scientific reports

Check for updates

OPEN

用于多模态情感分析的局部和全局表示的分层图对比学习

杜军、金建航、庄建、张成

多模态情感分析 (MSA) 旨在通过声音、视觉和文本线索对话语的整体情感进行回归或分类。然而，现有的大多数工作都集中在开发神经网络的表达能力以学习单个话语中多模态信息的表示，而没有考虑数据集的全局共现特征。为了缓解上述问题，在本文中，我们提出了一种新颖的 MSA 分层图对比学习框架，旨在探索多模态情感提取的单个话语的局部和全局表示以及它们之间复杂的关系。具体来说，对于每种模态，我们提取每种模态的离散嵌入表示，其中包括每种模态的全局共现特征。在此基础上，对于每个话语，我们构建两个图表：局部级别图和全局级别图，以解释特定级别的情绪含义。然后，采用两种图对比学习策略分别探索基于图增强的不同潜在表示。此外，我们设计了一种跨级别的比较学习，用于学习复杂关系的局部和全局潜在表示。

随着社交媒体的日益普及，文本、声音和视觉信息等多模态数据已成为个人和公众沟通的重要手段。在这种情况下，从多模态数据估计人类情绪倾向变得越来越重要。因此，多模态数据的多模态情感分析（MSA）已成为多媒体内容理解（MCU）和自然语言处理（NLP）领域的热门话题。其已广泛应用于工业界和学术界，如社交媒体分析、对话系统、电子商务推广和人机交互等。

为了有效地理解多模态信息，早期的 MSA 工作尝试通过基于张量的特征融合或基于注意力的特征融合来融合来自不同模态的信息。此外，一些基于表示学习的方法旨在对模态之间的一致性和可变性进行建模，以提取模态之间的情感线索，或者考虑多模态序列数据与图模型的融合和对齐。研究人员专注于图神经网络，并提出了分层图对比学习框架，以探索用于提取的模态内和模间表示的复杂关系。他们还开发了全局和局部融合神经网络，聚合全局和局部融合特征来分析用户情绪。此外，他们还使用语言方法从多模态建模中提取序列特征，并通过隐马尔可夫模型表示情感关联。尽管当前的工作取得了有希望的进展，但他们通常侧重于通过单个实例内的多模态数据融合多模态表示，这忽略了单个实例具有特定的全局共现特征。如何更有效地利用跨实例的特征共现并捕获数据的全局特征仍然是一个巨大的挑战。

在本文中，我们研究如何捕获多模态数据的全局特征并对全局特征进行显式建模，从而使高度相关的模态表示能够显式链接以学习多模态情感信息。为了实现这一目标，我们提出了层次图对比学习（HGCL-LG），它构建了一个基于比较学习的网络来实现多层次的信息探索。具体来说，由于离散变分自动编码器（dVAE）可以将不同的样本映射到一个公共的离散嵌入空间中，因此我们假设该嵌入空间包含之间的全局信息

1 山东师范大学物理与电子学院，中国山东。云南大学民族学与社会学学院，云南，中国。邮箱：sdnuspe@163.com

样品。因此，我们使用 dAVE 来获取每个模态的嵌入空间。在此基础上，我们构建了局部图和全局图，并设计了三种比较学习：局部图对比学习、全局图对比学习和跨层图对比学习。通过三种比较学习方法，我们的模型充分学习了局部信息和全局信息中的情感特征以及两者之间的复杂关系。此外，我们还引入了一种自适应图增广策略，该策略可以自动进行节点增广，据我们所知，这是该策略首次应用于MSA任务。

简而言之，我们的工作贡献可概括如下：

- 我们从一个新颖的角度来处理 MSA 任务，它对全局和局部信息进行显式建模，以利用全局和局部信息的潜在表示和情感关系。
- 我们设计了一个新的层次图对比学习（HGCL-LG）框架，用于提取局部级别和全局级别的情感关系。
- 在基于图对比学习的 MAS 任务中，我们引入了一种自动图增强策略来探索更好的多模态图结构。
- 对 CMU-MOSI 和 CMU-MOSEI 数据集的性能评估表明，与几个竞争基线相比，所提出的框架具有优越性和鲁棒性。

本研究的其余部分结构如下。“相关工作”部分主要介绍了两个方面的研究：多模态情感分析和对比学习。“方法论”部分详细描述了所提出的 HGCL-LG 架构，并描述了分层图对比学习的训练过程。“实验”部分介绍了实验设置、基线模型描述，并进行了HGCL-LG与基线模型的对比实验，以及消融实验和实验结果可视化。最后，“结论”部分总结了所有研究结果并得出结论。

