Deep audio and video emotion detection

by

Masih AMINBEIDOKHTI

THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR A MASTER'S DEGREE
WITH THESIS IN INFORMATION TECHNOLOGY ENGINEERING
M.A.Sc.

MONTREAL, JULY 20, 2020

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Patrick Cardinal, Thesis Supervisor
Department of Software Engineering and IT, École de technologie supérieure

Mr. Marco Pedersoli, Thesis Co-supervisor
Department of Systems Engineering, École de technologie supérieure

Mr. Pierre Dumouchel, President of the Board of Examiners
Department of Software Engineering and IT, École de technologie supérieure

Ms. Sylvie Ratté, External Independent Examiner
Department of Software Engineering and IT, École de technologie supérieure

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON JUNE 30, 2020

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisors Professor Patrick Cardinal and Professor Marco Pedersoli for their patience, motivation, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis. It is an honor and a privilege to be their student.

Beside my advisors, I would like to thank the rest of my thesis committee: Professor Pierre Dumouchel and Professor Sylvie Ratté for their insightful comments and encouragement, but also for the hard questions which incented me to widen my research from various perspectives.

I thank my friends in the Livia Lab. In particular, I am grateful to Professor Eric Granger for his useful input at various points through the course of this research.

My sincere thanks also go to my managers at Teledyne Dalsa Stephane Dalton, and Bruno Menard, who encouraged me throughout my research career and gave access to the research facilities.

Last but not the least, I would like to thank my parents, my brother and sister and my friends Parsia Shahini, Sina Mirzaeifard and Mirmohmmad Saadati for providing me with unfailing support and continuous encouragement throughout writing this thesis. This accomplishment would not have been possible without them.

Détection d’émotions audio et vidéo profondes

Masih AMINBEIDOKHTI

RÉSUMÉ

Les êtres humains utilisent principalement deux méthodes de communication pour réussir leurs interactions sociales Cowie *et al.* (2001). La première, la plus évidente, la parole permet de transmettre explicitement les messages pour une grande variété de situations, l’autre, l’emotion humaine, est plus subtile et transmet des messages implicites sur les personnes eux-mêmes. Au cours des dernières années, avec l’avancement de la technologie, l’interprétation du premier canal est devenue de plus en plus facile. Par exemple, les systèmes de traitement de la parole peuvent facilement convertir la parole en texte ou les systèmes de vision par ordinateur peuvent détecter un visage dans une image. Le deuxième canal de communication n’est pas encore aussi bien maîtrisé. L’un des éléments clés de l’exploitation de la communication implicite est l’interprétation de l’émotion humaine, qui une tâche assez difficile, même pour les humains. Pour résoudre ce problème, les travaux antérieurs sur la reconnaissance des émotions se sont appuyés sur des caractéristiques faites à la main en incorporant la connaissance du domaine dans le système sous-jacent. Cependant, au cours des dernières années, les réseaux neuronaux profonds se sont avérés être des modèles efficaces pour s’attaquer à une variété de tâches.

Dans cette thèse, nous explorons les effets de l’application de méthodes d’apprentissages profonds à la tâche de la reconnaissance des émotions. Nous démontrons ces méthodes en montrant que les représentations obtenus sont plus riches et atteignent une précision supérieure à celle des techniques traditionnelles. De plus, nous démontrons que nos méthodes ne sont pas liées aux tâches de reconnaissance des émotions et que d’autres catégories de tâches telles que la classification multi-étiquettes peuvent aussi bénéficier de nos approches.

La première partie de ce travail se concentre uniquement sur la tâche de la reconnaissance des émotions par la vidéo en utilisant uniquement des entrées visuelles. Nous montrons qu’en exploitant l’information des aspects spatiaux et temporels des données d’entrée, nous pouvons obtenir des résultats prometteurs. Dans la deuxième partie, nous portons notre attention sur les données multimodales. Nous nous concentrons en particulier sur la manière de fusionner les données multimodales. Nous introduisons ensuite une nouvelle architecture qui incorpore les meilleures caractéristiques des architectures de fusion précoce et tardive.

Mots-clés : Informatique Affective, Reconnaissance des Emotions, Mécanismes d’Attention, Réseaux Neuronaux Convolutionnels, Fusion Multimodale

Deep audio and video emotion detection

Masih AMINBEIDOKHTI

ABSTRACT

Human beings rely on two capacities for successful social interactions Cowie *et al.* (2001). The first is more obvious and explicitly conveys messages which may be about anything or nothing and the other is more subtle and transmits implicit messages about the speakers themselves. In the last few years with the advancement of technology, interpretation of the first channel becomes more feasible. For instance, speech processing systems can easily convert a voice to text or computer vision systems can detect a face in an image. The second channel is still not as well understood. One of the key elements for exploiting the second one is interpreting human emotion. To solve the problem, earlier works in emotion recognition have relied on hand-crafted features by incorporating domain knowledge into the underlying system. However, in the last few years, deep neural networks have proven to be effective models for tackling a variety of tasks.

In this dissertation, we explore the effects of applying deep learning methods to the emotion recognition task. We demonstrate these methods by learning rich representations achieve superior accuracy over traditional techniques. Moreover, we demonstrate our methods are not bound to emotion recognition task and other classes of tasks such as multi-label classification can get benefit from our approaches.

The first part of this work focuses only on the task of video-based emotion recognition using only visual inputs. We show that by exploiting information from the spatial and temporal aspects of input data we can get promising results. In the second part, we move our attention to multimodal data. Particularly we focus on how to fuse multimodal data. We introduce a new architecture that incorporates the best features from early and late fusion architecture.

Keywords: Affective Computing, Emotion Recognition, Attention Mechanisms, Convolutional Neural Networks, Multimodal Fusion

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF ABREVIATIONS

HCI	Human-Computer Interaction
CCC	Concordance Correlation Coefficient
CNN	Convolutional Neural Network
DNN	Deep Neural Network
GAP	Global Average Precision
RNN	Recurrent Neural Network
MSE	Mean Square Error
ReLU	Rectified Linear Unit
LGBP-TOP	Local Gabor Binary Patterns from Three Orthogonal Planes
LPQ-TOP	Local Phase Quantization from Three Orthogonal Planes
LBP	Local Binary Patterns
SDM	Supervised Descent Method
SGD	Stocastic Gradient Descent
LSTM	Long Short Term Memory Networks
MLE	Maximum likelihood Estimation
MFCCs	Mel-Frequency Cepstrum Coefficients
HOF	Histogram of Oriented Gradients
CRF	Conditional Random Field
eGeMAPS	Extended Geneva Minimalistic Acoustic Feature Set

SVM	Support Vector Machine
LLD	Low-Level Descriptors
ECG	Electrocardiogram
EDA	Electrodermal Activity

INTRODUCTION

Motivation

Would it be helpful if the computer could have the ability to respond to your frustration and sadness? Or your smartphone comfort you if you got upset after getting a call? Or your smart home adjusts the mood around you after you have had a bad day at work — without being asked?

It may seem absurd, but the idea that humans and computers must learn to coexist and interact is nothing new. Human-Computer Interaction (HCI) surfaced in the 1980s with the advent of personal computing, just as personal computers in homes and offices in society-changing numbers. Since then as technological systems become more sophisticated, the need for including human interaction during the process has become more apparent. One way to improve this interaction is by designing accurate emotion recognition systems. Emotion is a key ingredient to human experience and widely affects human decision-making, yet technologists have largely ignored emotion and created an often frustrating experience for people.

Generally, an emotion recognition system is constructed from two major components as follow:

1. An interface between the system and the human for capturing the required information.
2. Processing the input information and make decision.

The first component is a bridge between a system and a human. Its role is to capture the human current state. This state comprises information such as facial expressions, gestures, voice and physiological behavior of humans to name a few. Each of the devices from which this information is extracted has its own sophistication. The details about these devices are out of the scope of this dissertation and we assume that we have access to the required information.

The second component is more of interest to us. Here we are given the raw inputs from the previous component and we have to manipulate them to achieve our goal. We call each of the input, a modality (e.g. audio signals) hence the system which processes them is a multimodal system. Based on (Baltrušaitis *et al.* (2018)) work, for constructing a successful multimodal emotion recognition system, the five following challenges need to be tackled:

1. **Representation** The first fundamental challenge is to find an efficient and effective way to represent each of the multimodal data. The encoded representations have to carry a sufficient amount of redundancy of multiple modalities (effectiveness), yet not too much that could lead to frustration of the system (efficiency). The heterogeneity of multimodal data makes it challenging to construct such representations. For example, language is often symbolic while audio and visual modalities will be represented as signals.

2. **Translation** The second challenge relates to the mapping between each modality. Because there may be multiple correct ways to represent a modality, it is challenging or sometimes not possible to find one perfect translation.

3. **Alignment** The third one is because there is a temporal dimension in the input data. The extracted information from the users is gathered sequentially through time, therefore, there is an extra dependency within sub-elements of each modality. This inter-relationship makes it more challenging to identify the direct intra-relations between sub-elements from two or more different modalities.

4. **Fusion** The fourth challenge addresses the ways multiple modalities combine together. The information coming from different modalities may have varying predictive power and noise topology, with possibly missing data in at least one of the modalities.

5. **Co-learning** The last challenge only accounts for learning systems. In the learning systems, the ultimate goal is to improve the performance of the system through data with

respect to some criterion Goodfellow *et al.* (2016). In the multimodal system, the learned knowledge should be transferable to other modalities so that learning from one modality can help a computational model trained on a different modality.

Throughout this dissertation, we try to address a few of these challenges using deep learning techniques. In chapter 2 we address the representation and the alignment challenge by introducing two components to the convolutional neural network architecture. In chapter 3 we try to tackle the fusion and co-learning challenge by presenting a new method to combine features from different modalities and train the whole system end to end.

Contributions

In this dissertation, we highlight how deep learning methods, when applied to emotion recognition, by learning rich representations achieve superior accuracy over traditional techniques (Wöllmer *et al.* (2013)). Moreover, we show that our methods are not bound to emotion recognition and in other tasks such as multi-label classification can be utilized to get better performance.

Our contributions are twofold. First, we consider the task of video-based emotion recognition using only visual inputs. We train a convolutional neural network (CNN) on a facial-expression database and present two simple strategies to improve the performance of emotion recognition in video sequences. We demonstrate visually that using our approach, one can detect the most important regions of an image that contributed to the particular emotion. We address the temporal aspect of the input data by introducing a simple method for aggregating frames over time. In contrast to more complex approaches such as recurrent neural networks for the temporal fusion of the data, our temporal pooling method achieves not only incorporate the relationship between each frame but also select most important frame of a video. This work is based on our publication (Aminbeidokhti *et al.* (2019)) which was selected as the best paper at ICIAR 2019 conference.

Second, we focus our attention on multimodal inputs. In particular, we design a multimodal fusion mechanism that incorporates best practices from well-known fusion models including early and late fusion. We also show that this method is not only useful for the emotion recognition task and can be applied to other problems as well.

Organization

The rest of this dissertation is organized as follows: Chapter 1 describes the relevant background on different representations of emotion as well as previous work on hand-crafted feature representations. We give a brief overview of deep learning architectures particularly convolutional neural network (CNN) and recurrent neural networks (RNN) and training methods associated with deep learning models. In chapter 2, we address the sequential data and demonstrate how reducing the complexity of the model in terms of architecture can lead to better performance and interpretability. In chapter 3, we utilize the existing approaches for multimodal fusion by applying a simple trick and demonstrate the improved performance on different tasks. Finally, chapter 4 concludes the dissertation and presents directions for future work.

CHAPTER 1

BACKGROUND

1.1 Emotion Representation

Models of emotion are typically divided into two main groups, namely discrete (or categorical) and dimensional (continuous) ones (Stevenson *et al.* (2007)) Discrete models are built around particular sets of emotional categories deemed fundamental and universal. Ekman et al. Ekman (1992) propose to classify the human facial expression resulting from an emotion into six basic classes (happiness, sadness, anger, disgust, surprise and fear). These emotions were selected because they have unambiguous meaning across cultures.

In contrast, dimensional models consider emotions to be composed out of several low-dimensional signals (mainly two or three). The two most commonly used dimensions are referred to as Valence (how positive or negative a subject appears), Arousal (represents the excitation rate). A third dimension has been added by Mehrabian (1996), the dominance, which depends on the degree of control exerted by a stimulus.

Russell (1980) suggests that all Ekman's emotions (Ekman (1992)) and compound emotions could be mapped in the circumplex model of affect. Furthermore, this two-dimensional approach allows a more accurate specification of the emotional state, especially by taking its intensity into account. This relationship is shown visually in Figure 1.1.

