Video Sentiment Analysis in the Wild

Boyang Tom Jin
Stanford University
tomjin@stanford.edu

Leila Abdelrahman*
University of Miami
lxa215@miami.edu

Cong Kevin Chen
Stanford University
ckchen95@stanford.edu

Amil Khanzada
Stanford University
amilkh@stanford.edu

Abstract

Determining the emotional sentiment of a video remains a challenging task that requires multi-modal, contextual understanding of a situation. In this paper, we describe our entry into the EmotiW 2020 Audio-Video Group Emotion Recognition Challenge to classify group videos containing large variations in language, people, and environment, into one of three sentiment classes. Our approach consisted of independently training models for different modalities, including a ResNet-LSTM end-to-end scene sentiment classification model, OpenPose multi-layer perceptron pose model, a mel-spectrogram CNN-LSTM-based audio sentiment classification model, and an FER-based facial pipeline network. Fully-connected fusion ensembling was used to combine the modalities to achieve the best validation accuracy of 63.6%, which was 11.5 percentage points higher than that of the baseline ensemble. The audio modality had the best independent classifier performance with an F1-score of 57.7%, while the scene modality was most important in the ensemble, increasing the ensemble F1-score by 3.1 percentage points.

1. Introduction

The human ability to perceive the emotional context of a situation is an essential aspect in understanding how the world operates. As such, sentiment recognition has an important role in society, especially in interpersonal interactions. Specifically, emotions can play a significant role in social experiences such as relationships, business decisions, and learning.

The prospect of using artificial intelligence to automatically detect and analyze emotions has recently become more popular, due to applications in fields such as surveillance, healthcare, and robotics. For instance, deep CNN models have been applied on facial data to automatically detect forms of depression, which can help reduce missed clinical diagnoses [1].

*not enrolled in CS231n

Moreover, having emotional awareness is key to understanding group dynamics and social cohesion. While early affect research was primarily focused on individuals, there has been a push in recent years for analysis of the "group affect" within natural (i.e., "in the wild") settings. [2]. Analysis performed on these real-world situations offers far more practical benefits than those done in controlled settings, and translates well to real-world settings such as videos captured on mobile phones.

In this paper, we describe our entry into the EmotiW 2020 Audio-Video Group Emotion Recognition grand challenge, which attempts to classify "in the wild" group videos into one of three emotion categories: positive, neutral, or negative.

This dataset is challenging. The videos come from a diverse set of real-world situations where subjects can be found in entirely different environments, lighting conditions, and video quality. While a traditional approach involves independently analyzing each face within the video [3], this does not sufficiently capture the social context within the scene, which is dependent on a variety of contextual clues such as the collective interactions between individuals. A person crying at a wedding may contribute different sentiments than a person crying after a fall. Even local factors, such as the relative positions of people in videos or level of face occlusion, can influence the perception of the mood of the video [4].

To address the complexity of the task, we use top-down and bottom-up approaches to train and ensemble multiple independent models along the following tracks: (1) **Scene:** Extend pre-trained CNN models with LSTM layers to perform full-frame video sentiment analysis. (2) **Pose:** Train CNN-LSTM models on extracted pose keypoints. (3) **Audio:** Train CNN-LSTM models on extracted audio features and spectrograms to detect sentiment patterns. (4) **Facial Expression:** Extend pre-trained CNN models with LSTM layers to classify emotions of extracted faces.

Preprocessing is used to extract relevant features for each modality, such as the relative face locations within each frame. The outputs of the models for each modality are ensembled together to perform a final sentiment prediction.

Our **baseline** for comparison is the highest-reported accuracy for group-level video sentiment detection by Sharma *et al.* [5]. The authors reported validation accuracies of a transfer learning Inception-V3 approach to sentiment classification of 52.1% with video alone and 50.2% with both audio and video.

2. Related Works

While affective computing and sentiment analysis originated in the field of NLP, real-time affect prediction on multi-modal videos is a relatively new and challenging field. Because videos are high-dimensional in nature, there are many feature modalities such as audio, object detection, pose analysis, and facial expression which can be leveraged for powerful classifiers. Moreover, unlike text and image processing, video analysis requires heavy computational resources to achieve meaningful results.

By far, one of the most studied forms of sentiment analysis in images is facial expression recognition (FER). Goodfellow *et al.* [6] was one of the first to implement deep learning for emotion recognition on the FER2013 dataset, which includes facial images from the wild, similar to those in the EmotiW dataset. Zhang *et al.* [7] showed how statistical methods such as wavelet entropy and fuzzy support vector machines could attain about 75% accuracy on the emotion recognition task. Further work by Khanzada *et al.* in 2020 demonstrated that current state-of-art accuracies of 75.8% could be obtained by ensembling five-layer CNNs and ResNet-based transfer learning networks [8].

Audio signal processing has been studied extensively and has yielded insights into pure acoustic and multi-modal affect detection. The openSMILE library [9] was developed to extract low-level acoustic features from raw waveforms and was used by Poria *et al.* in 2017 for classifying sentiment in the MOSI dataset [10]. The authors reported the success of Long-Short Term Memory Recurrent Networks (LSTM-RNNs), specifically bi-directional LSTMs in video sentiment analysis. Individual modalities such as audio, video, and textual transcriptions were assembled together through concatenation and fed into a final recurrent fusion layer for multi-modal classification. These methods yielded an overall accuracy of about 80%. Deep learning methods to classify audio sentiment include SoundNET [11] and VGGish [12]. As of 2019, the highest performing model for audio classification is the OpenL3 embedding model devised by Cramer *et al.* in 2019 [13].

Emotions are also heavily expressed through posture and kinesthetics. In 2017, Hussain *et al.* [14] highlighted how poselets could be extracted for unimodal sentiment analysis of group images. Guo *et al.* [15] showed how the distribution of skeletal axial points in images was a strong determining factor of overall sentiment.

Building on this challenging work, Majumdar *et al.* [16] showed in 2019 how variational autoencoders (VAEs) could be used to extract latent representations of raw features that were then fed into a classifier. Most recently, in 2020 Wang *et al.* [17] demonstrated how features such as audio and text or text and video could be combined together before being input into an attention-based transformer.

