Masked Facial Expression Classification

By Nemat Shaikh, Zayn Khan, Ayyan Mumtaz

1. Introduction

The COVID 19 pandemic has brought about changes in our lives especially with the widespread adoption of face masks. While masks are important for health they make it difficult to interpret emotions through expressions, which are essential for human connection and communication in various situations like healthcare and education. The challenge lies in maintaining communication when crucial emotional cues are covered by masks.

Current facial recognition technologies struggle with identifying emotions when only parts of the face are visible due to mask wearing. Our research focuses on developing learning models that can accurately recognize basic emotions like happiness, sadness, anger, surprise, disgust and fear using only the upper portion of the face such as the forehead and eyes. By utilizing facial reconstruction strategies like cGAN and generative CNN's we hope to greatly increase classifying accuracy. This project involves adapting existing models to effectively interpret data and overcome the obstacles posed by obscured facial features.

2. Background and Related Work

2.1. Alzubi 2021

Alzu'bi et al. (2021) is a study discuss the implementation of deep learning techniques, specifically Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs), to enhance Masked Face Recognition (MFR). This comprehensive survey details the generic MFR pipeline widely adopted in recent research, covering image preprocessing, feature extraction, face detection and localization, face unmasking and restoration, and identity matching and verification. The study identifies critical advancements and ongoing challenges in the field, emphasizing the need for improved learning models to enhance MFR system performance under varied conditions and test sets. The potential of hybrid neural networks in learning concurrent tasks such as mask detection and face reconstruction is highlighted as crucial for the accuracy of MFR systems. This insight is pivotal for adapting these techniques to our masked facial expression classification project, as a holistic overview of the existing landscape of MFR was very helpful in determining where we might begin. Even though this study only covers masked facial recognition, and has nothing to do with emotional classification, it is still integral to our project as it explores deep learning techniques in regards to faces with reduced visual information. The study also helped us design our project as the challenges mentioned such as dataset variation and algorithm complexity were integral to account for, on our own approach. The study goes on about the effectiveness of MFR systems is significantly dependent on the diversity of the datasets used, which should ideally include real-world images with various types of mask, and that MFR methods that do not assume cooperative subjects who are directly facing the camera are incredibly more difficult and complex problems to solve. Along with this, it points to how the complexity of deep learning-based MFR algorithms often results in high computational demands, which is impractical for real-time or resource-limited applications. So solutions that balance computational efficiency and performance are crucial for the practical deployment of MFR systems. Considering our limitations of working on Google Colab and not being able to have the time our resources to train complex MFR algorithms, we used a simple approach. We focused specifically on emotion recognition from the upper half of the face, which remains visible despite mask-wearing. Our Kaggle dataset is very visible and simple, with 'cooperative subjects'. We adapt conventional deep learning methods to better handle the partial visibility of faces, incorporating strategies to optimize model performance under these constrained conditions.

2.2. Magherini 2022

Magherini et al. (2022) utilize the AffectNet dataset to develop a CNN-based model that categorizes emotions based on visible upper facial regions, achieving a high accuracy rate of 96.92%. This study is directly relevant to our project, which focuses on recognizing emotions from masked facial expressions. Seeing the high accuracy of CNN in their model for categorizing emotions based on just the upper facial regions was motivating for us to utilize it in our efforts. By adapting the CNN architectures tested in this study, we can refine our model to better recognize and classify emotions even when key facial features are obscured by masks. Their approach applied sophisticated training techniques including class weighting, Adam optimizer, and data augmentation to balance the dataset and enhance model performance, which inspired us to use data augmentation and advanced training mechanisms like Batch Normalization and ReLU activation to ensure robustness and prevent overfitting. Their utilization of advanced neural network architectures and transfer learning techniques such as ResNet50v2 and InceptionResnetV2 for handling specific tasks within their framework inspired us to employ ResNet50, VGG16, and Inception V3 in our own approach.