Several large databases of face images have been collected and annotated according to the emotional state of the person. The RECOLA database (Ringeval *et al.* (2013)) was recorded to study socio-affective behaviors from multimodal data in the context of remote collaborative work, for the development of computer-mediated communication tools. In addition to these recordings, 6 annotators measured emotion continuously on two dimensions: arousal and valence, as well as social behavior labels on five dimensions. SFEW Dhall *et al.* (2011), FER-13 Goodfellow *et al.* (2013) and RAF Li *et al.* (2017) propose images in the wild annotated in

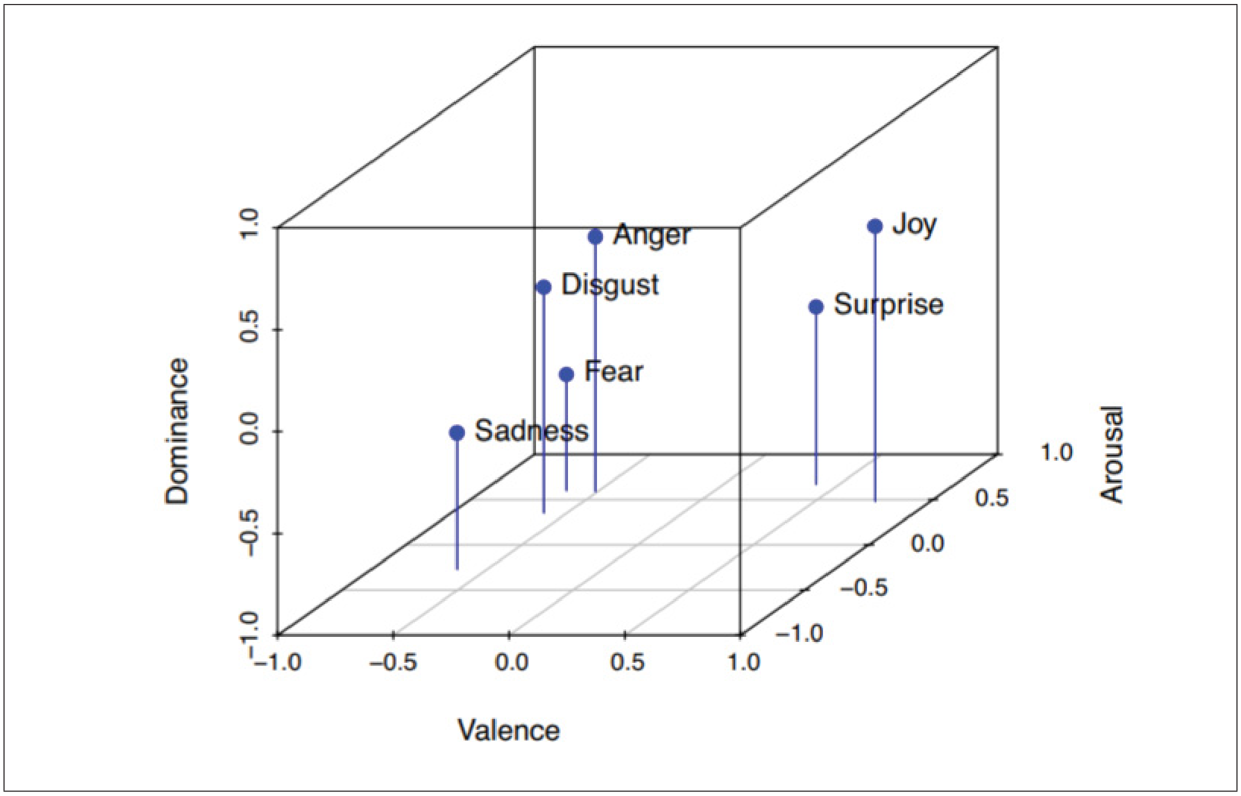

Figure 1.1 Circumplex of affect with the six basic emotions displayed (adapted from Buechel & Hahn (2016)). Affective space spanned by the Valence, Arousal and Dominance, together with the position of six Basic Emotions. Ratings are taken from Russell et al. (1977, p. 14)

basic emotions; AFEW Dhall *et al.* (2012) is a dynamic temporal facial expressions data corpus consisting of close to real-world environment extracted from movies annotated in discrete emotions.

1.2 Problem Statement

For defining the emotion recognition problem, we define the pattern recognition problem in general. A pattern is a representative signature of data by which we can take actions such as classifying the data into different categories (Bishop (2006)). Pattern recognition refers to the automatic discovery of regularities in data through the use of computer algorithms. Occasionally, a pattern is represented by a vector containing data features. Given the general definition

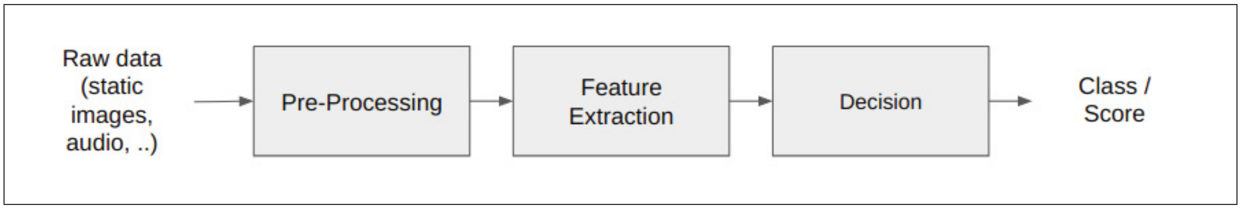

Figure 1.2 Basic steps of an emotion recognition system. For generality and avoiding targeting one specific task, we call the last step decision step

of pattern recognition, we can define the task of recognizing the emotion as discovering regularities in the psychological and the physical state of the human. In general, an emotion recognition system can be described by three fundamental steps, namely, Pre-Processing, Feature Extraction, and Decision. Figure 1.2 provides a general scheme for pattern recognition.

In emotion recognition tasks, we usually deal with raw data such as raw video inputs or static images. Given the emotional state of the human mind is expressed in different modes including facial, voice, gesture, posture, and biopotential signals, the raw data carries some unnecessary information. This extra information not only confuses the model but sometimes can lead to a nonoptimal result. In the Pre-Processing step, we extract useful cues from the raw data before applying further steps. For example, facial expression–based emotion recognition requires the extraction of the bounding box around the face. The Feature Extraction process involves transforming the raw features into some new space of variables where the pattern is expected to be easier to recognize. In general, the main goal of the Decision component is to map the Feature Extraction results in the designated output space. For instance, for the emotion recognition task, based on the previous section we saw that there exists more than one representation for emotions. For recognizing categorical emotion which contains a finite number of discrete categories, the Decision module simply classifies the extracted features into one of several object classes. Whereas for the dimensional representation of emotion, the Decision module does a regression task and outputs a continuous value.

Next, we define the task of emotion recognition in a learning algorithm framework. The focus of the emotion recognition task is on the problem of prediction: given a sample of training examples $(x_1, y_1), \ldots, (x_n, y_n)$ from $\mathbb{R}^d \to \{1, \ldots, K\}$, the model learns a predictor $h_\theta : \mathbb{R}^d \to \{1, \ldots, K\}$ defined by parameters θ, that is used to predict the label y of a new point x, unseen in training. The input feature, x_i, has D dimensions and for each label, y_i, there are K number of classes. The predictor h_n is commonly chosen from some function class $\mathscr{H}$, such as neural networks with a certain architecture, optimized with empirical risk minimization (ERM) and its variants. In ERM, the predictor is a function $h_\theta \in \mathscr{H}$ that minimizes the empirical (or training) risk $\frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i; \theta), y_i) + \Omega(\theta)$, where $\Omega(\theta)$ is a regularizer over model parameters, ℓ is a loss function; negative cross entropy loss $\ell(y', y) = -y \log y'$ in case of classification. Here y' is the predicted value from the model. The goal of machine learning is to find h_n that performs well on new data, unseen in training. To study performance on new data (known as generalization) we typically assume the training examples are sampled randomly from a probability distribution P over $\mathbb{R}^d \to \mathbb{R}$, and evaluate h_n on a new test example (x, y) drawn independently from P. The challenge stems from the mismatch between the goals of minimizing the empirical risk $\frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i; \theta), y_i) + \Omega(\theta)$ (the explicit goal of ERM algorithms, optimization) and minimizing the true (or test) risk $\mathbb{E}_{(x,y) \sim P_{true}}[\ell(h(x; \theta), y)]$ (the goal of machine learning). Intuitively, when the number of samples in the empirical risk is high, it can approximate well the true risk.

So far we defined the task of emotion recognition in terms of a classification problem. As we described in section 1.1, there is another model for mapping the emotional state which is continuous. I such cases we usually cast the task as a regression problem. The regression task has two major differences compared to the classification task. Firstly, the learning algorithm is asked to output a predictor $h_\theta : \mathbb{R}^d \to \mathbb{R}$ defined by parameters θ, that is used to predict the numerical value given some input. Secondly, the loss function that usually is minimized is the squared loss $\ell(y', y) = (y' - y)^2$. The same architectures and learning procedures can apply to both of the tasks.

1.3 Deep Learning

In this work, we use special class of machine learning models to approximate a predictor for the emotion recognition task called deep learning methods. We begin by describing the deep feedforward networks also known as feedforward neural networks, or multilayer perceptrons (MLPs) (Goodfellow *et al.* (2016)). These models are called feedforward because there are no feedback connections in which outputs of the model are fed back into itself and information flows directly from the input through the intermediate layers and the output. The goal of a feedforward network is to approximate some unknown function f^*. A feedforward network learns a mapping $y = f^*(x;\theta)$ by optimizing the value of the parameters θ that best describes the true underling function f. A deep neural network is usually represented as the composition of multiple nonlinear functions. Therefore, $f(x)$ can be expressed in the form of $f(x) = f^3(f^2(f^1(x)))$. Each of f^i is the layer of the neural network, in this case, f^1 is called the first layer of the network and so on. The depth of the network is given by the overall length of the chain. Consider some input $x \in \mathbb{R}^N$ and its corresponding output $h \in \mathbb{R}^M$, the equation for one layer of the network is defined as:

$$h = g(W^T x + b) \tag{1.1}$$

the network using weights matrix $W \in \mathbb{R}^{M \times N}$ and bias vector $b \in \mathbb{R}^M$ first linearly transform the input to the new representation and then apply nonlinear function g often rectified linear unit (ReLU). The premise of a feedforward network is at the end of the learning process each layer captures a meaningful aspect of the data and by combining them together the model can make a decision. Figure 1.3 provides a general view for one layer of a neural networks.

Among several architectures of a neural network, there are two specialized architectures which are more common in the field: Convolutional Neural Networks (CNNs) (LeCun *et al.* (1998)) and Recurrent Neural Networks (RNNs) (Rumelhart *et al.* (1988)).

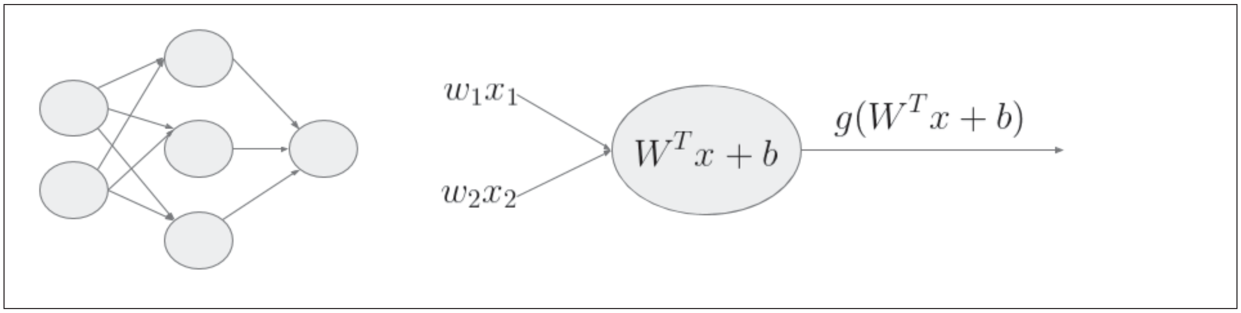

Figure 1.3 Left: A 2-layer Neural Network (one hidden layer of 3 neurons (or units) and one output layer with 1 neuron), and two inputs. Right: Mathematical model of one neuron from a hidden layer

Convolutional neural networks are a specialized kind of neural network which makes the network more suitable for tasks specific to the grid-like topology. Examples include images that can be seen as a 2-D grid of pixels and also time-series data which can be thought of as 1-D data acquired through time intervals. A convolutional network uses a convolution operation in place of the affine transformation in neural network layers. Instead of regular matrix weights for each layers, a convolutional network takes advantage of the grid-like topology of the data and defines sets of weights as a filter (or a kernel), that is convolved with the input. The way the kernels are defined is especially important. It turns out that we can dramatically reduce the number of parameters by making two reasonable assumptions. Firstly, parameter sharing which states that a feature detector (such as vertical edge in case of image data) that is useful in one part of an image is probably useful in other parts of the image as well. Secondly, sparse connectivity. This is accomplished by defining the kernel smaller than the input. As a result, in each layer, the output value in the hidden layer depends only on a few input values. Reducing the number of parameters helps in training with smaller training sets, and also it is less prone to overfitting. Figure 1.4 provides a general architecture of a convolutional neural networks for an image input.

