AMSA: Adaptive Multimodal Learning for Sentiment Analysis

JINGYAO WANG, Institute of Software Chinese Academy of Sciences, University of Chinese Academy of Sciences, China
LUNTIAN MOU*, Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Beijing Institute of Artificial Intelligence, Faculty of Information, Beijing University of Technology, China
LEI MA, TIEJUN HUANG, and WEN GAO, Peking University, China

Efficient recognition of emotions has attracted extensive research interest, which makes new applications in many fields possible, such as human-computer interaction, disease diagnosis, service robots, etc. Although existing work on sentiment analysis relying on sensors or unimodal methods performs well for simple contexts like business recommendation and facial expression recognition, it does far below expectations for complex scenes, such as sarcasm, disdain, and metaphors. In this article, we propose a novel two-stage multimodal learning framework, called AMSA, to adaptively learn correlation and complementarity between modalities for dynamic fusion, achieving more stable and precise sentiment analysis results. Specifically, a multiscale attention model with a slice positioning scheme is proposed to get stable quintuplets of sentiment in images, texts, and speeches in the first stage. Then a Transformer-based self-adaptive network is proposed to assign weights flexibly for multimodal fusion in the second stage, and update the parameters of the loss function through compensation iteration. To quickly locate key areas for efficient affective computing, a patch-based selection scheme is proposed to iteratively remove redundant information through a novel loss function before fusion. Extensive experiments have been conducted on both machine weakly labeled and manually annotated datasets of self-made Video-SA, CMU-MOSEI, and CMU-MOSI. The results demonstrate the superiority of our approach through comparison with baselines.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; *Machine learning*; • **Human-centered computing** → *Human computer interaction (HCI).*

Additional Key Words and Phrases: Sentiment analysis, multimodal fusion, self-adaptive mechanism, Transformer, patch-based selection.

1 INTRODUCTION

Since social multimedia data containing multiple modalities of content, such as visual expression, textual content, and audio description, has been increasingly popular on various social sites (e.g., Twitter, Instagram, and WeChat), sentiment analysis of it can uncover people's attitude and attention to some important and hot topics [1–7]. For example, companies can gain a more complete picture of market interest in their products [3, 4]. Governments are interested in understanding the willingness of the people through Twitter or Weibo [2]. What's more, multimedia data gives machines the basis for learning human emotions and analyzing high-level behaviors through sentiment

*This work was supported in part by the Natural Science Foundation of China under Grant 61672068. (Corresponding author: Luntian Mou.)

Authors' addresses: Jingyao Wang, jingyao_wang0728@163.com, Institute of Software Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China; Luntian Mou, ltmou@pku.edu.cn, Beijing Key Laboratory of Multimedia and Intelligent Software Technology, Beijing Institute of Artificial Intelligence, Faculty of Information, Beijing University of Technology, Beijing, China; Lei Ma, lei.ma@pku.edu.cn; Tiejun Huang, tjhuang@pku.edu.cn; Wen Gao, wgao@pku.edu.cn, Peking University, Beijing, China.

1551-6857/2022/12-ART $15.00
https://doi.org/10.1145/3572915

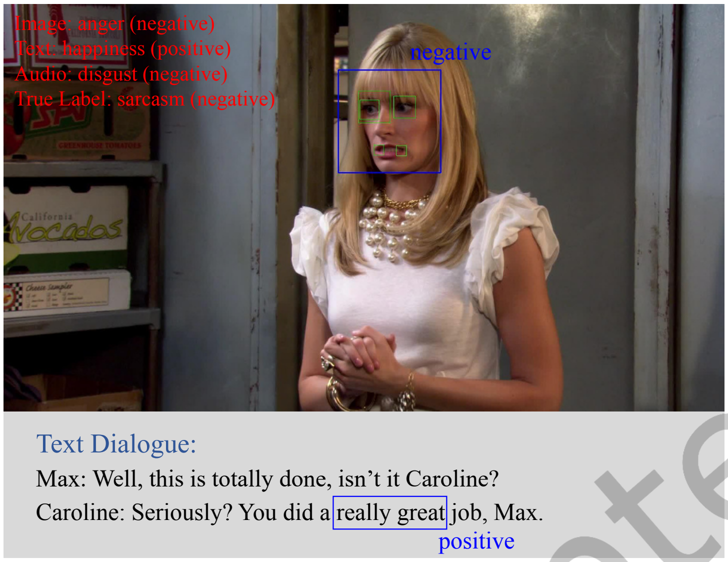


Image: anger (negative)
Text: happiness (positive)
Audio: disgust (negative)
True Label: sarcasm (negative)
negative
Text Dialogue:
positive

Max: Well, this is totally done, isn't it Caroline?

Caroline: Seriously? You did a really great job, Max.

Fig. 1. An example of an image and corresponding text description in a black comedy. The facial expression and textual message of this lady reflect negative and positive emotional states, respectively, which can easily leads to confusion. Our approach focuses on these inexplicable emotional expressions and is capable to accurately detect the negative emotional tendency in the "sarcasm" context of this data.

analysis [1]. Therefore, how to integrate multiple modalities of data (e.g., image, text, and audio), and precisely detect them for affective computing has become a very important and prominent research task [5, 8–10].

Sentiment analysis has gained widespread attention since it was proposed, and the trend increases from year to year. Traditional works mainly analyze unimodal data, such as image, text, and audio, for affective computing [11–15]. And some works perform affective computing by capturing and analyzing biological signals such as EEG or ECG based on sensors [16, 17], which helps figure out mood change and mental activities for interdisciplinary research. But with increasing needs for complex emotional computing, traditional detection of unimodal data, which are unpredictable and lagging for dynamically changing emotions, has been unable to meet the expectation.

Therefore, sentiment analysis of multimodal data has attracted increasing interest in recent years. Early researches usually perform multimodal fusion [18] through directly combining multiple sources of raw features by the early fusion [4, 19] or aggregating decisions from multiple sentiment classifiers by the late fusion [20–22]. However, the former may generate a large number of redundant input vectors, resulting in the expansion of the calculation scale, while the latter can hardly capture the correlation among different modalities. In that case, many deep fusion-based and adaptive methods [23–27] have been proposed for multimodal sentiment analysis. Pandeya et al. [27] capture all acoustic and visual emotional clues included in music videos through information-sharing methods incorporated into multimodal representations. Zadeh et al. [26] assign an LSTM (long short-term memory network) function to learn view-specific interactions in isolation and identify the Delta-memory Attention Network (DMAN) for multimodal sentiment analysis. However, experiments show that the methods mentioned above are not effective to discover the sentimental features, while extremely time-consuming to train deep fusion models.

Fortunately, the correlated features and spatiotemporal interactions of multimodal data provide clues for sentiment analysis. Firstly, there exists a close correlation among different modalities. As can be seen in Figure 1, the facial expression of the specific regions in the scene is closely related to the words “Caroline” (the female character’s name). If it can be parsed precisely to achieve joint encoding, the expression of the multimodal data will be easy to excavate. Secondly, different words and regions have different sentimental strengths for affective computing. For example, the region of facial expression is much more important than the background, and the words “great” or “truly” contains more emotional information. It’s necessary to capture the key parts of multimodal data and give proper weights for fusion. Recently, the attention mechanism has been proven to be beneficial to extract salient features from sequential data for further analysis [28–30], which can be used to solve plenty of visual or textual tasks. Therefore, the important fragments and regions can be easily highlighted by an attention-based model for more efficient prediction and further analysis. Thirdly, the combination of different modal information expresses different central emotions. As shown in Figure 1, the visual content gives us a “sadness” or “anger” impression, while the textual description (emotion-related words like “great” and “really”) represents “grateful” and “happiness”. Combining the two modalities, it can be easily judged that the context expresses "sarcasm" or "disdain". The modalities complement each other for sentiment presentation.

In this article, we propose a novel self-adaptive multimodal sentiment analysis approach named AMSA, which is designed to take advantage of the correlation and complementarity among modalities (e.g., image, text, and audio) and adaptively weighting them for affective computing. In particular, to capture and learn the close correlation among the three modalities, a multiscale attention model with a slice positioning mechanism is proposed to capture word-related visual and audio features, while locating emotional regions in the input video stream. To better exploit the complementary information, a Transformer-based self-adaptive mechanism is designed to learn the features that have been extracted and located by adaptively assigning appropriate weights for propagation through iterative refinement. To eliminate redundant information, a patch-based selection scheme is also proposed to find the emotional regions, such as the facial area of images and emotional adjectives. The main contributions are summarized as follows:

- A two-stage adaptive multimodal learning framework for sentiment analysis called AMSA is proposed to investigate the problem of leveraging correlated and complementary information in images, text, and audio, with a special focus on complex contexts, such as sarcasm, disdain and metaphors.
- A Transformer-based self-adaptive mechanism is proposed to both learn the fine-grained correlation among different modalities and iteratively update the assigned weights for fusion.
- A patch-based selection scheme is proposed to dynamically filter the most emotional fragments by calculating the gradient. To prevent important boundaries from being filtered out, we design a novel loss function to quickly update scores for key regions and the data filtered out by the current loop.
- A dataset containing multiple complex emotions is produced for multimodal affective computing, called Video-SA. The videos of this dataset are collected from real society or online series and expanded by automatic data enhancement.
- Extensive experiments have been conducted on both manually annotated and machine weakly labeled datasets, which demonstrates the superiority of our approach.

The remainder of the article is structured as follows. The related literature is reviewed in Section 2. We introduce the proposed approach in detail in Section 3. Next, the experimental results are elaborated in Section 4. Finally, the conclusions and future work will be described in Section 5.

