Fusical: Multimodal Fusion for Video Sentiment

Boyang Tom Jin*
tomjin@stanford.edu
Stanford University

Leila Abdelrahman*
lxa215@miami.edu
University of Miami

Cong Kevin Chen
kevincong95@berkeley.edu
University of California, Berkeley

Amil Khanzada
amil@berkeley.edu
University of California, Berkeley

Abstract

Determining the emotional sentiment of a video remains a challenging task that requires multimodal, 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 end-to-end approach consists of independently training models for different modalities, including full-frame video scenes, human body keypoints, embeddings extracted from audio clips, and image-caption word embeddings. Novel combinations of modalities, such as laughter and image-captioning, and transfer learning are further developed. We use fully-connected (FC) fusion ensembling to aggregate the modalities, achieving a best test accuracy of 63.9% which is 16 percentage points higher than that of the baseline ensemble.

CCS Concepts

• **Computing methodologies** → **Ensemble methods**; **Scene understanding**; **Neural networks**; **Computer vision**; • **Human-centered computing** → **Human computer interaction (HCI)**;

Keywords

Multimodal Sentiment Classification; Neural Networks; Affective Computing; Laughter; Word Embeddings; Computer Vision; Pose; Facial; FER; Emotion; Ensemble; Image-Captioning; Fusion

ACM Reference Format:
Boyang Tom Jin, Leila Abdelrahman, Cong Kevin Chen, and Amil Khanzada. 2020. Fusical: Multimodal Fusion for Video Sentiment. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3382507.3417966

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 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 convolutional neural network (CNN) models have been applied on facial data to automatically detect forms of depression, which can help reduce missed clinical diagnoses [1]. Moreover, emotional awareness is key to understanding group dynamics and social cohesion. While early affective computing research focused on individuals, there is 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 better to real-world settings such as videos captured on mobile phones.

In this paper, we discuss our entry into the EmotiW 2020 Audio-Video Group Emotion Recognition grand challenge as part of the ACM Conference on Multimodal Interaction. Here, we attempt to classify "in the wild" group videos into one of three emotion categories: positive, neutral, or negative. The videos come from a diverse set of real-world situations where subjects can be found in entirely different environments, lighting conditions, and video qualities. 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, as well as objects in the scene. A person joyfully crying at a wedding invokes a different sentiment than a person painfully crying after an injury. 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: (i) **Scene:** transfer learning for full-frame sentiment analysis. (ii) **Pose:** CNN-LSTM model on extracted pose keypoints. (iii) **Audio:** CNN-LSTM model on extracted audio features and embeddings to detect sentiment patterns. (iv) **Laughter:** LSTM model transfer learning for laughter classification. (v) **Facial Expression:** CNN-LSTM model transfer learning to classify emotions of extracted faces. (vi) **Image Captioning:** GPT2-Transformer-based classifier on generated captions. We preprocess the videos to extract relevant features for each modality, such as the individual face locations within each frame. The outputs of the models for each modality are ensembled together to perform a final sentiment prediction.

*Both authors contributed equally to this research.

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3417966

Our **baseline** for comparison is the highest-reported accuracy for group-level video sentiment detection by Sharma et al. [5]. The authors report 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. The baseline accuracy on the test set is 47.9%.

2 Related Works

While affective computing and sentiment analysis originates from NLP, real-time affect prediction on multimodal videos is a relatively new and challenging task. Because videos are high-dimensional in nature, there are many feature modalities such as audio, pose analysis, and facial expressions which can be leveraged for powerful classifiers. Moreover, unlike text and image processing, video analysis requires heavy computational resources to achieve meaningful results. Poria et al. [6] implemented computationally-expensive 3D convolutions in analyzing videos from the MOSI dataset [7]. For further scene analysis, image captioning methods with attention, such as those by Vinyals et al. [8] and Xu et al. [9], generate sentence captions for images, but these methods have yet to be applied for frame-level sentiment analysis in video.

By far, one of the most studied forms of sentiment analysis in images is facial expression recognition (FER): Goodfellow et al. [10] first implemented deep learning for emotion recognition on the FER2013 dataset, which includes facial images from the wild, similar to those in the EmotiW dataset. Moreover, Zhang et al. [11] showed how statistical methods such as wavelet entropy and fuzzy support vector machines attain about 75% accuracy on the emotion recognition task. Further work by Khanzada et al. [12] in 2020 obtained current state-of-art accuracies of 75.8% by ensembling five-layer CNNs and ResNet-based transfer learning networks.

