IDS 705 Project Proposal

Evaluating the Boundaries and Limits of Data Augmentation Techniques with Facial Emotion Recognition Image Data

Team Members:
Ahmed Ibrahim
Echo chen
Pragya Raghuvanshi
Suzanna Thompson

Team Number: 8

1 Introduction:

Provide a description of the problem and the value in finding a solution, motivate your reader as to why they should care about your problem or question.

The problem that this project aims to solve is twofold. Firstly, we are addressing the issue of limited data. The performance of machine learning models is heavily dependent on the amount of data available for training. With limited data, models may not learn effectively and can result in poor performance. One solution to this problem is data augmentation, which involves generating additional synthetic data to increase the amount of training data.

This project explores the use of various data augmentation techniques to create synthetic data. We would also investigate the boundaries and limits. It is essential to understand the impact of different augmentation techniques on model performance and the types of data that are suitable for each technique. For example, some augmentation techniques may be more effective for images with specific features, while others may not be suitable for certain types of data.

Furthermore, it is crucial to evaluate the impact of over-augmentation on model performance. Over-augmentation can lead to a model that is overfitting to the training data and may not generalize well to new, unseen data. Therefore, it is essential to understand the trade-off between the amount of augmented data and model performance.By investigating the boundaries and limits of different data augmentation techniques, this project can provide insights into the most effective techniques for generating synthetic data, which can improve the accuracy of machine learning models in various applications

Secondly, the project aims to address the challenge of accurately identifying emotions from facial images. This has become an increasingly difficult task in recent times due to the widespread use of masks, glasses, and other facial coverings. In addition, people often cover their faces with their hands or hair, making it even harder to detect facial expressions. This problem can have significant implications in fields such as healthcare, marketing, and human-computer interaction.

Therefore, this project has significant value in finding a solution to these challenges. By exploring the use of synthetic data and investigating the impact of facial occlusion on emotion classification accuracy, we can develop more effective machine learning models that can accurately classify emotions even when there are facial coverings. The outcomes of this project can lead to advancements in various fields that rely on facial emotion recognition, such as mental health, customer satisfaction, and human-robot interaction.

2 Goal/Objective:

Provide a precise statement about the goal of the project. Is it a problem you’re trying to solve? Are you trying to develop a specific application? Is there a question you’re trying to answer? This should be precise and unambiguous and act as the guiding premise behind your work

Overall Goal: The overall goal of the project is to investigate the effect of facial occlusion on emotion classification accuracy, and to determine the minimum amount of synthetic data required to generate accurate synthetic images using data augmentation techniques.

Sub-Goals:

- Sub-Goal 1: Build a model that can classify emotions on facial images.
- Sub-Goal 2: Creating synthetic data using types of data augmentation, for example GANs,color space transformations, kernel filters etc. Varying the size of the dataset to determine the optimal amount of data required to generate synthetic images that can effectively classify emotions. Investigate the boundaries and limits of different techniques.
- Sub-Goal 3: Create different synthetic occlusion data by varying the percentage of facial occlusion in the dataset using the FaceExtraction GitHub repository.
- Sub-Goal 4: To combine synthetic occlusion image data with synthetic facial emotion data.
- Sub-Goal 5: To vary the percentage of facial occlusion and determine the point at which the emotion classifier breaks.

Overall, achieving these sub-goals will enable the project to achieve its overarching goal of investigating the effect of facial occlusion on emotion classification accuracy and determining the minimum amount of synthetic data required to generate accurate synthetic images.

3. Background

Data Augmentation refers to the data space solution of the problem of limited and imbalanced data. Limited data is a problem as gathering enormous amounts of data and labeling it is a herculean task. In addition, class imbalance, where the dataset is skewed w.r.t the majority, can affect the performance of our model heavily. Also, objects in realistic settings exhibit considerable variability, so to learn to recognize them it is necessary to use much larger training sets. Hence, data augmentation techniques like data warping and data oversampling are heavily explored to solve these issues. These augmentations artificially inflate the training dataset size by either data warping or oversampling. Data warping augmentations transform existing images such that their label is preserved. This encompasses augmentations such as geometric and color transformations, random erasing, adversarial training, and neural style transfer. Oversampling augmentations create synthetic instances and add them to the training set. This includes mixing images, feature space augmentations, and generative adversarial networks (GANs). Hence, data oversampling can be very useful for problems with class imbalance.

