# reSenseNet: Ensemble Early Fusion Deep Learning Architecture for Multimodal Sentiment Analysis *

Shankhanil Ghosh[1], Chhanda Saha[1], Nagamani Molakathala[1], Souvik Ghosh[2], and Dhananjay Singh[3]

[1] School of Computer and Information Sciences, University of Hyderabad, India
{20mcmb04,20mcmb01, nagamanics}@uohyd.ac.in
[2] Heritage Institute of Technology, India
souvikg544@gmail.com
[3] Hankuk University of Foreign studies, South Korea
dsingh@hufs.ac.kr

**Abstract.** Multimodal sentiment analysis is an actively emerging field of research in deep learning that deals with understanding human sentiments based on more than one sensory input. In this paper, we propose reSenseNet, an ensemble of early fusion architecture of deep CNN and LSTM for multimodal sentiment analysis of audio, visual, and text data. ReSenseNet consists of feature extraction, feature fusion, and fully connected layers stacked together as a three-layer architecture. Instances of the generalized reSenseNet architecture have been experimented on several variants of modalities combined together to form different variations in the test data. Such a combination has produced results in predicting average arousal and valence up to an F1 score of 50.91% and 35.74% respectively.

**Keywords:** Multimodal deep learning · Sentiment Analysis · Feature Fusion · LSTM · Arousal · Valence · Deep Learning.

## 1 Introduction

Diagnosis of mental health issues is a big challenge in Human-Computer Interaction research. This research is focused this problem and attempts to find technological solutions towards the same by developing novel multi-modal deep learning methods for sentiment analysis tasks. It is inspired by the proposition by World Health Organization (WHO) saying "no health without mental health", which clearly mentions the importance of mental health in people's lives. Approximately 792 million people worldwide suffered from a mental health disorder,

* This research was supported under the India-Korea Joint Programme of Cooperation in Science & Technology by the National Research Foundation (NRF) Korea (2020K1A3A1A68093469), the Ministry of Science and ICT (MSIT) Korea and by the Department of Biotechnology (India) (DBT/IC-12031(22)-ICD-DBT)

according to Hannah Ritchie and Max Roser [4] (2017). It is slightly more than one in ten people globally, and that number is increasing with time. Researchers worldwide are trying to develop solutions towards various problems associated with mental health problems. On the other side, a massive volume of opinionated data recorded in digital form available for analysis in today's world.
The authors Soleymani et al. [1] have discussed the various challenges and opportunities in the domain of multimodal sentiment analysis. Using sentiment analysis techniques, it is possible to automate the extraction or classification of sentiments from opinionated data or reviews. Sentiment analysis techniques can analyze the reviews and opinions and classify them according to different classes of sentiment. To quantify the emotions while someone is reviewing, two variables called arousal and valence can be used. Arousal represents the state of excitement of the speaker, and valence represents the pleasantness of the sentiment. It helps us in identifying the speaker's emotions and sentiments. Since a significant portion of today's data is available in multiple modalities, recent research attempts to combine different modalities where it results in better accuracy.

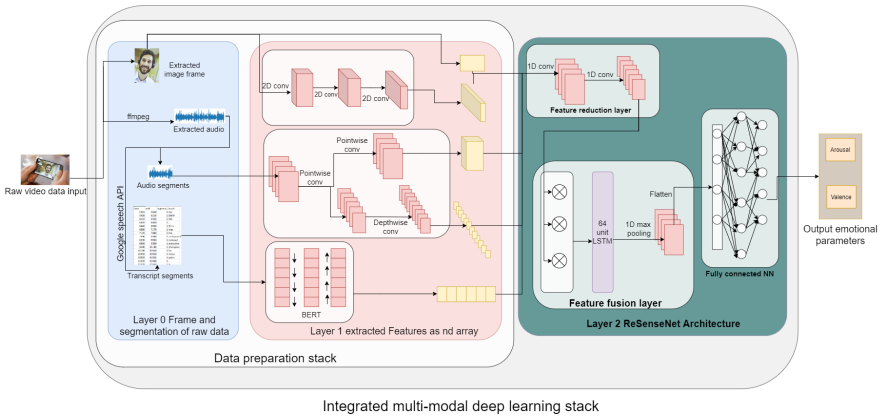



Fig. 1: Integrated deep learning stack for multi-modal sentiment analysis from raw video input. Our proposed reSenseNet is a part of this integrated stack

### 1.1 Research contribution

This work predicts advanced intensity classed of emotional parameters. These parameters are arousal and valence. The work uses segmented audio-visual-textual data. An integrated deep learning stack has been developed as shown in Figure 1. The novelty of the proposal lies in the deep CNN and LSTM based Early fusion architecture. This architecture is called reSenseNet.