Audio signal processing has been studied extensively and has yielded insights into pure acoustic and multimodal affect detection. audEERING© open-sourced the openSMILE library [13] to extract low-level acoustic features from raw waveforms; Poria et al. [6] used this library to classify sentiment in the MOSI dataset. The authors reported the success of Long-Short Term Memory Recurrent Neural Networks (LSTM-RNNs), specifically bi-directional LSTMs, for audio and video modes in video sentiment analysis. Deep learning methods to classify audio sentiment include SoundNET [14] and VGGish [15]. As of 2019, the highest performing model for audio classification is the OpenL3 embedding model devised by Cramer et al. [16]. Other aspects that relate to audio are specific signatures in the clip, such as laughter. Laughter detectors, such as those built by Ideo [17], generate binary output given a video input, but lack applications for sentiment analysis. Furthermore, emotions are heavily expressed through posture and kinesthetics: in 2017, Hussain et al. [18] highlighted how poselets could be extracted for unimodal sentiment analysis of group images. Guo et al. [19] reported how the skeletal axial point distribution in images was a strong determining factor for overall sentiment.

For ensemble classifiers and related fusion methods, Poria et al. [6] concatenated modalities such as audio, video, and textual transcriptions and fed these into a final recurrent fusion layer for multimodal classification. Lian et al. [20] combined transcriptions, facial expression, and audio for a beam-search fusion classifier. In another fusion approach, Hu et al. [21] implemented a supervised scoring ensemble for concatenating class-wise scoring activations for modalities, but this was limited to only audio and visual features. This was a hierarchical approach that used successive layers of supervised classifiers for a final prediction. In 2019, Majumdar et al. [22] showed how variational autoencoders (VAEs) extract latent representations of raw features that can then be fed into a classifier. We emphasize that most prior work is limited to audio, transcription, scene, and pose modalities, and there is a persistent lack of novel modality integration.

3 Datasets

3.1 EmotiW Dataset

We used the training and validation dataset from the Group Emotion Recognition subtask of the 2020 EmotiW Challenge. Sharma, Ghosh, and Dhall [23] compiled 1,004 five-second clips from YouTube videos of diverse everyday human interactions, spanning: protests, military actions, birthday parties, and talk shows. The videos were recorded in various languages including English, Chinese, and Arabic. Since our model is independent of verbatim transcriptions, it is language-agnostic and serves to highlight the universal nature of how emotions manifest in human interactions.

The videos were pre-divided amongst training (63%), validation (19%), and test (18%) groups. Each sliced video clip was labeled corresponding to the general sentiment conveyed: positive, neutral, or negative. Clips had slight variations 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.

4 Approach

4.1 Ensembling

Our best ensemble model took a fully-connected (FC) early fusion ensembling approach: individual models, frozen after training, were chopped at selected layers. Their exposed hidden layers were concatenated together and fed into a multi-layer perceptron with two FC layers. The first hidden layer was followed by a ReLU nonlinearity and the second by a softmax classifier. The hidden layers selected for concatenation were often the bidirectional LSTM layers, as all modalities had a time-dependent component. The FC component was then trained with a categorical cross-entropy loss function before yielding the final classification. Figure 1 describes in greater detail the extraction and concatenation before input into a non-linear dense network. We emphasize that our method introduces the implementation of two novel modalities for sentiment classification: laughter detection and image captioning.

Equation 1 illustrates the loss function we use on our final dense network. While each modality has its own categorical cross-entropy loss, we compute the final fusion's loss function using the concatenated hidden layers as input and the sentiment classification as output.