2 RELATED WORK

Affective computing, as an emerging interdisciplinary research field, has been studied broadly and the exploration has never stopped in the decade since 1995 when it was proposed [31]. Emotion modeling has been generalized into two categories: discrete-based and dimensional-based [32]. Early approaches explore mainly for positive or negative affective tendencies and could only make binary determinations based on typical emotion descriptors such as verbal adjectives using Word2Vec and simple CNNs [33, 34]. The breakthrough of deep learning on ImageNet in 2012 provided support for more complex and large-scale affective computing. The achievements in this field have demonstrated broad potential in various aspects of society, such as Transportation (Autonomous Driving), Medicine (Disease diagnosis) and Entertainment (Interactive Games) [35, 36].

Current existing algorithms mostly rely on sensors or unimodal methods for sentiment computation, mainly include three modalities: image, text, and audio [37–41], and have obtained good results in areas such as commercial recommendations. However, with the development of multimedia and sensor technologies, the information contained in the data has become more and more complex. The unimodal algorithms have been verified to have problems of overfitting, outlier bias, and low computational efficiency in complex contexts such as sarcasm and self-deprecation [31, 42]. By contrast, sentiment analysis based on biological signals (such as ECG and EEG) completes the acquisition through wearable devices, which is mainly used in the field of medical industry [43]. But their hardware-dependent nature leads to high research costs and difficulties in generalization.

The emergence of multimodal learning provides an opportunity for further development of affective computing [44]. Compared with unimodal approaches, multimodal affective computing can capture more advanced features. It uses machine learning to correlate and characterize multimodal data, while deeply exploring the complementarity between modalities to achieve effective fusion, and thus improve accuracy.

Traditional works for multimodal affective computing use the data-level fusion (early fusion) strategy [4, 19] to directly combine multiple sources of features, or the decision-level fusion (late fusion) strategy [20–22] to reconcile the sentiment calculations of different classifiers. Mai et al. [45] conduct fusion hierarchically for visual, acoustic, and language so that both local and global interactions are considered for a comprehensive interpretation of multimodal embeddings. Hazer-Rau et al. [46] build a multimodal dataset in which the joint sentiment is classified via subjective feedback and checked for quality issues. Xu et al. [47] employ an attention mechanism to learn the alignment between audio frames and text words, aiming to produce more accurate multimodal feature representations. Poria et al. [20] propose a temporal convolutional neural network to capture sentimental features from voice tone, speech, and facial expressions. Then multimodal sentiment is predicted with a multiple kernel learning (MKL) classifier. However, the early fusion easily results in the expansion of the calculation scale due to the generation of a large number of redundant input vectors, while it is difficult to capture the correlation between different modalities from late fusion.

Due to the shortcomings of traditional fusion, many works based on deep fusion [25, 48–52] have been proposed to predict multimodal sentiment with the development of deep learning [4, 53–55]. For example, Xie et al [48] input multiple parts of semantic uncertainty into the GPT-2 model to calculate the surprise value of the joke for humor recognition. Wang et al [49] propose a pre-trained neural network model to predict soft labels, which is used to learn distributed representations of emotion categories. Basiri et al [50] build an attention-based two-way CNN-RNN deep model (ABCDM) to achieve precise business sentiment analysis. By using independent two-way LSTM and GRU layers, they calculate the time information flow in two directions and extract the context to predict emotions. Chen et al. [51] propose bi-sense emoji embeddings with an attention-based LSTM for Twitter sentiment analysis. You et al. [25] develop a cross-modality consistent regression (CCR) model to utilize both the state-of-the-art visual and textual sentiment analysis techniques. A consistent KL divergence loss is used to enforce an agreement for sentiment prediction of different modality features. To further fill this gap at the intersection of aspect-level and multimodal sentiment analysis, MIMN [52] is put forward to supervise the textual and visual information, and learn both the interactive influences between cross-modality data and the self influences in single-modality data. However, the accuracy and complexity of sentiment computing still have much room for improvement.

So far, not much research has been conducted on employing the adaptive methods or attention strategy to exploit discriminative features for image-text-audio sentiment analysis. Most multimodal works are focused on two modalities and prefer training models using the weight update strategy set in advance. Yang et al. [56]

combine Multi-view Attentional Network (MVAN) with a memory network that is continually updated to obtain the deep semantic features of image-text. They also build a large-scale image-text emotion dataset (i.e., labeled with different emotions), called TumEmo based on the proposed approach. VistaNet [57] uses visual content as an alignment combining textual and visual components to highlight salient aspects for sentiment analysis, while focusing on the important sentences of a document using attention. Xu et al. [58] propose a deep semantic network to extract sentiment-important words for affective computing, which uses an attention-based LSTM guided by visual features. There are certainly some methods that exist to perform sentiment computation based on the three modalities, but mostly for sequential learning instead of dynamic Learning using adaptive methods or attention strategy. Nguyen et al. [59] intend to incorporate the concepts of fuzzy logic into the deep learning framework and present a novel convolutional neuro-fuzzy network to extract high-level emotion features from text, audio, and visual modalities. A Transformer-based method [60] is presented for sentiment analysis that encodes representation from a transformer and applies deep intelligent contextual embedding to enhance the quality of tweets. MFN [26] is a multi-view sequential learning model, which is designed for the joint sentiment analysis of texts, videos, and audio. This proposed approach has two forms of interactions between different views: view-specific interactions and cross-view interactions.

Compared to those works, how to learn three modal features (*i.e.*, image, text, and audio) and their relationships adaptively, and boost the performance for affective computing is the central theme in our approach. With the development of sensing and multimedia technologies, it is increasingly easy to obtain multimodal data such as video, image, text, and audio. Multimodal learning can describe the same object with multiple sources of heterogeneous forms and consistent intrinsic semantics, while describing the features of the object in a particular perspective separately for different modal forms. This technique is able to overcome the limitation of sample data and perform a more refined feature representation. Therefore, this study will explore a framework with both robustness and generalization based on multimodal affective computing for complex and dynamically changing scenarios such as sarcasm and mental illness.

3 METHOD

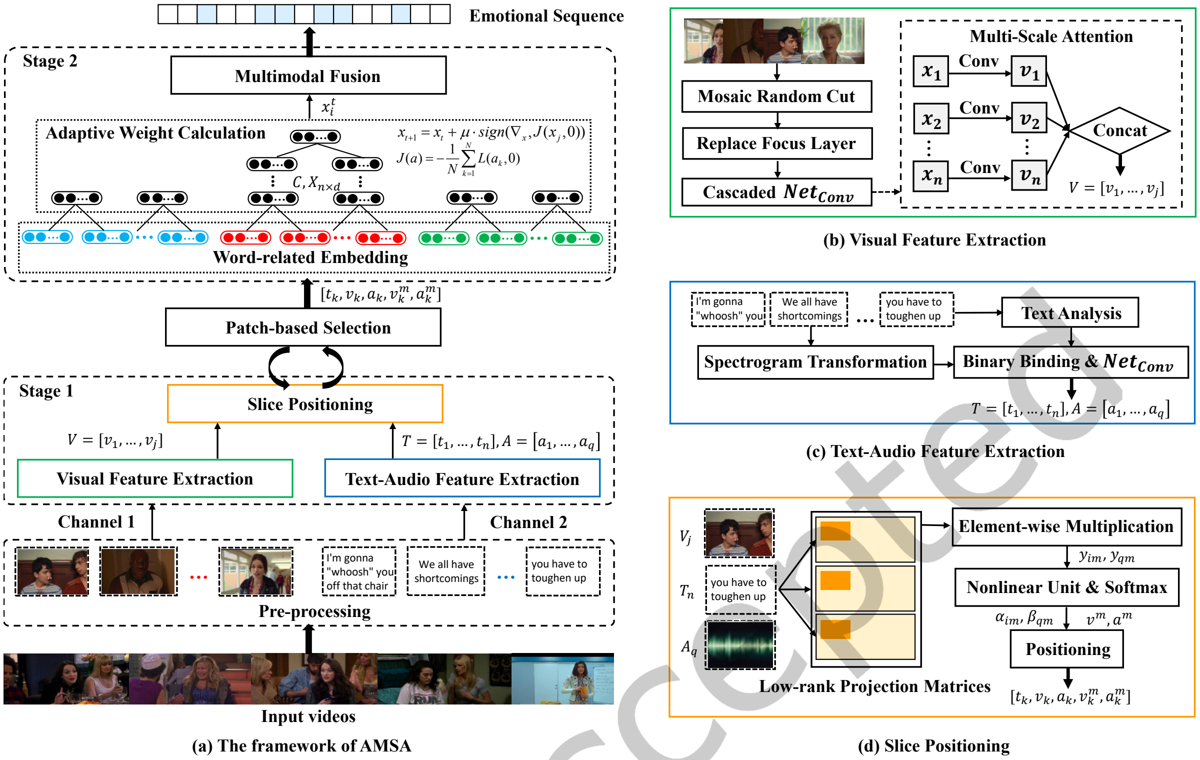
In this section, we illustrate an overview of the proposed framework and then detail the innovative components of the model for image-text-audio sentiment analysis, especially for complex contexts.

3.1 Architecture Overview

The affective computing of social media data has been broadly studied. In this work, we focus on mining the correlation and complementarity among the modalities of image, text, and audio, while adaptively assigning weights for multimodal fusion, which helps achieve a more effective sentiment analysis. Figure 2 shows the proposed self-adaptive multimodal sentiment analysis framework (AMSA), which is divided modularly into the following two stages.

In the first stage, to efficiently distinguish the emotional features of the three modalities (image, text, and audio), a multiscale attention model is proposed to extract word-related visual features and word-related audio features. After acquiring the features, a novel slice positioning mechanism is presented to integrate the three modal information and their correlations into a quintet to be sent to the fusion stage.

For the second stage, a novel Transformer-based self-adaptive network is proposed to perform multimodal fusion, especially for complex contexts. It will learn and assign proper weights for elements in the quintets where the emotional data of three modalities will be integrated as a sequence at the same time. The dynamic weight assignment is performed based on the heterogeneous data with the confidence of each element in the previous stage, and the obtained weights will be added into iterations for sentiment analysis until the result becomes steady.