[4] https://ourworldindata.org/mental-health

1. ReSenseNet architecture to perform multimodal sentiment analysis on segmented audio-visual-textual data.
2. Empirical experiments are performed on different instances of the architecture to predict emotional variables arousal and valence have been done. Performance measuring metric is F1 score.
3. Finally, two instances to predict arousal and valence is proposed. They have an F1 score of 50.91% and 35.74% respectively across three modalities.

## 2 State of Art in sentiment analysis

This research motivation is inspired by Kaur et al. [2], who propose a search-based stacking model that collectively exploits multiple base learners for human-activity analysis. Further exploration of sentiment analysis methods in context of deep learning application and multimodal feature fusion have been performed. Soleymani et al. [1] comprehensively presented "sentiment" and "sentiment analysis" in context of multimodal sentiment analysis is well summarized. Zhang et al. in [3] also proposed a brief survey about the application of deep learning in this context. Morency et al. [4] addresses opinion harvesting from large-scale multimodal raw data, proof-of-concept from joint model which integrates audio, visual, textual features to identify sentiments from web resources.

### 2.1 Deep learning research in Sentiment analysis

Rosas et al. in [5] have applied deep learning techniques to perform sentiment analysis on Spanish online videos, also have shown that deep multimodal features provide better accuracy than the singular features. Yadav et al. [6] have reviewed different deep learning techniques applied in sentiment analysis and solved different difficulties faced during this task. Another application can be seen in [7] where a deep learning-based framework for the multimodal sentiment analysis has been proposed, which gets a better result. [7] have also stated that by combining the audio, visual, and text feature, they got 10% improvement. Poria et al. [8] have proposed a decision level fusion framework used in deep learning techniques for a multimodal sentiment analysis task with a margin of 10–13% and 3–5% accuracy on polarity detection and emotion recognition, respectively.

### 2.2 Multimodal deep learning in sentiment analysis

The Multimodal Sentiment Analysis in Real-life Media Challenge (MuSe) 2020 [9] focused on the task of sentiment recognition, topic engagement, and trustworthiness detection. Three sub-challenge named MuSe-Wild, MuSe-Topic, and MuSe-Trust were proposed for teams to participate. Similarly, the 2nd Multimodal Sentiment Analysis Challenge (MuSe 2021) [10] focused on multimodal sentiment recognition of user-generated content and in stress-induced situations. This challenge compared multimedia processing and deep learning methods for

automatic audio-visual, biological, and textual-based sentiment and emotion-sensing under a standard experimental condition set. Four sub-challenges named MuSe-Wilder, MuSe-Sent, MuSe-Stress, and MuSe-Physio were proposed under MuSe 2021 challenge. The data for the challenge was provided by [11], which was collected from YouTube and manually annotated. Ghosh et al. [12] have proposed one data acquisition tool[5] which can be used to collect multimodal data for such tasks.
One significant task by Stappen et al. [13] have described about unifying a wide range of fusion methods and proposed the novel Rater Aligned Annotation Weighting (RAAW), which aligns the annotations in a translation-invariant way before weighting and fusing them based on the inter-rater agreements between the annotations. Strappen et al. [14] have also proposed a topic extractor on video transcripts, which uses neural word embeddings through graph-based clustering. This research also uses the MuSe-CaR dataset [11].

### 2.3 Multi-modal feature fusion methods

One primary technological method needed in this research is multimodal feature fusion methods, which allows us to fuse features across different sources and modalities into one single feature vector. Our inspiration came from some of the recent research papers on feature fusion methods. [15] have proposed a model which starts its task by eliminating the noise interference in textual data and extracting more essential image features, after which the feature-fusion part based on attention mechanism learns internal features from the text and images data through symmetry. The model then applies the fusion features to the sentiment classification tasks. Majumder et al. in [16] have proposed a novel feature fusion strategy, which proceeds hierarchically, first fusing the modalities two in two and then fusing all three modalities. Zadeh et al. in [17] solve the problem of multimodal sentiment analysis as an instance of modeling intra-modality and inter-modality dynamics and propose Tensor Fusion Network, which learns both such dynamics end-to-end. The proposed approach has been designed to make it worthwhile for the volatile nature of spoken language in online videos, voice, and gestures.