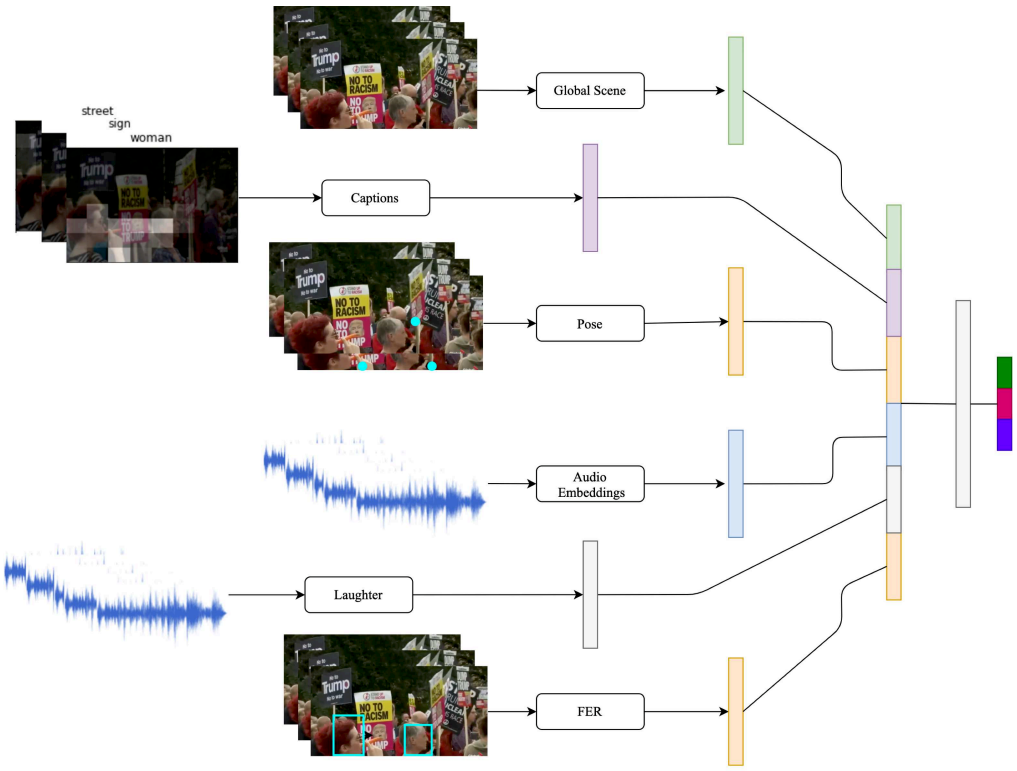

Figure 1: Our model adds a non-linear dense classifier that is absent in traditional late fusion voting methods. We extract hidden layers from the frozen, pretrained independent modalities and concatenate and insert them into a dense layer for further training and final predictions. The fusion includes a novel implementation of image captioning and laughter detection model embeddings for sentiment classification.

$$CCE(y, p) = -\sum_{i=1}^{N} y_i \log(p_i) \tag{1}$$

Here, N is the number of validation/test examples. y is a one-hot matrix of size $N \times C$ classes representing the target label of each example. p is a matrix of size N by C representing the predicted score of each class for each example. y_i and p_i are labels and scores corresponding to the ith example.

4.2 Preprocessing

4.2.1 Scene We obtained approximately 38,000 training frames and 11,000 validation still frames from the video dataset by extracting every 10th frame. Depending on the video frame rate, this corresponded to two to three frames per second (FPS) or 10-15 frames for the entire video clip. Each frame was resized with bilinear interpolation to 480×320 pixels. To standardize downstream training, we kept only the first 12 extracted frames of each video. We discarded extraneous and missing video frames, representing them as zero-arrays.

4.2.2 Pose Using the OpenPose library [24], we extracted 2D multi-person pose keypoints from the preprocessed 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. We excluded facial and foot keypoints, as they lacked consistent detection. The 13 remaining body keypoints were normalized between zero and one while preserving relative distances between them. If a frame contained more than one person, we averaged the keypoints of all people in the frame. We finally augmented this data by the number of detected people within each frame to produce twelve 27-dimensional embeddings for each sample, detecting an average of 6.2 people per frame. On average, we only extracted 72% of the available keypoints for each detected person, as people were often occluded within the video clips.

4.2.3 Audio To address preprocessing the audio, we relied on the OpenL3 model [25], which was pre-trained on millions of audio segments from the AudioSet database [26]. From the audio extracted from each video, we generated 6,144 embeddings with a hop size of 0.5, which corresponded to about 2 FPS. Because all the video files are roughly five seconds in length, this translates to 11 embedding frames per video. We zero-padded samples shorter than this cutoff and normalized sets of frames on a per-sample basis.

4.2.4 Laughter We used a VGGish model pretrained on the AudioSet data to extract 128-dimensional embeddings from each audio clip [27], exclusively for the laughter detection model.

4.2.5 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 [28], which employed a pre-trained dlib ResNet network with 29 convolutional layers [29]. This preprocessing extracted a total of 65,000 faces from the training videos.

4.2.6 Image Captioning We used an attention-based model, pre-trained for 20 million steps on Microsoft COCO [30], to generate caption sentences; we tokenized these sentences and converted them into 140-unit sized embeddings for input into an NLP-inspired sentiment classification model.

