reasoned that in this case, "The details provided by this particular caption feels very irrelevant and confusing. I am not able to understand its connection with the text". This rightly highlights the failure of current captioning systems for meme images.

In the case of our system (c), reading out the face emotion caption generated by our model, along with the meme text is significantly more useful for the visually impaired users than the other systems; as they were not only able to understand the meme but also appreciate the conveyed humor. All of the five users who are active on social media agreed that they would readily make use of such an automated Reaction meme captioner if made available. They also point out a key area for improvement - the need for more visual context. One of their popular queries was, "Can I know which movie scene or character is being shown?" We plan to investigate these two scopes of improvement in our future work. In conclusion, our field study strongly indicates that our emotion captions are very helpful for the visually impaired users to understand Reaction memes for the first time without any manual assistance.

6 CONCLUSION

We introduced a new, challenging problem of describing the visual content in meme images to the visually impaired users on social media. We identified the shortcoming of existing assistive methods for meme images and emphasized the need for specialized methods that can describe the specific, abstract visual features present in memes in a vivid manner. We proposed to caption the face emotion present in Reaction memes, as this visual aspect is quite prominent in the online meme ecosystem. To this end, we collected a dataset of meme face emotion captions and devised an effective two-stage approach to generate rich emotion captions for meme faces. Finally, we validated the strong utility of our approach with the help of visually impaired users who are active on social media. Our research opens a wide range of future directions in the unexplored space of meme accessibility and understanding.

One of the promising directions is to conduct a more extensive study with a large, diverse group of users with visual impairment, and further understand their unique challenges with meme images. A good starting point would be to deploy a complete, working application of our current emotion captioning model and deliver meme content to the visually impaired users through this "meme social network". The continuous user feedback through the application will help the research community gain valuable insights into this complex problem.

ACKNOWLEDGMENTS

We thank the Youth4jobs.org foundation for helping us conduct the field study with the visually impaired users.

[8] https://www.youth4jobs.org/

REFERENCES

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*. Springer, 382–398.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*.

[3] Jyoti Aneja, Aditya Deshpande, and Alexander Schwing. 2018. Convolutional Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5561–5570.

[4] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.

[5] Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 279–283.

[6] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 333–342.

[7] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. 2018. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 71–79.

[8] Maria Claudia Buzzi, Marina Buzzi, Barbara Leporini, and Fahim Akhter. 2010. Is Facebook really" open" to all?. In *Technology and Society (ISTAS), 2010 IEEE International Symposium on*. Citeseer, 327–336.

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

[10] Justin Cheng, Lada A Adamic, Jon M Kleinberg, and Jure Leskovec. 2016. Do cascades recur?. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 671–681.

[11] Barbara Dancygier and Lieven Vandelanotte. 2017. Internet memes as multimodal constructions. *Cognitive Linguistics* 28, 3 (2017), 565–598.

[12] Abhimanyu Dubey, Esteban Moro, Manuel Cebrian, and Iyad Rahwan. 2018. MemeSequencer: Sparse Matching for Embedding Image Macros. *arXiv preprint arXiv:1802.04936* (2018).

[13] Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion* (1999), 45–60.

[14] Paul Ekman and WV Friesen. 1978. Facial action coding system: A technique for the measurement of facial action. *Manual for the Facial Action Coding System* (1978).

[15] C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. 2016. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5562–5570.

[16] Ran R Hassin, Hillel Aviezer, and Shlomo Bentin. 2013. Inherently ambiguous: Facial expressions of emotions, in context. *Emotion Review* 5, 1 (2013), 60–65.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[18] Davis E King. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research* 10, Jul (2009), 1755–1758.

[19] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out* (2004).

[20] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 5988–5999.

[21] Kate M Miltner. 2014. âĂIJThereâĂŹs no place for lulz on LOLCatsâĂİ: The role of genre, gender, and group identity in the interpretation and enjoyment of an Internet meme. *First Monday* 19, 8 (2014).

[22] Ali Mollahosseini, David Chan, and Mohammad H Mahoor. 2016. Going deeper in facial expression recognition using deep neural networks. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–10.

[23] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. 2017. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *arXiv preprint arXiv:1708.03985* (2017).

[24] Asaf Nissenbaum and Limor Shifman. 2018. Meme Templates as Expressive Repertoires in a Globalizing World: A Cross-Linguistic Study. *Journal of Computer-Mediated Communication* 23, 5 (2018), 294–310.

[25] Zizi Papacharissi. 2016. Affective publics and structures of storytelling: Sentiment, events and mediality. *Information, Communication & Society* 19, 3 (2016), 307–324.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

[27] James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology* 39, 6 (1980), 1161.

[28] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 806–813.

[29] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 2556–2565.

[30] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[31] Lilian Weng, Filippo Menczer, and Yong-Yeol Ahn. 2014. Predicting Successful Memes Using Network and Community Structure.. In *ICWSM*.

[32] Adrienne Wood, Magdalena Rychlowska, Sebastian Korb, and Paula Niedenthal. 2016. Fashioning the face: sensorimotor simulation contributes to facial expression recognition. *Trends in cognitive sciences* 20, 3 (2016), 227–240.

[33] Shaomei Wu and Lada A Adamic. 2014. Visually impaired users on an online social network. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 3133–3142.

[34] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service.. In *CSCW*. 1180–1192.

[35] Yu Zhong, Walter S Lasecki, Erin Brady, and Jeffrey P Bigham. 2015. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2353–2362.