Emotional Sequence
Stage 2
Multimodal Fusion
x_i^t
Adaptive Weight Calculation
$x_{t+1} = x_t + \mu \cdot sign(\nabla_x, J(x_j, 0))$
$J(a) = -\frac{1}{N}\sum_{k=1}^{N} L(a_k, 0)$
$C, X_{n\times d}$
Word-related Embedding
$[t_k, v_k, a_k, v_k^m, a_k^m]$
Patch-based Selection
Stage 1
Slice Positioning
$V = [v_1, \dots, v_j]$
$T = [t_1, \dots, t_n], A = [a_1, \dots, a_q]$
Visual Feature Extraction
Text-Audio Feature Extraction
Channel 1
Channel 2
I'm gonna "whoosh" you off that chair
We all have shortcomings
you have to toughen up
Pre-processing
Input videos
Mosaic Random Cut
Replace Focus Layer
Cascaded Net_{Conv}
Multi-Scale Attention
x_1
x_2
x_n
Conv
v_1
v_2
v_n
Concat
$V = [v_1, \dots, v_j]$
Text Analysis
Spectrogram Transformation
Binary Binding & Net_{Conv}
$T = [t_1, \dots, t_n], A = [a_1, \dots, a_q]$
V_j
T_n
A_q
Element-wise Multiplication
y_{im}, y_{qm}
Nonlinear Unit & Softmax
α_{im}, β_{qm}
v^m, a^m
Positioning
Low-rank Projection Matrices
$[t_k, v_k, a_k, v_k^m, a_k^m]$

(a) The framework of AMSA

(b) Visual Feature Extraction

(c) Text-Audio Feature Extraction

(d) Slice Positioning

Fig. 2. The framework of the proposed AMSA. (a) describes the framework which mainly contains two stages. The multimodal data pre-processed from input videos will be divided into two channels for feature extraction and obtaining stable quintets. The Transformer-based self-adaptive network in stage 2 will perform the weight calculation and complete the fusion which is shown in Figure 3. The blue, red, and green nodes in the figure represent the three modalities of text, image and audio respectively, and the model will calculate the optimal weights of them. The final result is an emotional sequence containing the confidence score of 15 contexts emotions. (b), (c), (d) illustrate the processing details of the three modules (Visual Feature Extraction, Text-Audio Feature Extraction and Slice Positioning) in stage 1, respectively.

3.2 Multiscale Attention Model with Slice Positioning

The attention mechanism can flexibly concentrate on specific emotional regions to capture high-level features. It has been proven to be beneficial to many multimodal tasks in recent years [28, 61–64]. For the sentiment analysis tasks of this article, it is necessary to dynamically extract correlation and complementarity among multiple modalities—image, text, and audio. Therefore, a novel multi-scale attention model with a slice positioning mechanism is proposed for feature extraction and complementarity analysis of multimodal data, which is the first stage of the proposed AMSA.

The calculation of fine-grained correlation and complementarity using this model requires two steps: feature extraction for two channels image-text and audio-text, and sentimental regions combination based on slice positioning mechanism. We design a novel network unit called Net_{Conv} based on a convolutional neural network with batch normalization etc. The network contains several basic modules that can perform fast filtering of information such as images and make sentiment-related judgments based on the $Softmax$ function.

Algorithm 1: Multi-scale Feature Extraction and Learning

Input: I_0 (Input video data containing multiple modal information).

Parameters: i (iterations), i_{Max} (iteration upper limit: 80k), i_{stable} (i_{stable}-th iteration when results are stable), ε (detection deviation lower limit), j (subscript of emotional sequence), $\{V, T, S\}$ (image, text, audio sequence from I_0), $\{v_k, t_k, s_k\}$ (k-th element of $\{V, T, S\}$), X_i (sentiment feature map after the i-th iteration), $\sum_{j=0}^{n} g_{ij} \in G$ (G: sentiments sequence after slice positioning, g_j: the result of the jth emotion at the i-th iteration), $Net1$, $Net2$ (discriminator for image and text-audio), Net (slice positioning mechanism).

Output: O (Emotion detection results after optimization).

```
/* INITIALIZATION                                                              */
Initialize $O_0 \Leftarrow 0\ (normal),\ j \Leftarrow 0,\ i \Leftarrow 0,\ X_0 \Leftarrow \emptyset,\ G \Leftarrow \emptyset,\ \{v\} \Leftarrow \emptyset,\ \{t\} \Leftarrow \emptyset,\ \{a\} \Leftarrow \emptyset$;
/* CALCULATION                                                                 */
while $\sum_{j=0}^{n} g_{ij} \in G$ and $i \in [1, i_{Max}]$ do
    read current;
    $I^{'} \Leftarrow$ get and fill the $I_0$;
    $I_1, I_2, I_3, I_4, I_5, I_6 \Leftarrow$ two stream feature extraction pyramid $I^{'} \cap (X_i \Leftarrow (X_{i-1} \Leftarrow g_i))$;
    $\{I\} \Leftarrow I_1, I_2, I_3, I_4, I_5, I_6$;
    repeat
        define $X_i$;
        repeat
            $\{v_i\} \Leftarrow Net1\,(\{I\}),\ \{t_i, a_i\} \Leftarrow Net2\,(\{I\})$;
            /* SLICE POSITIONING                                                   */
            $g_i \Leftarrow Net\,(\{v_i\}, \{t_i, a_i\})$;
        until all element values are obtained;
    until obtain stable multi-scale feature fusion results;
    while $X_i - X_{i-1} < \varepsilon$ do
        $X_i \Leftarrow Net\,(\{v_i\}, \{t_i, a_i\})$;
    end
    Calculate $\sum_{j=0}^{n} g_{ij} \in G$ based on patch-based selection mentioned in Section 3.4;
    Update and Extract $g_i$;
    for each $g_i = \sum_{j=0}^{n} g_{ij} \in G$ do
        /* UPDATE FEATURE MAP                                                  */
        Update $X_i \Leftarrow$ Integrate $(X_i \cup g_i)$;
    end
    $O \Leftarrow X_{i_{stable}}$;
end
```

The extraction of visual features is the operation of the first channel. Given the image-text pair, the area ratio calculation is first used to perform Mosaic random cropping on small-frame faces for enhancement. Then the expanded data will be put into the backbone, which is formed by stacking four basic network modules—Net_{Conv}. For higher-level visual features extraction, we replace the Focus layer with the Stem block and use a smaller kernel, which increases the generalization ability of the network and reduces the complexity. In order to calculate the face part as comprehensively as possible, we additionally set up a multi-scale output layer and added a

128-stride as the output block. On this basis, the visual region features $V = \left\{v_1, ..., v_i, ..., v_j\right\} \in \mathbb{R}^{j\times d}$ is captured through the pre-trained cascaded deep CNN of backbone, where j is the number of regions and d is the feature dimension of each region.

For the feature extraction of text and audio in the second channel, the dynamic time warping and Fourier transform outline texts and audios of the input videos. After acquiring the audio, the clutter is filtered out using wavelet transform, which will be extracted as a speech spectrogram based on DTW. Since we get pure data about text and audio, the backbone of the second channel for feature extraction is only one basic module–Net_{Conv}. The pre-trained word embedding is employed to obtain textual features $T = \{t_1, ..., t_m, ..., t_n\} \in \mathbb{R}^{n\times e}$ and audio features $A = \left\{a_1, ..., a_p, ..., a_q\right\} \in \mathbb{R}^{q\times f}$, where n is the number of words, q is the length of speech spectrogram, e and f are the embedding dimensions. Specifically, the e for texts is set to three levels of "good" (high tone), "neutral" (flat tone) and "bad" (low tone), while f refers to the context-free modeling [65] model released by MIT to detect the "upward" and "downward" segments in the audio.

All of the extracted features are located and combined through learning the attention of each element, which is the slice positioning mechanism. To learn the attention over regions from each word, the three modalities are first fused for the computation of the attention scores using three low-rank projection matrices, which can project the three feature vectors (i.e., image region vector V_j, word vector T_n and audio vector A_q) into a z-dimensional common space. Then two fused vectors y_{im} and y_{qm} will be calculated using element-wise multiplication:

$$y_{im} = \left(U^T v_i \odot V^T t_m\right),\ U \in \mathbb{R}^{d\times z},\ V \in \mathbb{R}^{e\times z} \tag{1}$$

$$y_{qm} = \left(W^T a_p \odot V^T t_m\right),\ W \in \mathbb{R}^{f\times z},\ V \in \mathbb{R}^{e\times z} \tag{2}$$

where $\odot$ means the Hadamard product of two vectors, and the fused vectors y_{im} and y_{qm} are fed into a nonlinear process with Softmax to make the final elemental attention score determination of each modality into "good", "fair" and "bad". The attentive strength will be regulated with attention scores, and the final average weighted confidence in the sentiment of the features, image-text v^m and audio-text a^m is calculated as follows. The α_{im} and β_{pm} are normalized attention scores of v^m and a^m calculated with Softmax, where $w^T \in \mathbb{R}^c$ and $b \in \mathbb{R}^1$ are weight and bias parameters, respectively.

$$v^m = \sum_{i=1}^{j} \alpha_{im} \cdot v_i,\ \alpha_{im} = \frac{exp\left(\sigma\left(w^T y_{im} + b\right)\right)}{\sum_{i=1}^{j} exp\left(\sigma\left(w^T y_{im} + b\right)\right)} \tag{3}$$

$$a^m = \sum_{p=1}^{q} \beta_{pm} \cdot a_p,\ \beta_{pm} = \frac{exp\left(\sigma\left(w^T y_{pm} + b\right)\right)}{\sum_{p=1}^{q} exp\left(\sigma\left(w^T y_{pm} + b\right)\right)} \tag{4}$$

The final result will also be normalized and calculated as shown below:

$$\hat{y} = argmax_{y\in\tau^l} p\left(y|X, M\left(X, N\right)\right) \tag{5}$$

τ represents the set of tags K of all staging results in the sentence ($K = 3$); l is the sentence length; $\sigma(\cdot)$ is sigmoid function that $\sigma(x) = \frac{1}{1+e^{-x}}$; $\hat{y}$ is the optimal result of the model. X and M respectively represent the input sentence and the stable model after iteration.