3. Datasets

3.1. EmotiW Dataset

The primary dataset used for our training and validation is from the Group Emotion Recognition subtask of the 2020 EmotiW Challenge. Sharma, Ghosh, and Dhall [18] compiled 1,004 YouTube videos of diverse everyday human interactions, including, but not limited to: protests, wars, birthday parties, sporting events, award ceremonies, and talk shows. The videos were recorded in various languages including English, Chinese, Arabic, Russian, and others. Since our model does not depend on verbatim transcriptions, it is language agnostic and serves to highlight the universal nature of how emotions manifest in human interactions.

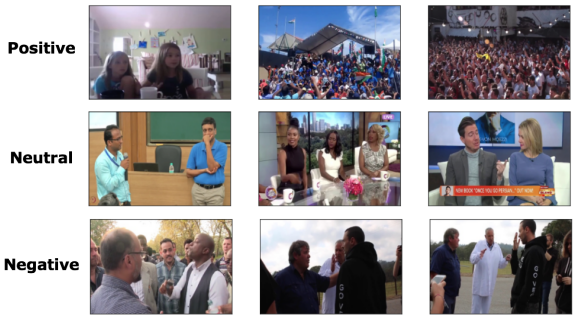


Figure 1: **Sample Dataset Clips** Video clips from each of the three sentiment categories

A subset of these compiled videos was split into 5-second video clips for the purposes of the EmotiW contest, where they were pre-divided amongst training (n=2661) and validation (n=766) groups. Each sliced video clip received a label corresponding to the general sentiment conveyed: positive, neutral, or negative. Clips had slight variation in frame rates, but generally consisted of 125-150 frames. The training set class distribution was 30.1% positive, 34.7% neutral, and 35.2% negative, while the distribution of the validation set was 39.4% positive, 36.6% neutral, and 24.0% negative.

3.2. Flickr

To augment the training data, we also downloaded 10,000 images from Flickr using curated keywords that represented human actions associated with positive and negative sentiments, such as “protest” and “parade”. Only im-

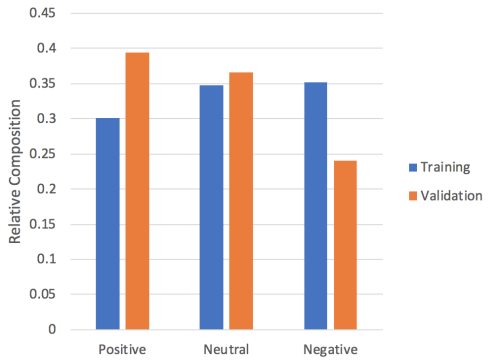


Figure 2: **Class Distribution** The training and validation datasets were not evenly distributed: the validation set had more positive and fewer negative samples.

ages containing people were kept and the images were resized to 480 x 320 pixels.

4. Preprocessing

4.1. Scene

Approximately 38,000 training frames and 11,000 validation still frames were obtained from the video dataset by extracting every tenth frame. Depending on the video frame rate, this corresponded to 2-3 frames per second or 10-15 frames for the entire video clip. Each frame was resized to 480 x 320 pixels. To standardize downstream training, we kept only the first 12 extracted frames of each video. Extraneous video frames were discarded and missing video frames were represented with zero-arrays.

4.2. Pose

Using the OpenPose library [19], we extracted 2D multi-person pose keypoints from the preprocessed scene frames. For each person, there were up to 25 keypoints available, each representing the spatial coordinates of a different facial or body part within the originating frame. Facial and foot keypoints were filtered out, as they were not consistently detected. With the 13 remaining body keypoints, normalization was performed to restrict keypoints between 0 and 1 while preserving relative distances between body keypoints. If a frame contained more than one person, the keypoints of all people in the frame were averaged. This data was finally augmented by the number of detected people within each frame to produce twelve 27-dimension embedding for each sample.

An average of 6.2 people was detected per frame. However, on average only 72% of available keypoints were extracted for each detected person. This was because people were often occluded within the video clips.

4.3. Audio

For audio preprocessing, we leveraged the OpenL3 model [20], which was pre-trained on millions of audio segments from the AudioSet database [21]. From the audio portion of each video, we extracted 6,144 embeddings with a hop size of 0.5, which corresponded to about 2 frames per second. Because all the video files were roughly 5 seconds in length, this translated to 11 embedding frames per video. Samples shorter than this cutoff were zero-padded and sets of frames were normalized on a per-sample basis. Finally, these audio-segments were converted to mel-spectrograms.

4.4. Facial Extraction

For input to our downstream facial emotion classification models, we extracted all faces from EmotiW's training and validation videos using the Face Recognition library [22], which employs a pre-trained dlib ResNet network with 29 convolutional layers [23]. This preprocessing extracted a total of 65,000 faces from the training videos.

5. Modalities

Each modality's models were independently trained and validated on the preprocessed EmotiW dataset to perform sentiment classification. The best performing model from each modality as determined by the lowest categorical cross-entropy loss on the validation dataset was then used for the final ensemble model training and validation.

5.1. Scene

The scene model chosen for the ensemble used a multi-headed approach. A ResNet-50 model, pre-trained on ImageNet, was applied directly on each extracted frame in a time-distributed approach. The output of this backbone was fed into three 2D convolutional-LSTM layers, each with 10 filters and a 3 x 3 kernel-size. The three separate outputs were concatenated together and fed back through a fully-connected layer with L2 weight regularization of 0.001 and a softmax classifier to perform the final sentiment prediction. The Adam optimizer was used with a learning rate of 0.01.

This model configuration was selected after performing a hyperparameter grid search on both the regularization factor and the output dimensionality of the conv-LSTM layer.

5.2. Pose

The pose model consisted of a bi-directional LSTM followed by a fully-connected layer and a softmax classifier. L2 regularization of magnitude 0.01 was applied on each layer.