## 3 The ReSenseNet architecture

The reSenseNet architecture for multi-modal feature fusion for sentiment analysis task is a part of a three-layered integrated deep learning stack for multimodal sentiment analysis. The architecture uses an Early fusion mechanism to fuse various features across various modalities. ReSenseNet is made of three significant layers: the feature reduction layer, the Early Fusion Layer, and the Fully connected neural network. The detailed description of the ReSenseNet architecture is shown in Figure 2.

[5] https://intellispeechscis.web.app/

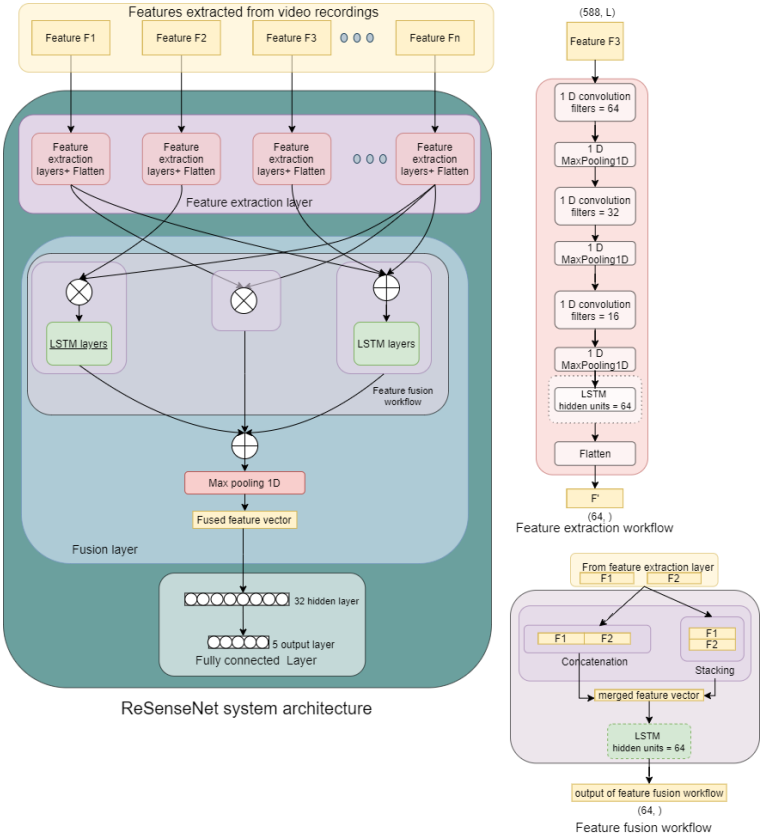



Fig. 2: A detailed description of a deep-learning stack that integrates reSenseNet to predict arousal and valence (left). The resenseNet architecture is to the right. The figure also includes description of the feature reduction workflow and the feature fusion workflow. The optional layers are denoted in dotted boxes.

### 3.1 Feature reduction Layer

The intial layer of the reSenseNet layer is the feature reduction layer. Out of various methods of feature reduction, two techniques have been considered: consecutive convolution layers (followed by a max-pooling layer), and Long-short-term memory, or LSTM layers.
The model initially performs an 1D convolution on all the modal features. The intuition behind this is to build reduced feature maps for timestamped data. As mentioned previously, each audio, video, and text feature matrices are of size $(588, F_s)$, where the first dimension of the feature vectors represents the timestamp dimension (after zero padding). The convolution layers are followed by 1D Max pooling layers along the feature-length. The model applies these layers repeatedly over the features until it achieves a feature vector of sufficient length. In the end, the features are flattened to produce a feature vector of length 160 for each feature.
Every instance of the reSenseNet architecture contains the convolution layers. In some instances of the reSenseNet, the authors have experimented with an additional 64-unit LSTM layer, which is applied to the output of the convolution layer. The output of this LSTM layer is a feature vector of length 64.

### 3.2 Fusion layer

The fusion layer is the most important portion of the reSenseNet architecture which performs early fusion on the reduced feature maps. The output of the feature reduction layer is sent as an input to the Fusion layer. Three different approaches to performing feature fusion have been proposed. These are simple concatenation, concatenation with LSTM early fusion and stacking with LSTM early fusion. Exhaustive experimentation on different types of fusion techniques have been performed to make architecture level decisions on ReSenseNet.

$$F_k^{\prime(2n,1)} = F_i^{(n,1)} || F_j^{(n,1)} \tag{1a}$$

$$F_k^{\prime(n,2)} = F_i^{(n,1)} || F_j^{(n,1)} \tag{1b}$$

$$\forall F_i \epsilon F_1, F_2, \dots F_n$$

$$lstm_i = LSTM_{64}(f), \forall f \epsilon F_1, F_2 \dots F_n \tag{2}$$

1. In the simple concatenation fusion method, multiple feature vectors are concatenated against one another (as shown in equation 1a and feeding it to the fully connected layer.
2. Concatenation with LSTM early fusion is a technique where certain features are concatenated (Equation 1a) and then passed through 64-unit LSTM layer, as shown in as described in Equation 2.
   Furthermore, the output of the LSTM layers is again passed to a stacking sub-layer and then passed into a pooling layer. The output is flattened and sent into the final fully connected layer.