4.3 Modalities

To perform sentiment classification, we trained and validated each modality's models independently on the preprocessed EmotiW dataset. We determined the best performing model from each modality by the lowest categorical cross-entropy loss on the validation dataset, and used the best models for the final ensemble model training and validation.

4.3.1 Scene The scene model chosen for the ensemble had a multi-headed approach: the different heads provide redundancy, allowing the model to use different aspects of the subspace produced by the base model. We applied a ResNet-50 model, pre-trained on ImageNet, directly to each extracted frame in a time-distributed approach, and fed this backbone's output into three 2D convolutional-LSTM layers, each with 10 filters and a 3×3 kernel-size. The three separate outputs were concatenated together and fed back through a fully-connected layer with L_2 weight regularization of e^{-3} and a softmax classifier to perform the final sentiment prediction. We used the Adam optimizer with a learning rate of e^{-2}, and selected this model's configuration after performing a hyperparameter grid search on both the regularization factor and the output dimensionality of the conv-LSTM layer. Alternative models using the above architecture, but with VGG-19 and Inception-v3 as the base models, were also trained for comparison.

4.3.2 Pose The pose model consisted of a bi-directional LSTM followed by a dense layer and a softmax classifier. We applied L_2 regularization of magnitude e^{-2} on each layer, before performing a hyperparameter grid search on the regularization factor and the number of LSTM output units. We also used the Adam optimizer with a learning rate of e^{-2}.

4.3.3 Audio The audio model was comprised of a CNN bidirectional LSTM architecture: it has three CNN layers, with 64, 512, and 512 feature map sizes, respectively. This was then followed by two bi-directional LSTM layers. We fine tuned the learning rate hyper-parameter using random search in the range of e^{-4} to e^{-6}, and determined the optimal learning rate to be $4e^{-5}$.

4.3.4 Laughter To produce a single sigmoid output score [17], we applied an LSTM-based laughter detection model pre-trained using 10-second-long, re-balanced AudioSet laughter clips on the pre-extracted VGGish embeddings. These output scores were later used as input features for the ensemble layer. No further fine tuning was performed.

4.3.5 Facial Extraction Recognizing faces as important features for classifying overall video sentiment and taking inspiration from Ghosh et al. [31], we designed a pipeline model architecture to account for the predicted emotions of all faces extracted from the video frames. For the base model, all extracted faces larger than 40×40 pixels were resized to 48×48 pixels, converted to grayscale, and fed through an Xception-based FER network [32]. The output consisted of seven predicted emotion probabilities, one for each face, which resulted in a matrix of shape $(F, 7)$, where F was 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 dense layer with eight neurons and ReLU activation. This was finally followed by a dense layer with three neurons and a softmax classifier. For training, the Adam optimizer was used with a learning rate of $5e^{-3}$.

4.3.6 Image Captioning A GPT-2 transformer-based model pre-trained on the WikiText-103 dataset was used as the base model [33]. Following the methods described by Ruder et al. [33], we utilized the BERT base cased tokenizer on the pre-generated image captions. Two additional dense layers were added with output dimensions of 16 and 3, respectively. These last two layers were fine tuned over nine epochs using a learning rate of $6.5e^{-5}$.

5 Experiments

5.1 Ensembling

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

5.1.1 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. For weighting, the scene modality output was given more weight, by virtue of independently outperforming the pose predictor. For soft voting, we attempted to weight modalities equally (denoted by Soft in Table 4) as well as with the aforementioned weighting methods with the following weights: (i) **Scene**: 0.4, (ii) **Pose**: 0.1 (iii) **Audio**: 0.4, (iv) **FER**: 0.05, (v) **Image Captioning**: 0.05. Hard voting used equal weights for all modalities. Note that laughter was excluded in voting ensembles as the modality could not independently predict sentiment.

5.1.2 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 dense layers was the best performing ensembling method; the FC approach automatically resolved the weights and determined the weighted contribution for each modality.

5.1.3 Variational Autoencoder The same concatenated hidden layers used by the FC ensembling approach were fed as input into a VAE, generating latent z encodings of reduced dimensions. We experimented with the number of encoded dimensions, ultimately settling on 20. We then fed these z encodings into a fully-connected classifier.

5.2 Ablation Study

After determining that the FC early fusion model performed the best, we individually ablated each modality to determine which ones impacted classification performance the most as shown in Table 1. Figure 2 illustrates how modal ablations influenced model decision making in corresponding confusion matrices. This study revealed that the best performing models fused audio, scene, image captioning, laughter, and either the pose or FER modality. The five-modality model using pose achieved a validation accuracy of 64.0%.