相关作品

近年来，多模态情感分析因其多模态数据中生动有趣的信息而在多媒体界引起了广泛的关注，下面我们主要介绍在没有跨实例信息的传统MSA模型上的相关工作以及我们提出的方法。

多模态情感分析

MSA 的目标是通过声音、视觉和文本线索对话语的整体情绪进行回归或分类。 TFN 和 LMF 等模型使用基于张量的方法来获得话语的联合表示。 MSAF 设计了一种加权跨模态注意力机制来探索跨模态交互。 MAMN采用多级注意力图网络在多模态融合之前过滤噪声，并捕获多粒度特征之间的一致和异构相关性以进行多模态情感分析。

这些方法已被应用于提取欧氏结构数据的特征并取得了巨大成功。这些方法在图数据等非欧几里得结构数据上的性能仍然不能令人满意。图神经网络（GNN）被提出来处理图结构数据以捕获节点之间的交互。多模态图将顺序学习问题转化为图学习问题，可以有效地学习更长的模内和模间时间依赖性。 TGCN引入图卷积网络来获取特定于模态的语义信息，作者设计了一个两阶段注意力融合网络来融合特定于模态级别和跨模态级别的特征。

上述方法在MSA中表现出了优异的性能。然而，这些模型都是用来探索单个实例中多模态信息之间的关系，不存在对跨实例信息的额外处理。我们提出了一种新颖的基于图的方法来学习跨实例的关系。

对比学习

我们的工作还涉及对比学习。对比学习（CL）最初被提出作为一种自监督学习方法，用于解决监督信号缺乏的问题。 CL 通常需要有效的数据增强作为基础。 MISA 学习每种模态的模态不变和模态特定表示，以改进融合过程。 MMCL 被提出来同时捕获模态内和模态间动态。与图网络的结合是对比学习的另一个新应用。图网络可以对节点之间的关联进行建模，图结构上的数据增强是可行和可操作的。常见的增强方法包括节点或边的添加和删除、节点或边的表示的屏蔽等，这些方法通常不能很好地适应输入数据或保留原始语义结构。因此，为了探索更合适的图结构，受 的启发，我们通过自动删除和屏蔽图中的节点来应用图增强，从而导出相对于源的多种多样但相似的图结构。

方法论

在本节中，我们首先从任务制定开始。然后，我们详细介绍了我们提出的 HGCL-LG。我们的 HGCL-LG 的架构如图 1 所示。最后，我们描述了分层图对比学习的训练过程。

