计算机科学
ISSN: 2790-0622(print)
雙清學術出版社

Contents lists available at www.qingpress.com
Journal homepage:qingpress.com/zh-cn/journals/25

基于跨模块融合的多模态情感分析

张敬伟 [1*] 李彦 [2]

1. 张敬伟，天津师范大学,电子与通信工程学院；天津市无线移动通信与无线电能传输重点实验室
2. 李彦，天津师范大学
*. 通讯作者：张敬伟

摘要：理解表达的情感和情绪是多模态情感分析的两个关键因素。人类语言通常是多模态的，包括视觉觉，语音以及文本三个模态，而每个模态又包含众多不同信息，比如文本模态包括基本的语言符号、句法和语言动作等，语音模态包括：语音、语调以及声音表达等。视觉模态包括姿态特征、身体语言、眼神以及面部表达等信息。因此如何高效融合模态间信息便成为当下多模态情感分析领域的一个热点问题。为此，文章提出一种基于跨模块融合网络模型。该模型利用 LSTM 网络作为语言、视觉模态的表示子网络，同时利用改进升级的 Transformer 模型的跨模块融合对两种模态信息进行有效融合；为了验证文章中提出的模型的效果，在 IEMOCAP 和 MOSEI 数据集上进行了仔细评估，结果表明，该模型针对情感分类的准确度有所提高。

关键词：Transformer 模型，多模态情感分析，跨模块融合

Multimodal Sentiment Analysis Based On Cross-Module Fusion

Abstract: Understanding expressed emotion and sentiment is critical in multimodal sentiment analysis. Human language is usually multimodal, including visual, speech, and textual modalities, and each modality contains much different information, such as textual modalities include basic speech symbols, syntax, and speech actions. Speech modalities include speech, intonation, and vocal expressions, and visual modalities include gesture features, body language, eyes, and facial expressions. Therefore, how to efficiently fuse inter-modal information has become a hot issue in the field of multimodal sentiment analysis nowadays. To this end, we propose a cross-modal fusion network model. The model uses the LSTM network as the representation sub-network of language and visual modalities. In contrast the cross-module fusion of the improved Transformer model is use the two modal information effectively. To

通讯作者简介：张敬伟 (1997-)，男；专业：信息与通信工程；研究方向：多模态情感分析

E-mail:3334783822@qq.com 或 274433799@qq.com

联系电话：13164376609(微信同号)；

单位：天津师范大学电子与通信工程学院，天津，300387

天津市无线移动通信与无线电能传输重点实验室，天津，300387

Article history: Received D February 7, 2023 Accepted February 19, 2023 Available online February 20, 2023
To cite this document: 张敬伟, 李彦(2023). 基于跨模块融合的多模态情感分析.计算机科学, 第 3 卷, 第 1 期, 9-18 页.
Doi: https://doi.org/10.55375/cps.2023.3.2

verify the effectiveness of the proposed model, we carefully evaluated on IEMOCAP and MOSEI datasets and the results show that the model for sentiment classification has improved accuracy.

Keywords: Transformer model, Multimodal sentiment analysis, Cross-module fusion

随着近年来社交媒体的空前发展以及配备高质量摄像头的智能手机的出现，我们见证了电影、短视频等多模态数据的爆炸式增长。理解表达的情感和情绪是多模态情感分析(Multimodal Sentiment Analysis,MSA)的两个关键因素，但从多媒体中预测情感状态仍然是一项具有挑战性的任务。情感识别任务已经存在于不同类型的信号上，典型的是音频、视频和文本。这些数据大部分都包含了视觉(图像)、听觉(声音)、文本(转录的文字)在内的三种模态(Wei Han et al. 2021)。如何利用这些不同模态的数据来高效且充分的挖掘其中所蕴含的情感信息已经成为现在自然语言处理(Nature Language Processing, NLP)领域中的一个热门研究主题。在多模态情感分析领域一直存在两大难题：(1)如何在输入数据中对较优的数据进行提取并表示；(2)对来自不同模态的数据进行有效融合，使得神经网络可以充分学习到这些数据中包含的情感信息。为此研究者们提出了各种可行性比较高的方法。文献(Baltrusaitis T et al. 2019)中将多模态分析领域的研究一共划分为五类：多模态数据翻译、多模态数据表示、多模态数据对齐、多模态数据融合和共同学习。通过所有模式捕捉对话的上下文，对话中当前的说话者和听者，以及通过适当的融合机制捕捉现有模式之间的相关性和关系，Shenoy 和 Sardana 提出了一种循环神经网络架构(Shenoy A& Sardana A，2020)，试图考虑所有提到的缺陷，并跟踪对话的上下文，对话者状态，以及对话中说话人所传达的情感。

