Abstract art generation using speech emotions

Harsha Tolani, Shivani Kalamadi, and Sourin Chakrabarti

University of California, Davis

Abstract. Translation of speech to image directly without text is an interesting and useful topic in various applications in computer science like computer-aided design, creation of an artistic interpretations, human-to-computer interaction, etc. In this project, we aim to analyze emotions from speech and use the predicted emotion along with the speech text context to generate an image. This generated image could be used as thumbnails for songs or generating abstract art for physically challenged people. We have used the VQGAN model pre-trained on the Imagenet dataset as a generator and a pre-trained CLIP model as a preceptor for generating the image based on emotion and speech. For predicting the emotions, we have compared various neural networks such as networks based on LSTMs, CNN networks, and finally an MLP network. The RAVDESS emotional speech recognition dataset was used for training. The results show that the LSTM model performs the best among the three for predicting emotions and the VQGAN+CLIP give accurate context-aware images. The code for this project can be found in - `https://colab.research.google.com/drive/15AI0qbNLf0MxHwQg6bSS2qyGs4pxn66y?usp=sharing`

Keywords: VQGAN · CLIP · LSTM · CNN · MLP.

1 Introduction and Literature Review

Nowadays, one of the primary areas of research is done on text-to-image generation. But this method couldn't take into account the emotion expressed by the speaker. A person's emotional state can be easily guessed based on his/her facial expression and speech as a modality. Even for human beings, judging the emotions of the speaker from the speech at times might not be naive. At the same time, multiple emotions could be well expressed through speech. So we will focus on building a model which takes input as speech from the user, analyzes the emotion, and accordingly displays the image asked for by incorporating the emotions of the user recognized through the speech input. This in turn will give a much more personalized experience for the users which is not possible by merely taking text input through a text prompt.

This problem statement appears in many domains such as biomedicine and space technology. To tend to these modern technological demands, we can utilize Digital Image Generation using GANs that is economical and at the same time shows promising results.

Speech emotion recognition as an independent topic is a widely researched area in the field of machine learning. Although most of the work exists by converting the speech to text and then using NLP for emotion recognition, we are using the features of the speech audio to determine the emotion. This can be especially helpful in cases where the same text can be used to describe two different emotions.

In this project, we aim to use deep learning and a GAN-based approach to take speech as input from the user, analyze the emotions associated with it and accordingly generate the artwork which has been demanded by the user which will provide a personalized experience. The RAVDESS dataset is used for training the speech emotion recognition models. We have compared LSTM, CNN, and MLP-based neural networks for the above task. The result is fed to a VQGAN(Vector Quantized Generative Adversarial Network) [4] and CLIP(Contrastive Language Image-Pre-Training) [5] based model to generate the image. The two models work in tandem. VQGAN works on image generation from the prompt whereas CLIP scores the image on how well it matches the prompt. This combined effort guides the model to generate accurate images without the help of any image reference.

Producing artwork from text is widely researched, such as in [1] where the authors have used image generation and style transfer models with an improved GAN-based network to generate the images. They used classification information to compute the loss for the GAN which improved the model's performance.

Similarly, in [2] the authors have used VQGAN+CLIP to generate clip-artwork from a text prompt and compared the work to existing technologies such as CNN-based approaches like Alexander Mordvintsev's DeepDream and Pix2Pix. A discrete codebook was used which provides an interface among these designs and a patch-based discriminator that empowers strong density while ensuring the retention of high perceptual quality.

Training a text-to-image generator model involves gathering gigantic volumes of text-image records, which is not very practical to gather. In article [3], the authors have presented CLIPGEN, which uses a trained CLIP model to generate language-image priors, which further helps in self-supervised image generation. A set of unlabeled images converted to a VQGAN codebook are used as directives for the generator.

2 Proposed Methodology

This section of the paper dives deep into the approach followed by the generation of the images from speech. There are two major components in our approach. Our first aim is to predict the emotion from the speech. Speech audio will be considered as input for the model. The speech is converted into text with the help of python's SpeechRecognition library. This text is not used for emotion recognition, but only as a prompt for the image generation model. The complete architecture can be seen in Figure 4.

For emotion recognition, we have compared three different models. The LSTM (Long Short Term Memory) model uses LSTM layers to learn the long-term dependencies between the audio features. The second network used for classification is based on CNN. Finally, an MLP-based approach was also trained to establish the baselines.

The RAVDESS(The Ryerson Audio-Visual Database of Emotional Speech and Song) [6] dataset was used for training the models. We have only used the audio-only speech files from the dataset. This trimmed the dataset to 1440 files: 60 trials per actor x 24 actors = 1440. Each of the 1440 files has a unique filename which has a numerical identifier of 7 components. These identifiers represent the following features: Modality, Emotion, Intensity, Statement, Repetition, and Actor.

