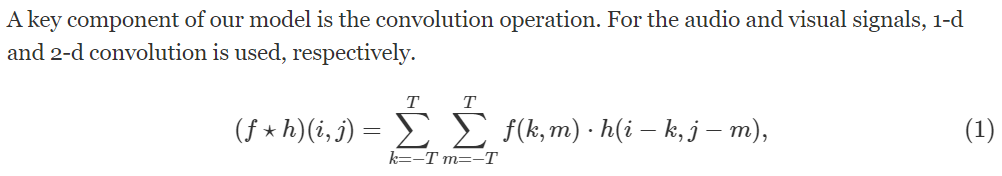
## 1-End-to-End Multimodal Emotion Recognition Using Deep Neural Networks

·CNN to extract features from the speech, deep residual network of 50 layers for the visual modality.

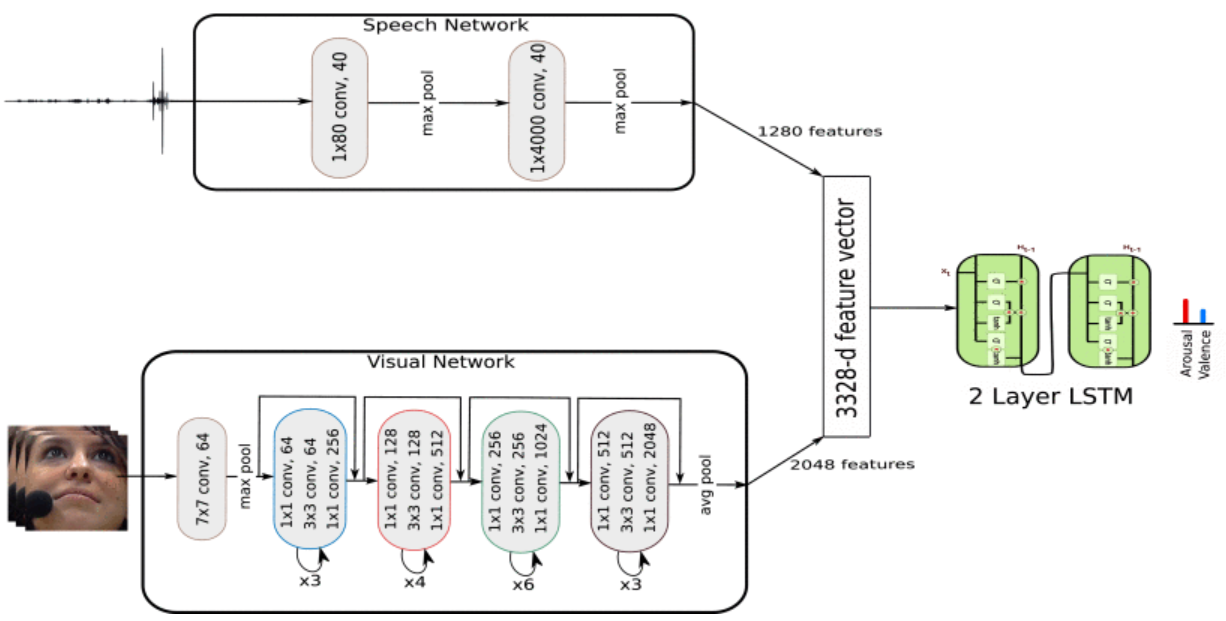
·LSTM to make the algorithm insensitive to outliers while being able to model the context.(对异常值不敏感，同时能够对上下文进行建模)

·trained in an end-to-end fashion

（Finally, our model is subsequently trained with backpropagation by maximizing the concordance correlation loss）



where f(x) is a kernel function whose parameters are learnt from the data of the task in hand.



**A. Visual Network:** deep residual network (ResNet) of 50 layers

**Input**；

传统用Scale Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG)

本文用pixel intensities（像素强度）from the **cropped** faces of the subject's video



where x and y are the input and output of the layer k, F(xk,{Wk}) is the residual function to be learned and h(xk) can be either an identity mapping or a linear projection to match the dimensions of the function F and the input x.