3.1 Data Augmentation on basic image manipulations

Data augmentations like data warping, geometric transformations, color space transformations, kernel filters, mixing images and random erasing can massively inflate the dataset size with a large set of potential outcomes. There has been a lot of research in this regard and how they affect the model's performance. Usually, the safety of data augmentations is generally associated with its ability to preserve labels. A non-label preserving transformation could potentially strengthen the model's ability to output a response indicating that it is not confident about its prediction. However, achieving this would require refined labels [See Reference] post-augmentation. This may be out of scope for our project and we will stick to label preserving augmentations.

The Image Net Classification by AlexNet CNN architecture [Reference] uses simple Data Augmentation techniques of data warping like cropping patches from original image, horizontal flipping, and using PCA augmentation to inflate the dataset and reduce the error rate of the model by over 1%. There have been studies which give a possible comparison of performance of these different augmentations, which might be considered for evaluation of performance [See Reference]

3.2 Augmentations using Deep Learning

This survey [See Reference] talks about augmentation using deep learning use approaches like feature space augmentation and adversarial training methods. Feature space augmentations can be implemented with auto-encoders if it is necessary to reconstruct the new instances back into input space. It is also possible to do feature space augmentation solely by isolating vector representations from a CNN. Adversarial training is a framework for using two or more networks with contrasting objectives encoded in their loss functions. Another exciting strategy for Data Augmentation is generative modeling. Generative modeling refers to the practice of creating artificial instances from a dataset such that they retain similar characteristics to the original set. The principles of adversarial training discussed above have led to the very interesting and massively popular generative modeling framework known as GANs, which we would explore in the project going forth.

3.4 Facial Emotional Detection

There has been a lot of work in the past dealing with emotion classification and issues like lower accuracy in the same due to imbalanced distribution of certain classes [See Reference]. Here, a transfer learning approach is used for Deep CNN architectures.

3.5 Data Augmentation in Emotion Detection

It is a difficult task to classify images with multiple class labels using only a small number of labeled examples, especially when the label (class) distribution is imbalanced. Emotion classification is such an example of imbalanced label distribution. This paper [See Reference] that we will follow very closely, talks about data augmentation methods using generative adversarial networks (GAN). It can complement and complete the data manifold and find better margins between neighboring classes.

3.6 Occlusion Data

A facial occlusion refers to the partial or complete obstruction of a person's face. A facial occlusion can be caused by someone's hair, class, mask, other people, or any other occluding object. Facial occlusion causes a myriad of problems in computer vision with facial detection and in some medical applications There is a lot of research into the field of facial detection given a facial occlusion. In the following sections we will do a partial and non-exhaustive review of the work done so far.

3.6.1 Local Feature Methods

Local feature methods use local regions and descriptors to represent facial features. Local features on a face include the nose, eye, and mouth, as well as more granular features like the eyebrows and cheeks. These methods include and are not limited to scale-invariant feature transform (SIFT), speeded up robust feature (SURF), and local binary pattern (LBP). In future iterations of this report we will provide more detailed descriptions of each of these methods, including advantages and disadvantages. [See reference]

3.6.2 Template Matching Methods

Template matching methods compare the parts of the face that are occluded with a set of reference faces in order to find a best match. These methods typically use a similarity score to evaluate the percent difference in facial occlusion. These methods include and are not limited to eigenface, fisherface, and local binary pattern histogram (LBPH). [See reference]

3.6.3 Deep Learning Methods

Deep learning methods use deep neural networks to train a model to recognize specific and common features within an occluded face. The model works by understanding what a face looks like first, and then classifies if an image has a face and what parts of the face are occluded. Deep learning based methods include but are not limited to convolutional neural networks (CNN), recurrent neural networks (RNNs), and generative adversarial networks (GANs). [See reference]
In future iterations of this report we will provide more detailed descriptions of each of these methods, including advantages and disadvantages.