6 Results

6.1 Quantitative Results

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. Because our models and methods can be applied to a variety of real-world contexts, each requiring different metrics, it is important for us to deeply understand their nuanced behavior.

Table 1: Performance on the validation dataset after removing modalities from the FC early fusion ensemble architecture shown in Figure 1. The following modalities are ablated from the model.

Modality Ablated	Precision	Recall	F1 Score
Scene	.620	.605	.604
Pose	.653	**.639**	**.637**
Audio	.617	.606	.605
Laughter	.650	.632	.629
FER	**.654**	**.639**	**.637**
Caption	.653	.635	.633

6.2 Test Set Results

Our model achieved an overall validation accuracy of 64.0% and test accuracy of 63.9%. This significantly outperformed the current published baselines of 52.1% [5] on the validation data, as well as 47.9% on the test data. The entire breakdown of the classification accuracy for all submissions is shown in Table 2.

The best model consisted of a fully-connected classifier applied on an embedding produced from the concatenated outputs of five unimodal classifiers: audio, scene, pose, image captioning, and laughter. Note that Submission 1 was optimized for the best validation loss. Submission 2 was optimized for the best validation loss, but also included a boosting layer to reclassify positive and neutral

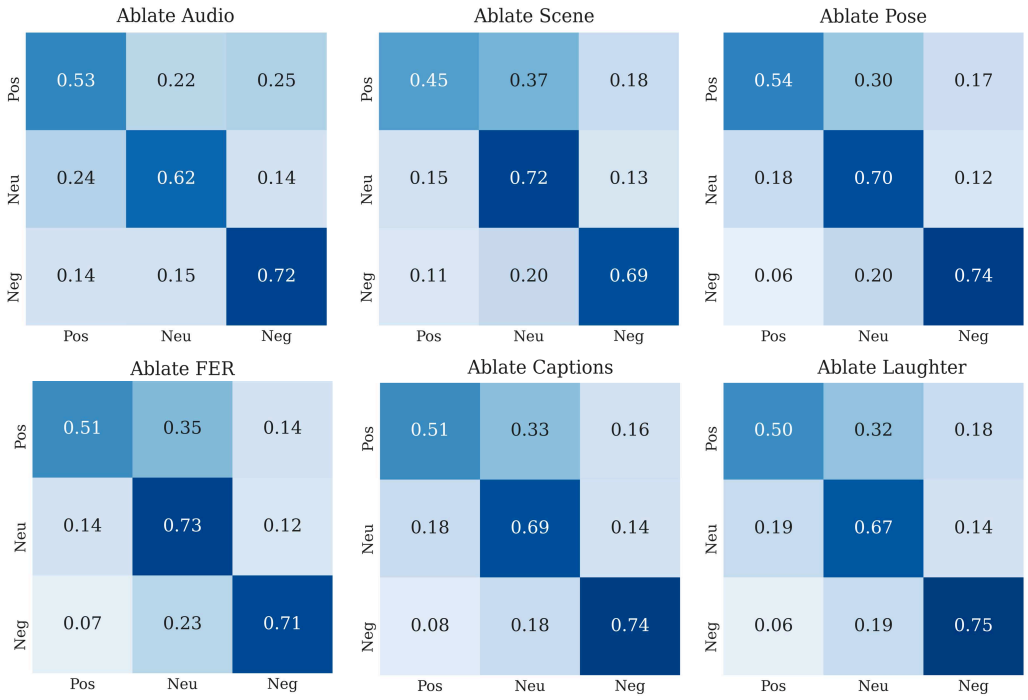

Figure 2: Confusion matrices for all ablations on the fusion architecture in Figure 1, where only one modality is removed per matrix. The matrix titles indicate the ablated modality. The positive class accuracy is lower when the scene modality is removed. Ablating audio leads to lower neutral class accuracy.

Table 2: Submission results on the test dataset. We present all of our five submissions with the individual class accuracies, along with global accuracy.