3. In the Stacking with LSTM early fusion technique, the first level of feature fusion is performed by stacking feature pairs $(F_i, F_j)$ on top of one another, to produce features $F'_1, F'_2, \ldots F'_N$, which are of size $(L, 2)$ each. Stacking is mathematically defined in Equation 1b. Each of these stacked feature matrices are then passed into $N$ separate 64-unit LSTM layers. The output of the LSTM layers are passed through a stacking layer and max-pooling layer after which it is sent to the fully connected.

### 3.3 Fully connected layer

The final layer of the reSenseNet architecture is where the feature fusion layer's output is passed into a fully connected neural network, with one hidden layer containing 32 hidden units and five output units. Specific details of the fully connected neural network in the later sections.

## 4 Dataset and Data preparation

### 4.1 The dataset and the multi-modal features

The dataset used in this research the MuSe-CaR dataset[11] which is a collection of 40+ hours of YouTube videos of car reviews. Data annotation of continuous values like arousal, valence for this dataset have been done by human annotators. This research uses the pre-extracted features that are provided by the MuSe-CaR dataset. In the audio features, eGeMAPS [18] and VGGish [19] have been considered for the experimentation. EGeMAPS and VGGish are feature vectors of lengths 88 and 128 for a specific timestamp for each audio segment. In the visual/facial modality, Xception [20] , Facial Action Units and VGGface [21] have been considered. Xception, FAU, and VGGface have feature vectors of length 2048, 35, and 512, respectively, for each audio segment at a certain timestamp. For the text modality, BERT [22] feature is used which has a feature vector of length 768.

In future mentions VGGish will be used as $v'_f$, eGeMaps as $eGe$ , Xception as $X$, VGGFace as $v_f$, FAU as $au_f$ and BERT as $B_T$

### 4.2 Data Preprocessing

The data was available to us in the form of raw audio-visual data and pre-extracted feature vectors. The features that are being used for this study are pre-extracted from the audio, video, and text data in the MuSe-Car dataset. Each audiovisual-textual sample from the dataset is divided into smaller segments of variable length. Each segment is several timestamps long. The valence and arousal are annotated against each segment, meaning that each segment has one annotation of valence and arousal. Hence, one unit of data is made of multiple feature vectors, thus forming a feature matrix. However, these feature matrices are of variable dimension because of the variable-length segments. Thus, the

feature matrices are preprocessed to make them uniform length. For all samples that have segment length lesser than 588, zero paddings have been applied to the right of the feature matrix, and for those samples whose segment length is greater than 588, the feature matrix have been truncated . Hence, at the end of preprocessing, the feature matrix of each segment is of size $(588, L_f)$, where $L_f$ is the feature-length. $L_f$ is 35, 4096, 2046, 88, 512, 128, 2048 for Facial Action Units, DeepSpectrum, BERT, eGeMaps, VGGFace, VGGish and Xception respectively. All these features were padded and stored in separate h5 files for further usage. The model reads the feature matrices from the disk in small batches and sends them to the model for training and evaluation. The preprocessing procedure is described in Algorithm 1.

**Algorithm 1** Given a feature matrix file of $i^{th}$ video data, generate the pre-processed feature matrix

**procedure** PREPROCESS($F_i$, featureName)
    outputFile $\leftarrow OPEN(featureName.H5)$
    totalSegments $\leftarrow$ total number of segments in that feature matrix file.
    $F' \leftarrow emptymatrixwithshape(totalSegments, 588, L_F)$
    **for** s in range(0, totalSegments) **do**
        $S \leftarrow F_i[s]$ ▷ $|S| = (L_S, L_F)$, where $L_S$ is the length of the segment, and $L_f$ is the feature length.
        **if** $L_S < 588$ **then**
            $S' \leftarrow zeroPadding(S, extraWidth = (588 - L_S))$
        **else if** $L_S > 588$ **then**
            $S' \leftarrow truncate(S)$ ▷ Making $|S'| = (588, L_F)$ by padding or truncating
        $F'.append(S')$
    outputFile.write( F' )

## 5 Performance Analysis

### 5.1 Experimental setup

The models have been trained using the Keras functional API, set for a maximum of 20 epochs with a batch size of 32. An early stopping mechanism has been set in place with a patience of 15 epochs, and the metric used in this case is the training validation F1 score. The model was compiled using Adam optimizer with Categorical cross-entropy as the loss function. The entire dataset was split with an 80-20 ratio, with 20 percent of the data (randomly chosen) was used for validation. The fully connected neural network has one hidden layer of 32 units and an output layer of 5 units (for five levels of arousal/valence). Dropout layers (with dropout rate = 0.2) and L1 and L2 regularizers have been deployed to prevent overfitting. L1 and L2 regularization factor is 0.000001 and 0.00001.