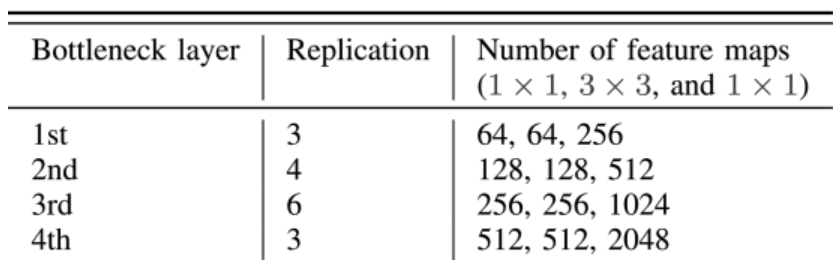
**ResNet-50：**

7×7 convolutional layer with 64 feature maps

↓

max pooling layer of size 3×3

↓

4 Bottleneck Architectures：

↓

shortcut connection/ an average pooling layer

**B. Speech Network**

在辅助语言学paralinguistics领域，首先提取声学特征，然后传递到机器学习算法。

但我们专注于**在一个共同训练的模型中**学习特征提取和回归步骤来预测情绪。

**Input**：96000-dimensional vector (preprocess raw waveform(16kHz, μ=0 σ=1) then segment to 6s sequences)

↓

Temporal Convolution: 从高采样频率的信号中提取更精细尺度的信息，通过size=80 window=5ms F=40 space time finite impulse filters

↓

Max Pooling across time. 每个滤波器的脉冲响应通过一个half-wave rectifier，然后通过对每个脉冲响应进行size=2的池化来降采样到8Hz

↓

Temporal Convolution. 提取语言信号的长期特征和粗糙度(不规则性)，通过M=40 space time finite impulse filters of 500 ms window (size of 4000)

↓

Max Pooling Across Channels. 减少信号的维数，同时保留对卷积信号的必要统计。通过大小为10的池在通道域中执行max-pooling。

↓

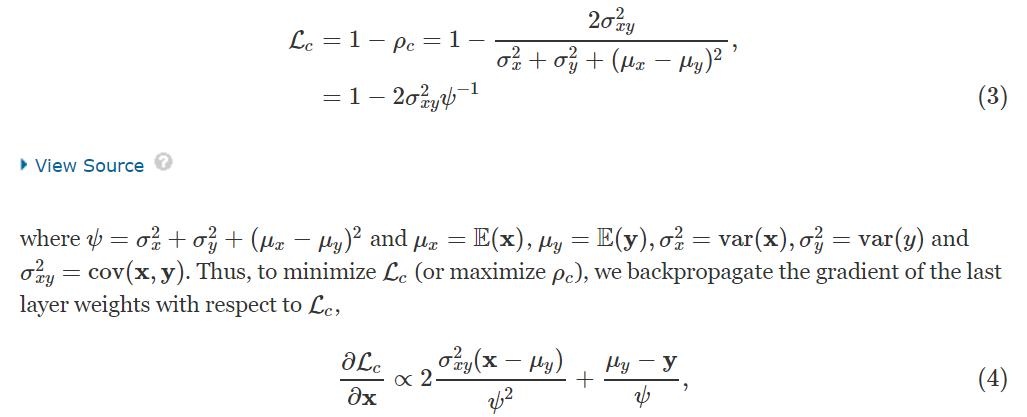
Dropout. 正则化避免过拟合，概率为0.5

CNN网络对原始信号进行操作以从中提取特征。为了考虑语音的时间结构，我们使用两个LSTM层，每个层在CNN的顶部有256个单元。

**C. Objective Function ？？？**

评估agreement level between the predictions of the network and the gold-standard derived from the annotations

we propose to **include** **the** **metric** used to evaluate the performance **in the objective function (Lc)** (a cost function) used to train the networks.