The slice positioning mechanism is also used to locate calculated emotional features and bind the elements by calculating the positions where the three modal features appear. Based on this, it forms a multi-pair quintet $\left\{t_k, v_k, a_k, v_k^m, a_k^m\right\}$ (k from 1 to the end) containing sentiment features.

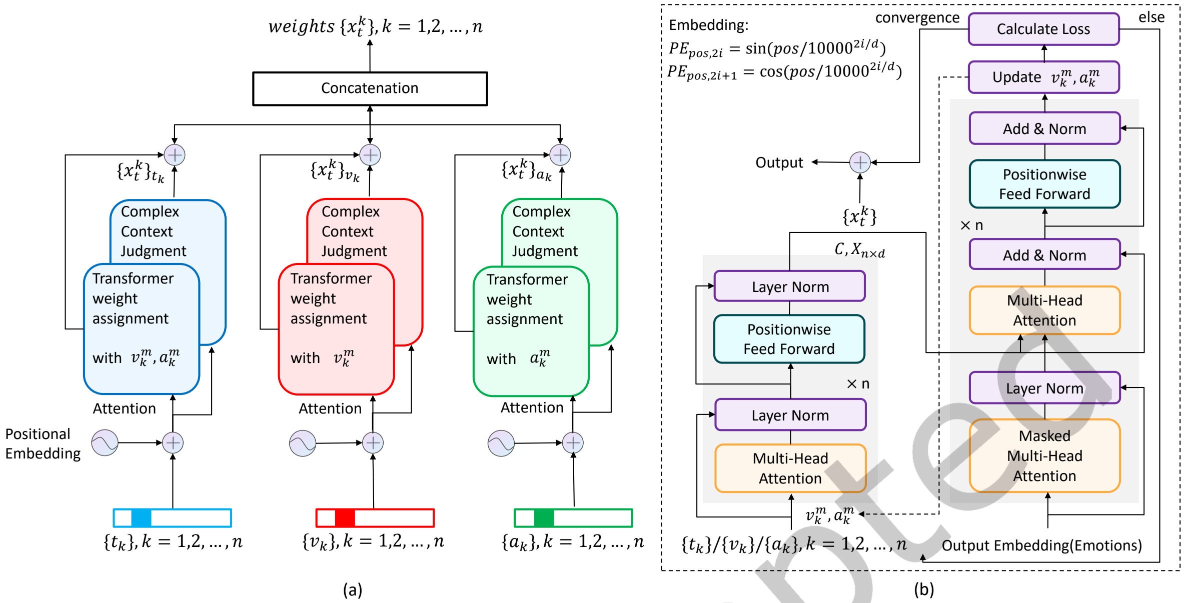

weights $\{x_t^k\}, k = 1,2,\ldots,n$
Concatenation
$\{x_t^k\}_{t_k}$
$\{x_t^k\}_{v_k}$
$\{x_t^k\}_{a_k}$
Complex Context Judgment
Transformer weight assignment with v_k^m, a_k^m
Transformer weight assignment with v_k^m
Transformer weight assignment with a_k^m
Attention
Positional Embedding
$\{t_k\}, k = 1,2,\ldots,n$
$\{v_k\}, k = 1,2,\ldots,n$
$\{a_k\}, k = 1,2,\ldots,n$
(a)
Embedding:
$PE_{pos,2i} = \sin(pos/10000^{2i/d})$
$PE_{pos,2i+1} = \cos(pos/10000^{2i/d})$
convergence
Calculate Loss
else
Update v_k^m, a_k^m
Add & Norm
Positionwise Feed Forward
× n
Add & Norm
Multi-Head Attention
Layer Norm
Masked Multi-Head Attention
Output
$\{x_t^k\}$
$C, X_{n\times d}$
Layer Norm
Positionwise Feed Forward
× n
Layer Norm
Multi-Head Attention
v_k^m, a_k^m
$\{t_k\}/\{v_k\}/\{a_k\}, k = 1,2,\ldots,n$
Output Embedding(Emotions)
(b)

Fig. 3. The workflow of Transformer-based self-adaptive weight assignment. (a) shows a brief flow of the mechanism, and (b) explains the details of the model.

3.3 Transformer-based Self-adaptive Mechanism

Emotions in complex contexts are difficult to determine through the tendency of different modalities, e.g. "anger" image + "happiness" text + "disgust" audio = "sarcasm", while their internal links are also immeasurable. Therefore, a Transformer-based self-adaptive mechanism is proposed to dynamically assign weights to the heterogeneous results for multimodal fusion. Compared to the existing strategy of fixed weight fusion, it not only enhances the interpretability of the model, but also eliminates the single-modal measurement errors and the limitations of CNN-based methods to the greatest extent by considering the change of emotions.

The Transformer-based self-adaptive model consists of 6 Encoders and 6 Decoders, and Figure 3 shows the details of Transformer-based self-adaptive weight assignment. Its input is multi-pair vectors $\left\{t_k, v_k, a_k, v_k^m, a_k^m\right\}$ obtained in the slice positioning scheme, in which the input represents the modality information of the k-th fragment while the original input data input is a sequence of length n after removing redundancy. Specifically, the elements in these vectors include text feature labels (good, neutral, bad), visual feature labels (anger , disgust, fear, happiness, sadness, and surprise), audio feature labels (upward, flat, downward), while image-text features and audio-text features contain tendency information (same, opposite). The role of its encoder and decoder is to discriminate variability in the temporal dimension through mapping the input quintet to the hidden layer, then enabling re-adjustment after sentiment calculation based on the updated sequence. As shown in Figure 2, we will obtain a sequence containing the image, text, and audio information with sentiment fusion weights x_i. Ultimately AMSA can use the weighting of this sequence to discriminate the final results.

After 6 encoder blocks, we will get the encoding information matrix C of all words in the sentence, and the word vector matrix is represented by $X_{n\times d}$. Next, the decoder will train word embedding and position embedding of the word-related elements, which represent the positions v_k^m, a_k^m in the sentence, respectively. Instead of using the RNN structure, our Transformer uses the global information so that the relative or absolute position of the word in the sequence is preserved. The calculation for Embedding is as follows:

$$PE_{pos,2i} = sin\left(pos/10000^{2i/d}\right) \tag{6}$$

$$PE_{pos,2i+1} = cos\left(pos/10000^{2i/d}\right) \tag{7}$$

where i is the element index of the text t_i^m in the current quintuple, and the Decoder will calculate the position of the feature v_i^m and a_i^m corresponding to the next word t_{i+1}^m according to the currently positioned t_1^m to t_i^m. And pos denotes the position of the t_i^m in the sentence; d denotes the dimension of PE (as in word Embedding); $2i$ denotes the even dimension, and $2i+1$ denotes the odd dimension (*i.e.*, $2i \leq d$, $2i+1 \leq d$).

The constructed transformer-based method will be used to update the weights by iterative refinement. Firstly, we set the initial weight $x = 1/n; \sum_{i=1}^{n} x_i = 1$ for each modality. The sum of the weights equals 1, and n represents the number of modalities ($n = 3$ in this article: image, text, and audio).

Next, we use the compensation iteration method for calculation. This step is designed according to the confidence of results calculated in stage 2 and results for final fusion. The following formula shows how to modify the weights.

$$x_{t+1} = x_t + \mu \cdot sign\left(\nabla_x, J\left(x_i, 0\right)\right) \tag{8}$$

$$J(a) = -\frac{1}{N}\sum_{k=1}^{N} L\left(a_k, 0\right) \tag{9}$$

among them, x_{t+1} and x_t represent the calculation results of $(t+1)^{th}$ round and $(t)^{th}$ round. $L(\cdot,\cdot)$ is the Mean Square Error (MSE), and $J(a)$ is the loss function we use to optimize the weight (μ: the learning rate–hyperparameters obtained by experiments; ∇_x: the current gradient; a_k: the confidence of the kth sentiment category; N: Top K categories).

For multimodal fusion (back-end fusion), based on the Transformer sequence encoding, we use multiple cross-modal attention blocks and self-attention to encode multimodal sequences for classification: we construct a cross-modal transformer for each modality in the previous stage and augment the representation of a single modality with a fused representation.

3.4 Patch-based Selection Scheme

Since emotion is changeable, it is necessary to continuously process heterogeneous information with high speed to analyze emotions dynamically. The amount of calculation need updating will reach a terrifying scale. Therefore, a novel patch-based selection scheme is proposed to filter out disturbing areas and dynamically search for key regions based on the cumulative gradient heat map.

As shown in Figure 4, a correction cell is added to the blocks according to the gradient of each iteration, and it will be multiplied by a mesh network into a mesh-like patch, which will be added to the original stream. The $loss_{keyArea}$ is created to calculate the region score.

$$loss_{keyArea} = -mse\left(conf, 0\right) \tag{10}$$

where mse means Mean Squared Error, and $conf$ represents the confidence of emotion. For different modalities, the physical meaning of $conf$ is different. For example, it refers to the confidence of the final proposal for images, while refers to the confidence of Top K (K=3) elements of selected output word vector for texts. In this way, we can find the influence of the current retained segment, and also realize the targeted deletion of redundant segments through the manipulation of obstacles as shown in Figure 4. For example, a text sentence -" My name is..., the weather is so nice today, isn't it? " retains "nice" after processing.