This adoption was crucial for leveraging pre-existing models to enhance our system's capability in understanding and reconstructing facial features. Another really important aspect of this study is the iterative refinement of models, as it utilized the InceptionResnetV2 architecture and trained the model iteratively on recategorized emotions to improve accuracy. This dual-phase training inspired us to consider an approach to refine our model progressively.

2.3. Hosen 2022

The paper titled "Masked Face Inpainting Through Residual Attention UNet" by Md Imran Hosen and Md Baharul Islam presents a novel deep learning approach for inpainting facial images that have been obscured by masks. This is particularly relevant for systems that rely on facial recognition, which may be compromised when faces are partially covered, as has become common due to health safety measures like mask-wearing during the COVID-19 pandemic The study introduces a specialized model that incorporates residual blocks and attention mechanisms within a UNet architecture. The residual blocks help address the vanishing gradient problem often encountered in deep networks by ensuring that gradients can flow through the network during training, which also stabilizes training. The attention mechanisms focus the model on relevant regions of the input, improving the efficiency and speed of the model. The authors also provide a comprehensive evaluation, showing that their model outperforms existing methods in terms of both the quality of inpainting and computational efficiency. This model is trained and tested using the CelebA dataset, a large-scale face dataset with diverse images of celebrities, modified to include simulated masks for training purposes. The model demonstrates high fidelity in restoring facial images, effectively removing masks and reconstructing occluded parts of the face with a high degree of detail. This inspired us to include an attempt at the U-Net convolutional neural network (CNN) for our own approach in the project. U-Net is a type of deep learning model that's particularly effective for tasks like image segmentation, where the goal is to identify and delineate multiple parts or objects within an image. It is most commonly known for its original purpose and the purpose of it's development– image segmentation. In this study, U-Net has been modified and enhanced with residual blocks and attention mechanisms to perform the specific task of inpainting—filling in masked or occluded parts of facial images. For the masked face emotion classification, we could potentially utilize U-Net for a similar sort of image reconstruction, as we will only have access to one half of the face which will finally help us use classify the emotions.

3. Methods

3.1. Pre-Processing

We use a dataset of 36,000 48x48 pixel labeled grayscale photos for our study. This Kaggle dataset is incredibly helpful due to its wide range of facial expressions, which gives our analysis a solid basis. Our strategy modifies the dataset to focus on the upper half of the face, given the present global health environment where face masks are common. In order to accomplish this, we reduce the photographs to a size of 24 by 48 by starting at the first 24 pixels in the vertical direction. This particular tweak is important because it enables us to focus on facial elements that are usually visible above a mask, such as the forehead and eyes, which helps to address the problem of identifying emotions based on partial facial information.

3.2. Initial Classification

In our research we started off by using a three layer neural network (CNN) to identify emotions, from cropped images of the upper part of faces. This standard CNN model was created to understand and categorize emotions based on these images serving as a starting point for our project. The reason for choosing this CNN approach was to establish a basis for comparison to assess the performance of advanced models developed later in our study. Notably this same CNN model was also utilized towards the end of our project for emotion classification working with face images that were reconstructed from the top half inputs combined with the generated lower half. By sticking with this CNN model throughout our project we were able to compare outcomes with those achieved after implementing more sophisticated modeling techniques thus giving us a clear indication of any enhancements in emotion recognition accuracy resulting from our subsequent improvements in reconstructing facial features.

3.3. Initial CGAN

After establishing the foundation with our CNN approach we delved into exploring how Generative Adversarial Networks (GANs) could enhance our models ability to reconstruct the part of faces. GANs consist of two networks. A generator and a discriminator. They compete against each other in training. The generator creates data instances resembling a given dataset while the discriminator assesses their authenticity. In this case we utilized a Conditional GAN (CGAN) using the face as information to influence the generation process. This method theoretically allows for precise and controlled reconstructions of the face. However our initial outcomes did not align with our expectations as the reconstructions often lacked the accuracy and detail, for emotion recognition. The decision to employ CGAN stemmed from wanting to utilize the features of the face to guide the reconstruction process more efficiently ensuring that the generated lower face would better match with the visible expressions and contours of the upper half. Nonetheless encountering difficulties in achieving high quality reconstructions underscored the importance of refining our GAN training procedures and implementing conditioning mechanisms to fully harness CGANs potential, in facial reconstruction tasks.