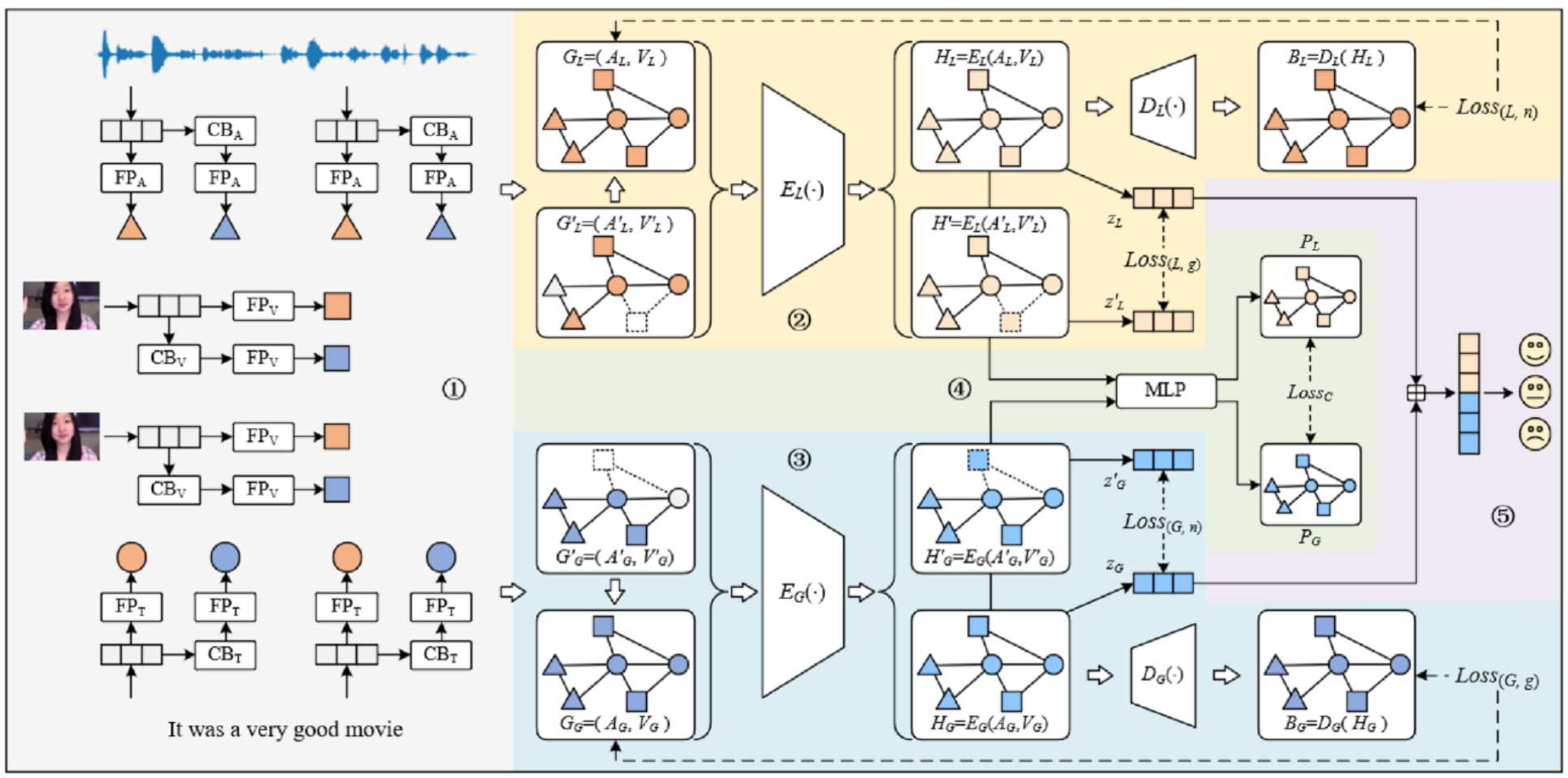


图 1.我们提出的 HGCL-LG 框架的总体架构。该模型由五个主要模块组成：①图构建，②局部图对比学习，③全局图对比学习，④跨级图对比学习，⑤融合与情感预测。

任务设置

形式上，假设有一个由文本 t 以及视频中相应的图像帧 v 和音频 a 组成的样本，多模态情感分析 (MSA) 旨在预测情感分数 y，该分数是从 - 3.0 到 3.0 的常数，对于每个样本。此外，根据情感得分 y，我们可以识别情感极性（即，如果 y > 0，则为正；如果 y = 0，则为中性；如果 y < 0，则为负）。

图构建

本节描述如何为每个多模式实例构建局部和全局图。
原始多模态序列特征是直接从一个话语样本中提取的，不考虑与数据集中其他样本的关系，我们将其定义为局部序列特征。相反，考虑数据集中样本之间关系的序列特征被定义为全局特征。

创建码本
dVAE可以从数据集中学习一个嵌入空间，这个嵌入空间包括数据集的全局共现特征。
我们以声学模态为例来解释创建码本的过程。首先，给定原始声学序列特征 X，可以定义为：

$$X = \{a|i = 1, . 。 。 , T\} \in R \quad (1)$$

其中 ai 表示序列特征的第 i 个向量。 Ta 是序列长度，da 是表示向量维度。然后，dVAE将训练集中所有样本的声学序列特征作为输入，得到声学码本 CB：

$$CB = \{cb|k = 1, .。 。 , k \in R\} \quad (2)$$

在哪里 cb 表示声学码本的第 k 个向量，并且 k 表示离散空间的大小。最后，按照用同样的方法，我们得到文本码本 CB 和视觉密码本 CB.