However, although our operation eliminates most of the redundant information, it is easy to delete boundaries during the process, which may lead to overall performance degradation. Therefore, it is necessary to consider the

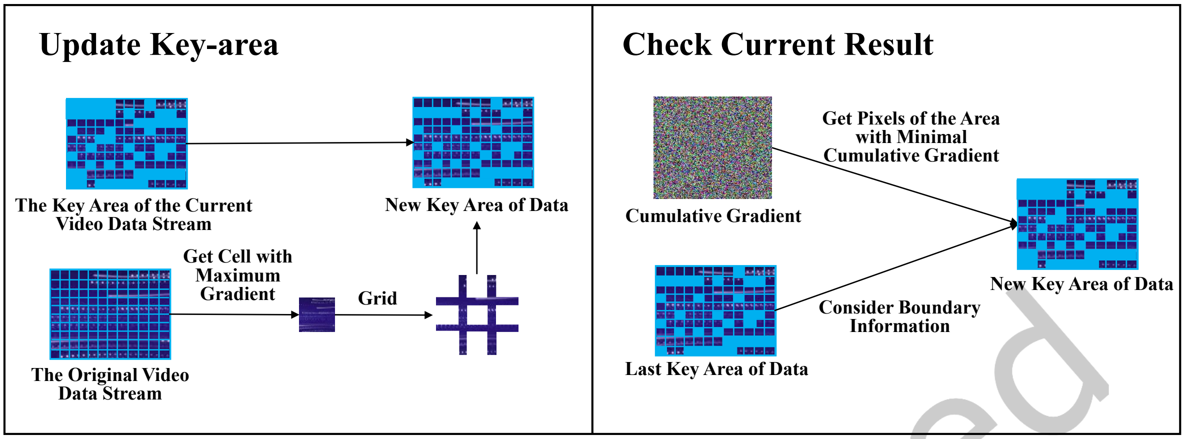

Update Key-area
The Key Area of the Current Video Data Stream
New Key Area of Data
Get Cell with Maximum Gradient
Grid
The Original Video Data Stream
Check Current Result
Get Pixels of the Area with Minimal Cumulative Gradient
Cumulative Gradient
New Key Area of Data
Consider Boundary Information
Last Key Area of Data

Fig. 4. The workflow of the proposed patch-based selection scheme. Update Key-area: it dynamically searches for the key area based on the cumulative gradient heat map and transforms it to participate in the convolution. Check Current Result: in order to prevent the loss of important pixels, the filtered redundant information is screened to determine whether it belongs to the boundary.

Algorithm 2: Patch-based Selection

```
Input: $G = (V; T, A)$ ($\{V, T, A\}$ are defined parameters in Algorithm 1: image, text, audio modality from $I_0$).
Parameters: $\theta_{patch} \Leftarrow \left\{\theta^l_{keyArea}, \theta^l_{edge}\right\}^L_{l=1}$ (the weight about Key area and boundary information).
Output: Key areas $\{\hat{y}_i\}$ of sentiment analysis after optimization.
/* INITIALIZATION                                                    */
Initialize $\theta^G_{patch}$, $\forall i, j, \theta_p \Leftarrow \theta_{patch}$, $\theta_k \Leftarrow \theta_{keyArea}$, $\theta_e \Leftarrow \theta_{edge}$;
/* CALCULATION                                                       */
for $l = 1, ..., L$ do
    /* KEY-AREA FEATURE UPDATE                                       */
    for $i = 1, ..., i_{Max}$ do
        $g^l_i \Leftarrow KeyAreaUpdate\left(\{g^l_{i=1}, [v^l_i; t^l_i, a^l_i]\}, \theta_k\right)$;
    end
    /* EDGE FEATURE UPDATE                                           */
    for $(i, j) = 1, ..., |i_{Max}, j_{Max}|$ do
        $g^l_{i,j} \Leftarrow EdgeUpdate\left(g^l_{i-1,j-1}, \theta_p\right)$;
    end
end
/* QUERY KEY-AREA PREDICTION                                         */
$\hat{y}_i \Leftarrow KeyArea - Edge2Pred\left(g^l_{i,j}\right)$;
```

boundary error. Firstly, an expansion operator [41] is carried out to expand the boundary and prevent the loss caused by convolution. Secondly, random boundary elimination is used in the process of subsequent iterative calculations. After that, the $Loss_{disapper}$ is calculated by removing the points whose change value is less than the average change value divided by 3.

$$Loss_{disapper} = -mse\left(conf_{edge}/conf, 0\right) \tag{11}$$

where $conf_{edge}$ epresents the confidence considering boundary, and $conf$ is the confidence without random boundary elimination. We repeat this process until reaching iteration upper limit I, which equals 3.5k in this paper, or the result is relatively stable. The details of the scheme can be referred to Algorithm 2.

Additionally, for the iterative refinement of this section, we design a novel loss function to help compute key regions and important boundary dynamically, which helps limit the iterative training and make the model approach the optimal result.

$$Loss = \alpha \cdot Loss_{keyArea} + \beta \cdot Loss_{disappear} \tag{12}$$

$Loss_{keyArea}$ and $Loss_{disapper}$ respectively represent the values of key regions and boundaries during the patch-based iteration mentioned above. The parameters α and β are used to adjust their specific weights, which are assigned and learned by calculating the cumulative gradient based on the Transformer-based self-adaptive mechanism (described in section 3.3).

4 EXPERIMENT

The proposed AMSA is comprehensively evaluated on several real-world datasets in this section. The experimental results are presented and detailed below.

4.1 Datasets

Three datasets obtained from social media websites (e.g., Youtube, Twitter) or self-collection are used for experiments, which are described below:

Video-SA dataset (Ours). This multimodal dataset for complex contexts contains 15 contexts under six basic emotions (anger, disgust, fear, happiness, sadness, and surprise). It consists of six representative contexts of basic emotions, such as laughing for happiness, crying for sadness, and nine complex contexts including sarcasm, weeping with joy, self-deprecating, melancholic, worried, jealous, ashamed, metaphor (positive), and metaphor (negative). The dataset has 1350 videos (30s, 600x360, 30fps), with 675 sentiment data and 675 neutral data, and is divided into two parts. The first part is collected from 5 volunteers in the 15 contexts (each person is filmed for 5 emotional and 5 neutral videos in each context, $5 \times 15 \times (5 + 5) = 750$ in total). The second part is collected from hit TV series (e.g., 2 Broke Girls, The Big Bang, etc.) with 600 videos in the 15 contexts (each containing 20 emotional and 20 neutral samples), such as the samples shown in Figure 1 and Figure 6. The details of our Video-SA can be found on: https://github.com/WangJingyao07/Video-SA-intro.

CMU-MOSEI dataset [66]. This is the world's largest multimodal dataset for sentiment analysis at the moment. More than 23,500 sentence utterance videos from over 1,000 online YouTube speakers are included. To establish sample balance, all data videos are randomly selected from various monologues or thematic videos and are gender-balanced. All of them are accurately written and punctuated. The CMU-MOSEI dataset is based on the CMU-MOSI dataset described below, but it is more complete and better suited to large-scale sample training.

CMU-MOSI dataset [67]. This dataset was previously released, albeit on a smaller size and with only sentiment labels. It has processed experimental data as well as some raw data, just like MOSEI. However, instead of captured photos, these are still videos. In the experiment, the CMU-MOSI dataset is used as the main dataset for small-batch sample training to analyze the correlation.

4.2 Baselines

We compare our AMSA with the following baselines, which include unimodal and multimodal methods.

DialogueCRN [68] focuses on dialogue emotion recognition (ERC), which examines the emotion in a dialogue scene. The program first employs a longitudinal LSTM to imitate the process of iterative cognition for each utterance. Then, after a series of iterations, it predicts emotions. We get ideas from the splicing of situation-level and speaker-level features.

DualGCN [69] performs fine-grained sentiment analysis using a dual-graph convolutional network model. It presents a SynGCN module with a self-attention mechanism and a regularizer to learn uncaptured semantic information, taking into account the complementarity of syntactic structure. Through comparison with DualGCN, we intend to uncover the optimizable fraction of AMSA in invisible feature analysis. The study also attempts to investigate the detection of both visible and hidden elements.

Visnet [70] faces the realistic issue of network emotion identification as a multimodal baseline method. It uses visual information as a way of alignment, and it employs attention mechanisms to highlight relevant sentences in the document. From two perspectives, visual features, and textual attention, it serves as a reference for us.

HAN-VGG and TFN-VGG [71] both are the most recent text sentiment analysis technologies, combining the HAN-ATT and VGG networks. HAN-ATT is built on a word encoder and does sentiment analysis using the document's hierarchical structure. In contrast to HAN-ATT, HAN-VGG adds a visual mechanism. TFN-VGG, unlike the previous model, employs the Tensor Fusion Layer and receives the final emotional label via the emotional reasoning sub-network. We also discuss HAN-various ATT's pooling and hyperparameter settings, as well as expand its variants to compare with our AMSA.

4.3 Experimental Setup

Combined with important indicators such as robustness and real-time performance, we design various experiments for comprehensive evaluation. The experimental settings are as follows:

Network Architecture. The framework consists of four aspects: a multiscale attention model with slice positioning, a Transformer-based self-adaptive network, a patch-based selection scheme, and iterative optimization. We use a convolutional neural network with multiple blocks for feature embedding as shown in Figure 5, incorporating 33% convolution, batch normalization, and the Leaky ReLU activation structure.

Train. The Adam optimizer is used to train the AMSA on a ten-card server. The NVIDIA Tesla P4 GPU is used to run our code. The learning rate is initially set to 1×10^{-5}. The weight decay lower limit is 5×10^{-7}. The mini-batch size of the meta-training task is set to 60, 40, and 20 for the subsequent multi-channel trials (3 channels, 6 channels, and 9 channels). We set the learning rate for the Video-SA dataset to halve every 10k, whereas it drops to 5k for CMU-MOSI. We chose a value of 20k for CMU-MOSEI. We also set the upper limit of the first iteration to 80k during the training process.