基于上述问题，受 Tsai(Tsai Yet al. 2019)和 Zhou(Yu Zet al.2019)的启发，本文提出跨模块融合网络(Cross-module converged networks)以下简称 CMN,CMN 基于两个阶段: 1)基于 LSTM 的独立序列阶段，其中模态特征分别计算；2)基于 Transformers。在文献(Tsai Yet al. 2019)中也使用基于 Transformer 的解决方案来编码他们的模型。首先，他们的最佳解决方案和得分报告都使用了视觉支持。其次，利用 Transformer 对每个模态进行交叉模态编码，相当于 6 个 Transformer 模块(每种模态对应两对)，而我们只使用两个 Transformer 模块(经深度联合模块处理后的模态，每模态使用一个)。CMN 的结构图如图 1 所示。

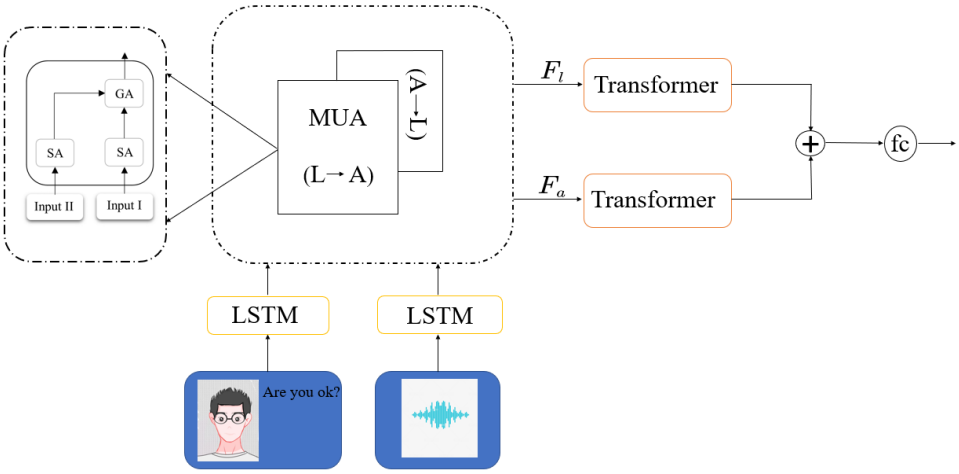

图 1 跨模块融合网络模型结构图

1 相关工作

1.1 多模态情感分析现状

在日常的生活中，人们利用各种社交平台例如抖音、快手、微博、YouTube 等，发表关于自己各种情

感的信息。以往的研究仅限于单一模态，尽管取得了一定的成功，但还是存在着识别率低、泛化性差等不足之处。这些问题的主要原因在于，无法充分利用一段多模态数据中的全部数据资源，导致模型无法充分学习到输入数据的特征信息，无法充分考虑到各种数据之间的互补关系。例如在图 2 中所示，观影者在欣赏完一部影视作品后发出评价“This video is sick”,仅通过一句文本描述，我们很难判断观影者的情绪状态，因为“sick”本身具有“呕吐、恶心”的含义另外在美国俚语中有“很酷”的含义，所以仅根据文本无法判断其情感倾向，同时说话声音很大也无判断是正向还是负向情感。

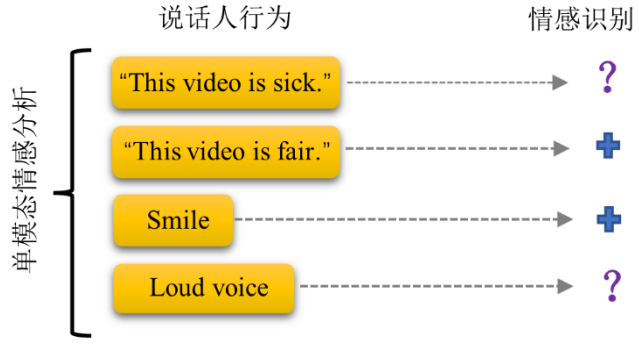

图 2 单模态情感融合图

但在图 3 中，当多个模态融合时，通过文本信息、面部表情以及声音特征就很容易判断出观影者的情感倾向。因此，多模态情感分析应运而生，正确且高效的识别和分析社交网站上各种数据所蕴含的情感信息，对于舆论引导、心理疾病的康复与治疗都有着较为重要的影响。在心理治疗方面，由于新冠肺炎疫情的影响，很多人会出现心理方面的疾病。心理医生可以通过远程视频方式与病人进行交流，并实时根据病人的面部表情、语音语调以及相关的文字信息来判定此时病人的心理状况。