构建本地图
为了利用局部特征中复杂的情感含义，我们基于原始序列特征构建了局部多模态图。
节点构建 如图 1 所示，每种模态的输入特征向量首先通过模态特定的前馈网络。这允许将不同模态的特征嵌入转换为相同的维度。然后，将位置嵌入（针对每种模态分别）添加到每个嵌入以对时间信息进行编码。该操作的输出成为图中的一个节点（图 2）。
边缘构建 之前的工作表明文本在 MAS 中起着最重要的作用，因此我们以文本为中心构建边缘。如图3所示，首先我们采用全连接方案来链接节点，

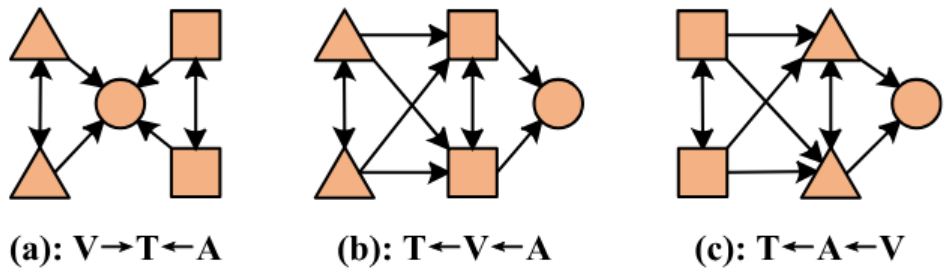


图 2. 三种边构造方式，圆圈表示文本节点，三角形表示音频节点，正方形表示视频节点。

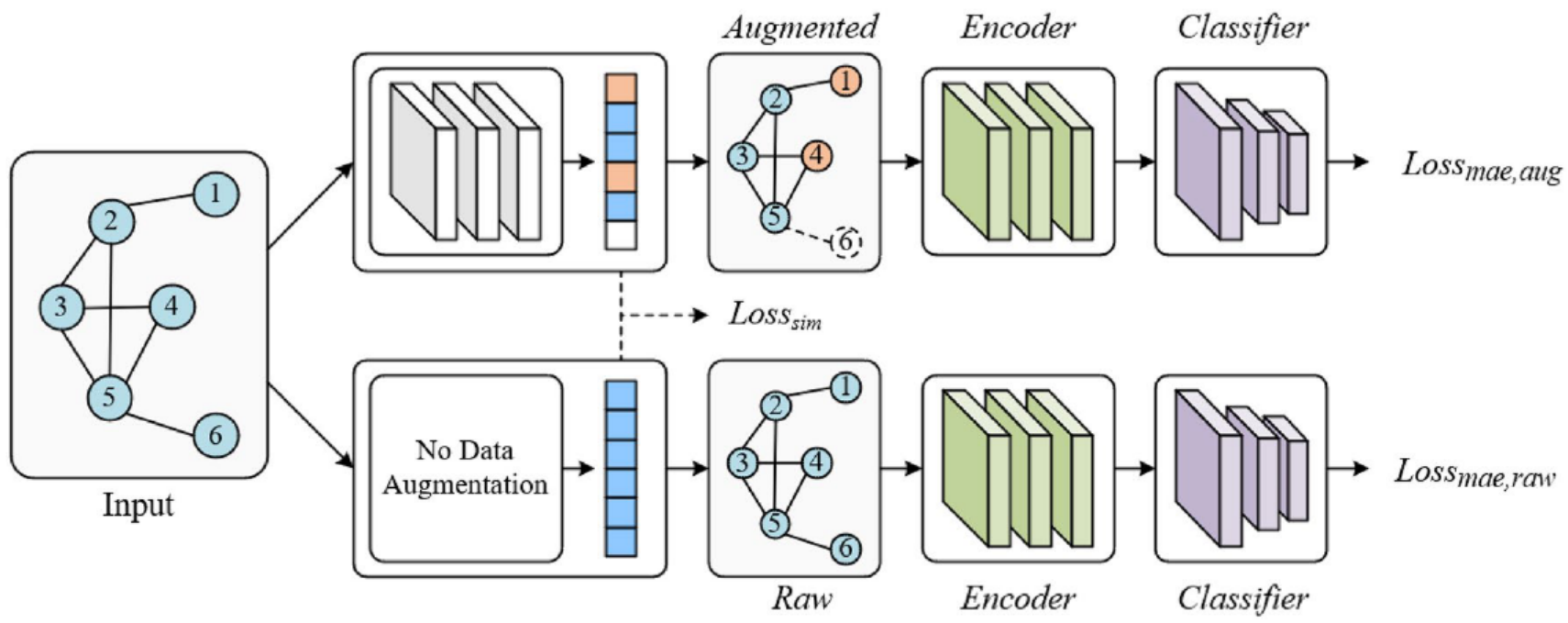


图 3. 自动图数据增强策略的架构。 GNN 层嵌入原始图以生成每个节点的分布。每个节点的增强选择是使用 gumbel-softmax 从中采样的。

来自相同的模式。然后，对于来自不同模态的节点，我们根据音频到文本和视频到文本标准连接这些节点。经过上述操作，我们可以得到局部图GL=(AL,VL)，其中AL表示邻接矩阵，VL为节点特征。