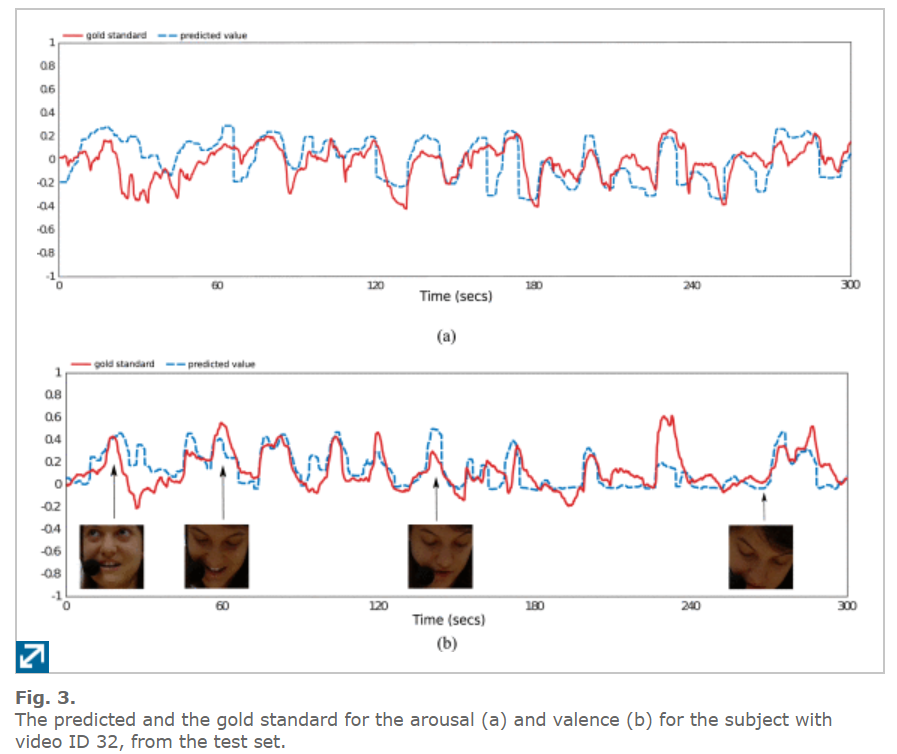
where all vector operations are done element-wise.

D. Network Training

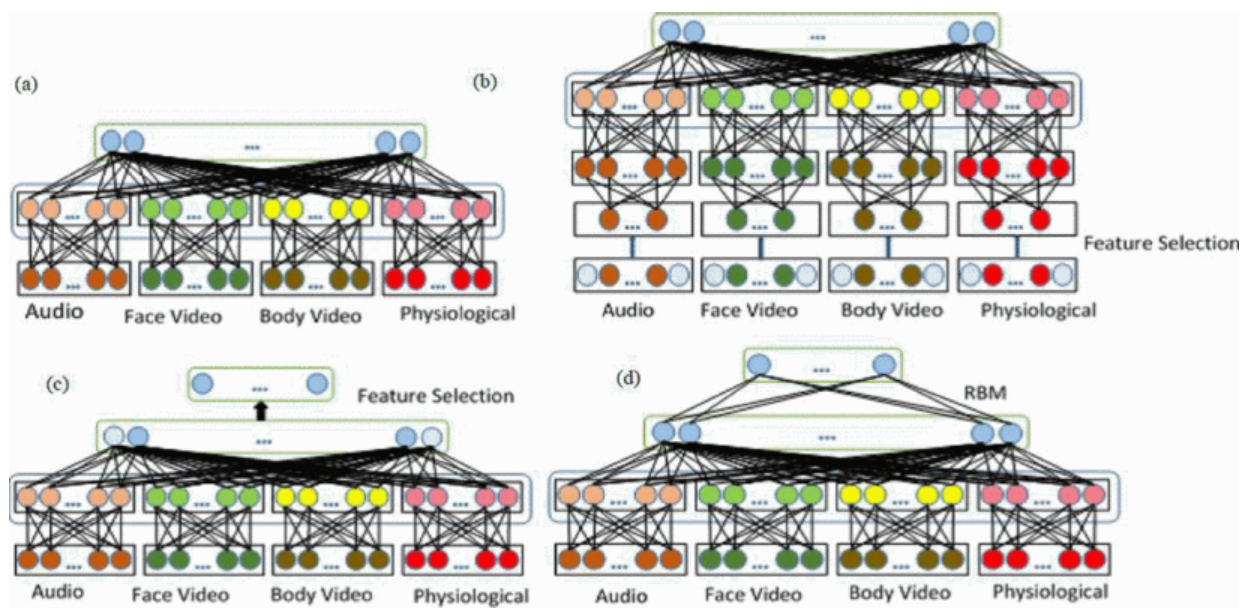
多模式网络。在训练了视觉和语音网络之后，LSTM层被丢弃，只考虑提取的特征。语音网络提取1280维特征，而视觉网络提取2048维特征。**这些连接在一起形成一个3328维特征向量，并馈送到每个具有256个单元的2层LSTM。**在基于Glorot的初始化之后初始化LSTM层的权重，并且利用单峰模型的权重初始化视觉和语音网络。最后，整个网络进行端对端训练。