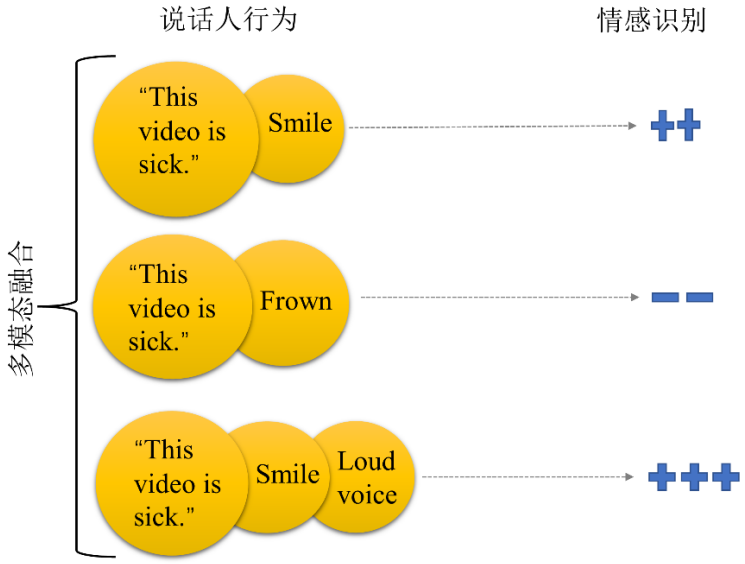

图 3 多模态情感融合图

在社交媒体方面，可以通过人们发布的微博、朋友圈中的图片和文字描述来识别该用户所表达的情感信息。针对某些热点事件，可以根据用户的表述来判断其所蕴含的情感信息，做到意见挖掘等。在心理治疗方面，由于新冠肺炎疫情的影响，很多人会出现心理方面的疾病,心理医生可以通过远程视频方式与病人进行交流，并实时根据病人的面部表情、语音语调以及相关的文字信息来判定此时病人的心理状况。

人类语言不仅具有口头语言，还具有非语言行为，例如视觉(面部属性)和声学(声调模式)(Gibson K R& Ingold T, 1994)。这些丰富的信息为我们能够充分理解人类行为和意图提供了更好的帮助(Manning C Det al. 2014)。然而，不同语言模式之间的异质性往往增加了分析人类语言的难度。例如，音频和视觉流的受体可能会随着接收频率的变化而变化，因此我们可能无法获得它们之间的最佳映射。皱眉可能与过去说过的悲观的话有关。换言之，多模态语言序列往往表现出“不对齐”的性质，需要推断跨模态的长期依赖，因此 Tsai 提出 Multimodal Transformer (MulT) (Tsai Yetal. 2019),这是一种端到端模型，扩展了标准 Transformer 网络(Vaswani A et al. 2017)，以直接从未对齐的多模态流中学习表示。先前分析人类多模态语言的工作是在跨语言、视觉和声学模态的多模态序列中推断表征的领域。与从图像和文本属性等静态领域学习多模态表示不同(Ngiam J et al. 2009;Srivastava N&Salakhutdinov R, 2012)。人类语言包含时间序列，因此需要融合时变信号(Liang P Pet al.2018; Tsai Y et al.2018)。早期的工作使用早期融合方法来连接来自不同模态的输入特征(Ngiam J et al. 2009; Lazaridou A et al. 2015)，与单一学习模式相比，表现出了更好的效果。最近，更先进的模型被提出来学习人类多模态语言的表示。例如 Gu(Gu Y etal. 2018)使用分层注意力策略学习多模态表示，Wang(Wang Y et al. 2018)提出了循环参与 V 变化嵌入网络(RAVEN)，通过分析词段中发生的细粒度的视觉和听觉模式来建模表达性非语言表征。此外，他们还试图通过改变非语言行为的词语表征来捕捉非语言意图的动态本质。Pham(Hai P et al. 2019)提出了一个模型，通过模态之间的循环翻译(MCTN)学习的联合表示，在各种单词对齐的人类多模态语言任务上取得了良好的结果。