A hyperparameter search was done on the regularization factor and the number of LSTM output units. We used the Adam optimizer with a learning rate of 0.01.

5.3. Audio

The audio model consisted of a CNN-bidirectional LSTM architecture. Three CNN layers, with 64, 512, and 512 feature map sizes respectively were used. This was then followed by two bi-directional LSTM layers.

Fine tuning of the learning rate hyper-parameter was performed using random search in the range of e-4 to e-6. The optimal learning rate was determined to be about 4e-5.

5.4. Facial

Recognizing faces as important features for classifying overall video sentiment and taking inspiration from Ghosh *et al.* [24], we designed a pipeline model architecture (see Figure 3) to account for the predicted emotions of all faces extracted from the video frames.

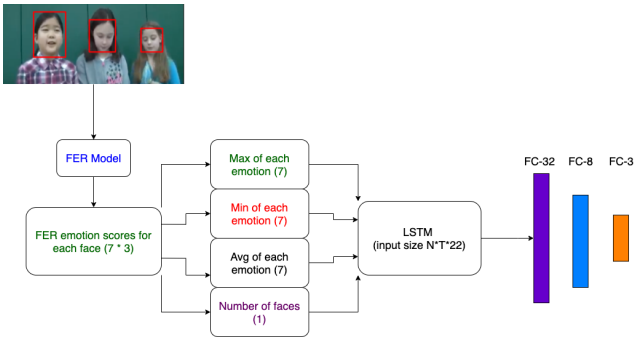


Figure 3: **Facial Mode Pipeline** Extracted faces are run through FER models before being fed to a sentiment classification network.

For the base model, all extracted faces larger than 40 x 40 pixels were resized to 48 x 48 pixels in grayscale and run through a Xception-based FER network [25]. The output consisted of seven predicted emotion probabilities, one for each face, which resulted in a matrix of shape $(F, 7)$, where F is the total number of faces in the given frame.

To standardize the embedding dimensions across videos, we applied the min, max, and average operations over each of the seven emotion classes on the F faces in each frame. We also included F as an additional feature, noting its importance in representing the nature of the interaction depicted.

The resulting 22 latent features for each frame were then passed through a bidirectional LSTM layer with 32 units, followed by a fully connected layer with 8 neurons and ReLU activation, and finally followed by a fully connected layer with 3 neurons and a softmax classifier. For training, Adam optimizer was used with a learning rate of 0.005.

6. Ensembling

Our best ensemble model took a fully-connected early fusion ensembling approach. Individual models were chopped at selected layers and had their exposed hidden layers concatenated together and fed into a multi-layer perceptron with two fully-connected layers, the first of which used a ReLU non-linearity and the second a softmax classifier. The hidden layers selected for concatenation were often the bidirectional LSTM layers, as all modalities had a time-dependent component.

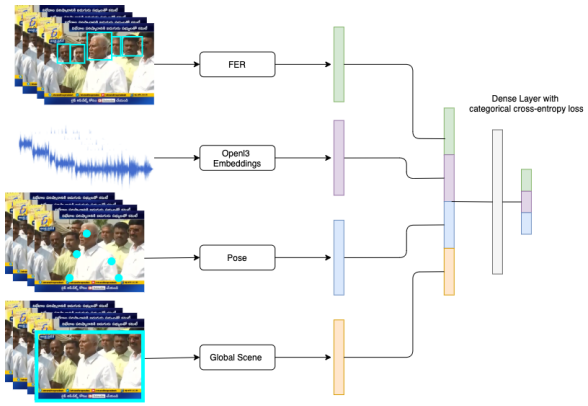


Figure 4: **Proposed Early Fusion Model** Our model adds a non-linear dense classifier that is not present in traditional late-fusion voting methods.

Figure 4 displays how layers were extracted early from individual modes and concatenated before input into a non-linear dense network.

7. Experiments

7.1. Individual Modalities

The following experiments were completed to evaluate the performance of different uni-modal models prior to selecting models to use in ensembling. Unless noted otherwise, the Adam optimizer along with the categorical cross-entropy loss function were used for all experiments.

7.1.1 Scene

- **Using Different Base Models**: We considered three base models: ResNet-50, VGG-19, and Inception-v3. The best Inception-v3 model had a validation accuracy of 42.0%, which was 15% lower than a comparable ResNet-50 model. The best VGG-19 model had a validation accuracy of 57.9%, which was similar to that of ResNet-50. However, the categorical cross-entropy validation loss of VGG-19 was significantly worse which led us to select ResNet-50 for further experiments.

- **Increasing Output Dimensionality**: We experimented with increasing the output dimensionality of each convolutional-LSTM layer in units of 10. Increasing the output dimension from 20 to 70 resulted in an approximately 4% gain in validation accuracy using one head, although this was not observed when using three heads.
- **Adding Attention**: We tried incorporating self-attention augmented convolutional layers as described by Bello *et al.* [26]. Although the authors noted their success in ImageNet classification, this type of attention incurs a memory cost of $O((HW)^2 N_h)$ (where H, W, and N_h are the height, width, and number of attention heads respectively) which became prohibitively expensive for our video data.
- **Increasing Number of Heads**: Inspired by the multi-head attention used in Transformers, we experimented with using between one and three convolutional-LSTM heads after the ResNet-50 layer. The number of heads did not seem to have an effect on validation accuracy, but the lowest validation loss achieved was with three heads.
- **Flickr-Augmented Data**: We attempted to first fine-tune a ResNet-50 model on our manually-curated Flickr dataset prior using it as part of a ResNet-LSTM model on the EmotiW dataset. The best validation accuracy achieved was 52.6%, which was approximately 5% lower than the best validation accuracy achieved with a standard ResNet-50 model trained on ImageNet weights.

7.1.2 Pose

- **Varying Output Dimensionality**: A hyperparameter search on the output dimensionality of the convolutional-LSTM layer was briefly performed. The best performance was observed with an output dimension of 64.
- **Keypoint-Frame Early Fusion**: The pose data was incorporated directly into each frame as the fourth channel. The first three channels were fed through a standard ResNet-LSTM model. The fourth channel was fed through a two-layer CNN, with a batch normalization layer in between. This was then concatenated with the ResNet-LSTM output. This resulted in a best validation accuracy of 56.9%, comparable to the best scene models.