The audio files were pre-processed to extract features. We extracted the Mel-frequency cepstral coefficients (MFCC) features (40 features), chromagram stft features (12 features), and the Mel spectrogram features (128 features) from the input audio. These features were used in training the models described above.

The LSTM network takes an input feature vector of 180 features and 5 time steps. The first layer is an LSTM layer with 256 nodes followed by another LSTM layer with 128 nodes. Then, we use a fully connected network with 2 layers of 2048 nodes each followed by an output layer of 8 nodes. We have used ReLU activation for the intermediate layers and softmax for the output. We have used Adam optimizer for training. The complete network can also be seen in Figure 1.

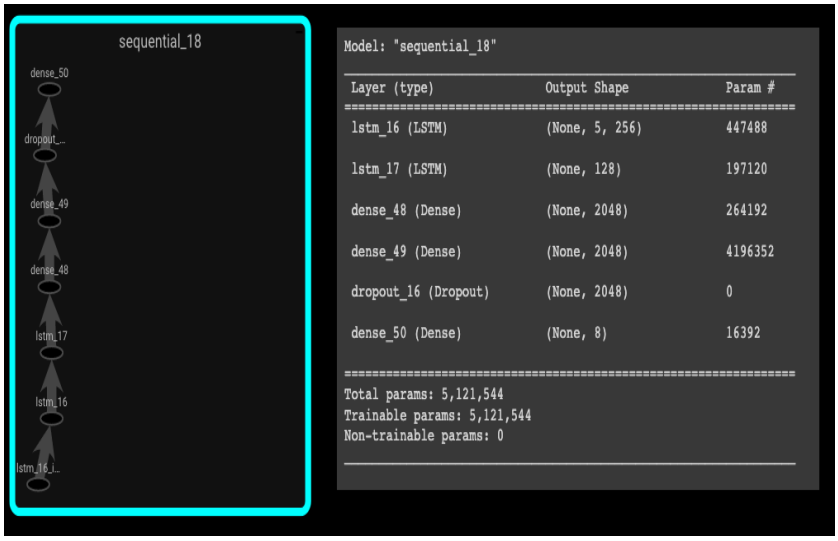

Fig. 1. Architecture of the LSTM network.

The MLP network consists of 2 fully connected layers of 256 and 128 nodes each with ReLU activation. This is connected to the output layer of 8 nodes with a Dropout layer with a probability of 0.3. Adam optimizer is used for training. The complete network can also be seen in Figure 2.

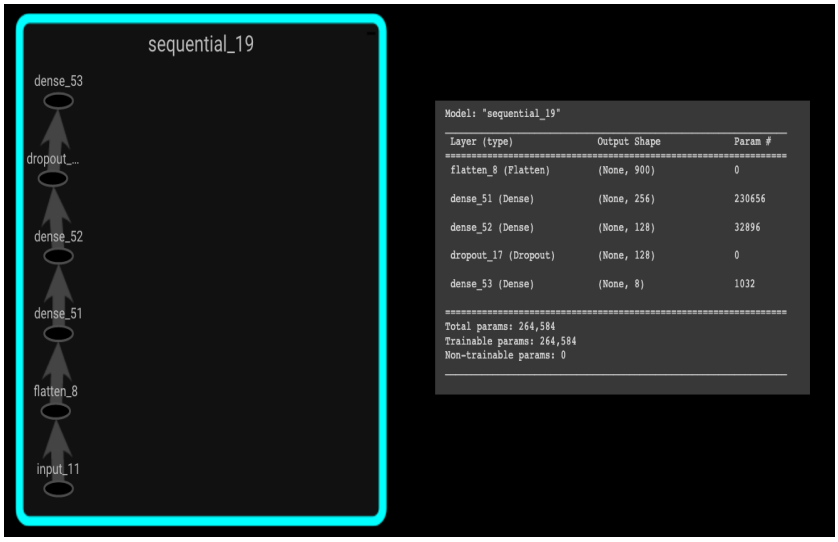

Fig. 2. Architecture of the MLP network.

The CNN network is made up of two convolutional layers followed by a fully connected network. The first convolutional layer takes an input of shape (5, 180,1) and has a kernel size of (4,4). Average pooling is done after the first layer with a pool size of (2,2). The second layer has a kernel size of (2,2). A similar average pooling is also done after the second layer. The fully connected network has 2 layers with 2048 and 1024 nodes respectively before an output layer of 8 nodes. Adam optimizer is used for training. The complete network can also be seen in Figure 3.