Evaluate. Three-way and six-way experiments were designed to perform preliminary validation. Each of the five test sets consists of 30 randomly selected queries, and performance is judged based on an average of 500 randomly created groups. In addition, we present a quantitative evaluation of accuracy and time fraction (computational speed) metrics. We executed a difficult 9-way experiment on Kaggle to demonstrate adaptation, especially in the case of different training and testing classes.

4.4 Results

Performance of AMSA. Sentiment analysis results of AMSA are shown in Figure 6. In this paper, a total of 6 basic emotional cores (anger, disgust, fear, happiness, sadness, and surprise) are selected, along with 18 contexts (typical scenes of laughing, crying, etc., as well as sarcasm that implies anger, self-deprecation that implies sadness, etc.). We use the collected Video-SA dataset to train the model and perform 10-fold cross-validation, and

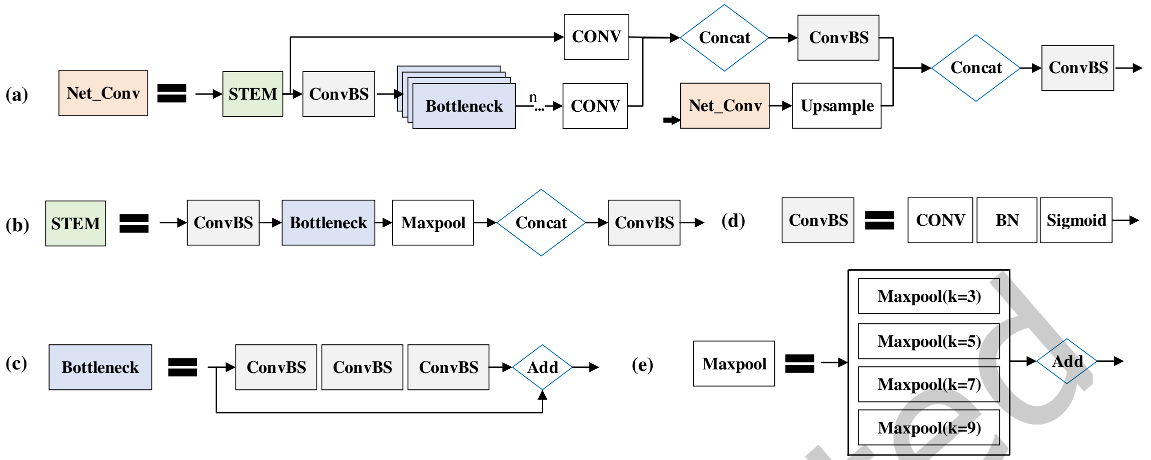


Fig. 5. The structure of Net_{Conv} used in AMSA. (a) shows the structure and composition of the network. (b), (c), (d), (e) are basic modules of Net_{Conv}.

then test the trained model in five conditions (dark unobstructed, bright unobstructed, natural unobstructed, natural light obscured by 10%, and natural light obscured by 30%) in the original (scenes used in training) and newly selected (unknown scenes) mentioned above. In Figure 6, the data we chose mostly for complex situations, such as sarcasm and metaphor, and listed the textual details included. The black labels below the data represent the true values, while the blue and red ones are for accurate and incorrect predictions, respectively. Through statistics, we found that AMSA's performance in complex contexts has a significant improvement compared to the baselines, averaging 9.6%. However, there is no significant breakthrough in the performance of rhetoric such as "metaphor" that the accuracy rate reached only 51%, and still much room for improvement.

In order to more comprehensively evaluate the performance of AMSA in different contexts, we conducted multi-fold experiments on the three datasets introduced in Section 4.1, including 5-fold, 8-fold, and 10-fold cross-validation. To ensure real-time performance, time limits (based on the average of 20 random tests) are added to the experiments, and the results of the experiments are averaged after removing the scores at both ends, as shown in Table 1. Figure 7 shows the confusion matrix for AMSA performance on the three datasets.

It is seen that the accuracy is above 70% on three distinct scale datasets. However, the CMU-MOSI and CMU-MOSEI results are lower, while the latter is significantly below the prediction. The scale of CMU MOSI is tiny, hence its ability to characterize unfamiliar situations is limited. For CMU-MOSEI with a much larger scale, maybe it is owing to a lack of computing capacity from the perspective of the computer platform that the calculation can not be completed within the time limit.

Comparison With Baselines. The baseline methods and their improvements described in section 4.2 on a horizontal scale are performed in this experiment, which uses multiple indicators, such as accuracy, recognition range, and computation speed, for evaluation. To undertake a more comprehensive analysis, we present a range of single-modal techniques. Naturally, all of the approaches have been altered to meet this experiment, such as image preprocessing, voice characterization, and so on.

As shown in Table 2, our method is superior in terms of detection performance, which means it can better balance accuracy and speed, even while working under time constraints. This occurrence illustrates its ability for adapting effectively to new surroundings. It has been discovered that several reinforcement learning-based efforts

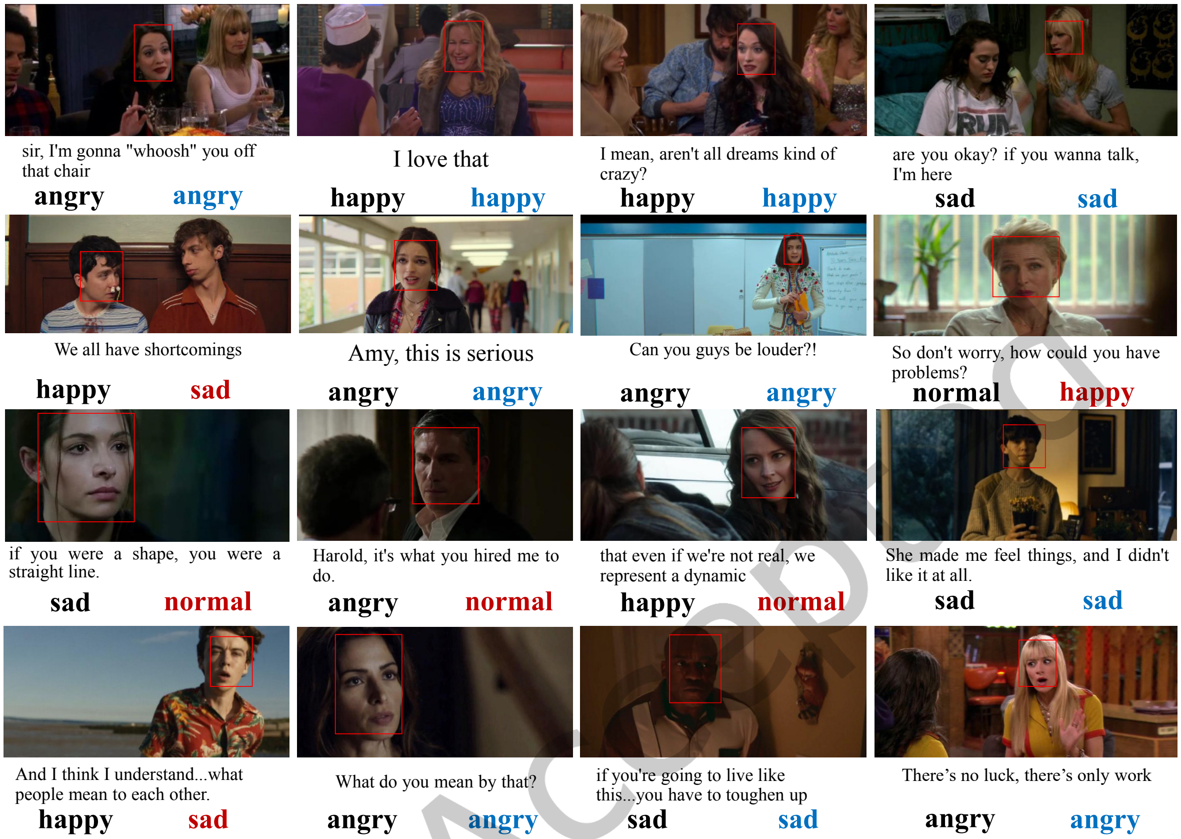

sir, I'm gonna "whoosh" you off that chair
angry angry
I love that
happy happy
I mean, aren't all dreams kind of crazy?
happy happy
are you okay? if you wanna talk, I'm here
sad sad
We all have shortcomings
happy sad
Amy, this is serious
angry angry
Can you guys be louder?!
angry angry
So don't worry, how could you have problems?
normal happy
if you were a shape, you were a straight line.
sad normal
Harold, it's what you hired me to do.
angry normal
that even if we're not real, we represent a dynamic
happy normal
She made me feel things, and I didn't like it at all.
sad sad
And I think I understand...what people mean to each other.
happy sad
What do you mean by that?
angry angry
if you're going to live like this...you have to toughen up
sad sad
There's no luck, there's only work
angry angry

Fig. 6. Sentiment analysis results of AMSA. We can recognize emotions in up to 20 different circumstances, such as sarcasm, self-deprecating, metaphor, etc. The truth value is represented in black, while the predicted feelings are in blue (correct) and red (wrong).