3.4. Generative CNN

The next attempt was using only CNN to generate the lower half of the face. We began by initially using simple architecture but achieving questionable results. After realizing limitations in detail accuracy and coherence from the CNN we progressively enhanced the complexity of our models. Moving towards an CNN with 256 nodes and additional layers was driven by our aim to improve the models capacity to capture finer facial details and enhance overall image quality. We also implemented training methods like data augmentation introducing variations such as rotations and shifts to better simulate real world conditions. Transitioning to a sophisticated CNN involved using techniques like Batch Normalization and ReLU activations for regularization purposes ensuring stable learning and preventing overfitting. These adjustments were essential for enabling the models to handle facial features nuances and variations especially when dealing with incomplete data due to mask obstructions. Each stage of elevating the model's intricacy was a tactic to improve our capacity for replicating facial expressions and boosting emotion recognition skills in difficult situations.

3.5. Transfer Learning

To improve our models ability to reconstruct hidden features and interpret emotions from visual data we utilized transfer learning methods. We took advantage of the established strengths of pretrained models, like ResNet50 VGG16 and Inception V3. These models were selected for their track record in handling image processing tasks. We fine tuned them to meet our specific requirements for facial reconstruction.

ResNet50 stands out for its network with 50 layers featuring 48 layers along with one MaxPool layer and one average pool layer. Its use of residual blocks aids in processing image data particularly in extracting features from the upper part of the face to predict the lower part. Implementing ResNet50 significantly enhanced our prediction accuracy and improved emotion recognition precision.

Similarly we enhanced the VGG16 model by incorporating layers while maintaining its architecture to leverage its robust pattern recognition abilities. This adjustment enabled the model to apply learned features in predicting facial regions based on upper face characteristics. To avoid overfitting we used dropout techniques. Closely monitored the models progress by stopping during training. This approach helped us keep both the training and validation losses low, confirming that the models accuracy improved over time.

When working with the Inception V3 model we took advantage of its capability to handle data. We kept its layers fixed to retain learned features while adding new layers for upscaling and prediction tasks specific to our project needs, notably the models input required 139x139 input which required resizing of our input images. The training of this modified model was carefully overseen with visual feedback mechanisms and strategic checkpoints to enhance performance. This was evident in the improvements seen in reconstructions during training sessions.

These strategic adjustments were based on insights, from our CNN models, which showed that although they were learning they struggled to capture and recreate the intricate features required for realistic facial reconstructions. By utilizing pretrained models and selectively training the dense layers while keeping lower layers frozen we focused on key aspects of our task. This approach accelerated the learning process and boosted model accuracy.

3.6. CGAN with ResNet

After applying transfer learning techniques using pretrained models, like ResNet50 VGG16 and Inception V3 we opted to boost our models capabilities by incorporating a GAN (Generative Adversarial Network) structure enhanced with ResNet blocks. This decision was driven by our goal to refine the quality of reconstructions to a point where the generated lower face portions would closely resemble images. The GAN framework, known for its ability to produce images, was viewed as an ideal solution for achieving these advanced visual results. By merging the GAN architecture with ResNet blocks we harnessed the power of learning to address common training challenges in deep networks, such as gradient disappearance thereby improving the stability and efficiency of both the generator and discriminator. This strategic approach enabled us to establish a standard for the authenticity of our generated images aiming to make AI generated face halves visually comparable to human faces. Our ultimate objective was to push the boundaries of what can be accomplished in facial feature reconstruction and emotion recognition technology.