1.2 Transformer 模型和 LSTM 神经网络

Transformer 网络(Vaswani A et al. 2017)首次被引入用于神经机器翻译(NMT)任务，是一种全新的、完全基于注意力机制的 Seq2Seq 模型，其中编码器和解码器一方都利用了自我注意力(Wang Y et al. 2018; Parikh A et al. 2016; Lin Z et al. 2017)Transformer，摒弃传统的卷积神经网络(CNN)和循环神经网络(RNN)，通过词嵌入和位置编码学习位置信息，在每一层自注意之后，编码器和解码器由一个额外的解码器子层连接，其中解码器针对目标文本的每个元素关注源文本的每个元素。除了 NMT，Transformer 网络也已成功应用于其它任务，包括语言建模(Dai et al. 2018; Baevski A& Auli M, 2018)、语义角色标注(Strubell E et al. 2018)、词义消歧(G Müller et al. 2018)和学习句子表征(Devlin J et al. 2018)等。

长短期记忆(Hochreiter S et al. 1997)(Long short-term memory, LSTM)是一种特殊的 RNN，主要是为了解决长序列训练过程中的梯度消失和梯度爆炸问题。相较于普通的 RNN, LSTM 能够在更长的序列中有更好的表现。相比 RNN 只有一个传递状态h^t，LSTM 有两个传输状态，一个c^t(cell state)，和另一个h^t(hiddenstate)。其中对于传递下去的c^t改变的很慢，通常输出的c^t是上一个状态传递过来的c^{t-1}加上一些数值，而 h^{t}则在不同节点下往往会有很大区别。

2 跨模块融合网络模块

2.1Self Attention

SelfAttention(以下简称 SA)结构中包含两个子层,分别是自注意力层和前馈神经网络层，每个子层搭配一个线性层,同时使用残差连接(He K et al. 2016)，将信息无损耗传递的更深来增强模型的拟合能力，结构图如 4 所示。我们用α代表任一模态，将模态特征向量F_α输入 SA 子层中。

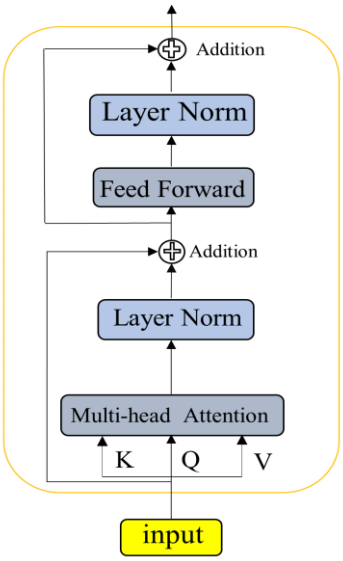

图 4Self Attention 结构图

定义 Query 为 $Q_\alpha = F_\alpha W_Q$，Key 为 $K_\alpha = F_\alpha W_K$，Value 为 $V_\alpha = F_\alpha W_V$，其中，$W_Q \in R^{d_\alpha \times d_k}$，$W_K \in R^{d_\alpha \times d_k}$，$W_V \in R^{d_\alpha \times d_v}$，然后计算 Query 与所有 Key 的点积,除以缩放因子 $\sqrt{d}$,接着应用 Softmax 函数来获得这些 Values 的注意权重，最后对 Q 和 K 学习到的注意力所有 Values 的 V 加权求和，得到下式：

$$F_\alpha = Att.(Q_\alpha, K_\alpha, V_\alpha) = Soft\max\left(\frac{Q_\alpha K_\alpha^T}{\sqrt{d_k}}\right)V_\alpha \qquad (1)$$

其中 $d_k = d_v = d_\alpha$ 分别为模态 α 向量的线性变换权重矩阵。

为了进一步提高被关注特征的表示能力，引入了由 h 个并行头组成的多头注意力层，每个头对应一个独立的缩放点积注意函数。输出模态特征表示下式：

$$F_\alpha = MHA\ (Q_\alpha, K_\alpha, V_\alpha) = Concat(head_1, head_2, \cdots, head_n)W^O \qquad (2)$$

$$head_i = Att.(Q_\alpha W_j^Q, K_\alpha W_j^K, V_\alpha W_j^V) \qquad (3)$$

其中 W_j^Q，$W_j^K, W_j^V \in R^{d_\alpha \times d_h}$，是第 j 个头的投影矩阵，$W^O \in R^{h * d_h \times d_\alpha}$，$d_h$ 是每个头部输出特征的维数，为了防止多头注意力模型变的太大，通常使用 $d_h = d_\alpha / h$。