Recurrent neural networks are a family of neural networks for processing sequential data. Consider a sequence of inputs of length T $(x^1,\ldots,x^T)$. RNNs have a state at each time point t, h^t, which captures all of the information of previous inputs $(x^1,\ldots,x^t)$. Then, when considering

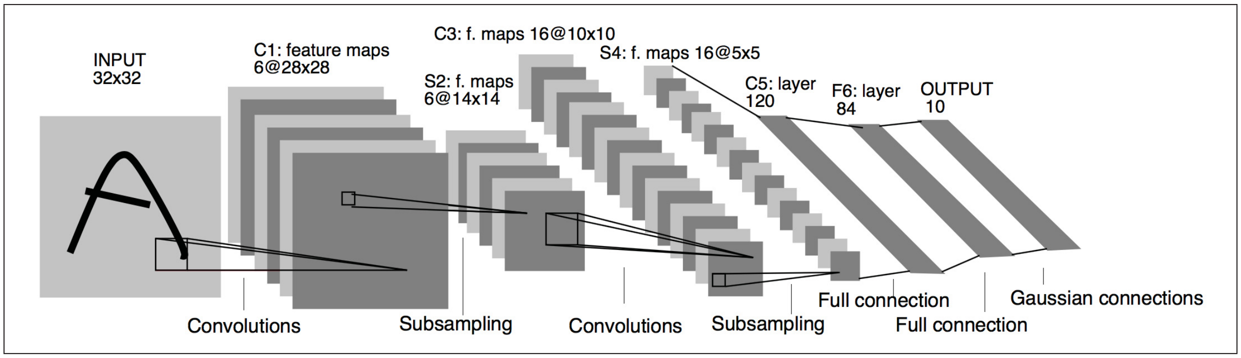

Figure 1.4 LeNet-5 architecture as published in the original paper Lecun et al. (1998, p. 7)

the input at the next time point, x^{t+1} , a new state, h^{t+1} , is computed using the new input x^{t+1} and the previous state h^t . At each time point t, the hidden state h^t can be used to compute an output o^t. Figure 1.5 provides a general architecture of a recurrent neural network. One drawback of the RNNs networks is the problem of the long-term dependencies (Bengio *et al.* (1994)). In cases where the gap between the relevant information and the place that it is needed is very large, RNNs become unable to learn how to connect the information. Long Short Term Memory networks (LSTM) Hochreiter & Schmidhuber (1997) has been proposed to mitigate the problem of long-term dependencies. LSTM units include a memory cell that can maintain information in memory for long periods of time. The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates.

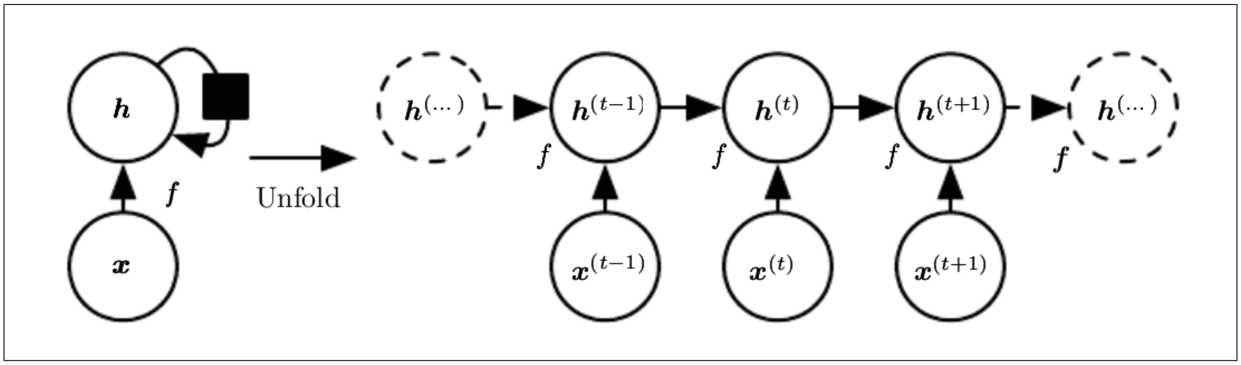

Figure 1.5 A recurrent neural network taken from Goodfellow *et al.* (2016). This recurrent network process the information from input x by incorporating it into the state h that is passed through time. On the left you can see the unfolded version of the network through time

One common criticism of deep learning methods is its requirement of a very large amount of data to perform better than other techniques. In practice, it is relatively rare to have a dataset of sufficient size for every task. Fortunately, there is a technique called transfer learning that helps to address the problem to some extent. The way that usually deep networks are used is to initialize the parameters of the network to some random number; usually, these initializations follow the carefully constructed procedures (Glorot & Bengio (2010)). It turns out the feature that the network learns during the training process can sometimes be used for other tasks as well. For instance, Zeiler & Fergus (2014) trained a large convolutional neural network model for image classification task and demonstrated that the learned features to be far from random, uninterpretable patterns. Figure 1.6 provides visualization on some of the learned features. As a result of these findings, it is common to pretrain a deep network, particularly CNN, on a very large dataset (e.g. ImageNet Deng *et al.* (2009), which contains 1.2 million images with 1000 categories), and then use the network either as an initialization or a fixed feature extractor for the task of interest. Transfer learning is not only useful for computer vision tasks, natural language processing tasks which usually models based on RNNs, are also take advantage of pretrained features from other datasets (Mikolov *et al.* (2013)).

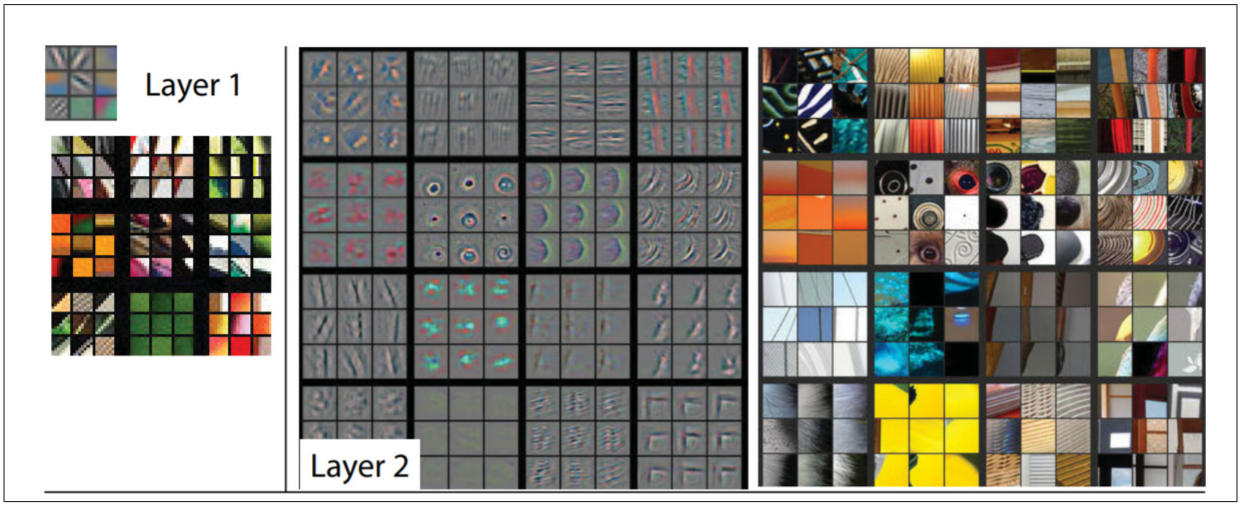

Figure 1.6 Visualization of features in a fully trained model taken from Zeiler et al. (2014, p. 4). For a random subset of feature maps, show the top 9 activations from the validation set. Each 9 activations projected back to pixel space using the deconvolutional method. The figure also shows the corresponding image patches for each feature map

1.4 Network Training: Optimization

After selecting the model based on the problem domain, the next step is to identify the parameter of the model by minimizing a certain cost function based on the training data. One important note is that compare to the traditional optimization algorithms, machine learning usually acts indirectly. In most machine learning scenarios, the final goal is to maximize the performance of the model on test data, by reducing the cost function as a proxy. Whereas in pure optimization the final goal is finding the minimum of the cost function itself. There are many estimation methods for identifying the model parameters. Maximum likelihood estimation (MLE) is one of the well-known principles for solving machine learning tasks. In Maximum Likelihood Estimation (MLE), we wish to make the model distribution to match the empirical distribution defined by training data.

Consider a set of m examples $\mathbb{X} = \{x^1, \ldots, x^m\}$ drawn independently from the true but unknown data-generating distribution $p_{data}(x)$. Let $p_{model}(x;\theta)$ be the probability distribution defined by neural network parameterized by θ. The MLE for θ is then defined as:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{m} p_{model}(x^i;\theta) \tag{1.2}$$

Because logarithm is monotonous therefore does not change its argmax, we can further simplify this product by taking the logarithm from the likelihood function and convert it to a sum:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{m} \log p_{model}(x^i;\theta) \tag{1.3}$$

Based on the MLE, we iteratively update the parameters θ using gradient descent:

$$\theta_{t+1} = \theta_t - \eta \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta \log p_{model}(x^i;\theta) \tag{1.4}$$

where η denotes the learning rate. Gradient descent is an iterative algorithm, that starts from a random point on a function and travels down its slope in steps (learning rate) until it reaches the lowest point of that function. Computing this summation exactly is very expensive because it requires evaluating the model on every example in the entire dataset. In practice, we can compute these summation by randomly sampling a small number of examples from the dataset, then taking the average over only those examples. This method is called minibatch stochastic gradient descent (SGD). In the context of deep learning, it has been observed in practice that when using a larger batch there is a degradation in the quality of the model, as measured by its ability to generalize (Keskar *et al.* (2016)). Based on the evidence they show that large-batch methods tend to converge to sharp minimizers of the training and testing functions and therefore sharp minima lead to poorer generalization.

1.5 Related Work

1.5.1 Emotion Recognition

The majority of traditional methods have used handcrafted features in emotion recognition systems. For instance, the spectrum, energy, and Mel-Frequency Cepstrum Coefficients (MFCCs) are widely exploited to encode the cues in the audio modality, while Local Binary Patterns (LBP) (Shan *et al.* (2009)), Histogram of Oriented Gradients (HOG) (Chen *et al.* (2014)), and Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) (Zhao & Pietikainen (2007)) are the representative ones to describe the facial characteristics in the video modality. They are designed based on the knowledge of human beings in the specific domain.

Since CNNs enjoy great success in image classification (Krizhevsky *et al.* (2012)) they have also been applied to emotion recognition tasks. In Kim *et al.* (2015) using an ensemble of multiple deep convolutional neural networks won the EmotiW 2015 challenge and surpassed the baseline with significant gains. If enough data is available, deep neural networks can generate more powerful features than hand-crafted methods.

Recurrent neural networks (RNN), particularly long short-term memory (LSTM) (Hochreiter & Schmidhuber (1997)), one of the state-of-art sequence modeling techniques, has also been applied in emotion recognition. Wöllmer *et al.* (2013) presented a fully automatic audiovisual recognition approach based on LSTM-RNN modeling word-level audio and visual features. Wöllmer *et al.* (2013) showed that compared to other models such as Support Vector Machine (SVM) (Schuller *et al.* (2011)) and Conditional Random Field (CRF) (Ramirez *et al.* (2011)), LSTM achieved a higher prediction quality due to its capability of modelling long range temporal dependencies.

Meanwhile, other techniques based on deep neural networks (DNN) have been successfully applied in extracting emotional related features. For speech-based emotion recognition, Xia *et al.* (2014) adds gender information to train auto-encoders and extracts the hidden layer as audio features to improves the unweighted emotion recognition accuracy. In Huang *et al.* (2014), convolutional neural networks (CNN) are applied in speech emotion task with novel loss functions to extract features.

Multi-task learning may also improve dimensional emotion prediction performance due to the correlation between arousal and valence. In Ringeval *et al.* (2015), two types of multi-task learning are introduced: one by learning each rater's individual track and the other by learning both dimensions simultaneously. Although it did not help for the audio feature based system, it improved the visual feature based system performance significantly.

Previous studies Ringeval *et al.* (2015) also provide some common insights:

1. It's highly agreed that arousal is learned more easily than valence. One reason is that the perception of arousal is more universal than is the perception of valence;
2. Audio modality is suitable for arousal prediction, but much less accurate for valence prediction;
3. Valence appears to be more stable than arousal using facial expression modality. Bio-signals are also good for valence assessment;

1.5.2 Attention and Sequence Modeling

The intuition behind attention is, to some extent, motivated by human visual attention. Human visual attention enables us to focus on a certain region with “high resolution” while perceiving the surrounding image in “low resolution”, and then adjust the focal point or do the inference accordingly. Similarly, during reading, we can explain the relationship between words in a sentence. In the same way attention in deep learning by providing a vector of importance weights, allows the model to focus on the important part of the input.

Attention was first introduced in the context of neural machine translation. Before that, the dominant way to process a sentence for machine translation was using the vanilla LSTM network. The problem with this method is, LSTM summarizes the whole input sequence into the fixed-length context vector with a relatively small dimension, therefore the context vector was not a good representation of the long sentences. Bahdanau *et al.* (2014) introduces the attention mechanism to mitigate this problem. The idea is, in the encoder-decoder model Sutskever *et al.* (2014) instead of having a single vector as the representation of the whole sequence, we can train a network to output the weighted sum over the recurrent network’s context vectors over the whole input sequence. Figure 1.7 provides a graphical illustration of the proposed attention model.