7.1.3 Audio

- **Uni-directional LSTM**: We first tried a unidirectional LSTM on the audio, but found that it did not perform as accurately as our bi-directional LSTMs. This was because bidirectional recurrent models provide more context, looking both at past and future time points with respect to the present input.
- **Varying the Learning Rate**: Random search has been shown to be optimal for hyperparameter optimization as compared to grid search. Thus, a normal distribution of rates between 1e-4 and 1e-6 was generated and learning rates within this distribution were chosen. The final initial learning rate as part of an exponential learning rate schedule was determined to be 4e-5.
- **Different Feature Embeddings**: In addition to OpenL3, we experimented with the openSMILE library to extract over 300 features [9]. Unfortunately, it was challenging to parallelize OpenSMILE, and the features did not lead to improved model performance. VGGish embeddings were also extracted, and while these features did lead to expected performance improvements credited to transfer learning, preprocessing was still sequential and slow. The best performing VGGish model achieved an accuracy of 51% [12]. The OpenL3 embeddings outperformed both aforementioned feature sets.

7.1.4 Facial

- **Varying Pre-Trained FER Models**: We chose the Xception model to run FER and extract latent facial features because it is lightweight yet still achieves a decent classification accuracy of 65% on the FER2013 dataset. We experimented with various other FER models, including state-of-the-art transfer learning models trained by Khanzada *et al.* [8]. However, they proved to be too computationally intensive, as they were much deeper and also required resizing input images to 224 x 224 pixels in RGB. It is worth noting here that color is not as relevant for FER, as facial emotions can be distinguished from facial structure alone. [8].
- **Increasing Extraction Frame-Rate**: Instead of sampling every 10th frame, we tried decreasing the step size to sample every 3rd or 5th frame. Although this provided more data for our networks to work with, it could not achieve significant classification improvements.
- **Removing Small Faces**: We found that faces with very low resolutions were often misclassified and also not relevant to the overall sentiment of the frame. To resolve this, we modified our pipeline to ignore faces smaller than 40 x 40 pixels. This also significantly sped up preprocessing and improved all classification metrics.

- **Varying FER Model Depth**: Recognizing our pipeline model as a latent facial feature extractor, moreso than an emotion classifier, we also experimented with removing the last few layers of our pretrained FER models, obtaining 252 facial features as opposed to just 7. However, these experiments did not significantly improve our overall model performance.

- **Varying Facial Model Architecture**: To address overfitting on the latter part of our pipeline, we experimented with using both one-way and bidirectional GRUs and vanilla RNNs and varied the number of fully-connected layers. These techniques did not yield improvements; instead, our experimentation showed that the best models had dense layers with fewer units.

7.2. Ensembling

In addition to the fully-connected ensembling described previously, we also tried various other approaches to ensemble the different modalities.

7.2.1 Late Fusion

Baseline late fusion voting decision classifiers were initially used to join the softmax predictions of each of the independent modalities. The confidence probabilities were either averaged (soft voting), majority selected (hard voting), or weighted across all modalities (where the scene modality output was given more weight, by virtue of independently outperforming the pose predictor).

7.2.2 Early Fusion

We also researched early fusion methods, which combined early hidden layers of the models. These alternative approaches added more complexity, but allowed us to contextualize different modalities with each other. Early fusion followed by fully-connected layers was demonstrated to be the best performing ensembling method.

7.2.3 Variational Autoencoder

VAE is a type of generative, unsupervised autoencoder that takes in features as inputs and compresses the input to a lower dimension latent space. It uses a reconstruction loss that optimizes for how well the model is able to regenerate the input features. Thus, the model learns to make optimal encodings that represent the input at lower dimensions.

The same concatenated hidden layers used by the fully-connected ensembling approach were fed as input into a VAE. This generated latent z encodings of reduced dimensions. We experimented with the number of encoded dimensions, ultimately settling on 20. To measure encoding performance, the sum of the reconstruction loss and the KL-sampling loss was used as the final cost function. We then fed these z encodings into a fully-connected classifier.

8. Results

8.1. Quantitative Metrics

In evaluating our models, we considered the accuracy, precision, recall, and the combined F1-score of each of the individual modalities and ensemble model on the validation videos in the EmotiW dataset.

Rather than focusing on a primary metric, we considered all measures in parallel. Because our models and methods can be applied to a variety of real-world contexts, each requiring different metrics, it was important for us to deeply understand their performance.

8.2. Independent Classifier Performance

Model	Precision	Recall	F1 Score
Scene	.547	.540	.541
Pose	.483	.512	.489
Audio	**.640**	**.580**	**.577**
FER	.396	.403	.348

Table 1: **Modality Comparison** Summary of the best model performance on the validation dataset for each modality.

8.2.1 Scene

When compared to other uni-modal models, the independent scene-based classifier excelled at predicting true positives and was the only model with a precision rate greater than 50% for that category as shown in Figure 8A. The scene model was also fairly unbiased and robust, as it was not skewed towards any one sentiment class. Nevertheless, it still had difficultly distinguishing neutral labels from positive and negative labels.

Despite being trained on full-frames without any context, the model appeared to focus predominantly on people as shown in Figure 5. Specifically, faces and hands were the areas that the model consistently leveraged in its predictions. Through our use of time-series data, we found the model would actually focus on different areas between adjacent frames, even if the frames contained similar content.

Unlike the facial modality, the scene model was invariant to the orientation of the body and face which meant it could use features from the scene even if the person had their back to the camera or parts of the body were occluded.