2.2Guide Attention

在 Guide Attention (以下简称 GA)模块中，我们考虑两个任意模态 α 和 β，每个模态对应的输出向量为 F_α，F_β(结构图，如图 5 所示)。

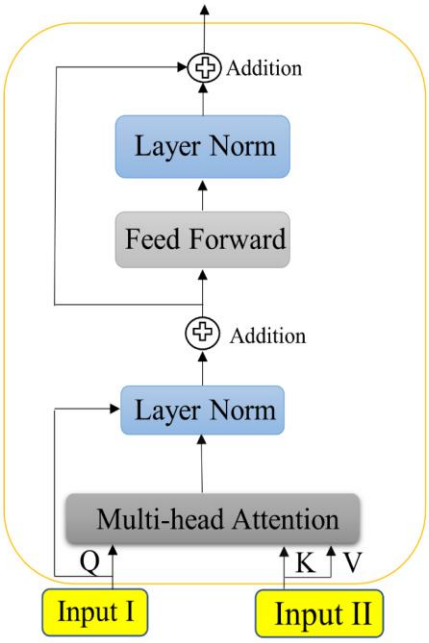

图 5Guide Attention 结构图

我们定义 Query 为$Q_\alpha = F_\alpha W_{Q_\alpha}$，Key 为$K_\beta = F_\beta W_{K_\beta}$，Value 为$V_\beta = F_\beta W_{V_\beta}$，其中$W_{Q_\alpha} \in R^{d_\alpha \times d_k}$，$W_{K_\beta} \in R^{d_\beta \times d_k}$，$W_{V_\beta} \in R^{d_\beta \times d_v}$，通过引导注意力可以使网络充分学习到来自其余模态的不同重要信息和互补信息。GA 定义公式如下：

$$GA_{\beta \to \alpha} = Soft\max\left(\frac{Q_\alpha K_\beta^T}{\sqrt{d_k}}\right)V_\beta = Soft\max\left(\frac{F_\alpha W_{Q_\alpha} W_{K_\beta}^T F_\beta^T}{\sqrt{d_k}}\right)F_\beta W_{V_\beta} \qquad (4)$$

2.3 深度模块化联合注意力层

深度模块化联合注意力层(Deep Modular Union Attention Layer，以下简称 MUA),每个 MCA 利用两个基本单元的模块组合，结构图如图 6 所示。

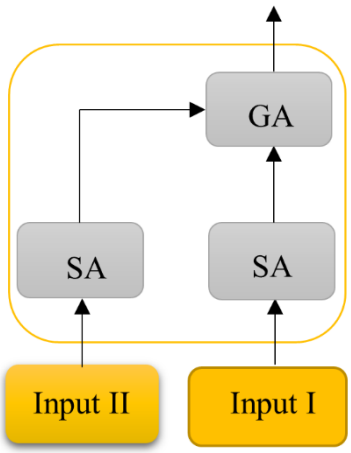

图 6 深度模块化联合注意力层结构图

3 数据集

3.1 IEMOCAP 数据集

IEMOCAP(Carlos Busso et al. 2008)数据集是参与者二元对话的多模态数据集。该数据集由 151 个录制对话的视频组成，每个对话有两个演讲者，整个数据集总共有 302 个视频。记录模态包含音频、视频和动作捕捉数据。情绪类别分为六种基本情绪(Ekman P, 1999)：快乐、悲伤、愤怒、惊讶、恐惧和厌恶以及三种持续的情绪维度：效价、唤醒和支配。

关于情绪维度的进一步说明。情绪效价分为正性的和负性的情绪，即对情绪属性的自我评估，效价指的是描述一个人对于一个事物的吸引(感兴趣)或排斥(厌恶)的程度。唤醒指的是生理或心理被吵醒或是对外界刺激重新产生反应。支配即表示一个人对自己情绪外在表现的控制程度。