With the success of attention in neural machine translation, people started exploring attention in different domains such as image captioning (Xu *et al.* (2015)), visual question answering (Chen *et al.* (2015)). Various forms of attention emerge as well. Vaswani *et al.* (2017) introduced a self-attention network that completely replaces recurrent networks with the attention mechanism. Jetley *et al.* (2018) deployed attention as a separate end-to-end-trainable module for convolutional neural network architectures. They proposed a model that highlights where and in what proportion a network attends to different regions of the input image for the task of classification. Parkhi *et al.* (2015), Cheng *et al.* (2016) introduced self-attention, also called intra-attention, that calculates the response at a position in a sequence by attending to all positions within the same sequence. Parmar *et al.* (2018) proposed an Image Trans-

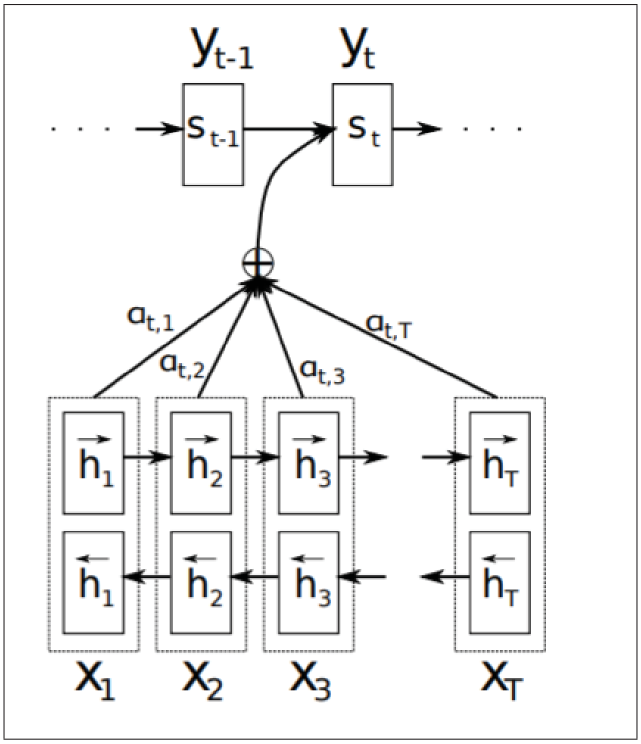

Figure 1.7 Additive attention mechanism. Taken from Bahdanau et al. (2014, p. 3)

former model to add self-attention into an autoregressive model for image generation. Wang *et al.* (2018) formalized self-attention as a non-local operation to model the spatial-temporal dependencies in video sequences. Zhang *et al.* (2018) proposed attention based on fully convolutional neural network for audio emotion recognition which helped the model to focus on the emotion-relevant regions in speech spectrogram.

For capturing temporal dependencies between video frames in video classification, LSTM have been frequently used in the literature (Chen & Jin (2015), Liu *et al.* (2018a), Lu *et al.* (2018). Sharma *et al.* (2015)) proposed a soft attention LSTM model to selectively focus on parts of the video frames and classify videos after taking a few glimpses. However, the accuracy on video classification with these RNN-based methods were the same or worse than simpler methods, which may indicate that long-term temporal interactions are not crucial for video classification. Karpathy *et al.* (2014) explored multiple approaches based on pooling local spatio-temporal features extracted by CNNs from video frames. However, their models display only a modest improvement compared to single-frame models. In the EmotiW 2017 challenge Dhall *et al.* (2017), Knyazev *et al.* (2018) exploited several aggregation functions (e.g., mean, standard deviation) allowing the incorporation of temporal features. Long *et al.* (2018) proposed a new architecture based on attention clusters with a shifting operation to explore the potential of pure attention networks to integrate local feature sets for video classification.

1.5.3 Multimodal Learning

Modalities fusion is another important issue in emotion recognition. A conventional multimodal model receives as input two or more modalities that describe a particular concept. The most common multimodal sources are video, audio, images and text. Multimodal combination seeks to generate a single representation that makes easier automatic analysis tasks when building classifiers or other predictors. There are different methods for combining multiple modalities. One of the focuses is on building the best possible fusion models e.g. by finding at which depths the unimodal layers should be fused (typically early vs. late fusion). In early fusion, after extracting the features from each unimodal systems, we usually concatenate different modalities and feed it to the new model for the final prediction. Late fusion is often defined by the combination of the final scores of each unimodal branch (Lahat *et al.* (2015)). Late fusion is often defined by the combination of the final scores of each unimodal branch. For audiovisual emotion recognition, Brady *et al.* (2016) achieved the best result on AVEC16 challenge (Valstar *et al.* (2016)) by using late fusion to combine the estimates from individual modalities and exploit the time-series nature of the data, while Chen *et al.* (2017) followed an early fusion hard-gated approach for textual-visual sentiment analysis and achieved state-of-the-art sentiment classification and regression results on Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis (CMU-MOSI) dataset (Zadeh *et al.* (2016)). To take advantage for both worlds, Vielzeuf *et al.* (2018) introduced a central network linking the unimodal networks. Pérez-Rúa *et al.* (2019) reformulated the problem as a network architure search and proposed multimodal search space and exploration algorithm to solve the task in an efficient yet effective manner. In Arevalo *et al.* (2017) the author proposed the Gated Multimodal Unit model whose purpose is to find an intermediate representation based on an expert network for a given input.

However, more fine-grained multimodal fusion models have been extensively explored and validated in visual and language multimodal learning. Shan *et al.* (2007) have studied synchronization between multimodal cues to support feature-level fusion and report greater overall accuracy compared to decision-level fusion. Bilinear models were first introduced by Tenenbaum & Freeman (2000) to separate style and content. Bilinear pooling computes the outer product between two vectors, which allows, in contrast to element-wise product, a multiplicative interaction between all elements of both vectors. As the outer product is typically infeasible due to its high dimensionality, Fukui *et al.* (2016), Kim *et al.* (2016), Yu *et al.* (2017) improved bilinear pooling method to overcome the issue.

CHAPTER 2

EMOTION RECOGNITION WITH SPATIAL ATTENTION AND TEMPORAL SOFTMAX POOLING

2.1 Introduction

Designing a system capable of encoding discriminant features for video-based emotion recognition is challenging because the appearance of faces may vary considerably according to the specific subject, capture conditions (pose, illumination, blur), and sensors. It is difficult to encode common and discriminant spatio-temporal features of emotions while suppressing these context and subject-specific facial variations.

Recently, emotion recognition has attracted attention from the computer vision community because state-of-the-art methods are finally providing results that are comparable with human performance. Thus, these methods are now becoming more reliable, are beginning to be deployed in real-world applications (Cowie *et al.* (2001)). However, at this point, it is not yet clear what is the right recipe of success in terms of machine learning architectures. Several state-of-the-art methods (Knyazev *et al.* (2018), Liu *et al.* (2018a)) originating from challenges in which multiple teams provide results on the same benchmark without having access training-set annotations. Although these challenges measure improvements in the field. one a drawback of challenges is that result focuses mostly on final accuracy of approaches, without taking into account other factors such as their computational cost, architectural complexity, quantity of hyper-parameters to tune, versatility, generality of the approach, etc. As a consequence, there is no clear cost-benefit analysis for component appearing in top-performing methods and often represent complex deep learning architectures.

In this chapter, we aim to shed some light on these issues by proposing a simple approach for emotion recognition that i) is based on the very well-known VGG16 network which is pre-trained on face images; ii) has a very simple yet performing mechanism to aggregate temporal information; and iii) uses an attention model to select which part of the face is the most important to recognize a certain emotion. For the selection of the approach to use, we show that a basic convolutional neural network such as VGG can perform as well or even better than more complex models when pre-trained on clean data. For temporal aggregation, we show that softmax pooling is an excellent way to select information from different frames because it is a generalization of max and average pooling. Additionally, in contrast to more complex techniques (e.g. attention), it does not require additional sub-networks and therefore additional parameters to train, which can easily lead to overfitting when dealing with relatively small datasets, a common problem in affect computing. Finally, we show that for the selection of the most discriminative parts of a face for recognizing an emotion, an attention mechanism is necessary to improve performance. For doing that, we built a small network with multiple attention heads (Lin *et al.* (2017)) that can simultaneously focus on different parts of a human face.

The rest of the chapter is organized as follows. In the next section our methods based on spatial attention and temporal softmax are presented. Finally, in our experimental evaluation, we show the importance of our three system components and compare them with other similar approaches.

2.2 Proposed Model

We now describe our method based on spatial attention and temporal softmax pooling for the task of emotion recognition in videos. We broadly consider three major parts: local feature extraction, local feature aggregation and global feature classification. The local feature extractor takes the video frame as its input and produces local features. Using the local features, the multi-head attention network computes the weight importance of each local image feature. The aggregated representation is computed by multiplying multi-head attention output and the local image features. This representation is then propagates through temporal softmax pooling to extract global features over the entire video. The overall model architecture is shown in Figure 2.1. The local feature extraction uses a pre-trained CNN, the spatial feature aggregation is implemented using an attention network, and the temporal feature classification uses a softmax

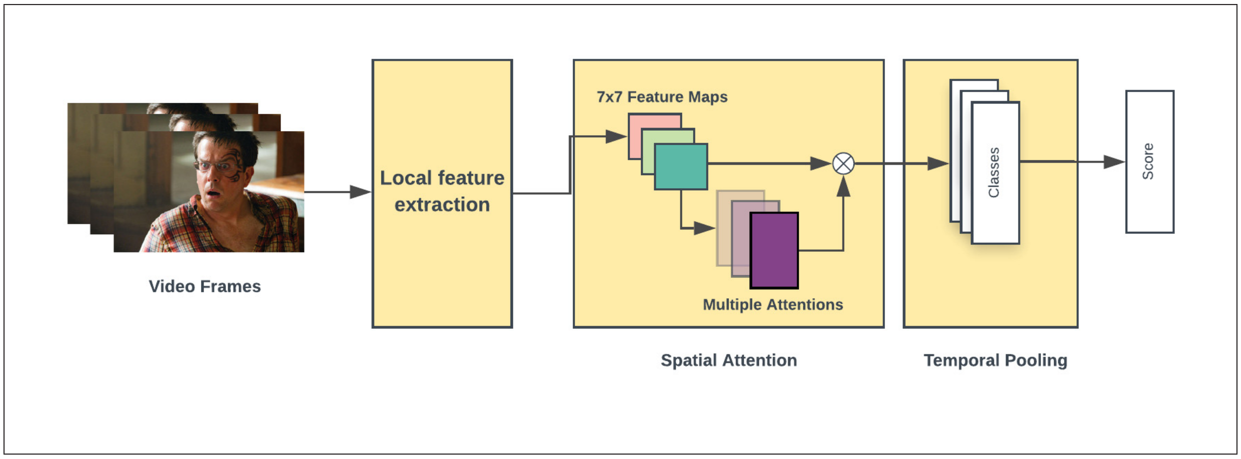

Figure 2.1 The Overview of the model pipeline

pooling layer. Given a video sample $\mathbf{S}_i$ and its associated emotion $y_i \in \mathbb{R}^E$, we represent the video as a sequence of F frames $[\mathbf{X}_{0,i}, \mathbf{X}_{1,i}, .., \mathbf{X}_{F,i}]$ of size $W \times H \times 3$.

2.2.1 Local Feature Extraction

We use the VGG-16 architecture with the pre-trained VGG-Face Model (Parkhi *et al.* (2015)) for extracting an independent description of a face on each frame in the video. For a detailed procedure of face extraction, see the experimental results in section 2.3. For a given frame $\mathbf{X}$ of a video, we consider the feature map produced by the last convolutional layer of the network as representation. This feature map has spatial resolution of $L = H/16 \times W/16$ and D channels. We discard the spatial resolution and reshape the feature map as a matrix $\mathbf{R}$ composed of L D-dimensional local descriptors (row vectors).

$$\mathbf{R} = VGG_{16}(\mathbf{X}) \tag{2.1}$$

These descriptors will be associated to a corresponding weight and used for the attention mechanism.

2.2.2 Spatial Attention

For the spatial attention we rely on the self-attention mechanism (Vaswani *et al.* (2017)), which aggregates a set of local frame descriptors $\mathbf{R}$ into a single weighted sum v that summarizes the most important regions of a given video frame:

$$v = a\mathbf{R}, \tag{2.2}$$

where a is a row vector of dimension L, which defines the importance of each frame region. The weights a are generated by a two-layers fully connected network that associates each local feature (row of $\mathbf{R}$) to a corresponding weight:

$$a = softmax(w_{s2} tanh(\mathbf{W}_{s1}\mathbf{R}^{\top})). \tag{2.3}$$

$\mathbf{W}_{s1}$ is a weight matrix of learned parameters with shape $U \times D$ and w_{s2} is a vector of parameters with size U. The softmax function ensures that the computed weights are normalized, i.e. sum up to 1.

This vector representation usually focuses on a specific region in the facial feature, like the mouth. However, it is possible that multiple regions of the face contain different type of information that can be combined to obtain a better idea of the person emotional state. Based on Lin *et al.* (2017), in order to represent the overall emotion of the facial feature, we need multiple attention units that focus on different parts of the image. For doing that, we transform w_{s2} into a matrix $\mathbf{W}_{s2}$ of size $R \times L$, in which every row represents a different attention:

$$\mathbf{A} = softmax(\mathbf{W}_{s2} tanh(\mathbf{W}_{s1}\mathbf{R}^{\top})). \tag{2.4}$$

Here the softmax is performed along the second dimension of its input. In the case of multiple attention units, the aggregated vector v becomes a matrix $D \times N$ in which each row represents a different attention. This matrix will be then flattened back to a vector v by concatenating the rows in a single vector. Thus, with this approach, a video is now represented as a $F \times (ND)$ matrix $\mathbf{V}$ in which every row is the attention based description of a video frame. To reduce the possible overfitting of the multiple attentions, similarly to Lin *et al.* (2017) we regularize $\mathbf{A}$ by computing L^2 norm of matrix $(\mathbf{A}\mathbf{A}^\top - I)$ and adding it to the final loss. This enforces diversity among the attentions and shown to be very important in our experiments for good results.