Note that the ResNet-50 base model tended to focus on specific individuals within a scene, even if there were other

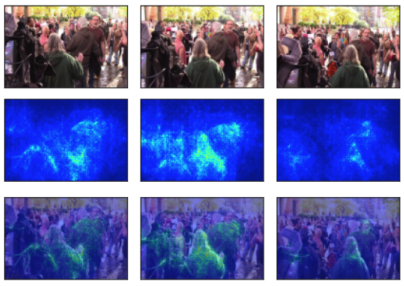


Figure 5: **ResNet-LSTM Saliency** Three sequential sample frames overlayed with a saliency map of the ResNet-LSTM model performed on a negative sentiment video clip.

people. In comparison, the Inception-v3 model showed strong feedback on multiple areas of the scene, not limited to people. For instance, Figure 11 shows strong backpropagation by Inception-v3 on the overhead lights and background signage in addition to multiple individuals in the crowd. In comparison, ResNet-50 focused predominantly on the foreground individuals.

8.2.2 Pose

Similar to the audio model, the pose model was fairly robust at predicting true negative sentiments but poor at predicting true positive sentiments. This model also in general had a lower validation accuracy - although this was not surprising given that there was great variation in the number of key points extracted within each video.

The pose model focused predominantly on upper body joints, specifically the hands and elbows as shown in Figure 6.

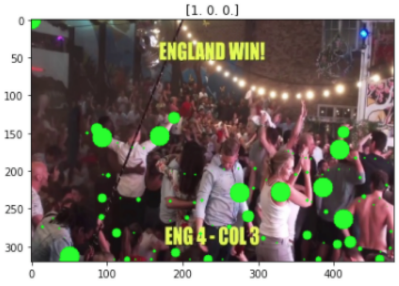


Figure 6: **Pose Saliency** Saliency map with respect to the pose keypoints of a positive sentiment video that was correctly classified. Each dot presents a particular keypoint, with larger dots indicating areas where the pose model had more focus.

8.2.3 Audio

In studying how the model performed, the model was very good at discriminating neutral and negative videos, but only achieved random-guess accuracy for positive sentiment, as shown in Figure 7. Moreover, when observing the features that the model was looking at, the model often made errors when there were regions of no sound. The model often misidentified positive spectrograms and later embeddings as neutral.

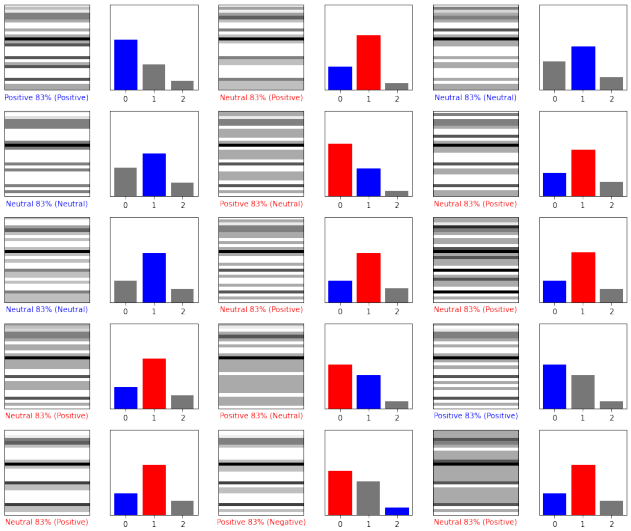


Figure 7: **Audio Error Analysis** The 2D-transformed embedding layers are juxtaposed with the classifier's softmax confidence. Red indicates misclassifications. Blue indicates correct classification. In plain text is the model prediction and in parenthesis is the true label. The model often made errors when there were samples with gaps with no audio. It had a pattern of confusing positive with neutral sentiment.

8.2.4 Facial

Interpreting results for the FER models we experimented with was not straightforward, as videos in the EmotiW dataset only provided overall sentiment labels and not emotion labels for individual faces. Through manual analysis of cropped faces and their predicted emotions, we found that our heavier, transfer learning based FER models seemed to perform best. However, we saw accuracy drop significantly when including too many faces with resolutions too low for proper classification (below 40 x 40).

Our facial network pipeline classified an overwhelming majority of videos as neutral, resulting in a 76% recall rate for those videos. However, it was only able to recall 10% of videos in the positive category, which was also the most prevalent type in the validation set. The facial network

struggled as a stand-alone modality and also did not significantly enhance the performance of the final ensemble network, although its inclusion did improve the recall rate of negative videos.

8.3. Ensembling Performance

Ensemble	Precision	Recall	F1 Score
Hard	.605	.581	.576
Soft	**.658**	.620	.608
Weighted	.632	.602	.593
FC-Fusion	.639	**.629**	**.626**
VAE	.271	.369	.276

Table 2: **Ensemble Comparisons** Performance on the validation dataset for different ensembling strategies. Each ensemble strategy used the same model outputs from all four modalities.

Our early-fusion technique yielded the best results with an F1 score of 0.626. Early fusion outperformed late-fusion methods because it contextualized features earlier and allowed the model to learn and backpropagate on concatenations of different modalities. Thus, the feature spaces were interdependent, allowing the model to learn how combinations of global scene, human pose, audio, and facial expressions all contribute to the sentiment of a video.

8.3.1 VAE

Independent modality hidden layers were fused and then fed into the encoder. The 20 z latent features were extracted from the encoder after 200 epochs of training and mapped into 2 dimensional space by the t-SNE [27] clustering method. The results of this unsupervised portion showed a distinct overlap between positive and neutral samples as illustrated in Figure 9.

8.4. Ablation Study

After determining that the fully-connected early fusion model performed the best, we individually removed each modality to determine which ones impacted classification performance the most as shown in Table 3. We also generated confusion matrices to display how modal ablations influenced model decision making, as shown in Figure 8B.

This study also revealed the best performing model which used only the audio, scene, and pose modalities to achieve a validation accuracy of 63.6%.

9. Discussion

Although the scene model was one of the strongest models, it still only achieved a validation accuracy of 54.6%.

Ensemble	Precision	Recall	F1 Score
Aud-Sce-Pose	**.641**	**.632**	**.630**
Sce-FER-Pose	.632	.629	.627
Sce-Aud-FER	.641	.629	.627
Aud-Pose-FER	.627	.602	.595

Table 3: **Ablation Comparisons** Performance on the validation dataset after removing different modalities from the fully-connected early-fusion ensemble.