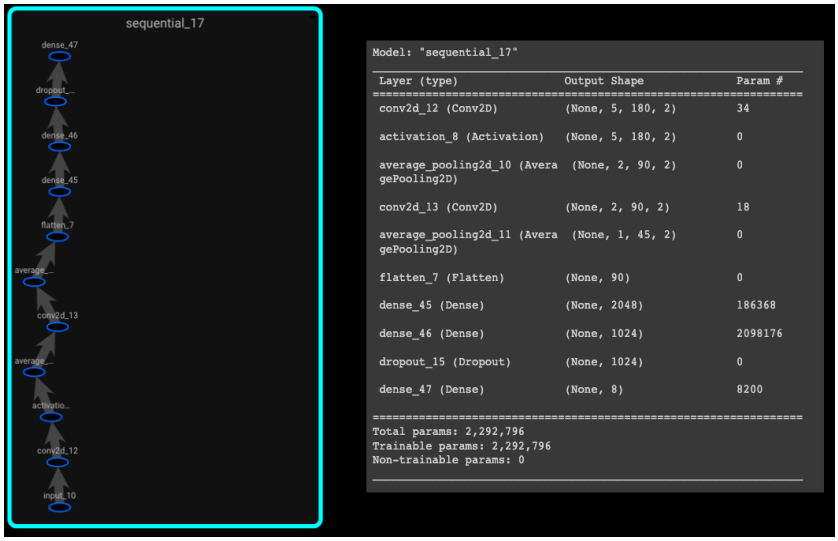

Fig. 3. Architecture of the CNN network.

Another important component of our system is image generation from the prompt. When used together, VQGAN acts as the generator whereas CLIP acts as the preceptor. The emotion detected earlier is used to provide a base image for the VQGAN to work on. The text prompt is given to the VQGAN to generate the image. We have used pre-trained models for VQGAN and CLIP. The VQGAN model being a GAN will be processing the input image dataset in a vector quantized format where clusters of different image vectors are encoded and the cluster center called a "codeword" is considered a representation of the image data. VQGAN will adapt to the textures and colors corresponding to the speech emotion being detected and outputs the final image consisting of colors and textures giving a more personalized experience in the process of artwork generation. The CLIP scores the image generated by VQGAN by using the image and text prompt and using the cosine similarity between the two. The code for VQGAN+CLIP was referred from [7] and we have added our own modifications to it.

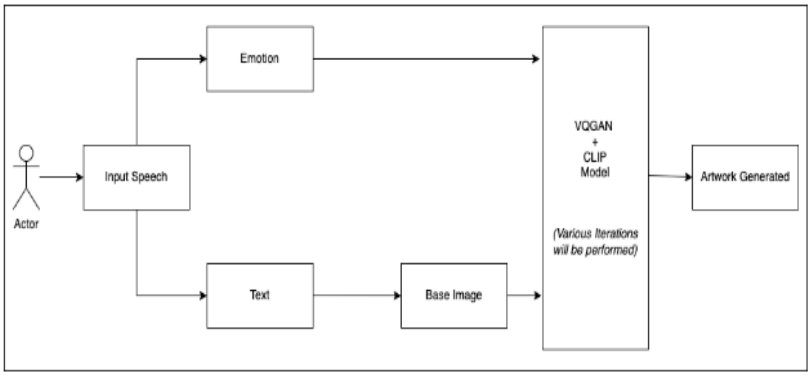

Fig. 4. System Architecture.

3 Experimentation and Results

We trained our model on Google Colab with approximately 12 GB of RAM and 15GB of GPU memory. The LSTM model was trained with a batch size of 32 and a learning rate of 0.001 for 100 epochs. It achieved the best accuracy among the 3 models with approximately 75% accuracy. The CNN model was trained with a batch size of 16 and a learning rate of 0.001 for 100 epochs. The MLP model was also trained with the same learning rate and batch size for 200 epochs. Both achieved an average accuracy of around 58%. The results are also presented in 1.The training accuracy and loss are plotted in Figures 5 and 6. The VQGAN+CLIP was trained for about 300 epochs and generated discernible images. The loss achieved was about 0.8 on average. The final image after 300 iterations can be seen for two cases in Figure 7.

Table 1. Accuracy for speech emotion detection

Model	**Accuracy**
MLP model	62%
CNN model	56%
LSTM model	75%

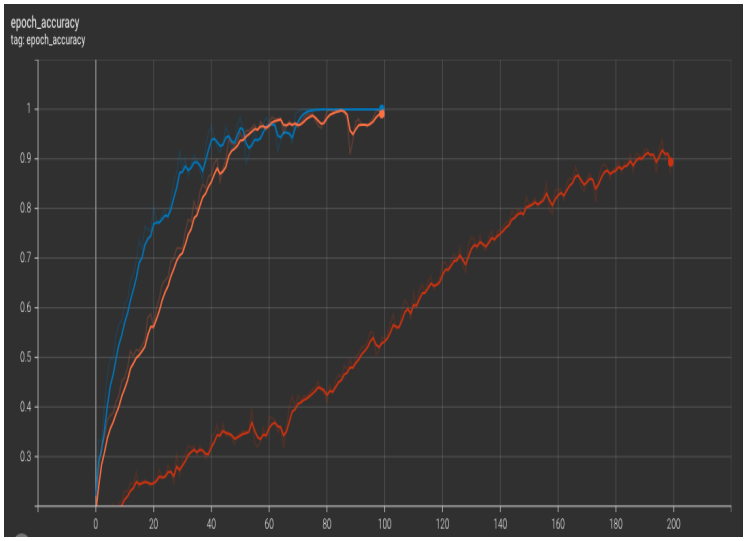

Fig. 5. Training accuracy vs epochs for all the networks. The blue lines correspond to the LSTM trainings, the orange lines represent CNN trainings and the red lines are for MLP.