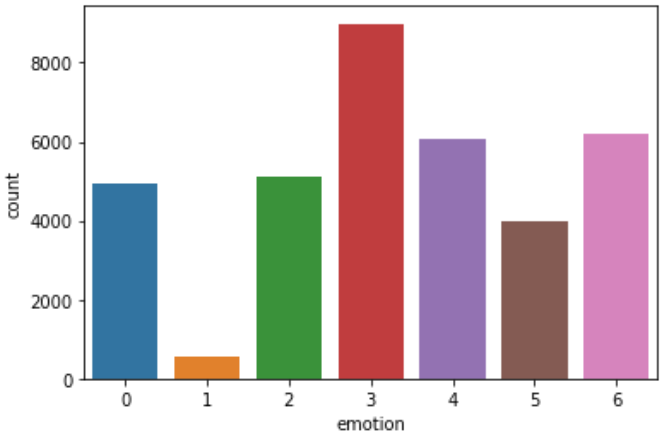
4 Data

Facial emotion dataset

Facial emotion datasets are crucial for training machine learning models to recognize and classify emotions expressed on human faces. These datasets provide large amounts of labeled images or videos of facial expressions that allow algorithms to learn patterns and features associated with different emotions. In short, facial emotion datasets are essential for the development of accurate machine learning models for facial emotion recognition, which have many practical applications in various industries. They also provide valuable insights into the nature of human emotions and their expression.

Here, we have collected a dataset from Kaggle which was originally adapted from a research paper titled "Eavesdrop the Composition Proportion of Training Labels in Federated Learning". Our primary intention of using this dataset is to accurately represent each face based on the emotion shown in the facial expression, into one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).

The training dataset consists of approximately 30000 examples, which contains two columns, "emotion" and "pixels". The "emotion" column contains a numeric code ranging from 0 to 6, inclusive, for the emotion that is present in the image. The "pixels" column contains a string surrounded in quotes for each image. The contents of this string are space-separated pixel values in row major order. The test dataset contains only the "pixels" column and our task is to predict the emotion column. The test dataset consists of roughly 3500 examples.



The distribution of the training dataset shows that in our training dataset, we have the highest count of label 3 emotions, which are ‘happy’ emotions, and the lowest emotion count is ‘Disgust’. The lowest one makes sense since most people don’t make disgusting faces frequently. The distributions of other emotions are quite even. The uneven distribution of class labels (a.k.a emotions) suggests that our machine learning algorithm might have some bias towards emotion 1 and 3 depending on their imbalanced proportions in respect to other classes.

Facial occlusion dataset

Face occlusion refers to the partial or complete obstruction of a person’s face by an object, another person, or the environment. This obstruction can prevent the viewer from seeing or perceiving the facial features and expressions of the occluded person. In machine learning, face occlusion is a challenge for facial recognition systems and other computer vision applications that rely on the analysis of facial features. When a person’s face is partially or completely occluded by an object such as a mask or sunglasses, it can hinder the accuracy of facial recognition algorithms.

The paper, “FaceOcc: A Diverse, High-quality Face Occlusion Dataset for Human Face Extraction” proposes a novel diverse, high-quality face occlusion dataset entitled FaceOcc, which contains all mislabeled occlusions in CelebAMask-HQ and complements some occlusions and textures from the internet. Together with the facial attribute masks in CelebAMask-HQ, the proposed dataset yields face masks and augmented data for training face extraction models. According To the paper, the dataset is validated by training a straightforward face extraction model.
The dataset consisted of only 500 images along with complementing internet occlusions consisting of cups, cigarettes, mask, mask face and microphone. The internet occlusions consists of about 40-50 images, approximately 10% of the training dataset. In addition, we have about 2300 celeb A HQ manually labeled face occlusions.