### 5.2 Evaluation on reSenseNet architecture

The proposed architecture have been used to build models to predict the sentiment variables, namely arousal, and valence. And for that purpose, a set of experiments have been designed, based on which two separate fine-tuned models have been proposed for each variable. The reSenseNet architecture have been evaluated with various combinations of modalities and fusion methods, along with an extensive hyperparameter search. The following set of evaluation experiments have been performed on the reSenseNet architecture

1. Modality search: To study which combination of modalities work best. Various combinations of features, keeping the hyperparameters and fusion method the same, were tested. The symbols A, V and T indicate audio, visual and text modalities respectively. Only concatenation was used as the fusion method. Learning rate $\alpha = 0.001$, Dropout frequency = 0.2, Regularizer parameter $L_1 = 10^{-5}$ & $L_2 = 10^{-6}$, were kept constant through the experiments. Given in Table 1.
2. Fusion method search: To study which specific combination of features work the best. Various fusions of features within the fusion layer were evaluated, keeping the features and hyperparameters the same. Three different fusion mechanisms (as described early in the paper) have been tested out in these set of experiments: Concatenation, Concatenation + LSTM, Stacking + LSTM. The various combinations of features in the A+V and A+V+T modality have been tested here. In the *Feature* column, the brackets indicate the way stacking/concatenation was performed. Learning rate $\alpha = 0.001$, Dropout frequency = 0.2, Regularizer parameter $L_1 = 10^{-5}$ & $L_2 = 10^{-6}$, were kept constant through the experiments. Given in Table 2
3. An extensive hyper-parameter search against the features and fusion method. Various hyperparameters for the model have been tested. The experiments were conducted against the A+V+T modality only because it gave better results in the past experiments, (Table 1. Two different fused features are used as model input, which are $(eGe + B_T) + (v_f + B_T) + (X + B_T)$ and $(eGe+B_T)+(v_f+B_T)+(eGe+au_f)+(X+au_f)$. The regularizer parameters have been kept constant, as $L1 = 10^{-5}$ & $L2 = 10^{-4}$. Given in Table 3

For evaluation, the entire annotated dataset is split into separate training and testing sub-dataset. 1641 data samples are used for training and 572 data samples for testing the models in the experiments as mentioned above. While splitting the datasets, it is ensured that there is no overlap between the train and test dataset, meaning that the test dataset is entirely unseen by the model. This ensures that the results provided by the model are entirely accurate.

### 5.3 Results and Discussion

The results of the experiment shows that the best results comes out when all the three modalities (audio, video, and text) are fused to predict the variables. From Table 1 there could be observed a significant increase in F1 scores in training and

Table 1: Evaluation table for modality search.

| | | **Arousal** | **Valence** |
|---|---|---|---|
| **Modality** | **Feature** | **Train/Test** | **Train/Test** |
| A | $v'_f$ | 45.21/31.23 | 20.26/9.14 |
| A | eGe | 46.32/33.15 | 22.13/15.23 |
| V | $au_f$ | 46.55/36.71 | 38.11/16.32 |
| V | $au_f + v_f$ | 50.45/44.97 | 50.00/29.08 |
| T | $B_T$ | **86.82/46.20** | **85.12/33.92** |
| A+V | $(au_f + v_f) + eGe$ | **50.15/44.93** | 51.11/29.12 |
| A+V+T | $au_f + eGe + v'_f$ | 43.21/28.12 | 95.21/27.40 |
| A+V+T | $(v_f + au_f) + (eGe + X) + B_T$ | **68.08/46.88** | **62.08/32.12** |