构建全局图

为了利用局部特征中复杂的情感含义，我们基于原始序列特征构建了局部多模态图。

我们获取在“构建码本”部分中的每种模态，它是包含数据集的全局共现特征的二维矩阵。因此，对于每个话语，我们使用相应的码本来映射每个模态的序列特征。与《教派》中相同。 3.2.1，我们使用声学模态解释这个映射过程。

$$X = \{CB_i \mid i = 1, \ldots, T\} \in R^{CB} \quad (3)$$

在哪里

$$id = \arg\min_k \left\| CB - x_i \right\|_2, \quad k = 1, \ldots, K \quad (4)$$

其中X是全局声学序列特征，CB_{id} 是 C 的第一个向量。 C 的大小和视频和文本的原始序列特征进行相同的操作，对文本和视频的原始序列特征进行相同的操作，得到文本全局序列特征X和视觉全局序列特征X。最后，我们使用相同的方法如“构建局部图”部分所示，构建全局多模态图 G^{GAVA} 代表邻接关系矩阵和 V 是节点特征，探索全局层面的信息交互。

层次图对比学习

本节由局部图对比学习、全局图对比和跨层图对比学习与融合以及情感预测四部分组成。以下各节讨论这三个部分的详细信息。

本地级图对比学习

为了探索多模态情感提取中的局部信息表示，我们设计了局部图对比学习。首先，给定一个局部图 $G^{A,V}$），一种自动图增强策略

（“分层图对比学习”一节）用于获得增强图G=（ ，A V ）。和 然后，图编码器（“自动图数据增强策略”部分）采用 G 和气体输入和 输出的潜在表示 G 和G。

$$H = \text{图形编码器}(G) \tag{5}$$

$$H = \text{图编码器}(G) \tag{6}$$

和哪里 H 。我们期望这些表示具有最终输出所具有的不变性。为此，我们分别考虑图神经网络中的编码器和解码器。遵循 Ji 等人的理论。 。对于编码器，我们引入了读出函数，即全局均值池，以考虑图级别的不变性。

$$z = \text{读数}(H) \tag{7}$$

$$z = \text{读数}(H) \tag{8}$$

其中 READOUT(•) 是读出函数，对于解码器，我们采用全连接层作为解码器以保持节点级别的不变性。

z z 代表全局 H

$$G = \text{解码器}(H) = (A, V) \tag{9}$$

基于此，给定小批量中的 N 个示例，我们设计了用于局部级图对比学习的损失函数：

$$\text{损失} = \text{损失} + \alpha L \tag{10}$$

$$\text{损失} = \frac{1}{N}\sum_{i=1}^{N} \left\| V - V \right\|_2 \tag{11}$$

$$\text{损失} = \frac{1}{N}\sum_{i=1}^{N} \left\| z - z \right\| \tag{12}$$

在哪里 Loss and Loss 分别表示节点级和图级自监督对比损失的比较损失。上标i表示mini-batch的索引值， V 解说的数量 第 i 个图中的节点，α 是调整平衡的超参数。

跨层次图对比学习

从局部和全局级别的图对比学习中，我们可以获得局部和全局潜在图表示。它们是同一样本的不同潜在表示，引用相同的情感信息。跨级图对比学习的目的是学习两个编码器，使得两种模态的嵌入在学习空间中彼此接近。在那里，我们定义了 Hand Has 一个正样本对。我们应用具有共享参数的非线性投影 MLP 将不同表示形式的嵌入转换到同一空间进行比较。

$$p = \text{MLP}(H) \tag{16}$$

$$p = \text{MLP}(H) \tag{17}$$

跨级图对比学习中的对比损失公式为：

$$\text{损失} = -\text{对数}\left[\frac{\sum_{i=1}^{N} \exp(\text{sim}(,)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(,)/\tau)}\right] \tag{18}$$

其中sim (•)是余弦相似度，τ是温度值。

融合与情感预测

两个表示的串联被视为融合结果，并被输入到一个简单的分类器中以做出情感强度的最终预测。

$$O = \text{连接}[z \quad z] \tag{19}$$

$$y = W \cdot \text{LeakReLU}(W \cdot \text{BN}(O) + b) + b \tag{20}$$

其中BN是BatchNorm操作，LeakyReLu用作激活。

模型训练
与图对比学习损失一起，模型的整体学习是通过最小化来执行的：

$$L = \frac{1}{N}\sum^{N} \left(y - \hat{y} \right) + \beta \text{损失} + \gamma \left(\text{损失} + \text{损失} g \right) \tag{21}$$