Figure 10a shows a case where a positive video was misclassified as negative. Here, the model focused on the raised hand position of one of the individuals. This misclassification could have been unduly influenced by the negative sentiment training videos showcasing the raised hands of protestors. In another example, Figure 10b shows a negative video misclassified as positive. In this case, although the video mostly featured a woman talking, the negative sentiment is due to a brief scene in the beginning of a girl crying. Due to the way that the video was sampled, most frames of this particular video clip did not include the girl, which may have misled the model.

We initially believed that augmenting our training dataset with Flickr photos would improve the overall performance of our scene models. However, the fine-tuning performed on this dataset actually reduced performance on the validation set. One possible reason for this is that the curated dataset used images rather than videos which meant that transfer learning had to occur twice, once to train on the Flickr dataset and another time on the EmotiW data. This likely caused the fine-tuned ResNet model to overfit during the first transfer learning and led to poorer performance on the second round of training.

For the scene modality, we used the ResNet-50 model as the base model because our experiments showed that this model had the lowest validation loss. This was in contrast to the baseline paper, which used an Inception-v3 model [5]. This might be one of the reasons why our model improves upon the baseline paper; we found that using Inception-v3 led to a 15% drop in performance. This gap can be explained by Figure 11, which shows that the Inception-v3 model activated strongly on multiple areas of the input image and was unable to focus on the important sentiment discriminators of a scene like the ResNet-50 model could. In this successfully classified example, the ResNet-50 model activated predominantly on the person with their arms up in celebration, which was a more important discriminator than the background signage or the lighting elements.

The performance of the pose model was limited to the expressiveness of the poses themselves, which explains the relatively lower F1 score of 0.489. Nevertheless, for many videos the pose model seemed to focus on hands and elbows, an example of which is shown in Figure 6. This

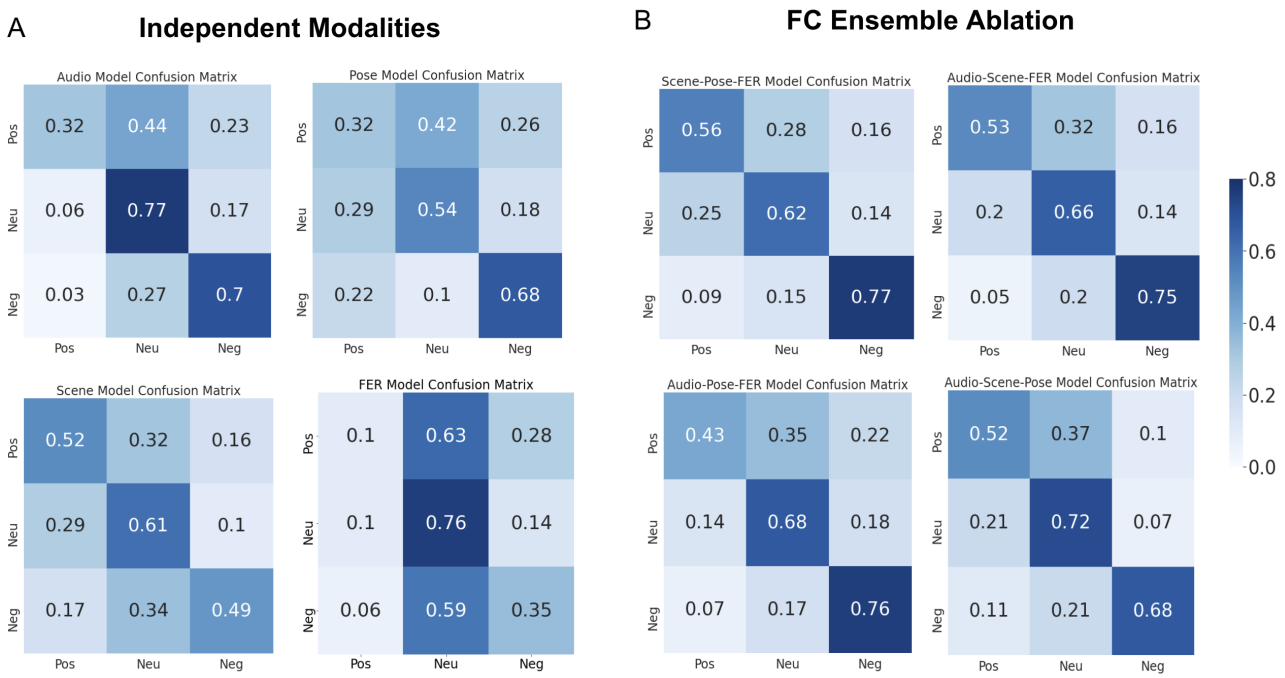


Figure 8: **Model Confusion Matrices**. Figure 8A displays the individual independent modality confusion with digits representing the accuracy per class. Figure 8B shows the results of ablating one modality from the fully connected ensemble, which had the overall best performance.

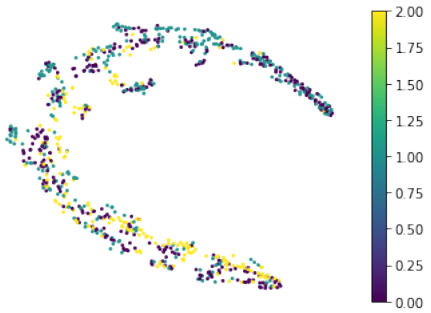


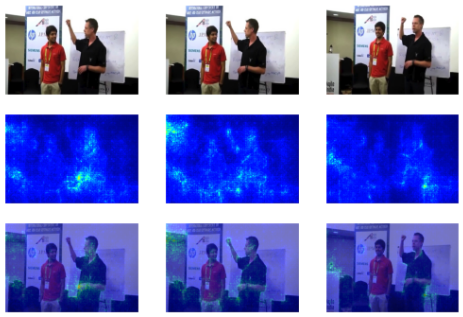
Figure 9: **VAE 2D t-SNE** In the latent space, there is a clear separation between negative samples shown in yellow and the positive and neutral samples as shown in green and purple, respectively. Positive and neutral overlapped in scope, which explains how our models often confused positive samples for neutral.

makes sense since the position of arms are an important indicator of emotions, with raised arms and protruded arms traditionally indicating joy and fear respectively [28].