Table 2: Evaluation table for fusion method search

| | | | **Arousal** | **Valence** |
|---|---|---|---|---|
| **Fusion method** | **Modality** | **Feature** | **Train/Test** | **Train/Test** |
| Concatenation | A+V | $(au_f + v_f) + eGe$ | 46.67/42.12 | 50.00/29.08 |
| | A+V+T | $au_f + eGe + v'_f$ | 43.21/28.12 | 45.67/27.40 |
| | A+V+T | $(v_f + au_f) + (eGe + X) + B_T$ | 68.08/46.88 | 62.08/32.12 |
| Concatenation + LSTM | A+V | $(au_f + v_f) + eGe$ | 50.45/44.97 | 50.11/29.08 |
| | A+V+T | $au_f + eGe + v'_f$ | 47.11/32.98 | 50.67/30.18 |
| | A+V+T | $(v_f + au_f) + (eGe + X) + B_T$ | 65.18/46.10 | 65.12/33.21 |
| Stacking + LSTM with Early fusion | A+V | $(au_f + v_f) + eGe$ | 50.45/44.97 | 49.22/29.18 |
| | A+V+T | $au_f + eGe + v'_f$ | 46.18/42.12 | 45.12/37.36 |
| | A+V+T | $(v_f + au_f) + (eGe + X) + B_T$ | 59.18/47.67 | 52.58/33.11 |
| | A+V+T | $(eGe + B_T) + (v_f + B_T) + (X + B_T)$ | 52.77/50.91 | 54.78/35.74 |
| | A+V+T | $(eGe + B_T) + (v_f + B_T) + (eGe + au_f) + (X + au_f)$ | 52.77/49.24 | 54.78/35.88 |

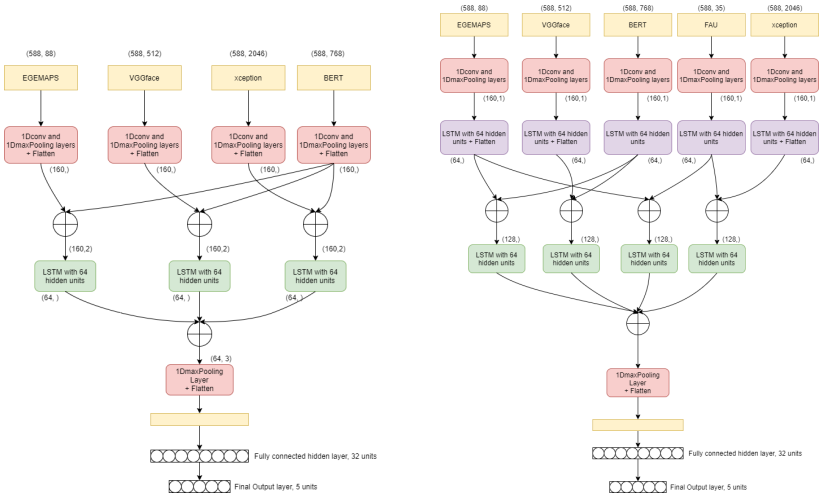
Table 3: Evaluation table for hyperparameter searching.

| | | | **Arousal** | **Valence** |
|---|---|---|---|---|
| **Input fused feature** | **Learning rate** | **Dropout** | **Train/Test** | **Train/Test** |
| $(eGe+B_T)+(v_f+B_T)+(X+B_T)$ | **0.001** | **0.2** | **52.77/50.91** | 54.78/35.74 |
| | 0.002 | 0.2 | 52.42/49.11 | 54.72/35.32 |
| | 0.001 | 0.4 | 53.21/49.21 | 55.11/35.04 |
| | 0.002 | 0.4 | 53.25/49.56 | 55.18/34.98 |
| $(eGe+B_T)+(v_f+B_T)+(eGe+au_f)+(X+au_f)$ | **0.001** | **0.2** | 52.77/49.24 | **54.78/35.74** |
| | 0.002 | 0.2 | 54.42/48.03 | 54.11/35.14 |
| | 0.001 | 0.4 | 53.98/48.74 | 53.18/35.01 |
| | 0.002 | 0.4 | 54.18/48.04 | 54.05/35.23 |

testing scenarios. For example, from Table 1, it can be seen that using modalities audio + video, there is jump in test F1 score from 33.15% to 44.97%. It must also be noticed that the text modality itself is potent for predicting arousal and valence because it scored an F1 score of 46.20% and 33.92% respectively for test datasets.