在哪里$\hat{y}$是模型的预测输出，y是真实标签，β和γ是超参数，控制不同损失的影响。

自动图数据增强策略

为了更好地探索图的结构，受 的启发，我们引入了自动图数据增强模型。

自动图数据增强框架
如图 3 所示，给定一个图 G，我们使用 GIN 层从节点属性中获取节点嵌入。

$$h = \text{杜松子酒}\left(h_v \right) \tag{22}$$

我们使用 n 个 GIN 层作为嵌入层，我们将 h 表示为第 n 层之后节点 v 的嵌入。
对于每个节点，我们使用嵌入的节点特征来预测选择某个增强操作的概率。每个节点的增强池是 drop、keep 和 mean-mask。我们使用gumbel-softamx从这些概率中进行采样，然后为每个节点分配一个增强操作。

$$f = \text{GumbelSoftmax}\left(h \right) \tag{23}$$

对于节点 v，我们有节点特征 x 、增强选择 fv 以及用于应用增强的函数 Aug(x, f)。然后通过以下方式获得节点 v 的增强特征 x：

$$x = \text{8月}(x, f) \tag{24}$$

最后一层 n 的维度被设置为每个节点可能的增强数量相同。因此，h表示选择每种增强的概率分布。 f 是通过 gumbel-softmax 从该分布中采样的one-hot 向量。

自动图数据增强的训练
根据InfoMin原理，一个好的对比学习正样本对应该最大化标签相关信息，并最小化它们之间的互信息（边缘相似性）。在此基础上，我们设计了一个训练过程（见图4）。对于标签相关信息，首先，我们使用图编码器（“融合和情感预测”一节）来融合节点之间的信息。

$$H = \text{图编码器}(G) \tag{25}$$

$$H = \text{图编码器}(G) \tag{26}$$

其中编码器表示原始图和增强图，然后H使用全局均值池化来获得每个图的图级表示（z 和 z ）。接下来，将 z 和 z′ 输入到两个前馈神经网络中以获得预测的情感分数。