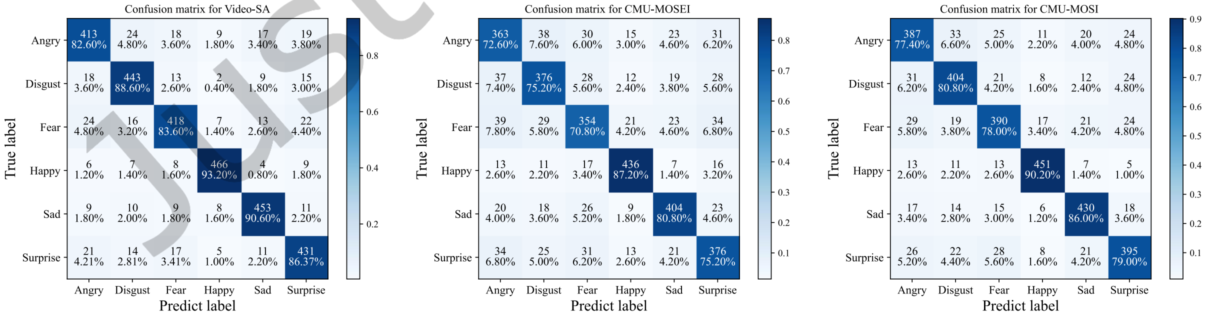


Fig. 7. The confusion matrix for AMSA performance on the three data sets (i.e., Video-SA, CMU-MOSEI and CMU-MOSI). The horizontal axis is the predict label, and the vertical axis is the true label. We randomly selected 500 samples for each emotion. The value of the elements on the diagonal of the matrix is the number and percentage of predicted correct samples in each emotion.

Table 1. Performance of AMSA on three datasets. The maximum and minimum value are eliminated. In this table, each dataset contains three rows of data: the minimum, the maximum, and the average.

Datasets		5-fold	8-fold	10-fold
Video-SA Dataset	MIN	0.791	0.813	0.845
	MAX	0.894	0.902	0.893
	AVERAGE	0.855	0.858	0.863
CMU-MOSEI Dataset	MIN	0.698	0.724	0.733
	MAX	0.765	0.778	0.771
	AVERAGE	0.720	0.741	0.757
CMU-MOSI Dataset	MIN	0.817	0.835	0.838
	MAX	0.845	0.863	0.854
	AVERAGE	0.824	0.836	0.829

Table 2. Comparative Experiment. We conducted more than 50 experiments in ten scenes, including 7 basic emotions (anger, disgust, fear, happiness, sadness, and surprise) and 15 complex situations (including sarcasm, self-deprecating, metaphor etc.). The endpoints are eliminated for each round, while the average is taken. The three indicators "Recall", "ACC" and "TS", indicate detection range, accuracy and calculation speed respectively

Datasets	Video-SA Dataset			CMU-MOSEI dataset			CMU-MOSI dataset		
Methods	**Recall**	**ACC**	**TS**	**Recall**	**ACC**	**TS**	**Recall**	**ACC**	**TS**
DialogueCRN	0.735	0.692	0.815	0.692	0.701	0.763	0.705	0.791	0.882
DualGCN	0.809	0.775	0.731	0.719	0.772	0.756	0.613	**0.855**	0.936
Visnet	0.782	0.633	**0.915**	0.760	0.714	**0.903**	0.697	0.762	0.917
HAN-VGG	0.692	0.597	0.804	0.712	0.723	0.836	0.678	0.595	0.910
TFN-VGG	0.676	0.645	0.795	0.736	0.682	0.828	0.635	0.631	0.906
HAN-Resnet	0.724	0.672	0.787	0.692	0.655	0.804	0.652	0.640	0.893
TFN-Resnet	0.713	0.634	0.813	0.677	0.653	0.822	0.664	0.573	0.905
AMSA	**0.871**	**0.786**	0.746	**0.836**	**0.745**	0.819	**0.732**	0.826	**0.947**

are also devoted to boosting machine generalization capacity. They concentrate on achieving sentiment analysis for unknown scenarios by examining scene relationships. These targets are essentially the same as ours, but instead of focusing on accuracy and balance, they prioritize detection range, which will be explored in the future.

Ablation Study. We have modularly wrapped the framework so that different modules can be activated or deactivated at any time to explore the interpretability of the study. As shown in Table 3, the self-adaptive mechanism and gradient-based optimization have a positive impact on model representation ability and robustness. The patch-based selection scheme successfully locates the main regions of emotional computing, resulting in improved overall real-time performance, while the gradient-based iteration can increase the detected range and accuracy to some extent when compared to multi-fold cross-validation.

Self-adaptive and balance assessment. Many discrimination scenarios are chosen for objective evaluation, while experimenting with various initial weights and adaptive allocation procedures. The effect is depicted in Figure 8. The self-adaptive mechanism not only improves the robustness shown from the above-mentioned

Table 3. Ablation study. We selected five main innovative modules, listed in the upper half of the table as independent variables. The check mark in the table means it is applied in the experiment. The experiment uses the three indicators mentioned above: "Recall", "ACC" and "TS".

Stabilizing gradient	✓	✓	✓	✓	✓	✓
Gradient-based optimization		✓		✓		✓
Patch-based selection scheme			✓	✓	✓	✓
Self-adaptive mechanism		✓			✓	✓
Novel loss function					✓	✓
Recall	0.687	0.756	0.690	0.742	0.777	**0.885**
ACC	0.572	0.693	0.641	0.705	0.794	**0.853**
TS	0.773	0.758	**0.843**	0.743	0.711	0.693

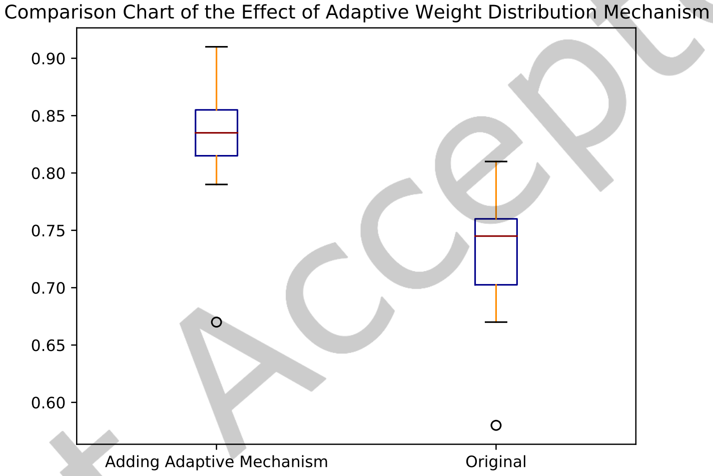


Fig. 8. The performance of self-adaptive mechanism. We conducted 10 rounds of testing with our AMSA and the same model without the adaptive weight calculation mechanism, respectively. For each round of testing we took the average of accuracy (5 tests/round) based on the 300 videos randomly selected from 15 emotional contexts (20 pieces/context). The horizontal axis in the figure represents the accuracy, the red line is the median result, and the circle represents the minimum value (also the outlier, which we think is due to the "metaphor" context).

trials, but also stabilizes the findings even when outliers are found in this experiment. However, due to the diversity of text semantics in "metaphor" contexts, our method still has outliers and it is difficult to obtain a large improvement in the lower limit of accuracy. Furthermore, we counted the time consumption of 300 samples in this experiment. It is discovered that 80 percent of the samples are computed to generate results in 3/4 of the average time. Following that, our tests reveal that the overall calculation time is lowered by 19.6 percent, while the accuracy under this time criterion (3/4 of the average time) is just 3.2 percent lower than without the threshold.

5 CONCLUSION

In this work, we investigate the problem of exploring complementarity and correlation between different modalities for image-text-audio sentiment analysis, especially for complex contexts (e.g., sarcasm, disdain, metaphors, etc.), and propose an adaptive multimodal sentiment analysis framework (AMSA). A multiscale attention model with slice positioning is proposed to extract higher-level text-related visual and audio features, while locating the sentimental elements. Then a patch-based selection scheme is proposed to dynamically calculate key regions and important boundaries based on gradients, which helps eliminate redundant information and speed up the calculation with a novel loss function. Finally, the Transformer-based self-adaptive mechanism is employed to adaptively assign weights and parameters for complex sentiment analysis in back-end fusion. Our approach is different from other methods that treat modalities equally and use settled hyperparameters. The experiments conducted on both manually annotated and machine-labeled datasets, including a self-made dataset Video-SA, demonstrate the superiority and effectiveness of our AMSA, which achieves an average accuracy of 79.57%. The detection running on NVIDIA Tesla P4 GPU is on average 5.2% faster than baselines.

The main limitation of this study lies in that the fine-grained correlation between modalities is analyzed inside the element pairs. However, there may be some closely related samples in the presence of contexts that may degrade the model performance. The experiments show that the performance of our model in rhetorical scenarios such as metaphors is hardly improved which may be due to the multiple meanings of language. In the future, we hope to find an appropriate method to fuse and learn this relationship between pre-sentiment and post-sentiment element pairs to complement AMSA. In addition, we would explore how to add biological signals (e.g., EEG, ECG, etc.) for more effective sentiment analysis.

REFERENCES

[1] Kashif Ahmad, Mohamed Lamine Mekhalfi, Nicola Conci, Farid Melgani, and Francesco De Natale. 2018. Ensemble of deep models for event recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 2 (2018), 1–20.

[2] Jessica Elan Chung and Eni Mustafaraj. 2011. Can collective sentiment expressed on twitter predict political elections?. In *Twenty-fifth AAAI conference on artificial intelligence*.

[3] Hang Cui, Vibhu Mittal, and Mayur Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In *AAAI*, Vol. 6. 30.

[4] Masoud Mazloom, Robert Rietveld, Stevan Rudinac, Marcel Worring, and Willemijn Van Dolen. 2016. Multimodal popularity prediction of brand-related social media posts. In *Proceedings of the 24th ACM international conference on Multimedia*. 197–201.

[5] Sicheng Zhao, Shangfei Wang, Mohammad Soleymani, Dhiraj Joshi, and Qiang Ji. 2019. Affective computing for large-scale heterogeneous multimedia data: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 3s (2019), 1–32.

[6] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. 2020. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5194–5202.