5. Methods

5.1 Emotion Classification

1. CNN for classification of emotion detection:
   CNN is a popular method used to extract features from images to detect emotions. A CNN is a DL algorithm which takes an input image, assigns importance (learnable weights and biases) to various aspects/objects in the image and is able to differentiate between images. The preprocessing required in a CNN is much lower than other classification algorithms. One role of a CNN is to reduce images into a form which is easier to process without losing features that are critical for good prediction. The main CNN operations are convolution, pooling, batch normalization and dropout. The objective of the convolution operation is to extract high level features such as edges from an input image. The convolution layer functions are as follows:
   - The first convolutional layer(s) learns features such as edges, color, gradient orientation and simple textures.
   - The next convolutional layer(s) learns features that are more complex textures and patterns.
   - The last convolutional layer(s) learns features such as objects or parts of objects.
   - The element involved in carrying out the convolution operation is called the kernel. A kernel filters everything that is not important for the feature map, only focusing on specific information. The filter moves to the right with a certain stride length till it parses the complete width. Then, it goes back to the left of the image with the same stride length and repeats the process until the entire image is traversed.
2. Transfer Learning: We also might want to explore transfer learning which works by training a network on a big dataset such as ImageNet and then using those weights as the initial weights in a new classification task. Typically, just the weights in convolutional layers are copied, rather than the entire network including fully-connected layers. This is very effective since many image datasets share low-level spatial characteristics that are better learned with big data.

5.2 Data Augmentation

1. Conducting basic image manipulations and data augmentation techniques that are label preserving. This will result in data inflation and we can check for which data augmentation measures lead to improvement in the model's performance.
3. GAN for synthetic images of emotions: Among the many applications of GAN, image synthesis is the most well-studied one, and research in this area has already demonstrated the great potential of using GAN in image synthesis. A Generative Adversarial Net consists of two neural networks, a generator and a discriminator, where the generator tries to produce realistic samples that fool the discriminator, while the discriminator tries to distinguish real samples from generated ones.

5.3 Data Augmentation for Occlusion

4. Random Erasing: This technique (See Reference) was specifically designed to combat image recognition challenges due to occlusion. Occlusion refers to when some parts of the object are unclear. Random erasing will stop this by forcing the model to learn more descriptive features about an image, preventing it from overfitting to a certain visual feature in the image. Aside from the visual challenge of occlusion, in particular, random erasing is a promising technique to guarantee a network pays attention to the entire image, rather than just a subset of it. Random erasing is a Data Augmentation method that seeks to directly prevent overfitting by altering the input space. By removing certain input patches, the model is forced to find other descriptive characteristics. This augmentation method can also be stacked on top of other augmentation techniques such as horizontal flipping or color filters.
5. Cycle GAN: CycleGAN introduces an additional Cycle-Consistency loss function to help stabilize GAN training. This is applied to image-to-image translation. It can do image-to-image transition between two unpaired image domains. In the paper referenced above, neutral class is used to generate images for minority classes like disgust and fear. We can use this structure to build occluded images from our base class, that is, neutral class.

6 Experiments

6.0.1 Facial Emotion Image Data Cleaning and Processing

To ensure that our model can perform well and is able to learn the data accurately, we will evaluate our data for any excessive noise, orientation issues, color discrepancies, outliers, or duplicates. This is important to the modeling process as if we train with low-quality inputs, then our results will be low-quality as well.

Noise

Although it is very difficult to quantify the amount of noise in an image in general, we will attempt to calculate a rough estimate and eliminate images with too much noise. To do this, we will use a histogram analysis, which is a graphical representation of the distribution of pixel intensities in an image. After a brief visual inspection of each histogram, if an image is found to have an anomalous distribution, we will investigate it further to see if it is a reasonable image to include in the data.

Since visual inspections are not accurate and not exhaustive, we are also considering using a Fourier transform on each image to evaluate the image's frequency components. After using a Fourier transform, we will (try to) check that each transform does not have random, high-frequency components that are not aligned with the original image numerically.

We are also considering using the signal-to-noise ratio (SNR), which is a metric that computes the ratio of the noise in an image relative to its signal. We will consider eliminating images that are below a threshold for SNR, as a low SNR can mean a high level of noise.

Orientation

To make sure that all images are roughly the same orientation, meaning that all faces are mostly front-facing, we will do a visual inspection of each image.

Color Discrepancies
The documentation of the images reports that all the images should be greyscale. To validate this supposition, we will check the image mode of each image using the PIL package in python. If images are found to not be in grey scale, we will convert them to grayscale using the PIL package as well.