From Table 2 in the fusion method search, the Early fusion Stacking + LSTM is outperforming all the other methods of feature fusion. In the concatenation feature, reSenseNet model achieved a highest F1 score of 46.88% and 32.12% on test dataset, for arousal and valence respectively. In that experiment, feature fusion has been performed as $(au_f + v_f) + eGe$. Using Concatenation + LSTM early fusion method, the highest F1 score was obtained as 46.10% and 33.21% for feature sets as $(v_f + au_f) + (eGe + X) + B_T$. However, some of the highest F1 scores were achieved by using stacking + LSTM early fusion. In that method, feature set $(eGe + B_T) + (v_f + B_T) + (X + B_T)$ achieved an F1 score of 50.91% for arousal and feature set $(eGe + B_T) + (v_f + B_T) + (eGe + au_f) + (X + au_f)$ achieved an F1 score of 35.88%, indicating that these feature sets, used with stacking + LSTM produce the best result.

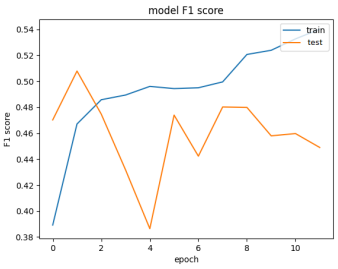
In Table 3, the best feature sets from the previous experiments have been considered. These are $(eGe + B_T) + (v_f + B_T) + (X + B_T)$ and $(eGe + B_T) + (v_f + B_T) + (eGe + au_f) + (X + au_f)$ and perform a hyperparameter search for Dropout frequency and learning rate. The best results came out with learning rate $\alpha = 0.001$ and Dropout frequency = 0.2. Hence, considering the best result from the experiments, a training F1 score of 52.77% and test score of 50.91% for arousal, and a training F1 score of 54.74% and a test F1 score of 35.88% for valence have been achieved.



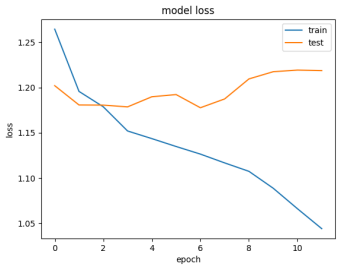

(a) reSenseNet-Arousal model architecture

(b) reSenseNet-Valence model architecture

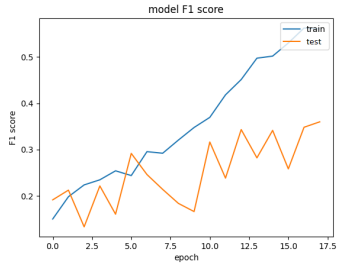
Fig. 3: reSenseNet-Arousal and Valence model architecture



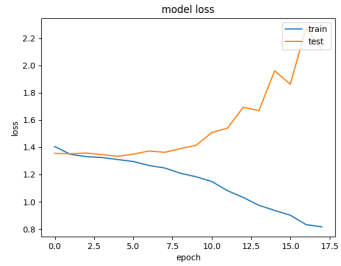

(a) F1 score training curve for Arousal




(b) Loss training curve for Arousal




(c) F1 score training curve for Valence




(d) Loss training curve for Valence

Fig. 4: reSenseNet-Arousal and valence training curves

Table 4: Final table describing results and specifications of ReSenseNet-Arousal and ReSenseNet-Valence

| | **Arousal** | **Valence** |
|---|---|---|
| Modality | A+V+T | A+V+T |
| Feature maps | $(eGe + B_T) + (v_f + B_T) + (X + B_T)$ | $(eGe + B_T) + (v_f + B_T) + (eGe + au_f) + (X + au_f)$ |
| Extraction method | CNN | CNN + LSTM |
| Dropout frequency | 0.2 | 0.2 |
| Learning rate | 0.001 | 0.001 |
| Regularziers L1/L2 | $10^{-5}$ & $10^{-4}$ | $10^{-5}$ & $10^{-4}$ |
| Train/Test | 52.77/50.91 | 54.78/35.88 |

Based on the above experiment results, the authors propose that the best model for predicting arousal is where the reSenseNet architecture uses the $(eGe + B_T) + (v_f + B_T) + (X + B_T)$ feature set with Stacking + LSTM early fusion method. It uses a 0.2 dropout frequency on the fully connected layer, with regularizer parameters $L_1 = 10^{-5}$ & $L_2 = 10^{-6}$, and learning rate of 0.001. The model needs to be trained for 20 epochs (which can early-stopped with patience of 15 epochs). This model is called reSenseNet-Arousal. Similarly, for predicting valence, the usage of feature map $(eGe + B_T) + (v_f + B_T) + (X + B_T)$ and $(eGe + B_T) + (v_f + B_T) + (eGe + au_f) + (X + au_f)$ with Stacking + LSTM early fusion, with a dropout frequency = 0.2, regularizer parameters $L_1 = 10^{-5}$ & $L_2 = 10^{-6}$ andlearning rate = 0.001 is suggested. The model is trained for maximum of 20 epochs with early stopping mechanism in place, for 15 epoch patience. This model is called reSenseNet-Valence. The structure of the models are given in and the structure is visualized in Figure 3, and the training curves for F1 score and loss are given in Figure 4. The details of the models are documented in Table 4. As it can be seen, the test F1 score for valence never reaches the same level as arousal, indicating that predicting valence in this case might be difficult. The authors assume that this might be due to nature of the dataset MuSe-Car, where the YouTubers (subject of the video dataset) maintain a certain level of valence (or pleasantness of voice) for their own audience. Also, the videos are about car reviews, where it is not expected to have a very big difference in valence. However, there are other datasets to train the models to predict valence in a better way.