3.7. Final Model and Classification

After examining models performance we decided to revisit the use of ResNet because of its outcomes. It was consistently producing high quality images. Encouraged by these results we decided to enhance the ResNet architecture by making it a bit more complex. Our choice was based on the idea that increasing the model's complexity could enhance its accuracy in generating reconstructions. By adding layers or deepening the structure our goal was to improve the models capability to capture and replicate finer facial details ultimately leading to better overall results, in our facial reconstruction projects. This strategic upgrade aimed at maximizing ResNets strengths to meet our project's requirements while pushing its performance boundaries.

We opted to stick with the CNN classifier we used initially in our project to maintain consistency and reliability, in our approach to emotion classification. This decision enables us to directly compare how well our initial model performs against the enhanced models we developed on. By keeping the classifier configuration. Including layers with 32, 64 and 128 filters, along with pooling and dropout for regularization. We aimed to evaluate the impact of our

improvements, such as refined training techniques and incorporating advanced architectures like ResNet on the models ability to classify emotions based solely on facial features above the nose.

Our objective was to achieve emotion recognition using the upper part of faces mirroring real world situations where masks may conceal lower facial expressions. Utilizing the model from start to finish allowed us to clearly measure any performance improvements showcasing the effectiveness of our development strategies and confirming that our model could handle scenarios where full facial expressions are not always visible.

4. Results

4.1. Initial Classification

The initial CNN classifier we trained achieved 57.31% on the training set and 36.9% on the test set. This was after modifying model complexity and experimenting with the optimal number of epochs to prevent overfitting. This was lower than our anticipated baseline for emotion classification accuracy, but provided a great margin for improvement.

4.2. Initial CGAN

To begin, we used a cGAN (Conditional Generative Adversarial Network). This was what we had anticipated would give us the greatest reconstruction results, however we were greatly disappointed. The generator consisted of a CNN architecture and even after modifying layer complexity, results were blurry and features were undefined. The discriminator loss was decreasing so the model as a whole was learning, but the produced images did not follow suit.

4.3. Generative CNN

In our second investigation, we delved into how (solely) Convolutional Neural Networks (CNNs) perform in reconstructing the lower part of faces based solely on the visible upper sections, across varying levels of model complexity. Initially, a simple CNN model used for generation demonstrated an ability to grasp characteristics but often produced blurry and unclear lower face reconstructions. As we progressed to more complex models, we observed improvements in the alignment and placement of facial features; however, these models still struggled to capture the sharp details necessary for complete facial recognition. Importantly, none of the models exhibited overfitting, as evidenced by slightly fluctuating validation losses alongside training losses, suggesting that they adapted well to the training data while maintaining a level of generalization to new data. This observation underscores the need for further adjustments and possibly integrating more advanced techniques to surpass these learning plateaus and achieve enhanced accuracy in facial reconstructions for the task at hand. Furthermore, the visual results of these experiments are crucial and often more telling than performance metrics alone.

4.4. Transfer Learning

ResNet50 was proficient in identifying facial features like the mouth, showing a consistent improvement trend in the loss graph over time. Nonetheless fluctuations in loss indicate a need for adjustments to ensure model stability and precision, the model was still learning.

When we turned to VGG16 it demonstrated effectiveness in reconstructing faces producing some results that closely resembled the face region. Despite an initial decline in training loss, a plateaued validation loss trend suggests potential overfitting issues.

On the other hand Inception V3 exhibited varying levels of accuracy in reconstruction, the 139x139 image input constraint led to blurry inputs and blurrier outputs as as result. While there was a decrease in training loss, an uptick in validation loss towards the end of training hinted at possible overfitting concerns.

In summary each model showcased strengths along with areas that could benefit from refinement, notably VGG16 and Inception V3 exhibited overfitting characteristics.