Outliers
An outlier in this context is defined to be an image that is not of a face or has no noted emotion label. If we find that an image does not have an associated emotion label, we will delete the image as we are not experts in classifying emotions. To validate that all images are of faces, we will use an existing model with sufficient accuracy, such as the Viola-Jones algorithm or the RetinaFace algorithm, with all of our images and filter out the images that are not classified as faces.

Duplicates
To ensure that there are no duplicate images in our dataset, we will use an image hashing algorithm, like average hashing, difference hashing, or perceptual hashing, and calculate the hashes for each image. Once each image has its hash, we will use either the Hamming distance or cosine similarity value and a high similarity threshold to compare if any two images are the same. If two images are found to have a high similarity, we will visually investigate if the two images are duplicates.

6.0.2: Classification of Facial Emotion Image Data

We will use a convolutional neural network (CNN) to create a classifier with our facial emotion image data. This classifier will take in a vectorized image and return what emotion the initial image was showing. We are considering an unsupervised pre-training on our CNN with the MS-Celeb-1M dataset. The MS-Celeb-1M dataset has over ten-million images of more than one-hundred-thousand celebrities. Pretraining with this dataset will be a first step for our model to recognize faces and the features associated with a face. Pretraining is an avenue that we are exploring because this classification model will serve as a baseline and accuracy metric for the rest of our project.
To evaluate the performance of this model, we will use accuracy, a precision and recall curve, a receiver operating characteristic (ROC) curve, and F1-score.

Accuracy
We are choosing accuracy as a performance metric as it is very interpretable and quantifies the number of correctly classified images.

Precision and Recall
We are choosing precision as a performance metric because it quantifies how well our model classifies emotions correctly. Precision divides the number of true positive classifications by the number of actual positive values. We are choosing recall as a performance metric because it quantifies how many of the classified positive values are correct. Because our outcome variable, emotion, is not binary, we will calculate precision and recall class-wise.

ROC Curve
We are choosing to plot a ROC curve to show performance because it is a visual representation of how often our model classifies positives correctly against how often how model classifies positives incorrectly. This is an

important metric to use because we seek for the model to classify each emotion well. Because our outcome variable, emotion, is multiclass, we will use the one-vs-all approach to avoid training a second classifier.

F1-score

We are choosing to use the F1 score as a performance metric because we value both precision and recall in our problem. Because our outcome variable, emotion, is multiclass, we will calculate the F1 score class-wise and average them. This approach is called the macro-averaged F1 score.

6.03: Create Synthetic Facial Emotion Image Data

To create synthetic facial emotion image data, we will replicate the work from Shorten, et. al and use a CycleGANs approach. To evaluate the performance of the model, we will give our synthetic facial emotion image data to the original facial emotion classifier mentioned above. We are considering using the dice coefficient and Jaccard coefficient to test how similar the synthetic data set is to the original dataset. These coefficients may also tell us to what extent we need to change the initial dataset in order to create synthetic data from it.

6.04: Vary the extent of which we limit our dataset and evaluate the ability of the model to faithfully synthesize new data

In this experiment, we are testing to see what the minimum size of our facial emotion image dataset is to generate synthetic image data. In that, we are seeking to understand if we only use 10% of the original dataset, how do those synthetic images compare to the case where we use 90% of the original dataset to generate the synthetic images. We will evaluate the performance of the GAN based on input size by classifying the synthetic images with our initial emotion classifier. If the classifier classifies the synthetic images well enough (which we will have to set a threshold for), then we can conclude that there is evidence to support that the sample size used is large enough to faithfully create synthetic image data.

We are also considering using an inception score, Frechet inception distance, and a visual Turing test to evaluate how well the GAN creates synthetic images.

We will also evaluate the model's performance using the following metrics:

1. Plotting of validation and training error: To check for signs of overfitting before data augmentation. After applying data augmentation, evaluate how the convergence changes(number of epochs, signs of overfitting etc.)
2. t-SNE Visualization: t-SNE is a visualization technique that learns to map between high-dimensional vectors into a low-dimensional space to facilitate the visualization of decision boundaries
3. Further metrics might be incorporated for better results as we move forth.