2.2.3 Temporal Pooling

After extracting the local features and aggregating them using the attention mechanism for each individual frame, we have to take into account frame features over the whole video. As the length of a video can be different for each example, we need an approach that support different input lengths. The most commonly used approaches are average and max pooling; however, these techniques assume that every frame of the video has the same importance in the final decision (average pooling) or that only a single frame is considered as a general representation of the video (max pooling). In order to use the best of both techniques, we use an aggregation based on softmax. In practice, instead of performing the classical softmax on the class scores, to transform them in probabilities to be used with cross-entropy loss, we compute the softmax on the class probabilities and the video frames jointly. Given a video sample $\mathbf{S}$, after feature extraction and spatial attention we obtain a matrix $\mathbf{V}$ in which each row represents the features of a frame. These features are converted into class scores thorough a final fully connected layer $\mathbf{O} = \mathbf{W}_{sm}\mathbf{V}$. In this way $\mathbf{O}$ is a $F \times E$ matrix in which an element $o_{c,f}$ is the score for class c of the frame f. We then transform the scores over frames and classes in probabilities with a softmax:

$$p(c, f|\mathbf{S}) = \frac{exp(o_{c,f})}{\sum_{j,k} exp(o_{j,k})}. \tag{2.5}$$

Figure 2.2 Sample images from AFEW database

In this way, we obtain a joint probability on class c and frame f. From this, we can marginalize over frames $p(c|\mathbf{S}) = \sum_f p(c,f|S)$ and obtain a classification score that can be used in the training process using cross-entropy loss:

$$\mathscr{L}_{CE} = \sum_i -\log(p(y_i|\mathbf{S}_i)). \tag{2.6}$$

On the other hand, the same representation can be marginalized over classes $p(f|\mathbf{S}) = \sum_c p(c,f|\mathbf{S})$. In this case, it will give us information about the most important frames of a given video. This mechanism looks very similar to attention, but it has the advantage to not require an additional network to compute the attention weights. This can be important in cases for which the training data is limited and adding a sub-network with additional parameters to learn could lead to overfitting. In this case, the weight associated to each frame and each class are computed as a softmax of the score obtained. Figure 2.3(d) shows the temporal importance of the selected frames. To make those values more meaningful they have been re-normalized between 0 and 100

2.3 Experiments

2.3.1 Data Preparation

We evaluate our models based on AFEW database, which is used in the audio-video sub-challenge of the EmotiW (Dhall *et al.* (2012)). AFEW is collected from movies and TV reality shows, which contains 773 video clips for training and 383 for validation with 7 various emotions: anger, disgust, fear, happiness, sadness, surprise and neutral. Sample images from the dataset are illustrated in Fig 2.2. We extract the frame faces using the dlib (King (2009)) detector for achieving effective facial images. Then faces are aligned to a frontal position and stored in a resolution of 256×256 pixels, ready to be passed to VGG16.

2.3.2 Training Details

Although hyperparameters selection is an essential part of the learning task, it is resource-intensive. We selected our hyperparameters according to the (Liu *et al.* (2018a)), and gradually changed them to meet our criterion. To overcome overfitting during training we sampled 16 random frames form the video clips. Before feeding the facial image to the network we applied data augmentation: flipping, mirroring and random cropping of the original image. We set weight decay penalty to 0.00005 and use SGD with momentum and warm restart (Loshchilov & Hutter (2016)) as optimization algorithm. All models are fine-tuned for 30 epochs, but we use a learning rate of 0.00001 for the backbone CNN parameters and 0.1 for the rest of the parameters.

2.3.3 Spatial Attention

Table 2.1 reports the accuracy (third column) based on the AFEW validation dataset. We compare our softmax-based temporal pooling with different configurations of attention by varying the number of attention units (second column) and the used regularization (third column) as described in sec. 2.3.2. Using just one attention unit does not helps to improve the overall per-

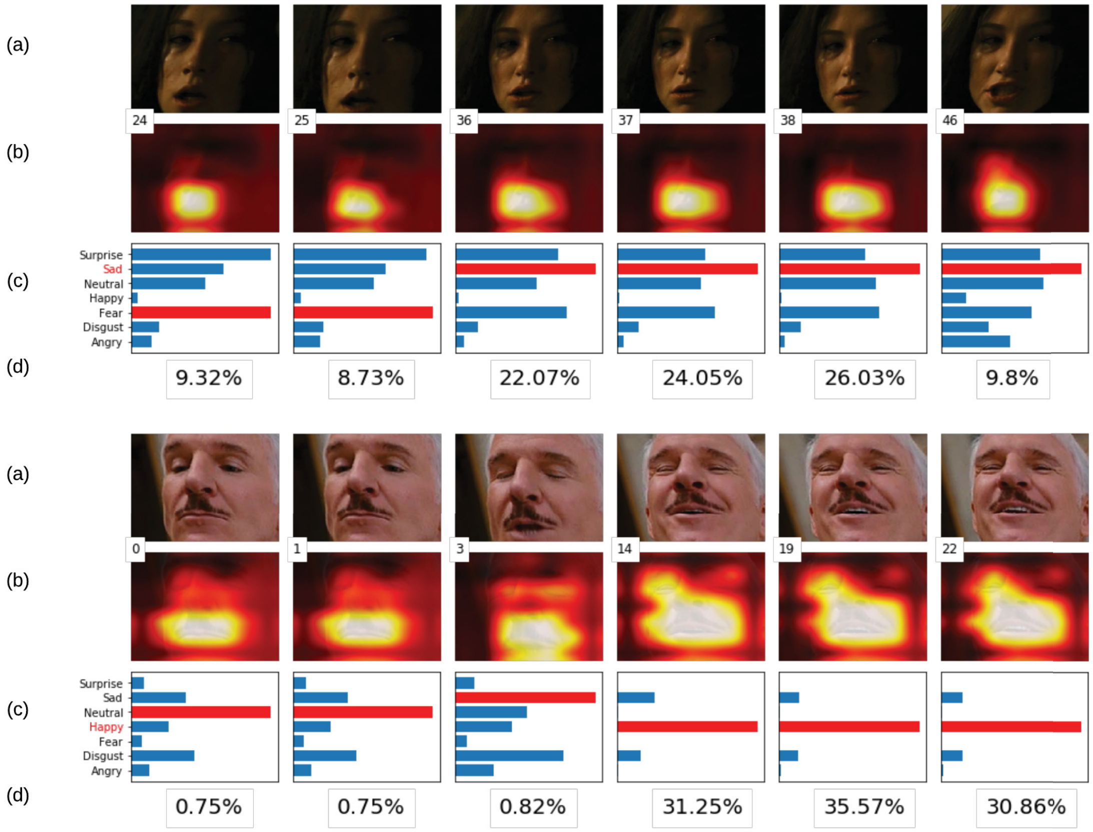

Figure 2.3 Video frames for a few time-steps for an example of sadness and anger. (a) Original video frames. (b) Image regions attended by the spatial attention mechanism. (c) Emotion probability for each frame. The red bar shows the selected emotion. (d) Temporal importance for selected frames

formance. This is probably due to the fact that a single attention usually focuses on a specific part of the face, like a mouth. However, there can be multiple regions in a face that together forms the overall emotion of the person. Thus, we evaluate our model with 2 and 4 attention units. The best results are obtained with 2 attention units and a strong regularization that enforces the models to focus on different parts of the face. We observe that, adding more than two attention units do not improve the overall performance. This is probably due overfitting. We also compare with our re-implementation of cluster attention with shifting operation (SHIFT) (Long *et al.* (2018)), but results are lower than our approach. Figure 2.3(b) demonstrate image regions attended by the spatial attention mechanism. Brighter regions represent the most important parts of the face to recognize a certain emotion for the attention. We show that the throughout the frames, model not only captures the mouth, which in this case is the most important part for detecting the emotion but also in the last three frames focuses on the eyes as well.

Table 2.1 Performance report of the proposed spatial attention block. TP is our temporal softmax, while SA is the spatial attention

Model	**# Att.**	**Reg.**	**ACC**
$\text{VGG}_{16}+TP$	-	-	46.4%
$\text{VGG}_{16}+TP+SHIFT$	2	-	45.0%
$\text{VGG}_{16}+TP+SA$	1	0	47.6%
$\text{VGG}_{16}+TP+SA$	2	0.1	48.9%
$\text{VGG}_{16}+TP+SA$	2	1	**49.0**%
$\text{VGG}_{16}+TP+SA$	4	0.1	48.3%
$\text{VGG}_{16}+TP+SA$	4	1	48.6%

2.3.4 Temporal Pooling

In this section we compare the performance of different kind of temporal pooling. The simplest approach is to consider each video sample i frame independent of the others $p(c|\mathbf{S}_i) = \prod_f p(c,f|\mathbf{X}_{f,i})$ and associating the emotion class c of a video to all its frames. In this case the loss becomes:

$$\mathscr{L}_{CE} = \sum_i -\log(p(c|\mathbf{S}_i)) = \sum_i \sum_f -log(p(c,f|\mathbf{X}_{f,i})), \tag{2.7}$$

which can be computed independently from each frame. In this way we can avoid keeping all the frames of a video in memory at the same. However, assuming that each frame of the same video is independent of the others is a very restrictive assumption and it is in contrast with the common assumption used in learning of identically independently distributed samples. We notice that this approach is equivalent to perform an average pooling (VGG+AVG) on the scoring function before the softmax normalization. This can explain the lower performance of this kind of pooling.

In Table 2.2 we report results of different pooling approaches. We report results for Liu *et al.* (2018a) in which they use VGG16 with an LSTM model to aggregate frames (LSTM). We compare it with a VGG16 model trained with average pooling (AVG) and our softmax temporal pooling (TP). Finally we also consider the model with our temporal pooling and spatial attention (TP+SP). Our temporal pooling performs slightly better than a more complex approach based on a recurrent network that keeps the memory of the past frames. We can further obtain a gain of almost 3 points by adding a spatial attention block (VGG+TP+SA). It is interesting to note that our model, even if not explicitly reasoning on the temporal scale, (i.e every frame is still computed independently, but then the scores are normalized with the softmax) outperforms a model based on LSTM, a state-of-the-art recurrent neural network. This suggest that for emotion recognition it is not really important the sequentiality of the facial postures, but the presence of certain key patterns.

Table 2.2 Performance report of the softmax temporal aggregation block

Model	**ACC**
$\text{VGG}_{16} + LSTM$ Liu *et al.* (2018a)	46.2%
$\text{VGG}_{16} + AVG$	46.0%
$\text{VGG}_{16} + TP$	46.4%
$\text{VGG}_{16} + TP + SA$	**49.0**%

2.4 Conclusion

In this chapter, we have presented two simple strategies to improve the performance of emotion recognition in video sequences. In contrast to previous approaches using recurrent neural networks for the temporal fusion of the data, in this chapter we have shown that a simple softmax pooling over the emotion probabilities, that selects the most important frames of a video, can lead to promising results. Also, to obtain more reliable results, instead of fusing multiple sources of information or multiple learning models (e.g. CNN+C3D), we have used a multi-attention mechanism to spatially select the most important regions of an image. For future work we plan to use similar techniques to integrate other sources of information such as audio.

CHAPTER 3

AUGMENTING ENSEMBLE OF EARLY AND LATE FUSION USING CONTEXT GATING

3.1 Introduction

Our perception of the world is through multiple senses. We see objects, hear sounds, feel texture, smell odors, and taste flavors. Modality refers to the way in which something happens or is experienced and multimodality refers to the fact that the same real-world concept can be described by different views or data types.

Multimodal approaches are key elements for many application domains, including video classification (Liu *et al.* (2018b)), emotion recognition (Brady *et al.* (2016)), visual question answering (Fukui *et al.* (2016)) and action recognition (Simonyan & Zisserman (2014)), to name a few. The main motivation for such approaches is to extract and combine relevant information from the different sources and hence make better decisions than using only one.

Existing multimodal approaches usually have three stages: (1) representing the inputs as semantically meaningful features; (2) combining these multimodal features; (3) using the integrated features to learn the model and to predict the best decision with respect to the task. Given the effectiveness and flexibility of deep neural networks (DNNs), many existing approaches model the three stages in one DNN model and train the model in an end-to-end fashion through back-propagation. The focus of this chapter is particularly on the second stage, i.e multimodal fusion.

Most existing approaches simply use linear models for multimodal feature fusion (e.g., concatenation or element-wise addition) at one specific layer (e.g., early layer or latest layer) to integrate multiple inputs feature. Since multimodal features distributions may vary dramatically, the integrated representations obtained by such linear models may not be necessarily the most optimal way to solve a given multimodal problem. In this chapter, we argue that considering features combination from all the hidden layers of independent modalities could potentially increase performance compared to only using a single combination of late (or early) features. Hence, this work tackles the problem of finding good ways to integrate multimodal features to better exploit the information embedded at different layers in deep learning models with respect to the task in hand.

In order to achieve this goal, we propose a new architecture that simultaneously trains a late and early fusion augmented with interconnected hidden layers. Our method not only gains from the flexibility of late fusion through designing individual models for each modality but also takes advantage of the joint representation of the low-level input through early fusion.

The rest of this chapter is organized as follows. In the next section, we provide the more formal definition of multimodal fusion, including early and late fusion technique. Next, we explain our architecture and methodology. In section 3.4, we present experimental results on two different tasks, i.e., multi-label classification and dimensional emotion recognition. Finally, we give final comments and conclusions.

3.2 Prerequisite

We first present the relevant work upon which our multimodal fusion approach is built. This section also establishes the notations used throughout the chapter.

As many others addressing multimodal fusion, we start from the assumption of having an off-the-shelf feature extractor for each of the involved modalities. Our proposed method is based on two existing types of popular multimodal approaches: early and late fusion. Therefore we begin first with notations and then describe the early and late fusion models followed by our new hybrid architecture. Early and late fusion are illustrated in Figure 3.1. Our proposed model is related to CentralNet (Vielzeuf *et al.* (2018)) in the sense that CentralNet also has mixed architecture of early and late fusion; However, the mechanism they use for interaction between two models is different.