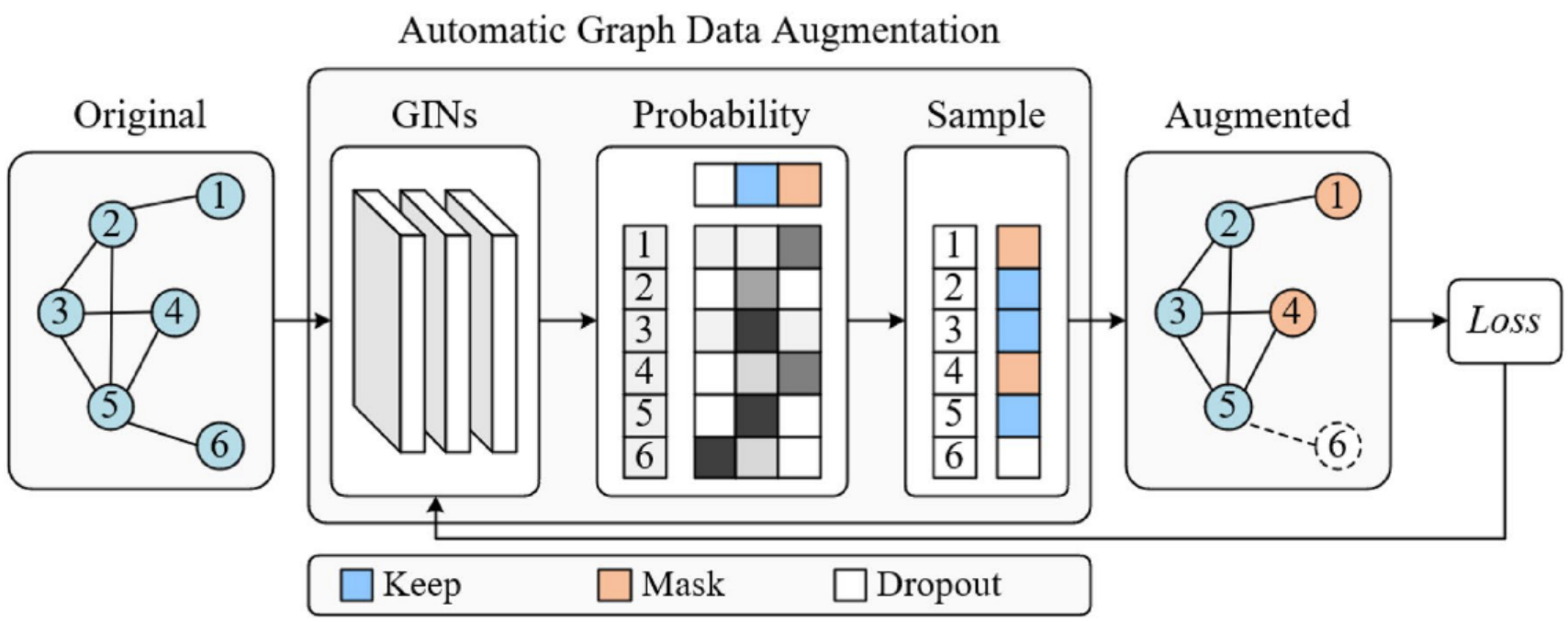


图 4. 自动图数据增强的训练。

$$y = W(Wz + b) + b \tag{27}$$

$$y = W(Wz + b) + b \tag{28}$$

其中W，W，W，W表示可学习权重，b，b，b，b表示可学习偏差。我们直接使用平均绝对误差（MAE）损失，损失函数计算如下：

$$L = \frac{1}{N}\sum_{i=1}^{N}\left[\left|\hat{y} - y\right| + \left|\hat{y} - y\right|\right] \tag{29}$$

对于互信息，在视图生成过程中，我们有一个采样状态矩阵 S 指示每个节点相应的增强操作。对于图G，我们将采样的增强选择矩阵表示为A，并定义所有“保留”的采样状态矩阵为A，然后我们制定相似性损失Las：

$$L = \text{模拟}(A, A) \tag{30}$$

其中sim(a,b)表示A和B之间的余弦相似度。模型的整体学习是通过最小化：

$$\text{损失} = L + L \tag{31}$$

图表示学习

基于我们的图结构，我们采用图注意力网络通过聚合来自具有不同权重的邻域的信息来更新图中的节点。具体来说，对于当前节点 v 和邻居节点 v，将它们连接起来，然后映射到标量作为注意力系数。

$$s = \text{泄漏 ReLU}\left(a\left[Wh \,\|\, Wh\right]\right) \tag{32}$$

其中a是权重向量，W是权重矩阵，‖是串联操作。然后通过softmax对所有邻居的注意力系数进行归一化。

$$a = \text{软最大值}_j(s) = \frac{\text{经验}}{\sum_{k\in N}\text{经验}} \tag{33}$$

其中 N 表示节点 i 及其邻居的集合。最后，节点 i 的表示通过邻居和自身表示的加权和进行更新，并应用多头注意力机制来稳定自注意力的学习过程。

$$h = \underset{j\in N}{\text{瓦时}} \tag{34}$$

其中 k 表示第 k 个注意力头。

$$L = \frac{1}{N}\sum^{N}\left|y - y\right| + \alpha L + \beta\left(L + L\right) \tag{35}$$

在哪里$\hat{y}$是模型的预测输出，y是真实标签，α、β和γ是超参数，控制不同损失的影响。

实验

该实验在由四个 NVIDIA GeForce RTX 3090 GPU 组成的高性能计算集群上进行，该集群提供了巨大的计算能力。集群通过高速网络互连，保证高效的数据通信和并行处理。

实验设置

数据集

在这项工作中，在两个公共多模态情感分析数据集 CMU-MOSI 和 CMU-MOSEI 上进行了实验。每个数据集的基本统计数据如表1所示。这里，我们对上述数据集进行简单介绍。

CMU-MOSI CMU-MOSI 数据集是最流行的 MSA 基准数据集之一。该数据集包含 2199 个简短的独白视频剪辑，取自 93 个 YouTube 电影评论视频。这些话语是用从 - 3（强烈负面）到 3（强烈正面）的情绪分数手动注释的。

CMU-MOSEI CMU-MOSEI 是 CMU-MOSI 的放大版。它具有与 CMUMOSI 相同的注释。在 CMU-MOSEI 中，有 16,326 个话语用于训练，1871 个话语用于验证，4659 个话语用于测试。