6.1.1: Facial Occlusion Image Data Cleaning and Processing

We will repeat each step in 6.01 with our facial occlusion image dataset.

6.1.2: Create Synthetic Facial Occlusion Image Data

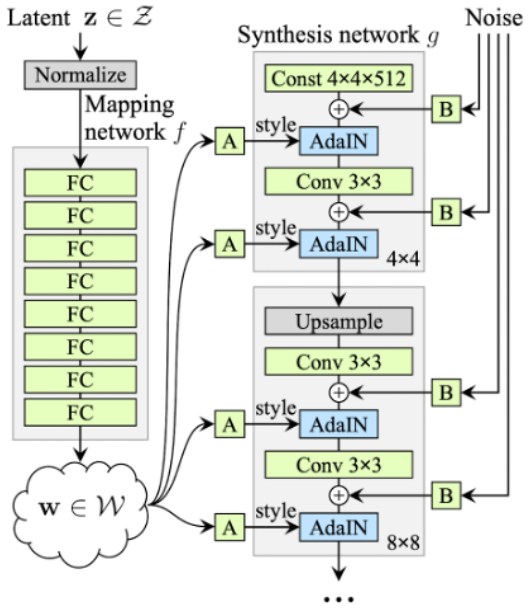
We will repeat the process in 6.03, excluding testing the synthetic images on a classifier. We are considering using the inception score, Frechet inception distance, and a visual Turing test for evaluating the performance of this particular synthetic image creation model.

6.1.3: Vary the extent of which we limit our dataset and evaluate the ability of the model to faithfully synthesize new data

We will repeat the process in 6.04 with our facial occlusion dataset.

6.2.1: Combine the synthetic facial occlusion images and the synthetic facial emotion images

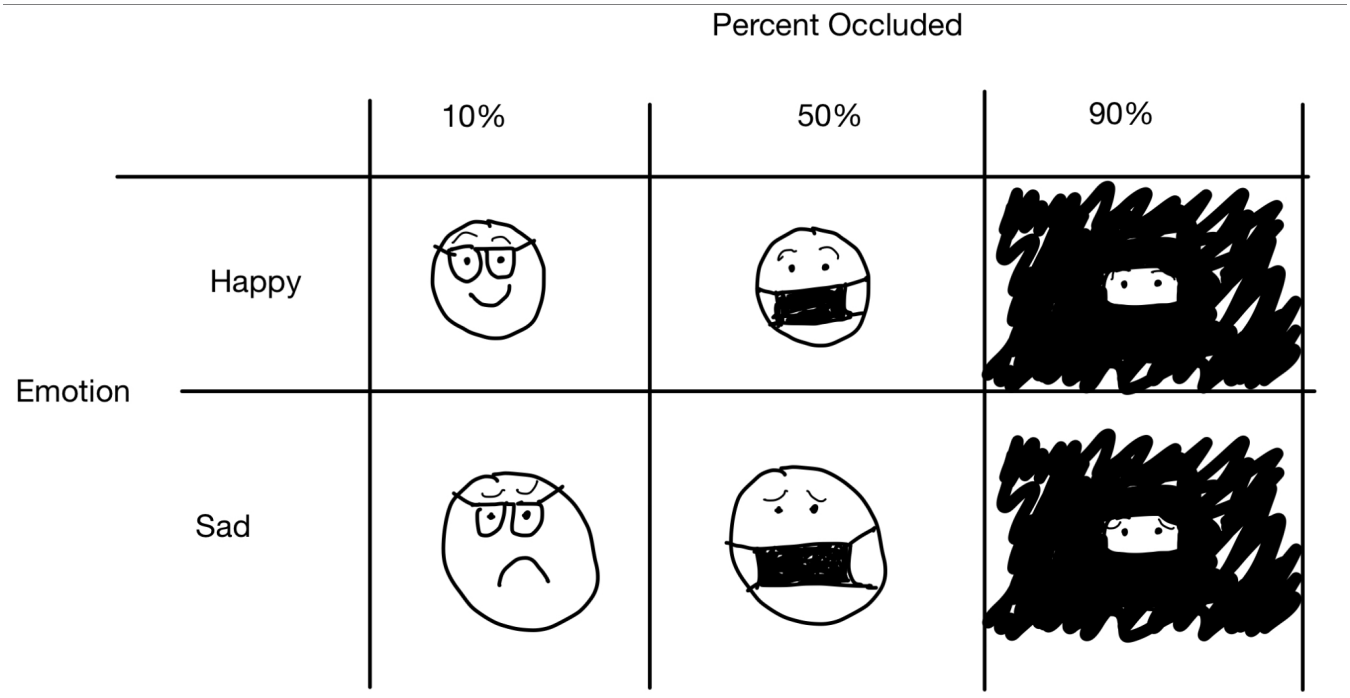
We will replicate the work of Kerras et. al, with their style generator architecture (shown below) to combine our synthetic emotion images and our synthetic occlusion images.