The audio modality performed the strongest, achieving an independent F1 score of 0.577. Although the EmotiW dataset is multilingual and scenes contain varying numbers of people, there are universal aspects of language and audio that correlate with sentiment. In this case, although the OpenL3 model used for extracting features was pre-trained primarily on music files, it performed well, as both music and human speech have prosody, or patterns of stress and intonation. Through transfer learning, we could translate the hidden features extracted from millions of training samples and fine-tune them on the EmotiW dataset.

Despite our extensive experimentation guided by seemingly reasonable assumptions, the facial modality had the lowest performance. A historically challenging problem, facial recognition in videos poses many roadblocks not present in static images. Micro facial variations, inconsistent video quality, and transition between emotions lead to highly variable predictions throughout videos, even for the same faces. Although we attempted to overcome these issues with temporal LSTMs, motion in videos made it tricky to correlate faces between frames. To complicate matters further, extraction libraries do not capture all faces; in our case, the Face Recognition library performed poorly on obscured, blurred, and slanted faces, which were very common in the EmotiW dataset.

Additionally, people often do not express their true emotions on their faces and emotions are expressed differently across cultures. These are difficult problems in our context, but potentially solvable with higher resolution faces. The



(a) Scene model where the predicted sentiment is negative and the actual sentiment is positive.

(b) Scene model where the predicted sentiment is positive and the actual sentiment is negative.

Figure 10: **Confusion Saliency Maps** The model made mistakes and paid attention to different aspects of the image, like the position of limbs and faces when confusing positive and negative sentiment.

correlation between facial emotion and overall video sentiment is further complicated by situations such as bullying, where positive facial emotions on more visible faces may not necessarily correlate with overall positive video sentiment.

In each modality, the model had difficulty distinguishing neutral sentiment from positive and negative sentiment. This is a challenging endeavor as the distinction between a neutral sentiment and a positive or negative sentiment is often a matter of context and human perception and bias. Many of the videos were from news clips, which naturally gravitate towards negative sentiment [29]. Additionally, five-second clips extracted from different parts of longer videos, were often labeled differently, despite having the same setting with the same people.

For these reasons, it is not surprising that our fully-connected late-fusion ensemble of modalities achieved the best validation accuracy: sentiment understanding is best provided through a combination of multiple modalities, just as a human might interpret a scene. This way, some modalities may compensate for the shortcomings of others. Our confusion matrices summarized this point. Removing the

Figure 11: **ResNet-50 vs Inception-v3** Saliency map comparisons between different base models used in the scene model.

scene classifier led to a 0.09 drop in the recall rate of positive samples. Similarly, when FER was removed, a 0.08 drop was observed in the recall rate of negative samples. Our fusion method is truly greater than the sum of its parts.

10. Conclusion

Video sentiment classification is a complex problem under active study, requiring investigation of various modalities and powerful computation. Running our models on low-cost cloud environments, we often faced RAM and disk space overflow, along with training runtimes of many hours. These challenges forced us to simplify our preprocessing techniques and models, while also taking advantage of multiple machines in parallel, threading, and GPU-enabled techniques.

Despite the many challenges we faced, our final late-fusion ensemble model of scene, pose, and audio modalities outperformed the baseline accuracy of 52.1% [5], achieving an overall validation accuracy of **63.6%**.

11. Future Work

For next steps, we intend to further refine and submit our work to the EmotiW Challenge, which will release a test dataset on June 15th, 2020. Given more time and computational resources, there are still many avenues left for us to explore.

For improving the facial pipeline, we intend to try approaches from other papers, experiment with different facial extraction libraries, retrain our FER models to better match the EmotiW dataset, leverage deeper FER models, integrate the relative distances between locations of faces, further tune hyper-parameters such as facial extraction frequency, and integrate models trained on facial emotion transition datasets such as the Extended Cohn-Kanade Dataset (CK+) [30].

We also intend to further explore our preliminary work on object sentiment classification using objects detected with YOLOv3 [31] and classified by sentiment analysis on their textual labels. Optimizing the VAE to tune the encoding number, as well as the quality of reconstruction is another area for research.

To improve ensembling, we propose to explore using hierarchical models, which combine early and late stage fusion and have been shown to produce promising results [32]. Boosting is also another good option to improve weaker classifiers within our ensemble.

After this, we hope to apply our findings to related research problems and contribute to the Affective Computing niche, potentially submitting our project to peer-reviewed conferences and journals. Applications ranging from video classification on social media (e.g. YouTube and Facebook) and privacy cameras to self-driving cars and psychiatric evaluations are all potential avenues for meaningful work. We have investigated several open video datasets, in particular the *Trending YouTube Video Statistics* dataset on Kaggle [33], which includes labels of video title, channel title, publish time, tags, views, likes and dislikes, description, and comment count. We believe our broad range of models can be applied, with modifications to all of these real world problems.

12. Code

We have open sourced our work under the MIT license for the benefit of academia. It can be accessed on GitHub: `https://github.com/kevincong95/cs231n-emotiw`

13. Contributions

Tom Jin

- Performed face, frame, and pose extraction.
- Developed scene and pose classifiers, including model experiments and error analysis.
- Created fully-connected fusion ensemble.

Leila Abdelrahman

- Developed audio classifier model.
- Supported development of facial models.
- Led work on all fusion methods, including baseline ensemble, VAE late fusion, and ablation studies.

Cong Chen

- Managed EmotiW competition procedures, including dataset collection and historical research.
- Developed facial sentiment model and latent feature extraction.

Amil Khanzada

- Managed project and logistics, defining overall direction and future work.
- Developed facial sentiment models, while leading integration and experimentation of FER models.
- Supported development of audio and ensemble models.