Model	Pos Acc (%)	Neu Acc (%)	Neg Acc (%)	Overall Acc (%)
Submission 1: Scene (ResNet), Pose, Audio, Laugh, Caption	55.3	**78.0**	53.0	**63.9**
Submission 2: Scene (ResNet), Pose, Audio, Laugh, Caption	**58.1**	74.4	53.0	63.2
Submission 3: Scene (ResNet), Pose, Audio, Laugh, Caption	54.4	74.1	**57.8**	63.5
Submission 4: Scene (ResNet), Audio, Laugh, Caption	55.8	68.6	53.9	60.4
Submission 5: Scene (ResNet), Audio, FER, Caption	54.8	71.5	57.4	62.4
Baseline	–	–	–	47.9

predictions. Submission 3 was optimized for the best validation accuracy.

6.3 Unimodal Classifiers

Table 3 summarizes the results of each of the unimodal classifiers, which were trained and validated independently. The laughter detection model was excluded from this experiment, as the model was trained to detect the binary laugh state, not the output sentiment.

Table 3: Summary of the best individual model performances on the validation dataset.

Model	Precision	Recall	F1 Score
Scene	.547	.540	.541
Pose	.483	.512	.489
Audio	**.640**	**.580**	**.577**
FER	.396	.403	.348
Image Caption	.523	.534	.506

6.4 Ensemble Classifiers

Our early fusion technique yielded the best results with an F1 score of 0.638 on the validation dataset. 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 are interdependent, allowing the model to learn the vital context associated with different modalities interacting together. The final dense model confusion matrix is illustrated by Figure 3.

Table 4: Performance on the validation dataset for different ensembling strategies. Each ensemble strategy uses the same unimodal model outputs for all modalities.

Ensemble	Precision	Recall	F1 Score
Hard	.575	.548	.543
Soft	.483	.466	.467
Weighted	.629	.590	.579
VAE	.282	.445	.341
FC-Fusion	**.655**	**.640**	**.638**

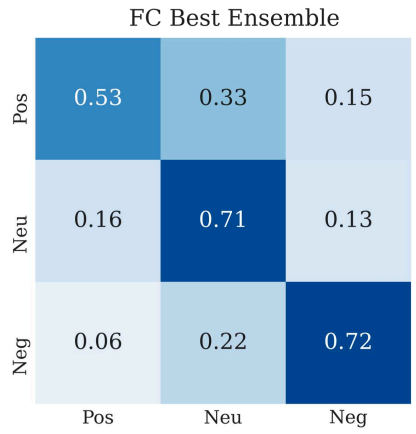

Figure 3: The confusion matrix of our best dense model on the validation set. It fuses all modalities except the VGG scene and FER classifiers. The model scores particularly high with neutral and negative classes, but struggles to differentiate positives from neutrals.

6.5 Qualitative Results

6.5.1 Unimodal Scene When compared to other unimodal models, the unimodal scene-based classifier excelled at predicting true positives and is the only model with a precision rate greater than 50% for that category as shown in Figure 3. The scene model is also fairly unbiased and robust, as it is not skewed towards any one sentiment class. Nevertheless, it still has difficultly distinguishing neutral labels from positive and negative labels. The saliency map in Figure 4 shows that the model predominantly leverages faces and hands for its predictions.

Unlike the facial modality, the scene model is invariant to the orientation of the body and face. Thus, it may use features from the scene even if the person has their back to the camera or parts of the body are occluded.

6.5.2 Unimodal Pose Similar to the audio model, the pose model is fairly robust at predicting true negative sentiments but poor at predicting true positive sentiments. This model also in general has a lower validation accuracy. This is not surprising, given that there is great variation in the number of keypoints extracted within each video. The pose model focuses predominantly on upper body joints, specifically the hands and elbows as shown in Figure 5.

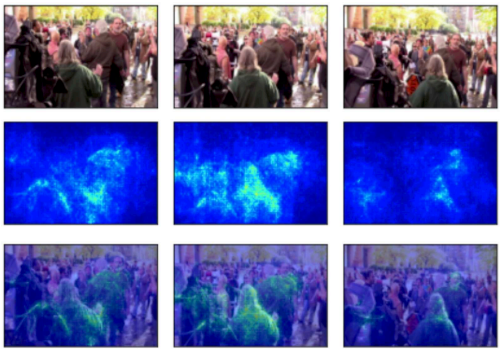

Figure 4: Three sequential sample frames overlayed with a saliency map of the ResNet-LSTM model performed on a negative sentiment video clip. The model often focuses on human figures and subjects, especially those in the foreground.

Figure 5: Saliency map with respect to the pose keypoints for a positive sentiment class, as shown by the bright green dots, with larger dots indicating areas where the model activates more.