[7] Luntian Mou, Chao Zhou, Pengfei Zhao, Bahareh Nakisa, Mohammad Naim Rastgoo, Ramesh Jain, and Wen Gao. 2021. Driver stress detection via multimodal fusion using attention-based CNN-LSTM. *Expert Systems with Applications* 173 (2021), 114693.

[8] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing* 65 (2017), 3–14.

[9] Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review* 53, 6 (2020), 4335–4385.

[10] Luntian Mou, Chao Zhou, Pengtao Xie, Pengfei Zhao, Ramesh C Jain, Wen Gao, and Baocai Yin. 2021. Isotropic Self-supervised Learning for Driver Drowsiness Detection With Attention-based Multimodal Fusion. *IEEE Transactions on Multimedia* (2021).

[11] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. 2014. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*. 47–56.

[12] Monisha Kanakaraj and Ram Mohana Reddy Guddeti. 2015. Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques. In *Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015)*. IEEE, 169–170.

[13] Ezgi Yıldırım, Fatih Samet Çetin, Gülşen Eryiğit, and Tanel Temel. 2015. The impact of NLP on Turkish sentiment analysis. *Türkiye Bilişim Vakfı Bilgisayar Bilimleri ve Mühendisliği Dergisi* 7, 1 (2015), 43–51.

[14] Shangfei Wang and Qiang Ji. 2015. Video affective content analysis: a survey of state-of-the-art methods. *IEEE Transactions on Affective Computing* 6, 4 (2015), 410–430.

[15] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, and Guiguang Ding. 2016. Continuous probability distribution prediction of image emotions via multitask shared sparse regression. *IEEE transactions on multimedia* 19, 3 (2016), 632–645.

[16] Nusrat J Shoumy, Li-Minn Ang, Kah Phooi Seng, DM Motiur Rahaman, and Tanveer Zia. 2020. Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals. *Journal of Network and Computer Applications* 149 (2020), 102447.

[17] Sudhanshu Kumar, Mahendra Yadava, and Partha Pratim Roy. 2019. Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction. *information fusion* 52 (2019), 41–52.

[18] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* 16, 6 (2010), 345–379.

[19] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 973–982.

[20] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2539–2544.

[21] Ningning Liu, Emmanuel Dellandréa, Liming Chen, Chao Zhu, Yu Zhang, Charles-Edmond Bichot, Stéphane Bres, and Bruno Tellez. 2013. Multimodal recognition of visual concepts using histograms of textual concepts and selective weighted late fusion scheme. *Computer Vision and Image Understanding* 117, 5 (2013), 493–512.

[22] Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3718–3727.

[23] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. 2017. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM international conference on multimodal interaction*. 163–171.

[24] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Mazumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Multi-level multiple attentions for contextual multimodal sentiment analysis. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1033–1038.

[25] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In *Proceedings of the Ninth ACM international conference on Web search and data mining*. 13–22.

[26] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[27] Yagya Raj Pandeya, Bhuwan Bhattarai, and Joonwhoan Lee. 2021. Deep-learning-based multimodal emotion classification for music videos. *Sensors* 21, 14 (2021), 4927.

[28] Linghui Li, Sheng Tang, Lixi Deng, Yongdong Zhang, and Qi Tian. 2017. Image caption with global-local attention. In *Thirty-first AAAI conference on artificial intelligence*.

[29] Jie Wu, Haifeng Hu, and Yi Wu. 2018. Image captioning via semantic guidance attention and consensus selection strategy. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14, 4 (2018), 1–19.

[30] Hongtao Xie, Shancheng Fang, Zheng-Jun Zha, Yating Yang, Yan Li, and Yongdong Zhang. 2019. Convolutional attention networks for scene text recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1s (2019), 1–17.

[31] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf. Fusion* 37 (2017), 98–125. DOI: http://dx.doi.org/10.1016/j.inffus.2017.02.003

[32] Junjun Chen. 2021. Refining the teacher emotion model: evidence from a review of literature published between 1985 and 2019. *Cambridge Journal of Education* 51, 3 (2021), 327–357.

[33] Kenneth Ward Church. 2017. Word2Vec. *Natural Language Engineering* 23, 1 (2017), 155–162.

[34] Vishwanath A Sindagi and Vishal M Patel. 2018. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters* 107 (2018), 3–16.

[35] Guozhen Zhao, Jinjing Song, Yan Ge, Yongjin Liu, Lin Yao, and Tao Wen. 2016. Advances in emotion recognition based on physiological big data. *Journal of Computer Research and Development* 53, 1 (2016), 80.

[36] ReadFace. 2020. ReadFace webpage on 36Kr. http://36kr.com/p/5038637.html. (2020).

[37] Srikumar Krishnamoorthy. 2018. Sentiment analysis of financial news articles using performance indicators. *Knowledge and Information Systems* 56, 2 (2018), 373–394.

[38] Xiaodong Li, Haoran Xie, Li Chen, Jianping Wang, and Xiaotie Deng. 2014. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems* 69 (2014), 14–23.

[39] Qianren Mao, Jianxin Li, Senzhang Wang, Yuanning Zhang, Hao Peng, Min He, and Lihong Wang. 2019. Aspect-Based Sentiment Classification with Attentive Neural Turing Machines.. In *IJCAI*. 5139–5145.

[40] Yanghui Rao, Jingsheng Lei, Liu Wenyin, Qing Li, and Mingliang Chen. 2014. Building emotional dictionary for sentiment analysis of online news. *World Wide Web* 17, 4 (2014), 723–742.

[41] Yuxiang Zhang, Jiamei Fu, Dongyu She, Ying Zhang, Senzhang Wang, and Jufeng Yang. 2018. Text Emotion Distribution Learning via Multi-Task Convolutional Neural Network.. In *IJCAI*. 4595–4601.

[42] Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4351–4360.

[43] Ngoc-Dau Mai, Boon-Giin Lee, and Wan-Young Chung. 2021. Affective Computing on Machine Learning-Based Emotion Recognition Using a Self-Made EEG Device. *Sensors* 21, 15 (2021), 5135. DOI: http://dx.doi.org/10.3390/s21155135

[44] Dhanesh Ramachandram and Graham W Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine* 34, 6 (2017), 96–108.

[45] Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 481–492.

[46] Dilana Hazer-Rau, Sascha Meudt, Andreas Daucher, Jennifer Spohrs, Holger Hoffmann, Friedhelm Schwenker, and Harald C Traue. 2020. The uulmMAC database—A multimodal affective corpus for affective computing in human-computer interaction. *Sensors* 20, 8 (2020), 2308.

[47] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li. 2019. Learning alignment for multimodal emotion recognition from speech. *arXiv preprint arXiv:1909.05645* (2019).

[48] Yubo Xie, Junze Li, and Pearl Pu. 2020. Uncertainty and surprisal jointly deliver the punchline: Exploiting incongruity-based features for humor recognition. *arXiv preprint arXiv:2012.12007* (2020).

[49] Xiangyu Wang and Chengqing Zong. 2021. Distributed representations of emotion categories in emotion space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2364–2375.

[50] Mohammad Ehsan Basiri, Shahla Nemati, Moloud Abdar, Erik Cambria, and U Rajendra Acharya. 2021. ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Future Generation Computer Systems* 115 (2021), 279–294.

[51] Yuxiao Chen, Jianbo Yuan, Quanzeng You, and Jiebo Luo. 2018. Twitter sentiment analysis via bi-sense emoji embedding and attention-based LSTM. In *Proceedings of the 26th ACM international conference on Multimedia*. 117–125.

[52] Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 371–378.

[53] Xiao Bai, Xiang Wang, Xianglong Liu, Qiang Liu, Jingkuan Song, Nicu Sebe, and Been Kim. 2021. Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognition* 120 (2021), 108102.

[54] Chen Wang, Xiang Wang, Jiawei Zhang, Liang Zhang, Xiao Bai, Xin Ning, Jun Zhou, and Edwin Hancock. 2022. Uncertainty estimation for stereo matching based on evidential deep learning. *Pattern Recognition* 124 (2022), 108498.

[55] Xinyu Ou, Hefei Ling, Han Yu, Ping Li, Fuhao Zou, and Si Liu. 2017. Adult image and video recognition by a deep multicontext network and fine-to-coarse strategy. *ACM Transactions on Intelligent Systems and Technology (TIST)* 8, 5 (2017), 1–25.

[56] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2020. Image-text multimodal emotion classification via multi-view attentional network. *IEEE Transactions on Multimedia* 23 (2020), 4014–4026.

[57] Quoc-Tuan Truong and Hady W Lauw. 2019. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 305–312.

[58] Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2399–2402.

[59] Tuan-Linh Nguyen, Swathi Kavuri, and Minho Lee. 2019. A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips. *Neural Networks* 118 (2019), 208–219.

[60] Usman Naseem, Imran Razzak, Katarzyna Musial, and Muhammad Imran. 2020. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems* 113 (2020), 58–69.

[61] Chaozhuo Li, Senzhang Wang, Yukun Wang, Philip Yu, Yanbo Liang, Yun Liu, and Zhoujun Li. 2019. Adversarial learning for weakly-supervised social network alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 996–1003.

[62] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 375–383.

[63] Jonghwan Mun, Minsu Cho, and Bohyung Han. 2017. Text-guided attention model for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31.

[64] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. 2018. Exploring human-like attention supervision in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[65] Luzi Sennhauser and Robert C Berwick. 2018. Evaluating the ability of LSTMs to learn context-free grammars. *arXiv preprint arXiv:1811.02611* (2018).

[66] Amir Zadeh and Paul Pu. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*.

[67] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).

[68] Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. *arXiv preprint arXiv:2106.01978* (2021).

[69] Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 6319–6329.

[70] Akmaljon Palvanov and Young Im Cho. 2019. Visnet: Deep convolutional neural networks for forecasting atmospheric visibility. *Sensors* 19, 6 (2019), 1343.

[71] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).