对于语音、视觉和多模态网络的重复层，我们将6 s序列分割为150个小的子序列，以匹配标注频率为40毫秒的粒度。

illustrates results for a single test subject from RECOLA：



## 2-Multimodal emotion recognition using deep learning architectures



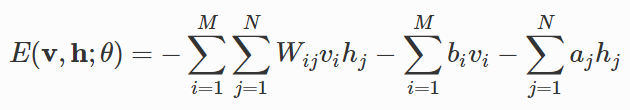
## 3-Emotion Recognition Using Multimodal Deep Learning

·Bimodal Deep AutoEncoder (BDAE)

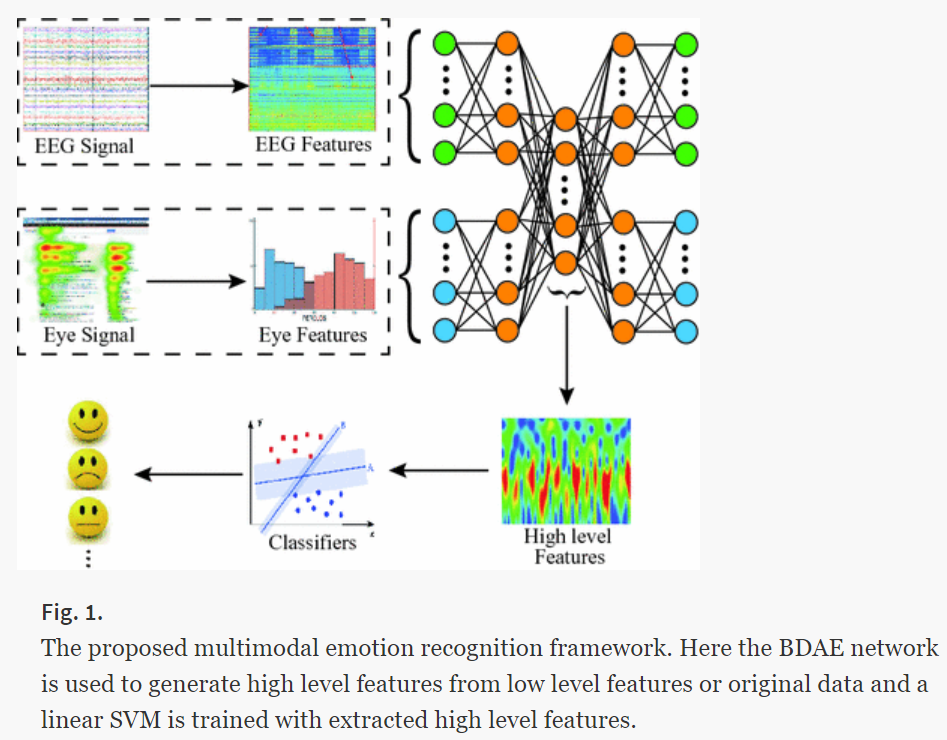
·mean accuracies of 91.01 % and 83.25 % on SEED and DEAP datasets, respectively.

·通过分析混淆矩阵confusing matrices，我们发现EEG和眼睛特征包含互补信息，BDAE网络可充分利用这种补充特性来增强情感识别。

核心原理：**Restricted Boltzmann Machine** (has a visible layer and a hidden layer)

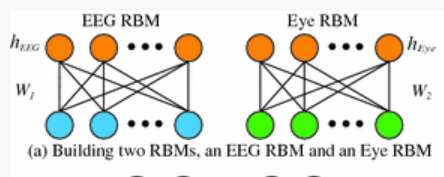


Wij is the symmetric weight between visible unit i and hidden unit j, and bi and aj are bias terms of visible unit and hidden unit, respectively.