6.5.3 Unimodal Image Captioning Interestingly, image captioning is the third best performing modality with an F1-score of 0.506 despite producing relatively generic, albeit correct, captions such as "a group of people sitting around a table". However, in the overall ensemble, we observe that the image captioning model often contributes the least to the overall prediction, as shown by Figure 6.

6.6 Modalities in Decision-Making

When performing a human reader study on the data, we note that certain modes weighted more greatly than others for making a final decision. To translate this cognitive aspect to computing, we generated vector "saliency" maps on the fusion layer of our model. The bar graphs demonstrated in Figure 6 plot the relative activation of each modality as a function of the input video. In this figure, even though the noise and scene play a role, the mere presence of a laugh in the video is the deciding factor for positive classification.

7 Discussion

Although the scene model is one of the strongest models, it still only achieves a validation accuracy of 54.6% independently. For

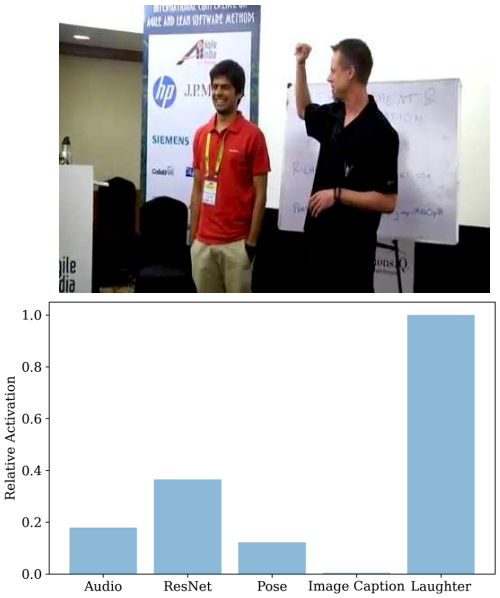

Figure 6: Relative activation for each modality as shown using the FC ensemble model on a positive-sentiment video. The laughter in the above video is the main deciding factor for positive classification, superseding the presence of raised hands or smiling.

the scene modality we use the ResNet-50 model as the base model because our experiments show that this model has the lowest validation loss. This contrasts with the baseline paper, which uses an Inception-v3 model [5]. This may help explain why our model improves upon the baseline paper; we found that using Inception-v3 led to a 15 percentage point drop in performance for the same architecture. The Inception-v3 model activates strongly on multiple areas of the input image and is unable to focus on the people and important sentiment discriminators of a scene like the ResNet-50 model could.

The performance of the pose model is limited to the expressiveness of the poses themselves, which explains the relatively lower independent 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 5. This is justified by the position of the arms, which is an important indicator of emotions, with raised arms and protruded arms traditionally indicating joy and fear, respectively [34].

The audio modality is the strongest individual modality, 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. While the OpenL3 model used for extracting embeddings was pre-trained primarily on music files, it performed well. Both music and human speech have prosody, or patterns of stress and intonation, and these universal features allow for high sentiment classification. Through transfer learning we translated the hidden features extracted from millions of music training samples and fine tuned them on the EmotiW dataset.

The low performance observed within the facial modality can be attributed to the difficulty of interpreting low resolution, obscured, and blurred faces, which are common in the EmotiW dataset. Moreover, micro facial variations and transition between emotions within a video can lead to highly variable predictions. Additionally, people often do not express their true emotions facially and feelings are expressed differently across cultures.

All modalities had difficulty distinguishing neutral from positive and negative sentiment. Even during our internal reader-study, the line between neutral and positive sentiment was blurred. Thus, to bolster the ensemble, we introduced the laughter modality as an additional feature. Despite only being a single-dimensional feature, laughter proved to be an important sentiment discriminator. This is unsurprising given that laughter is generally a clear indicator of a positive atmosphere. In fact, our ablation study showed that laughter is the third most important modality after audio and scene (see Table 1). Although we already have an audio modality, these results suggest that ensembled binary audio classifiers trained on specific emotional sounds such as laughter or cries can be used to improve specific shortcomings of a model, as shown in Figure 6. This is an open area for further research.

Distinguishing between a neutral sentiment and a positive or negative sentiment is often a matter of context and subjective human perception. Many of the videos are from news clips, which naturally gravitate towards negative sentiment [35]. The novelty of our model stems from our use of several non-traditional modalities such as laughter and image captioning. The underlying embeddings of image captions extracted from frames provide a condensed representation of the video itself, analogous to the embeddings extracted using VAEs. However, unlike VAE embeddings, these caption embeddings are easily interpretable. Although our model was hindered by captions that were factually-correct but vague, image captioning is a highly active research area with many researchers specifically looking to inject image captioning with sentiment terms [36, 37]. While the aim of this research is to produce more naturally sounding phrases, this opens a potential avenue of transfer learning in the sentiment classification space.