3.2 CMU-MOSEI 数据集

CMU-MOSEI(AmirAli Z et al. 2018c)数据集是新一代的 MOSI 数据集，在 CMU-MOSI 数据集的基础上进行扩展，包含超过两万条视频片段，每段视频的情感分数取值从区间-3 到+3。该数据集利用含有表达意见的在线视频，通过人脸检测算法分析了视频，并选择了只有一个人关注摄像头的视频。该数据集使用了一组 250 个不同的关键词来抓取视频，并保留每个视频最多 10 个片段，包括人工转录。然后手动管理数据集，只保留质量好的数据。在该数据集中，使用 4643 条视频进行测试，1869 条视频进行验证，16265 条视频进行训练。

4 实验设置与结果分析

4.1 实验设置

实验中我们使用 Adam(Kingma D& Ba J, 2014)优化器训练我们的模型,学习率为 1e-4，batch_size=32,本文的结果最多来自 10 个模型的平均预测，LSTM 的隐藏层维度大小设置为 512。

训练样本 16265 个，预测样本 4643 个。

学习的目标是：准确分析出每个样本所表达的情绪及对应维度。

4.2 结果分析

我们在两个情感识别数据集 IEMOCAP 和 MOSEI 上给出了研究结果。对于每个数据集，结果以数据集使用的流行指标的形式呈现。大多数情况下采用 F1 分数，有时也采用加权 F1 分数来考虑情绪或情绪类之间的不平衡。

我们首先在表 1 中比较了我们的模型变量在 IEMOCAP 上的精确度(Precision)、召回率(Recall)和未加权 F1 分数。

表 1IEMOCAP 的情绪任务结果(Prec.表示精确度、F1 为未加权的 F1 分数)

Model	Prec.	Recall	F1
CMN(L+A,Ours)	67.4	67.4	67.4
Mult(L+A+V)(Tsai Y et al. 2019)	——	——	71.5
RAVEN(Wang Y et al. 2018)	——	——	66.5
E1(L+A)(Sahu G, 2019)	56.6	57.3	55.7
E2(L+A)(Sahu G, 2019)	64.9	63.2	66.0

MOSEI 是一个相对较大规模的数据集。表 2 给出的是 MOSEI 2-sentiments 任务的结果。

表 2MOSEI 2-sentiments 任务的结果

Model	**A2**	**F1**
CMN(L+A,Ours)	42.5	42.5
Mult(L+A+V)(Tsai Y et al. 2019)	81.1	81.0
RAVEN(L+A+V) (Wang Y et al. 2018)	79.0	79.5
G-MFN(AmirAli Zadeh et al. 2018b)	76.9	77.0

由表中数据可以得出，在 IEMOCAP 数据集上，我们的模型相较于其他模型表现得更极其出色，但在 MOSEI 数据集上的测试结果表现得不尽人意，这里不再详细分析。在 IEMOCAP 数据集上，我们的模型在精确度上相较于 E1、E2 模型，分别提高了 19.1%、3.8%。在召回率上相较于 E1、E2 模型，分别提高了 17.6%、6.6%。在 F1 分数上，相较于 RAVEN、E1、E2 三个模型分别提高了 1.3%、21.0%、2.1%。以上优异的表现证明了本文所提模型达到了 MSA 中的最佳性能。

5 总结

本文提出的一种模型跨模块融合网络(Cross-module converged networks，CMN)。该网络模型首先基于 LSTM 的独立序列阶段，模态特征分别计算，之后基于 Transformer,不断提取模态间的重要特征信息和互补信息。在公开的数据集 IEMOCAP 和 MOCEI 上进行了对比试验，在 IEMOCAP 数据集上的表现达到了 MSA 领域中先进水平，各项指标均有所提高。

参考文献:

[1] Han W , Chen H , Poria S . Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal S entiment Analysis[J]. 2021.

[2] BaltrusaitisT , Ahuja C , Morency L P . Multimodal Machine Learning: A Survey and Taxonomy[J]. IEEE Transactions on Patter n Analysis & Machine Intelligence, 2017, PP(99):1-1.

[3] Shenoy A , Sardana A . Multilogue-Net: A Context Aware RNN for Multi-modal Emotion Detection and Sentiment Analysis in C onversation:, 10.18653/v1/2020.challengehml-1.3[P]. 2020.

[4] Tsai Y , Bai S , Liang P P , et al. Multimodal Transformer for Unaligned Multimodal Language Sequences[J]. 2019.