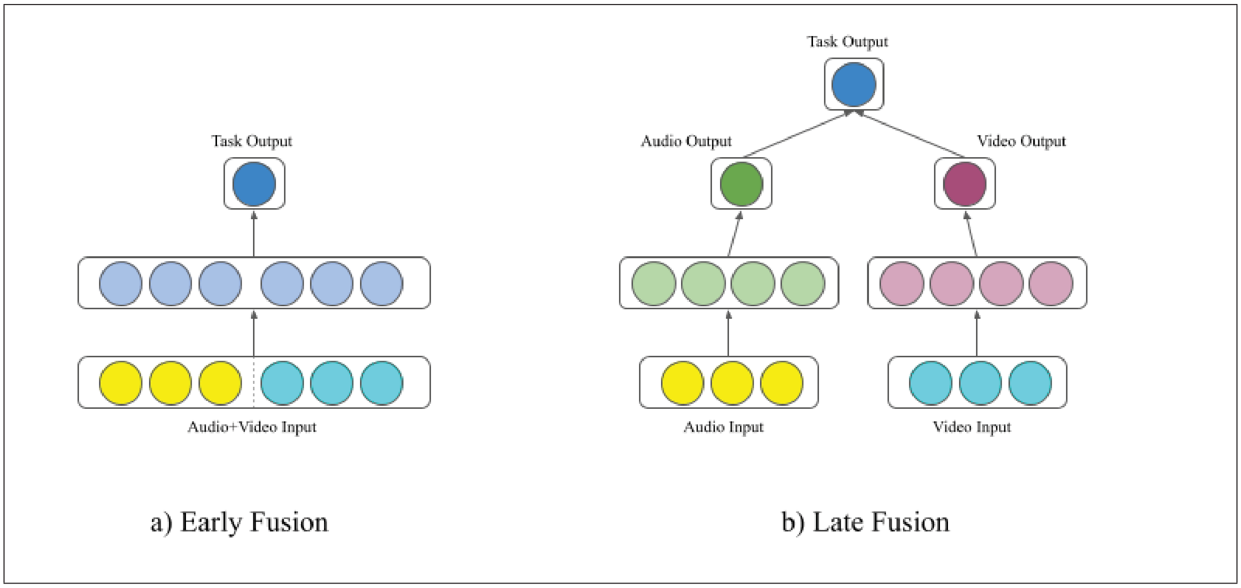

Figure 3.1 Early and late fusion block diagram. In this figure for simplicity, we show only one output but it can be generalized for multi-class classification as well

3.2.1 Notations

We use M as the total number of our modalities. Without loss of generality, in this chapter we assume that we use two modalities. One can easily apply the same method on more than two modalities. For each input modality we denote the extracted feature using an off-the-shelf feature extractor as a dense vector $v_m \in \mathbb{R}^{d_m}$ with dimension $d_m, \forall m = 1, 2, \cdots, M$. For instance, given $M = 2$ modalities in the video classification task, v_1 and v_2 are the extracted features of the frames and the encoded audio information from the targeted video accordingly.

3.2.2 Early Fusion

Early fusion methods create a joint representation of features from multiple modalities. In this method first, each unimodal system encodes the input feature to an intermediate representation. Next, the aggregation method F, combine unimodal representations into a single model. Then the single model is trained to learn the interactions between these intermediate representations of each modality. We denote the single model as h. The final prediction can be written as:

$$p = h(F(v_1, v_2)) \tag{3.1}$$

The aggregation function F can be a simple or weighted sum, bilinear product (Yu *et al.* (2017)) or concatenation. In the early fusion, the training pipeline is simple as only one model is involved. It usually requires the features from different modalities to be highly engineered and preprocessed so that they align well or are similar in their semantics. Furthermore, it uses a single model to make predictions, which assumes that the model is well suited for all modalities.

3.2.3 Late Fusion

In the late fusion methods, the combination happens only after getting scores from each unimodal systems. Then the model fuses the decision values using a fusion mechanism F such as simple or weighted score average (Natarajan *et al.* (2012)), rank minimization (Ye *et al.* (2012)) or Mixture-of-Experts models (Jordan & Jacobs (1994)). We denote each unimodal systems as $h_i, i \in 1,2$. The final prediction for late fusion can be written as:

$$p = F(h_1(v_1), h_2(v_2)) \tag{3.2}$$

Late fusion allows the use of different models on different modalities, thus allowing more flexibility. It is easier to handle a missing modality as the predictions are made separately. However, because late fusion operates on scores and not on the raw inputs, it is not effective at modeling signal-level interactions between modalities.

3.3 Proposed Model

In this section, we introduce our hybrid multimodal fusion architecture. We call it the Augmented Ensemble network because of the way it incorporates unimodal features into the early

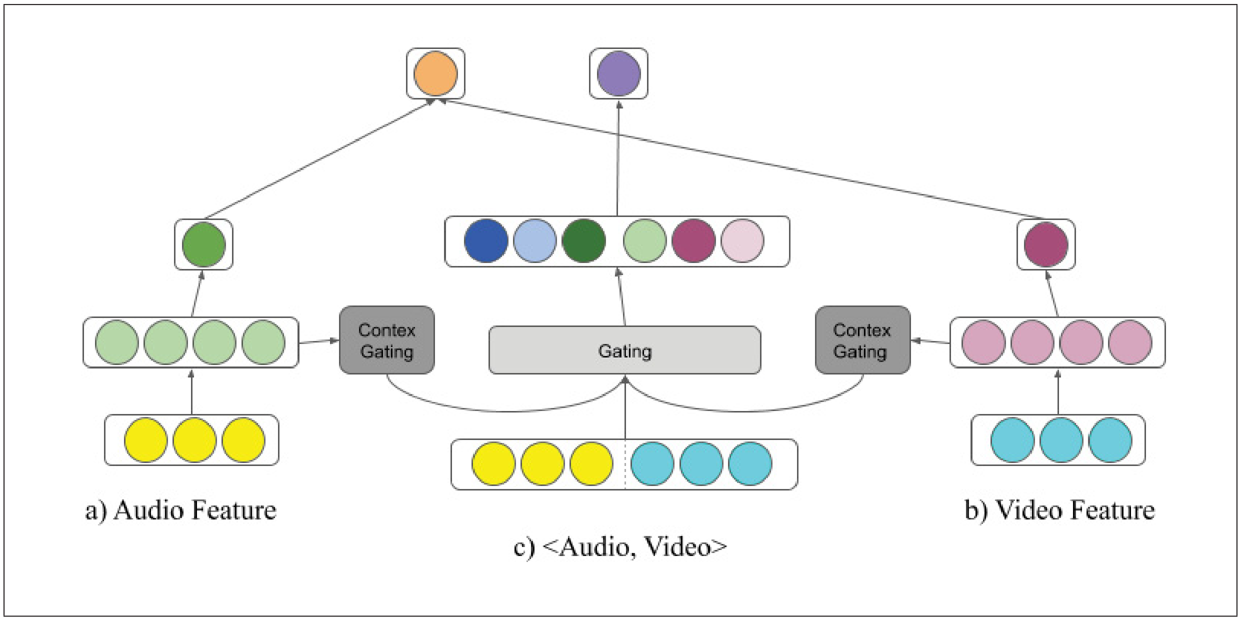

Figure 3.2 Block level diagram of the Augmented Ensemble multimodal method. The brightness of each neuron shows the strength of each feature before and after applying the context gating layer

fusion network features. Our multimodal fusion architecture for two modalities, i.e audio and video, is illustrated in Figure 3.2.

3.3.1 Augmented Ensemble Network

The model is essentially an ensemble of early and late fusion augmented by intermediate representations. First, for each modality, based on section 3.2 an early and late fusion network is constructed. For the early fusion input, we use concatenation of unimodals output representation. Next, at each layer for each unimodal network, the intermediate feature is enhanced by the Context Gating layer (3.3.2). Before feeding the extracted information to the early fusion network, we use attention mechanism to aggregate enhanced unimodal features and features generated by the previous layer of the early fusion network. The attention network focuses on learning the relationship between modalities. We hypothese that correlation between multiple modalities can be nontrivial and having a separate network to learn those lead to the overall improvement of the performance. Finally, the late fusion network processes the combined features. The final score is based on the sum of late fusion and early fusion score. The equations governing one layer of the proposed model can be written as:

$$s_1^i = \gamma(v_1^i; \theta_1^i) \tag{3.3}$$

$$s_2^i = \gamma(v_2^i; \theta_2^i) \tag{3.4}$$

$$s_{1,2}^i = [s_1^i, s_2^i] \tag{3.5}$$

$$g^i = Att([s_{1,2}^i, v_e^i]; \theta_{att}^i) \tag{3.6}$$

$$s_e^i = g_0^i * v_e^i + g_1^i * s_{1,2}^i \tag{3.7}$$

where i is the layer number, γ refers to the context gating module with θ_1^i, θ_2^i as the trainable parameters, Att with trainable parameter θ_{att}^i refers to the attention network used in the early fusion network. It is worth noting that, because we use concatenation throughout our early fusion network, the size of each hidden unit in early fusion is equal to the sum of each unimodal network hidden unit.

The CentralNet (Vielzeuf *et al.* (2018)) constructs a new network by taking the weighted sum of each unimodal hidden units and the early fusion hidden layer. Compared to our proposed model, in the CentralNet architecture, authors used scalar trainable weights to capture the correlation between different modalities. We argue that in each modality, some of the features are more important than the others for capturing this correlation, therefore a single scalar weight is not sufficient to capture the relationship between different modalities.

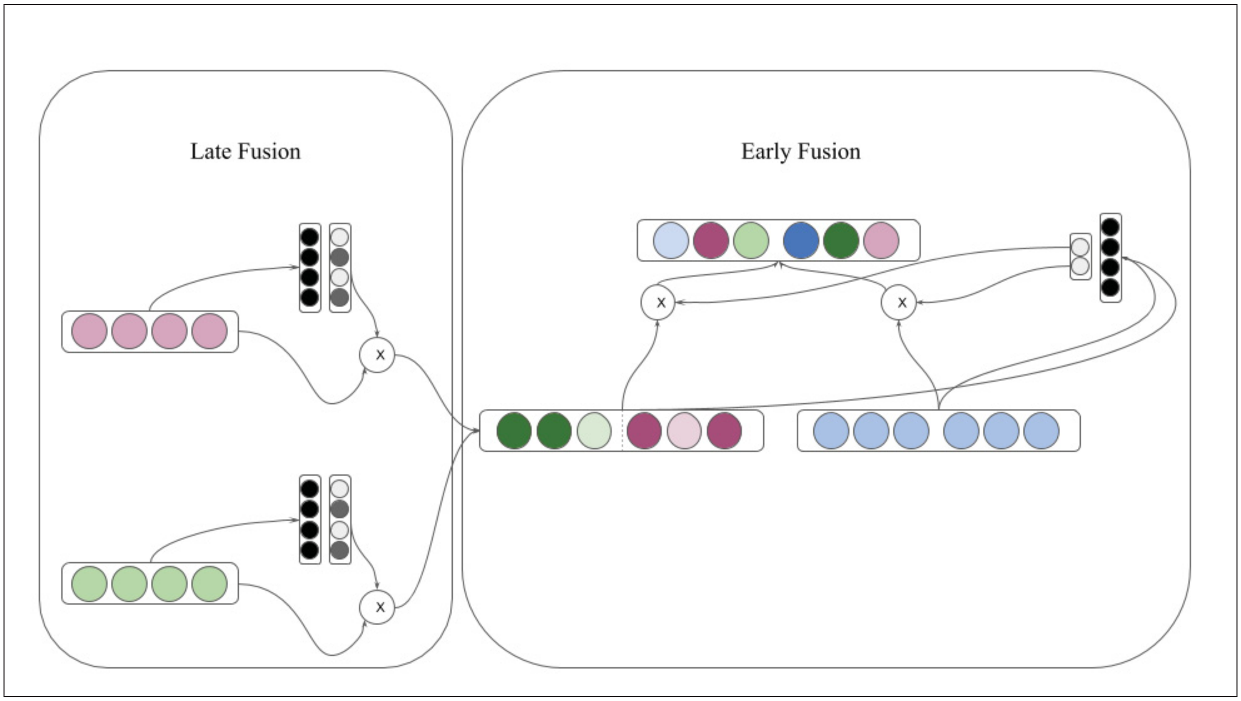

Figure 3.3 One layer of the Augmented Ensemble architecture

3.3.2 Context Gating

The context gating module transforms the input feature representation $\mathbf{x}$ into a new representation $\mathbf{s}$ as:

$$\mathbf{s} = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{x})) \circ \mathbf{x} \tag{3.8}$$

where $\mathbf{x} \in \mathbb{R}^n$ is the input feature vector, σ is the elementwise sigmoid activation, δ refers to the ReLU function, and $\circ$ is the element-wise multiplication. $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C \times \frac{C}{r}}$ are trainable parameters. The vector of weights $\sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{x})) \in [0,1]$ represents a set of learned gates applied to the individual dimensions of the input feature $\mathbf{x}$. Same as Hu *et al.* (2018), reduction ratio r is chosen to limit model complexity and aid generalization. The motivation behind this transformation is to recalibrate the strengths of different activations of the input representation through a self-gating mechanism. To fulfil this objective, it must learn a non-mutually-exclusive relationship since we would like to ensure that multiple features are allowed to be emphasised, opposed to one-hot activation. One layer of the model using the context gating layer is depicted in Figure 3.3.