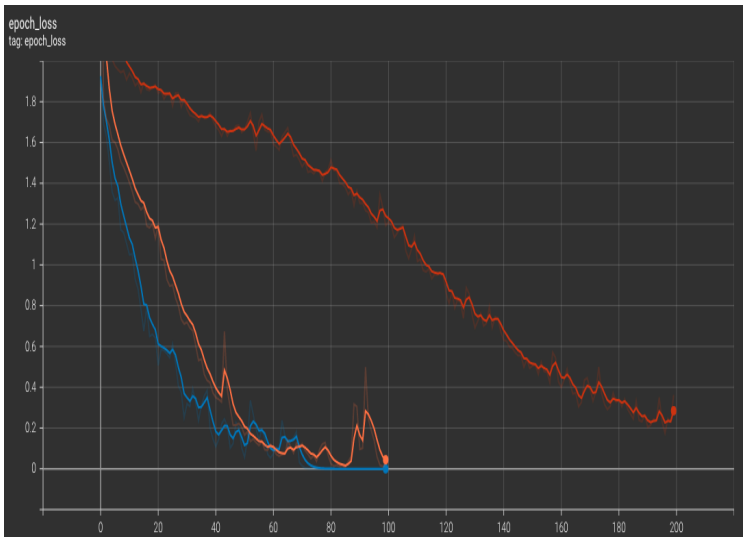

Fig. 6. Training loss vs epochs for all the networks. The blue lines correspond to the LSTM trainings, the orange lines represent CNN trainings and the red lines are for MLP.

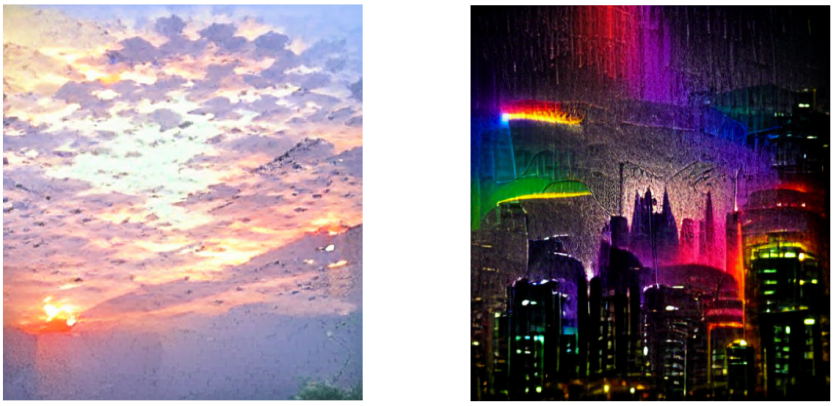

Fig. 7. Final image for "Morning sunrise" with happy emotion and "Dark rainbow city" with fearful emotion.

4 Conclusion and Future Scope

In this paper, we have proposed an image generation approach using speech emotion recognition. We have used VQGAN and CLIP models in tandem to generate an image from a text prompt from speech and emotions recognized from the speech spectrogram. We see that the achieved accuracy of the emotion recognition model was about 75%. Further work could be directed toward improving the accuracy of the model. Also, training the model for more emotions and for different languages would be an interesting and challenging topic.

References

1. Chen, Z., Chen, L., Zhao, Z. and Wang, Y., 2020, July. AI illustrator: Art illustration generation based on generative adversarial network. In 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC) (pp. 155-159). IEEE.
2. Yuan, T., Chen, X. and Wang, S., Gorgeous Pixel Artwork Generation with VQGAN-CLIP.
3. Wang, Z., Liu, W., He, Q., Wu, X. and Yi, Z., 2022. CLIP-GEN: Language-Free Training of a Text-to-Image Generator with CLIP. arXiv preprint arXiv:2203.00386.
4. Miranda, L.J., Taming the'Taming transformers' paper (VQGAN). ljvmiranda921. github. io.
5. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G., 2021, July. Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748-8763). PMLR.
6. Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.
7. `https://tuscriaturas.miraheze.org/wiki/Ayuda:Generar_im%C3%A1genes_con_VQGAN%2BCLIP/English`