Because of the subjective nature of the data, there is an element of bias even within our produced models: when the image captions were analyzed by word count, the words "men" and "man" appeared approximately four times more frequently within negative videos than positive videos. A simple two-tailed t-test highlighted that this difference is significant ($p < 0.05$). For these reasons, it is not surprising that our fully-connected late fusion ensemble of modalities achieves 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 summarizes this point. Removing the scene classifier led to a drop of nine percentage points in the recall rate of positive samples. Similarly, when FER is removed, we observe a drop of eight percentage points in the recall rate of negative samples. Our fusion method is truly greater than each of its parts.

8 Conclusion

Video sentiment classification is a complex problem under active study, requiring investigation of various modalities and powerful computation. In this paper, we use a unique combination of features extracted from different modalities to perform image sentiment classification, including novel modalities such as image captioning and laughter detection. Our final late fusion ensemble model of scene, pose, audio, laughter, and image captioning modalities outperformed the baseline validation accuracy of 52.1% and the baseline test accuracy of 47.9%, achieving an overall validation accuracy of **64.0%**, and an overall test accuracy of **63.9%**.

9 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.

This repository includes Colab notebooks for interactive demos, pre-trained model files, preprocessing, and training code.

10 Acknowledgements

We are grateful to Professor Ranjay Krishna and Christina Yuan for their mentorship and guidance throughout our project. We are grateful for the inputs of Dr. Pawan Nandakishore and Hala Khodr for guidance on literature and method refinement. We would also like to thank Vincent La, who helped us explore object detection and text-based sentiment analysis.

10.1 Preprocessing

We leveraged the Face Recognition library for facial extraction [28]. We worked extensively with the OpenL3 library [38] for audio embeddings. To extract poses from video frames, we used the OpenPose library [24].

10.2 Models

For development of our facial modality, we leveraged pre-trained FER models from Khanzada et al. [12] and Arriaga et al. [39]. The laughter model was adapted from Nat Stein [17]. The pre-trained caption generator from Beaumont and Kranthi [40] was used for fine tuning. Moreover, the sentence sentiment transformer was adapted from Ruder et al. [33]. For our VAEs models, we adapted methods from Kingma and Welling [41]. For saliency maps, we augmented base code from the keras-vis package from Kotikalapudi et al. [42] and the tf-keras-vis package from Keisen et al. [43] 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] 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.

[7] 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.

[8] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator, 2014.

[9] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2015.

[10] 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.

[11] 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.

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

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

[14] 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.

[15] 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.

[16] 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.

[17] Nat Steinsultz. Laugh detector. https://github.com/ideo/LaughDetection, 2018.

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

[19] Xin Guo, Luisa F. Polanía, and Kenneth E. Barner. Group-level emotion recognition using deep models on image scene, faces, and skeletons. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, page 603–608, New York, NY, USA, 2017. Association for Computing Machinery.

[20] Zheng Lian, Ya Li, Jianhua Tao, and Jian Huang. Investigation of multimodal features, classifiers and fusion methods for emotion recognition. *CoRR*, abs/1809.06225, 2018.

[21] Ping Hu, Dongqi Cai, Shandong Wang, Anbang Yao, and Yurong Chen. Learning supervised scoring ensemble for emotion recognition in the wild. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI '17, page 553–560, New York, NY, USA, 2017. Association for Computing Machinery.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] Dan Ellis, Shawn Hershey, Aren Jansen, and Manoj Plakal. Vggish. *URL https://github. com/tensorflow/models/tree/master/research/audioset/vggish*, 2019.

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

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

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[31] Shreya Ghosh, Abhinav Dhall, and Nicu Sebe. Predicting group cohesiveness in images. *CoRR*, abs/1812.11771, 2018.

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

[33] Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, 2019.

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

[35] 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.

[36] Alexander Patrick Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *Thirtieth AAAI conference on artificial intelligence*, 2016.

[37] Andrew Shin, Yoshitaka Ushiku, and Tatsuya Harada. Image captioning with sentiment terms via weakly-supervised sentiment dataset. In *BMVC*, 2016.

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

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

[40] Romain Beaumont and G. Kranthi. Pretrained-show-and-tell-model.

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

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

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