3.4 Experiments

To demonstrate the capability of our multimodal fusion method, we conducted experiments on two different tasks. The first task is multi-label video classification on Youtube-8M v2 dataset (Abu-El-Haija *et al.* (2016)) and the second task is dimensional, continuous, emotion recognition on RECOLA dataset (Ringeval *et al.* (2013)). On each task, the performance of the proposed method is compared to 5 different baselines. We fist conduct two comparative tests with single-modality input (only video or audio), to prove that the improvement of performance does not come from increasing parameters on each modality. We also implement vanilla late and early fusion based on the details provided in section 3.2. We verify the effectiveness of the Context Gating layer by comparing its performance with the ensemble of early and late fusion without augmented features. Finally, we compare out model with the CentralNet as well.

3.4.1 Youtube-8M v2 Dataset

We conduct experiments on the recently updated Youtube-8M v2 dataset (Abu-El-Haija *et al.* (2016)) with improved machine-generated labels and higher-quality videos. It contains a total of 6.1 million videos, 3862 classes, 3 labels per video averagely. Youtube-8M v2 dataset provides two sets of features, frame level and video level. Throughout the experiment we use video level features. Visual and audio features are pre-extracted per frame and averaged to a fixed-length representation (at the video-level). Visual features are obtained by Google Inception pretrained on ImageNet (Deng *et al.* (2009)), followed by PCA-compression to generate a vector with 1024 dimensions. The audio features are extracted from a VGG-inspired acoustic model described in (Hershey *et al.* (2017)). In the official split, training, validataion and test have equal 3844 tfrecord shards. In practice, we follow the same settings as (Liu *et al.* (2018b)), we use 3844 training shards and 3000 validation shards for training. We randomly select 200 shards from the rest of 844 validation shards (around 243 337 videos) to evaluate our model every 1000 training steps. Results are evaluated using the Global Average Precision (GAP) metric using the following expression:

$$GAP = \sum_{i=1}^{N} p(i)\Delta r(i) \tag{3.9}$$

where N is the number of final predictions, $p(i)$ is the precision, and $r(i)$ is the recall.

We implement our model using Pytorch (Paszke *et al.* (2017)) with the Google starter code[1]. All models are trained using Adam optimizer (Kingma & Ba (2014)) with an initial learning rate set to 0.01 and learning rate decay of 0.9 every 400000 training steps. The mini-batch size is set to 1024. We set the reduction ratio r to 4. We use cross entropy loss for maximizing the Global Average Precision (GAP). All models are trained for 10 epochs with early stopping. In order to observe timely model prediction, we evaluate the model on a subset of validate set every 1000 training steps.

Table 3.1 Performance comparison of our models versus other baseline models evaluated on the Youtube-8M v2 dataset

Model	**GAP**
Audio Only	40.1%
Video Only	71.1%
Early Fusion	83.8%
Late Fusion	83.2%
Ensemble	84.0%
CentralNet Vielzeuf *et al.* (2018)	83.9%
Augmented Ensemble (Ours)	**85.1%**

The detailed results of mutlimodal fusion on Youtube-8M v2 dataset are shown in Table 3.1. Firstly, the GAP performance of multimodal networks is far superior to single modality input

[1] https://github.com/google/youtube-8m

(Video Only or Audio Only). Secondly, we can observe the same performance for the early fusion and late fusion. Finally, the Augmented Ensemble method achieves 1.1% higher GAP compared with the simple fusion Ensemble baseline. The main reason is probably that the simple fusion Ensemble can not leverage high-order information across modalities. Our model achieves better performance compared to the CentralNet model. This suggests that the simple mechanism used in CentralNet for capturing the relationship between unimodal networks and the early fusion network is not sufficient.

3.4.2 RECOLA Dataset

The Remote Collaborative and Affective Interactions (RECOLA) dataset (Ringeval *et al.* (2013)) was recorded to study socio-affective behaviours from multimodal data in the context of computer supported collaborative work. Spontaneous and naturalistic interactions were collected during the resolution of a collaborative task that was performed in dyads and remotely through video conference. Multimodal signals, i.e., audio, video, electro-cardiogram (ECG) and electrodermal activity (EDA), were synchronously recorded from 27 French-speaking subjects. Even though all subjects speak French fluently, they have different nationalities (i.e., French, Italian or German), which thus provides some diversity in the expression of emotion. The dataset contains two types of dimensional labels (arousal and valence) which were annotated by six people. Each dimensional label ranges from $[-1, 1]$. The dataset is partitioned into three sets: train, development, and test, each containing nine subjects. The Audio-Visual Emotion Recognition Challenge 2016 (AVEC16) Valstar *et al.* (2016) uses data from RECOLA dataset.

The challenge organizers provide pre-extracted features for all of the modalities but in practice we use only features from audio and video modals. For video appearance and geometric based features are extracted. 84 appearance features are obtained using Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) (Almaev & Valstar (2013)) followed by PCA. 316 geometric features are obtained by tracking 49 facial landmarks with the Supervised Descent Method (SDM) (Xiong & De la Torre (2013)). Both appearance and geometric feature sets are interpolated by a piecewise cubic Hermite polynomial to cope with dropped frames. Finally, the arithmetic mean and the standard-deviation are computed on all features using a sliding window, which is shifted forward at a rate of 40ms. The audio features are extracted using the extended Geneva minimalistic acoustic feature set (eGeMAPS) (Eyben *et al.* (2015)). They also extract the acoustic low-level descriptors (LLD) with the openSMILE toolkit (Eyben *et al.* (2010)). Overall, the acoustic baseline features set contains 88 features.

In this experiments, we focus on predicting the arousal level.

We evaluate our methods by computing Concordance Correlation Coefficient (CCC) (Lawrence & Lin (1992)). The CCC tries to measure the agreement between two variables using the following expression:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{3.10}$$

where ρ is the Pearson correlation coefficient (PCC), σ_x^2 and σ_y^2 are the variance of the predicted and ground truth values respectively and μ_x and μ_y are their means, respectively. The strongest method is selected based on whichever obtains the highest CCC value. Compared with PCC, the CCC considers not only the shape similarity between the two series but also the value precision. This is especially relevant for estimating the performance of time-continuous emotion prediction models, as both the trends as well as absolute prediction values are relevant for describing the performance of a model. Figure 3.4 shows two sets of bi-vartiated, (x,y), data where in both of the sets PCC is equal to 1 but in only of them CCC is 1. The CCC metric falls in the range of $[-1,1]$, where $+1$ represents perfect concordance, -1 total discordance, and 0 no concordance at all.

We implement our model using Pytorch framework. All models are trained using SGD optimizer with an initial learning rate set to 0.1. The mini-batch size is set to 256. We use mean square error loss for maximizing the CCC. All models are trained for 50 epochs. We evaluate all of our experiments on the development set. For all investigated methods, we follow (Trige-

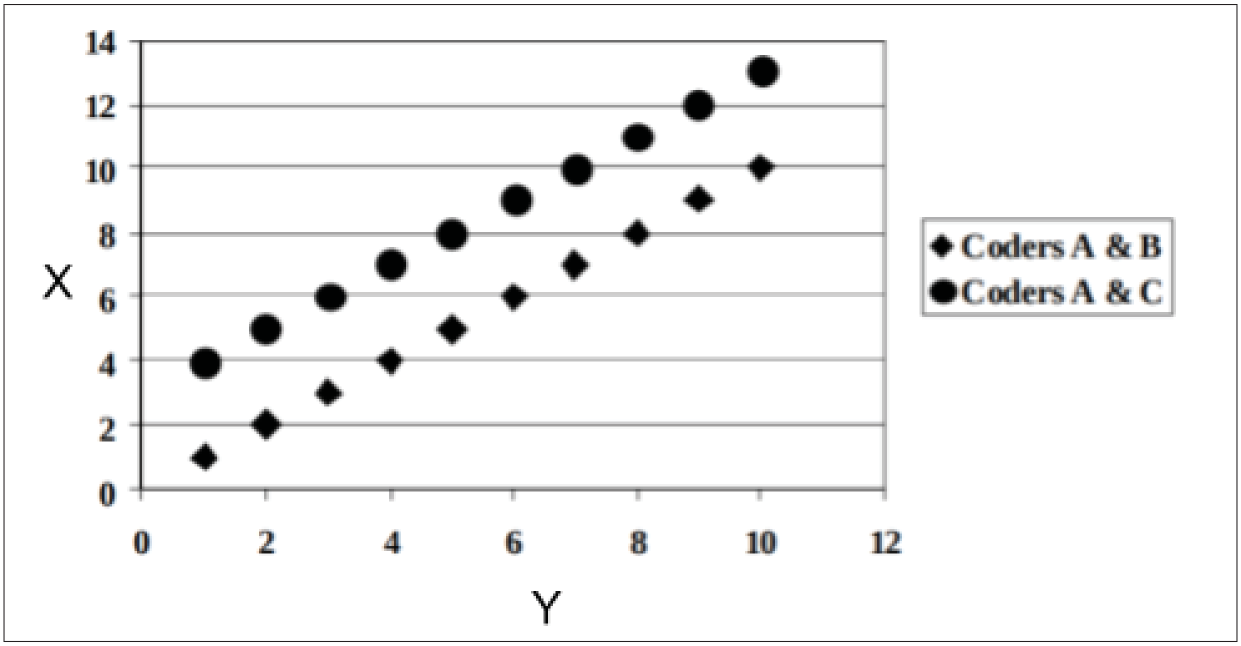

Figure 3.4 Coders A & B: PCC = 1.0, CCC = 1.0. Coders A & C: PCC = 1.0, CCC = .67. PCC provides a measure of linear covariation between set of scores, without specifying any degree of correspondence between the set.

orgis *et al.* (2016)) and a chain of post-processing is applied to the predictions obtained on the development set.

Table 3.2 Performance comparison of our models versus other baseline models evaluated on the RECOLA dataset. The reported CCC scores are for the arousal

Model	**CCC**
Audio Only	0.772
Video Only	0.444
Early Fusion	0.794
Late Fusion	0.783
Ensemble	0.802
CentralNet Vielzeuf *et al.* (2018)	0.820
Augmented Ensemble (Ours)	**0.823**

The detailed results of our experiments on the RECOLA dataset are shown in Table 3.2. Same as the previous experiments on the Youtube-8M v2 dataset, the multimodal methods achieve

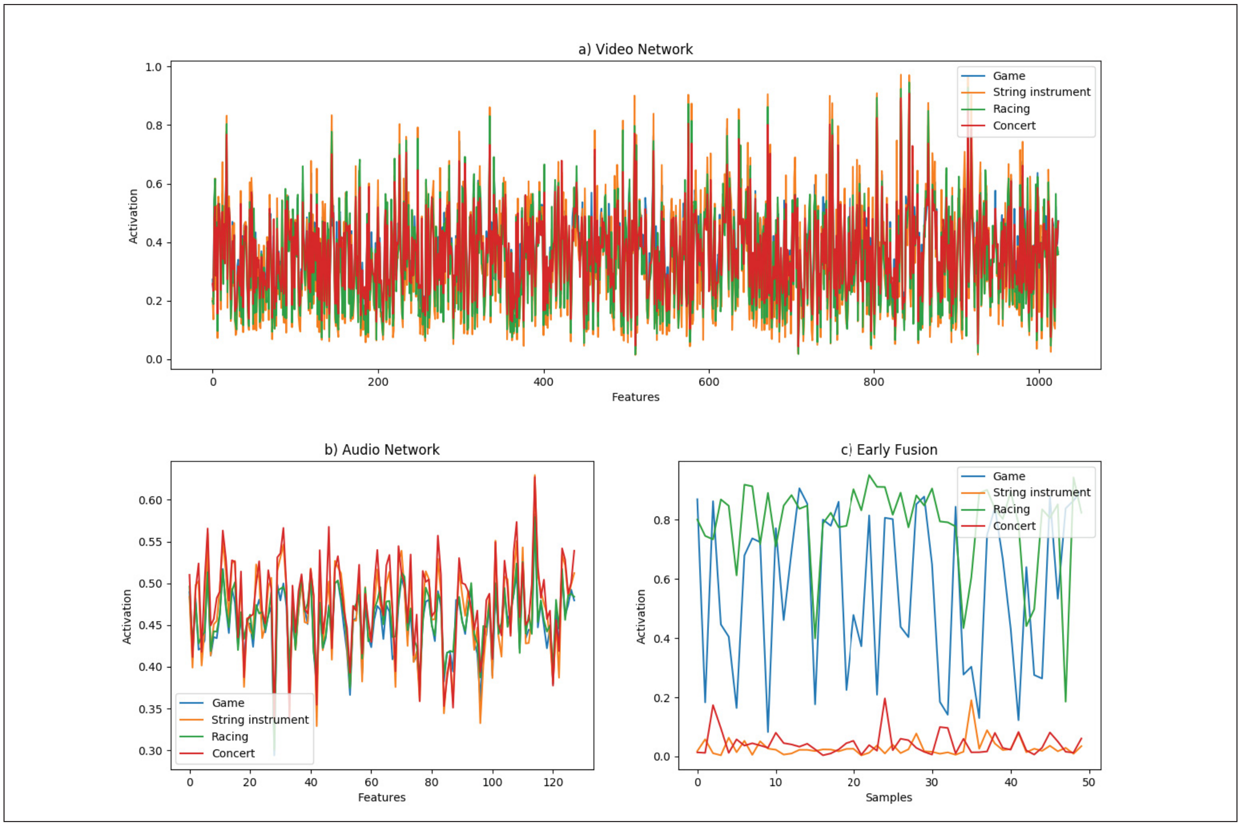

Figure 3.5 Activation produces by the context gating layer in the unimodal networks and attention unit in the early fusion network

superior performance compared to single modal techniques. We also demonstrate that our proposed model outperforms the rest of the baselines. Our model achieves comparable results with the CentralNet model.