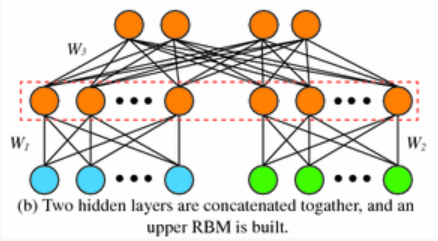
**1. Train the BDAE Network & Feature selection**

**Encoding part:**



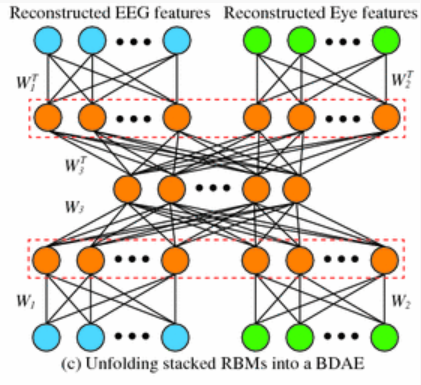
hEEG hEye: hidden layers

W: weight matrices权重矩阵



将hEEG和hEye连接在一起，作为上级RBM的可视层。

**Decoding part:**



将权重矩阵连接：W1, W2, and W3 and WT1, WT2, and WT3

最后用unsupervised back-propagation algorithm to fine-tune the weights and bias.

**2. Feature Extraction**

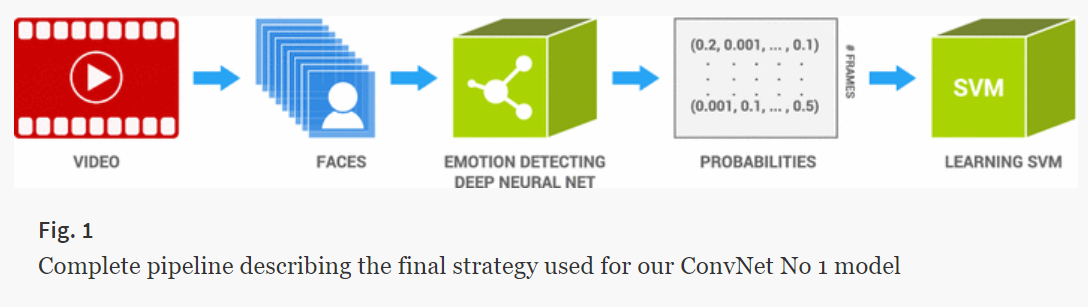
对于SEED数据集（15个人观看15段4min影片的采样率为1KHz的62通道EEG信号和眼球运动信号），从EEG数据中提取**功率谱密度（PSD）和微分熵（DE）**特征。这两种特征包含5个频带：δ（1-4Hz），θ（4-8Hz），α（8-14Hz），β（14-31Hz）和γ（31-50Hz） 。对于每个频带，提取的特征具有62个维度，并且总共有**310个维度**用于脑电特征。至于眼动数据，我们使用了与[10]相同的特征，共有41个维度，包括PSD和DE特征。提取的脑电特征和眼球运动特征然后重新调整到[0,1]，重新调整的特征被用作BDAE网络的输入。

对于**DEAP**数据集，我们直接使用下载的预处理数据作为BDAE网络的输入，以生成EEG信号和外围生理信号的共享表示。首先分离脑电信号和外周生理信号，然后将信号分割成63s。在分割之后，将同一时间段**（一秒）的不同频道数据组合**以形成BDAE网络的输入信号。然后，共享表示功能由BDAE网络生成。

2. Supervised training

3. Testing

## 4-



用额外数据集（FER dataset，TFD）训练神经网络

A random search procedure for determining the parameters of a linear per-class and per-model weighting was computed as described in Sect. 4.3, but for the AFEW4 (EmotiW 2014 challenge data).