(b) Style-based generator

6.2.2: Vary the amount of occlusion in the synthetic facial occlusion images

In this section of the project, we are going to vary the amount that the face is occluded and then create synthetic facial occlusion images. We will use the work of Yin, et. al to quantify the percent of facial occlusion.
In the image below, we see an example of what this may look like.



6.23: Feed the synthetic images into the classification model from section 6.02 and evaluate at what threshold of percent occlusion of the face reduces the accuracy significantly.

Once the synthetic occlusion facial images are generated, we will use them as input data for our facial emotion classifier and evaluate how well the synthetic images maintain or show emotion when the face is occluded. We are seeking to understand, at what percent of the occlusion can a facial emotion classifier no longer recognize emotion.

7 Roles

Describe the specific roles and responsibilities each team member is taking on for this project and is subject to change.

Echo
Develops and trains the emotion classification model using available facial image data.
Participates in the data pre-processing and hyperparameter optimization.
Evaluates the performance of the emotion classification model.

Suzanna
Develops and implements the GAN model for synthetic data generation.
Integrates synthetic data with real data for training and testing the model.
Participates in the data pre-processing and hyperparameter optimization.

Pragya
Investigates the impact of facial occlusion on emotion classification accuracy.
Creates different synthetic occlusion data by varying the percentage of facial occlusion in the dataset.
Combines synthetic occlusion image data with synthetic facial emotion data.

Ibrahim
Varies the percentage of facial occlusion to determine the point at which the emotion classifier breaks.Participates in the synthetic data generation and integration process.
Working together to write and edit the research paper based on the project's findings.

References

Include a list of references that you have already read (you should have read at least 5 prior to submitting your proposal to give you context for the project) or plan to read to further your knowledge of this problem.

https://openaccess.thecvf.com/content_CVPR_2019/papers/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.pdf

https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0

https://arxiv.org/abs/1711.00648

https://arxiv.org/pdf/1611.09961.pdf

https://arxiv.org/abs/2201.08425

https://dl.acm.org/doi/pdf/10.1145/3065386
https://ieeexplore.ieee.org/document/9799910

Other information

Note: this section is for information purposes and should not be included in your report

How to use this template

1. Click File > Make a copy and delete the instructions replacing text with your report content
2. Follow the formatting instructions contained in this document
3. Share your completed version of the file giving general access to “anyone with the link” and at least “commenter” privileges so that we can add comments and suggestions directly to the text

Length requirements

The proposal does not have any specific length requirements. You are welcome to reuse content written for the proposal for the final report, so if you have a well-written background section, for example, you can reuse that content in the final report saving your team time and effort.

Additional formatting requirements

See the **Final Report Template** for additional formatting requirements

Wednesday
Thursday
Proposal due friday

Part one:
Classify emotions
Use GANs to create synthetic data
Vary size of dataset → how much data do we need in order to create synthetic images well?

Part two:
Create synthetic occlusion data (https://github.com/face3d0725/FaceExtraction)
Vary percent facial occlusion (requires that % occluded is in the dataset).

This part requires part 1 completion: Combine synthetic occlusion image data and synthetic facial emotion data

Part three
Vary percent facial occlusion

At what percent facial occlusion does the emotion classifier break?

April 15th