14. Acknowledgements

We are very grateful to Dr. Fei-Fei Li and the teaching staff in CS231n for empowering us with the knowledge necessary to execute this complex project over the past two months. We appreciate Professor Ranjay Krishna and Christina Yuan for their mentorship and guidance throughout the quarter, giving us key pointers to be successful not only in this project, but in future endeavors. We would also like to thank Vincent La, who helped us explore using YOLOv3 to perform object detection and text-based sentiment analysis.

14.1. Preprocessing

We leveraged the Face Recognition library for facial extraction [22]. To extract relevant features for audio analysis, we worked extensively with the OpenL3 library [34]. To extract poses from video frames, we used the OpenPose library [19].

14.2. Models

For development of our FER models, we leveraged techniques and open source code from Khanzada *et al.* [8]. We also took advantage of a pre-trained FER model from Arriaga *et al.*'s Face Classification repository [35]. For our VAEs models, we adapted methods from Kingma and Welling [36].

14.3. Error Analysis

For saliency maps, we augmented base code from the keras-vis package from Kotikalapudi *et al.* [37] and the tf-keras-vis package from Keisen *et al.* [38] to work with LSTMs.

References

[1] Xiuzhuang Zhou, Kai Jin, Yuanyuan Shang, and Guodong Guo. Visually interpretable representation learning for depression recognition from facial images. *IEEE Transactions on Affective Computing*, 2018.

[2] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon. From individual to group-level emotion recognition: Emotiw 5.0. In *Proceedings of the 19th ACM international conference on multimodal interaction*, pages 524–528, 2017.

[3] Andrew C Gallagher and Tsuhan Chen. Understanding images of groups of people. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 256–263. IEEE, 2009.

[4] Abhinav Dhall, Roland Goecke, and Tom Gedeon. Automatic group happiness intensity analysis. *IEEE Transactions on Affective Computing*, 6(1):13–26, 2015.

[5] Garima Sharma, Shreya Ghosh, and Abhinav Dhall. Automatic group level affect and cohesion prediction in videos. In Nadia Bianchi-Berthouze, Julien Epps, Andrea Kleinsmith, and Picard Rosalind , editors, *International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) 2019*, pages 161–167, United States of America, 2019. IEEE, Institute of Electrical and Electronics Engineers. International Conference on Affective Computing and Intelligent Interaction Workshops and Demos 2019, ACIIW 2019 ; Conference date: 03-09-2019 Through 06-09-2019.

[6] Ian Goodfellow, Dumitru Erhan, Pierre Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, and Y. Bengio. Challenges in representation learning: A report on three machine learning contests. *Neural Networks*, 64, 07 2013.

[7] Y. Zhang, Z. Yang, H. Lu, X. Zhou, P. Phillips, Q. Liu, and S. Wang. Facial emotion recognition based on biorthogonal wavelet entropy, fuzzy support vector machine, and stratified cross validation. *IEEE Access*, 4:8375–8385, 2016.

[8] Amil Khanzada, Charles Bai, and Ferhat Turker Celepcikay. Facial expression recognition with deep learning, 2020.

[9] audEERING. Opensmile. `https://www.audeering.com/opensmile/`, 2020.

[10] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *ArXiv*, abs/1606.06259, 2016.

[11] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 892–900, Red Hook, NY, USA, 2016. Curran Associates Inc.

[12] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.

[13] J. Cramer, H-.H. Wu, J. Salamon, and J. P. Bello. Look, listen and learn more: Design choices for deep audio embeddings. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3852–3856, Brighton, UK, May 2019.

[14] Zeshan Hussain, Tariq Patanam, and Hardie Cate. Group visual sentiment analysis. *ArXiv*, abs/1701.01885, 2017.

[15] Xin Guo, Luisa Polania, and Kenneth Barner. Group-level emotion recognition using deep models on image scene, faces, and skeletons. pages 603–608, 11 2017.

[16] Navonil Majumder, Soujanya Poria, Gangeshwar Krishnamurthy, Niyati Chhaya, Rada Mihalcea, and Alexander Gelbukh. Variational fusion for multimodal sentiment analysis. *ArXiv*, abs/1908.06008, 2019.

[17] Zilong Wang, Zhaohong Wan, and Xiaojun Wan. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of The Web Conference 2020*, WWW '20, page 2514–2520, New York, NY, USA, 2020. Association for Computing Machinery.

[18] Roland Goecke Abhinav Dhall, Garima Sharma and Tom Gedeon. Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. ACM International Conference on Multimodal Interaction 2020, 2020.

[19] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d

pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[20] Jason Cramer, Ho Hsiang Wu, Justin Salamon, and Juan Pablo Bello. Look, listen, and learn more: Design choices for deep audio embeddings. In *2019 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2019 - Proceedings*, ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, pages 3852–3856. Institute of Electrical and Electronics Engineers Inc., May 2019. 44th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2019 ; Conference date: 12-05-2019 Through 17-05-2019.

[21] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[22] Adam Geitgey. Face recognition. `https://github.com/ageitgey/face_recognition`, 2020.

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

[24] Shreya Ghosh, Abhinav Dhall, Nicu Sebe, and Tom Gedeon. Predicting group cohesiveness in images. 12 2018.

[25] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. Real-time convolutional neural networks for emotion and gender classification. *CoRR*, abs/1710.07557, 2017.

[26] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3286–3295, 2019.

[27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[28] Daniel Bernhardt. *Emotion inference from human body motion*. PhD thesis, Citeseer, 2010.

[29] Ann Devitt and Khurshid Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 984–991, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[30] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101, 2010.

[31] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[32] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[33] Mitchell J. Trending youtube video statistics, Jun 2019.

[34] Music and Audio Research Laboratory NYU. Face classification and detection. `https://github.com/marl/openl3`, 2019.

[35] Octavio Arriaga. Face classification and detection. `https://github.com/oarriaga/face_classification/tree/master/trained_models`, 2017.

[36] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

[37] Raghavendra Kotikalapudi and contributors. keras-vis. `https://github.com/raghakot/keras-vis`, 2017.

[38] Keisen. tf-keras-vis. `https://github.com/keisen/tf-keras-vis`, 2020.