We then decided to modify the U-net model (not related to the transfer learnings above but just to see what results we could get) typically used for segmenting images to reconstruct the bottom half of faces based on their upper halves while focusing on preserving spatial relationships. While the U-Net successfully recognized the shapes and positions of attributes it faced challenges, with finer details resulting in blurry reconstructions. The training loss consistently decreased, indicating learning progress. Fluctuating validation loss highlighted difficulties, in generalizing to data.

4.5. CGAN with ResNet

Then, we employed a cGAN with ResNet blocks to generate quality reconstructions. The thought process was that the generated image would be better, so perhaps the GAN could capture the best of both worlds when it comes to GAN and transfer learning. The cGAN steadily improved during training iterations enhancing the quality of generated images. Visual comparisons revealed that AI generated reconstructions accurately captured feature positions and shapes but lacked the clarity and intricacies of images indicating areas where model refinement is needed.

4.6. Final Model and Classification

After testing out models we decided to focus on ResNet specifically, out of all transfer learning models with decent output images, it contained room for improvement as depicted by a lack of overfitting characteristics. Our analysis indicated that this model created top notch images that also appeared as contiguous with the real upper half of faces. The refined ResNet consistently generates images with orientation and positioning closely resembling real facial features in the lower sections. This generation accuracy suggests that our classifying accuracy may be greatly improved.

Once again we revisited the CNN architecture we had initially built for classification, this time stitching together the real top half of the face, with a reconstructed bottom half. Impressively, the training accuracy reached 80% and the test accuracy was 96.6%. This is undeniably a massive jump from having only an occluded top half as input, especially using the same exact model architecture.

5. Conclusion

Overall, we managed to fulfill our initial goal of getting good accuracy after reconstructing the bottom half of the face. This was done using our fine tuned ResNet to generate images in combination with our initial CNN classifying architecture. However there are several notions of improvement we could have pursued if we had the chance.

The first would be dynamically applying the mask to initial images. For now, we simply crop 48x48 images into 24x48 images. This is pretty easily done and accurate about 95% of the time. However, sometimes parts even above the nose get cut out, maybe half their eyes or even up to their forehead. This can be solved by algorithmically finding the center point for each face (middle of the nose). Either pretrained models, or other methods of doing this could be explored.

The second would be exploring alternative classification models. We used the same one at both the beginning and end, but this was to showcase the improvement in accuracy reconstruction has. If we wanted to truly maximize the accuracy, we could implement either more complex CNN architectures or perhaps utilize transfer learning. This would make more sense since the model at the end is getting a full 48x48 input meaning there are more features to detect and classify based upon.

The third and final improvement would be refining the reconstruction model even further. We did fine-tune the ResNet and make it more complex, but perhaps there are other methods that we didn't explore fully. Both attempts at using a CGAN (with CNN and ResNet generators) did not provide any meaningful results. The model architecture for these GAN's may have been flawed, or perhaps the approach itself is not feasible. Either way, with redefined architecture, and time to experiment, better reconstruction results may be achieved.

References

[1] A. Alzu'bi, F. Albalas, T. AL-Hadhrami, L. Bani Younis, and A. Bashayreh, "Masked Face Recognition Using Deep Learning: A Review," *Electronics*, vol. 10, no. 2666, Oct. 2021. [Online]. Available: https://doi.org/10.3390/electronics10212666

[2] R. Magherini, E. Mussi, M. Servi, and Y. Volpe, "Emotion recognition in the times of COVID19: Coping with face masks," in *Intelligent Systems with Applications*, vol. 15, no. 200094, June 2022. [Online]. Available: https://doi.org/10.1016/j.iswa.2022.200094

[3] Hosen, Md Imran, and Md Baharul Islam. "Masked face inpainting through residual attention UNet." 2022 Innovations in Intelligent Systems and Applications Conference (ASYU). IEEE, 2022. https://doi.org/10.48550/arXiv.2209.08850