## 6 Final discussion

In this paper, the authors have proposed the reSenseNet architecture, which is a novel deep learning based architecture for predicting emotional parameters (arousal and valence) in an sentiment analysis task across various modalities. Extensive tests are performed on the architecture and finally proposed 2 models based on the reSenseNet architecture, namely reSenseNet-Arousal and reSenseNet-Valence for predicting arousal and valence respectively. These models have scored an F1 score of 50.91% and 35.74% on test datasets.

## Acknowledgement

This research was supported by the National Research Foundation (NRF), Korea (2020K1A3A1A68093469) funded by the Ministry of Science and ICT (MSIT), RESENSE Lab HUFS.

## References

1. Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017. Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.
2. Maninder Kaur, Gurpreet Kaur, Pradip Kumar Sharma, Alireza Jolfaei, and Dhananjay Singh. Binary cuckoo search metaheuristic-based supercomputing framework for human behavior analysis in smart home. *The Journal of Supercomputing*, 76(4):2479–2502, Apr 2020.
3. Lei Zhang, Shuai Wang, and Bing Liu. Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253, 2018.
4. Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, page 169–176, New York, NY, USA, 2011. Association for Computing Machinery.
5. Verónica Pérez Rosas, Rada Mihalcea, and Louis-Philippe Morency. Multimodal sentiment analysis of spanish online videos. *IEEE Intelligent Systems*, 28(3):38–45, 2013.
6. Ashima Yadav and Dinesh Kumar Vishwakarma. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review*, 53(6):4335–4385, 2020.
7. Erik Cambria, Devamanyu Hazarika, Soujanya Poria, Amir Hussain, and R. B. V. Subramanyam. Benchmarking multimodal sentiment analysis. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 166–179, Cham, 2018. Springer International Publishing.
8. Soujanya Poria, Haiyun Peng, Amir Hussain, Newton Howard, and Erik Cambria. Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. *Neurocomputing*, 261:217–230, 2017. Advances in Extreme Learning Machines (ELM 2015).

9. Lukas Stappen, Alice Baird, Georgios Rizos, Panagiotis Tzirakis, Xinchen Du, Felix Hafner, Lea Schumann, Adria Mallol-Ragolta, Björn W. Schuller, Iulia Lefter, Erik Cambria, and Ioannis Kompatsiaris. Muse 2020 – the first international multimodal sentiment analysis in real-life media challenge and workshop, 2020.
10. Lukas Stappen, Alice Baird, Lukas Christ, Lea Schumann, Benjamin Sertolli, Eva Messner, Erik Cambria, Guoying Zhao, and Björn Schuller. The muse 2021 multimodal sentiment analysis challenge: Sentiment, emotion, physiological-emotion, and stress, 04 2021.
11. Lukas Stappen, Alice Baird, Lea Schumann, and Björn Schuller. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. *IEEE Transactions on Affective Computing*, -(-):–, 2021.
12. Shankhanil Ghosh, Naagamani Molakathaala, Chhanda Saha, Rittam Das, and Souvik Ghosh. Speech@scis:annotated indian video dataset for speech-face cross modal research (draft).
13. Lukas Stappen, Lea Schumann, Benjamin Sertolli, Alice Baird, Benjamin Weigel, Erik Cambria, and Björn W. Schuller. Muse-toolbox: The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox, 2021.
14. Lukas Stappen, Jason Thies, Gerhard Hagerer, Björn W. Schuller, and Georg Groh. Graphtmt: Unsupervised graph-based topic modeling from video transcripts, 2021.
15. Kang Zhang, Yushui Geng, Jing Zhao, Jianxin Liu, and Wenxiao Li. Sentiment analysis of social media via multimodal feature fusion. *Symmetry*, 12(12), 2020.
16. N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, 161:124–133, 2018.
17. Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis, 2017.
18. Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7:1–1, 01 2015.
19. Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.
20. François Chollet. Xception: Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.
21. Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.
22. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.