3.4.3 Visualization

To better understand the behavior of the context gating layer and attention unit, in this section we study example activations from our proposed model applied on Youtube-8M v2 Dataset and examine their distribution for different classes. We choose Youtube-8M v2 Dataset due to variations in its target classes. We sample images from the dataset that contains four classes that exhibit from the dataset and appearance diversity, namely game, string instrument, racing, and concert. We then draw fifty samples within each class from the validation set. For the context gating layers, we compute the average activations for each feature. For the attention unit used in the early fusion network, we compare the probability of choosing the late fusion features over the features coming from the previous layer of the early fusion network. The attention unit activation for video and audio network and the probability of the features coming from the previous layer of early fusion network are illustrated in Figure 3.5.

We make the following observations. First, in the audio network, it is clear that the classes which are more sensitive to audio e.g concert and string instrument on average, have bigger activations. Interestingly, the activations for each feature are different. This suggests that having one unit per feature in the context gating layer is crucial and implies one of the improvement factors over CentralNet which has only one scalar trainable parameter associated with each unimodal network. Second, in the early fusion network, compared to its previous layer features, the audio network activations have more contributions. This suggests that unimodal features can help to compensate for information loss. For the other two classes e.g game and racing, the effect is reversed. In the audio network, the activations corresponding to these classes are higher and their contribution to the early fusion network is lower. For the video network, even though the picks of the activations are almost the same as the audio network, the effect of the context gating layer is not as clear as the audio network.

3.5 Conclusion

This chapter investigates a novel approach for the fusion of multimodal information. It intends firstly to have the flexibility of the late fusion architecture by applying different models on different modalities, secondly, to learn rich features from a joint representation of the low-level signal feature through early fusion architecture. We also demonstrate the superior performance of the network compared to the baseline models by applying on two different tasks, i.e multi-class classification problem on the Youtube-8M dataset and dimensional emotion recognition on the RECOLA dataset.

CONCLUSION AND RECOMMENDATIONS

Summary

In this dissertation, we highlighted how deep learning methods, when applied to emotion recognition, by learning rich representations achieve superior accuracy over traditional techniques. Moreover, we demonstrated our methods are not bound to emotion recognition task and other classes of tasks such as multi-label classification can get benefit from our approaches.

First, we focused only on one modality and considered the task of video-based emotion recognition using only visual inputs. We addressed the spatial and temporal aspects of the input data using two mechanisms. For spatial dimension, we replaced the global pooling layer of CNN with a multi-head attention network. We demonstrated visually the multi-head attention can help the performance by selecting the most important regions of an image. For the temporal aspect of the data, we introduced the temporal softmax pooling layer which jointly learns the correlation among the frames and most predictive frames among the other frames within one video. We showed that, compare to the more complex mechanism such as RNN, the temporal softmax pooling over the emotion probabilities, that selects the most important frames of a video, can lead to promising results.

Second, we moved our attention to the multimodal data. Particularly we focused on multimodal fusion. By incorporating the best features from early and late fusion architecture, we introduced a new fusion method. We compared our method with different fusion approaches and noticed improvement across different tasks.

Future Work

In practice datasets related to affective computing, e.g. RECOLA Ringeval *et al.* (2013) and AFEW Dhall *et al.* (2012), are small and deep learning methods usually require vast amount of data to generalize well. Therefore, using pretrained networks always help to improve the final result. It is a good idea to pick a dataset that is related to the task in hand and use a previously trained deep modal on that dataset. For instance, in chapter 2, we used VGG-FACE network weights which previously trained to recognize celebrity's faces and noticed great boost in the final performance. One can further use other emotion recognition datasets as well.

In chapter 2, while we only demonstrated our result for visual features. It is possible by slightly changing the architecture configuration, make the network compatible with other modalities as well. Specify one can adopt CNN for inputs represented as signals such as audio and apply the same multi-head attention and temporal pooling mechanism on the extracted features.

In chapter 3, for multimodal fusion, we used off-the-shelf feature extractor. While these methods help to minimize the footprint of the model, their representations are fixed and the multimodal system is bounded to those extracted features. One can insert one learnable feature extractor for each unimodal system and train the multimodal system in the end-to-end fashion. This approach improves the flexibility of each unimodal system and also can take advantage of all of the best practices in deep models.

LIST OF REFERENCES

Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B. & Vijayanarasimhan, S. (2016). Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*.

Almaev, T. R. & Valstar, M. F. (2013). Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pp. 356–361.

Aminbeidokhti, M., Pedersoli, M., Cardinal, P. & Granger, E. (2019). Emotion Recognition with Spatial Attention and Temporal Softmax Pooling. *Image Analysis and Recognition*, pp. 323–331.

Arevalo, J., Solorio, T., Montes-y Gómez, M. & González, F. A. (2017). Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Bahdanau, D., Cho, K. & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Baltrušaitis, T., Ahuja, C. & Morency, L.-P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423–443.

Bengio, Y., Simard, P., Frasconi, P. et al. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2), 157–166.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Brady, K., Gwon, Y., Khorrami, P., Godoy, E., Campbell, W., Dagli, C. & Huang, T. S. (2016). Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 97–104.

Buechel, S. & Hahn, U. (2016). Emotion analysis as a regression problem—Dimensional models and their implications on emotion representation and metrical evaluation. *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pp. 1114–1122.

Chen, J., Chen, Z., Chi, Z. & Fu, H. (2014). Emotion recognition in the wild with feature fusion and multiple kernel learning. *Proceedings of the 16th International Conference on Multimodal Interaction*, pp. 508–513.

Chen, K., Wang, J., Chen, L.-C., Gao, H., Xu, W. & Nevatia, R. (2015). Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*, 255–268.

Chen, M., Wang, S., Liang, P. P., Baltrušaitis, T., Zadeh, A. & Morency, L.-P. (2017). Multimodal sentiment analysis with word-level fusion and reinforcement learning. *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 163–171.

Chen, S. & Jin, Q. (2015). Multi-modal Dimensional Emotion Recognition Using Recurrent Neural Networks. *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, (AVEC '15), 49–56. doi: 10.1145/2808196.2811638.

Cheng, J., Dong, L. & Lapata, M. (2016). Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. & Taylor, J. G. (2001). Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1), 32–80.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255.

Dhall, A., Goecke, R., Lucey, S. & Gedeon, T. (2011). Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 2106–2112.

Dhall, A., Goecke, R., Lucey, S. & Gedeon, T. (2012). Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *IEEE MultiMedia*, 19(3), 34–41. doi: 10.1109/MMUL.2012.26.

Dhall, A., Goecke, R., Ghosh, S., Joshi, J., Hoey, J. & Gedeon, T. (2017). From individual to group-level emotion recognition: EmotiW 5.0. *Proceedings of the 19th ACM international conference on multimodal interaction*, pp. 524–528.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4), 169–200.

Eyben, F., Wöllmer, M. & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S. et al. (2015). The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2), 190–202.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T. & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Glorot, X. & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep Learning*. MIT Press.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.-H. et al. (2013). Challenges in representation learning: A report on three machine learning contests. *International Conference on Neural Information Processing*, pp. 117–124.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B. et al. (2017). CNN architectures for large-scale audio classification. *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pp. 131–135.

Hochreiter, S. & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, 9(8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

Hu, J., Shen, L. & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141.

Huang, Z., Dong, M., Mao, Q. & Zhan, Y. (2014). Speech emotion recognition using CNN. *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 801–804.

Jetley, S., Lord, N. A., Lee, N. & Torr, P. H. (2018). Learn to pay attention. *arXiv preprint arXiv:1804.02391*.

Jordan, M. I. & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2), 181–214.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. & Fei-Fei, L. (2014). Large-scale Video Classification with Convolutional Neural Networks. *CVPR*.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Kim, B.-K., Lee, H., Roh, J. & Lee, S.-Y. (2015). Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 427–434.

Kim, J.-H., On, K.-W., Lim, W., Kim, J., Ha, J.-W. & Zhang, B.-T. (2016). Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul), 1755–1758.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Knyazev, B., Shvetsov, R., Efremova, N. & Kuharenko, A. (2018). Leveraging large face recognition data for emotion classification. *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pp. 692–696.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, pp. 1097–1105.

Lahat, D., Adali, T. & Jutten, C. (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9), 1449–1477.

Lawrence, I. & Lin, K. (1992). Assay validation using the concordance correlation coefficient. *Biometrics*, 599–604.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Li, S., Deng, W. & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2852–2861.

Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B. & Bengio, Y. (2017). A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Liu, C., Tang, T., Lv, K. & Wang, M. (2018a). Multi-Feature Based Emotion Recognition for Video Clips. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, (ICMI '18), 630–634. doi: 10.1145/3242969.3264989.

Liu, J., Yuan, Z. & Wang, C. (2018b). Towards good practices for multi-modal fusion in large-scale video classification. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0.

Long, X., Gan, C., De Melo, G., Wu, J., Liu, X. & Wen, S. (2018). Attention clusters: Purely attention based local feature integration for video classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7834–7843.

Loshchilov, I. & Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Lu, C., Zheng, W., Li, C., Tang, C., Liu, S., Yan, S. & Zong, Y. (2018). Multiple Spatio-temporal Feature Learning for Video-based Emotion Recognition in the Wild. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, (ICMI '18), 646–652. doi: 10.1145/3242969.3264992.

Mehrabian, A. (1996). Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4), 261–292.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pp. 3111–3119.

Natarajan, P., Wu, S., Vitaladevuni, S., Zhuang, X., Tsakalidis, S., Park, U., Prasad, R. & Natarajan, P. (2012). Multimodal feature fusion for robust event detection in web videos. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1298–1305.

Parkhi, O. M., Vedaldi, A. & Zisserman, A. (2015). Deep Face Recognition. *British Machine Vision Conference*.

Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, Ł., Shazeer, N., Ku, A. & Tran, D. (2018). Image transformer. *arXiv preprint arXiv:1802.05751*.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. (2017). Automatic differentiation in pytorch.

Pérez-Rúa, J.-M., Vielzeuf, V., Pateux, S., Baccouche, M. & Jurie, F. (2019). MFAS: Multimodal Fusion Architecture Search. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6966–6975.

Ramirez, G. A., Baltrušaitis, T. & Morency, L.-P. (2011). Modeling latent discriminative dynamic of multi-dimensional affective signals. *International Conference on Affective Computing and Intelligent Interaction*, pp. 396–406.

Ringeval, F., Sonderegger, A., Sauer, J. & Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pp. 1–8.

Ringeval, F., Eyben, F., Kroupi, E., Yuce, A., Thiran, J.-P., Ebrahimi, T., Lalanne, D. & Schuller, B. (2015). Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66, 22–30.

Rumelhart, D. E., Hinton, G. E., Williams, R. J. et al. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, 5(3), 1.

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6), 1161.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R. & Pantic, M. (2011). Avec 2011–the first international audio/visual emotion challenge. *International Conference on Affective Computing and Intelligent Interaction*, pp. 415–424.

Shan, C., Gong, S. & McOwan, P. W. (2007). Beyond Facial Expressions: Learning Human Emotion from Body Gestures. *BMVC*, pp. 1–10.

Shan, C., Gong, S. & McOwan, P. W. (2009). Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6), 803–816.

Sharma, S., Kiros, R. & Salakhutdinov, R. (2015). Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*.

Simonyan, K. & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, pp. 568–576.

Stevenson, R. A., Mikels, J. A. & James, T. W. (2007). Characterization of the affective norms for English words by discrete emotional categories. *Behavior research methods*, 39(4), 1020–1024.

Sutskever, I., Vinyals, O. & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, pp. 3104–3112.

Tenenbaum, J. B. & Freeman, W. T. (2000). Separating style and content with bilinear models. *Neural computation*, 12(6), 1247–1283.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B. & Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5200–5204.

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R. & Pantic, M. (2016). Avec 2016: Depression, mood, and emotion recognition workshop and challenge. *Proceedings of the 6th international workshop on audio/visual emotion challenge*, pp. 3–10.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, pp. 5998–6008.

Vielzeuf, V., Lechervy, A., Pateux, S. & Jurie, F. (2018). Centralnet: a multilayer approach for multimodal fusion. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 0–0.

Wang, X., Girshick, R., Gupta, A. & He, K. (2018). Non-local neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803.

Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B. & Rigoll, G. (2013). LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2), 153–163.

Xia, R., Deng, J., Schuller, B. & Liu, Y. (2014). Modeling gender information for emotion recognition using denoising autoencoder. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 990–994.

Xiong, X. & De la Torre, F. (2013). Supervised descent method and its applications to face alignment. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 532–539.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *International conference on machine learning*, pp. 2048–2057.

Ye, G., Liu, D., Jhuo, I.-H. & Chang, S.-F. (2012). Robust late fusion with rank minimization. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3021–3028.

Yu, Z., Yu, J., Fan, J. & Tao, D. (2017). Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. *Proceedings of the IEEE international conference on computer vision*, pp. 1821–1830.

Zadeh, A., Zellers, R., Pincus, E. & Morency, L.-P. (2016). MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Zeiler, M. D. & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European conference on computer vision*, pp. 818–833.

Zhang, Y., Du, J., Wang, Z., Zhang, J. & Tu, Y. (2018). Attention based fully convolutional network for speech emotion recognition. *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1771–1775.

Zhao, G. & Pietikainen, M. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6), 915–928.