[5] Yu Z , Yu J , Cui Y , et al. Deep Modular Co-Attention Networks for Visual Question Answering[J]. IEEE, 2019.

[6] Gibson K R , Ingold T . Tools, language and cognition in human evolution17[J]. Cambridge University Press, 1994.

[7] Manning C D ,Surdeanu M , Bauer J , et al. The Stanford CoreNLP Natural Language Processing Toolkit[C]// Proceedings of 52n d Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014.

[8] Vaswani A ,Shazeer N , Parmar N , et al. Attention Is All You Need[J]. arXiv, 2017.

[9] NgiamJ , Khosla A , Kim M , et al. Multimodal Deep Learning[C]// International Conference on Machine Learning. DBLP, 2009.

[10] Srivastava N ,Salakhutdinov R . Multimodal Learning with Deep Boltzmann Machines[J]. Journal of Machine Learning Research, 2012, 15.

[11] Liang P P , Liu Z , Zadeh A , et al. Multimodal Language Analysis with Recurrent Multistage Fusion[J]. 2018.

[12] Tsai Y , Liang P P , Zadeh A , et al. Learning Factorized Multimodal Representations[J]. 2018.

[13] LazaridouA , Pham N T , Baroni M . Combining Language and Vision with a Multimodal Skip-gram Model[J]. Computer ence, 2 015.

[14] Gu Y , Yang K , Fu S , et al. Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment:, 10.18653/v1/P18-1207[P]. 2018.

[15] Wang Y , Shen Y , Liu Z , et al. Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors[J]. 2018.

[16] Hai P , Liang P P , Manzini T , et al. Found in Translation: Learning Robust Joint Representations by Cyclic Translations betwee n Modalities[C]// 33rd AAAI Conference on Artificial Intelligence. 2019:6892-6899.

[17] Parikh A ,Tckstrm O , Das D , et al. A Decomposable Attention Model for Natural Language Inference:, 10.18653/v1/D16-1244 [P]. 2016.

[18] Lin Z , Feng M , Santos C , et al. A Structured Self-attentive Sentence Embedding[J]. 2017.

[19] Zihang Dai, Zhilin Y ang, Yiming Y ang, William WCohen, Jaime Carbonell, Quoc V Le, and RuslanSalakhutdinov. 2018. Transf ormer-xl: Languagemodeling with longer-term dependency.

[20] BaevskiA , Auli M . Adaptive Input Representations for Neural Language Modeling[J]. 2018.

[21] StrubellE , Verga P , Andor D , et al. Linguistically-Informed Self-Attention for Semantic Role Labeling[J]. 2018.

[22] G Müller, Rios M , Sennrich A , et al. Why Self-Attention? A Targeted Evaluation of Neural Machine Translation Architectures [J]. 2018.

[23] Devlin J , Chang M W , Lee K , et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2 018.

[24] HochreiterS ,Schmidhuber J . Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[25] He K , Zhang X , Ren S , et al. Deep Residual Learning for Image Recognition[J]. IEEE, 2016.

[26] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, AbeKazemzadeh, Emily Mower, Samuel Kim, Jean-nette N Chang, Sungbok Lee, an d ShrikanthSNarayanan. 2008. Iemocap: Interactive emotionaldyadic motion capture database. Language resources and evaluation, 42(4):335.

[27] Ekman P . Basic Emotions[J]. Handbook of Cognition & Emotion, 1999, 99(1):45-60.

[28] AmirAli Zadeh, Paul Pu Liang, SoujanyaPoria, Erik Cambria, and Louis-Philippe Morency. 2018c. Multimodal language analysis in the wild: CMUMOSEI dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Assoc iation for Computational Linguistics (Volume 1: Long Papers), pages 2236–2246, Melbourne, Australia. Association for Computat ional Linguistics.

[29] KingmaD , Ba J . Adam: A Method for Stochastic Optimization[J]. Computer Science, 2014.

[30] SahuG . Multimodal Speech Emotion Recognition and Ambiguity Resolution[J]. 2019.

[31] AmirAli Zadeh, Paul Pu Liang, SoujanyaPoria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusiongraph. In Proceedings of the 56th Annual Meeting ofthe Associ ation for Computational Linguistics (V olume 1: Long Papers), pages 2236–2246, Melbourne,Australia. Association for Computati onal Linguistics.