数据集	#Train	#Test	#Valid	#All
MOSI	1283	229	686	2198
莫塞伊	16,326	1871	4659	22,856

表 1. 基准 MSA 数据集的数据集基本统计。

评估指标

为了与基线进行全面比较，我们使用分类和回归的公共评估指标来展示我们提出的框架的性能，并进一步与基线进行比较：七类分类精度（Acc7）表明在 [− 3 范围内正确的情感标签预测， + 3]，二元分类 (Acc2) 和 F1 分数，平均绝对误差 (MAE) 计算预测标签和真实标签之间的平均绝对差，皮尔逊相关 (Corr) 测量预测偏差程度。

实施细节

我们的模型的结果采用使用不同随机种子的五次运行获得的平均结果，以获得稳定的结果。详细的训练设置如表2所示。此外，我们在训练时使用学习率调整策略来更新学习率。其中，α、β、γ是我们通过网格搜索找到的最合适的值。

基线

LMF 低秩多模态融合（LMF）是一种利用低秩权重张量在不影响性能的情况下提高多模态融合效率的方法。它不仅大大降低了计算复杂度，而且显着提高了性能。但它仍然存在一些缺点，如计算资源要求高、处理噪声和冗余的能力较弱、易受干扰等。

TFN 张量融合网络 (TFN) 利用张量融合层，其中笛卡尔积用于形成特征向量。因此，可以融合三种模态的信息来预测情绪。TFN的主要缺点包括计算复杂度高、对噪声和异常值敏感、对参数和模型结构的依赖、可解释性有限以及需要大量注释数据。

MISA 通过将样本的每种模态投影到两个子空间中，该方法学习模态不变的表示和特定的表示，然后将其融合以进行情感分析。

MulT Multimodal Transformer 通过定向成对跨模态注意力扩展了三组 Transformer，潜在地将流从一种模态适应到另一种模态。在使用过程中，应特别注意跨模态注意力机制的局限性以及部署和配置的复杂性。

自MM自监督多任务学习自动生成单模态标签，并通过多模态标签进行权重调整，以学习跨模态的一致性和差异。 Self-MM模型的缺点包括计算复杂度高、数据需求大、模态对齐挑战、泛化能力有限和可解释性有限。

TCM-LSTM 通过声学和视觉 LSTM 从不同的角度学习模态间动态，其中语言特征起主导作用。TCM-LSTM 模型的缺点包括计算复杂度高、参数调整困难、对初始状态敏感、倾向于局部最优以及对噪声和异常值的脆弱性。

MTAG模态时间注意力图（MTAG）能够进行融合和对齐。 MTAG 模型的缺点包括计算复杂度高、训练时间长、对噪声和异常值敏感、参数调整困难以及难以处理大规模图表数据。

范围	MOSI		莫塞伊	
	对齐	未对齐	对齐	未对齐
时代	30	30	15	15
批量大小	64	8	64	8
GAT 层	3	4	3	4
加特头	4	4	4	4
HGGL-LG LR	5e−4	1e−4	5e−4	1e−4
其他LR	1e−3	1e−3	1e−3	1e−3
辍学	0.3	0.3	0.3	0.3
α	0.1	0.1	0.1	0.1
β	0.01	0.01	0.01	0.01
γ	0.1	0.1	0.1	0.1

表 2. 培训设置详细信息。 LR学习率。

GraphCAGEGraph Capsule Aggregation (GraphCAGE)，利用基于图的神经模型和胶囊网络对未对齐的多模态序列进行建模。GraphCAGE 的缺点包括计算复杂度高、对数据质量和规模的严格要求以及需要大量标记数据。

与基线比较

我们在 CMU-MOSI 数据集上评估 HGCL-LG 模型，表 3 显示了实验结果。从结果中，我们观察到 HGCL-LG 在大多数情况下优于两个数据集上的所有基线模型，这验证了我们的方法在 MSA 任务中的有效性。这表明探索本地和全球层面的情绪影响对于提高 MSA 的绩效具有重要意义。通过T检验分析，我们发现两组数据的平均值存在显着差异（$p < 0.05$）。这表明该方法在CMU-MOSI和CMU-MOSEI上具有显着的测试结果。此外，我们提出的模型在对齐和未对齐数据集上都表现良好，但由于我们没有显式地对对齐数据进行建模，因此未对齐数据集上的结果比对齐数据集上的结果稍差，因此我们提出的模型在对齐和未对齐数据集上都表现良好，但由于我们没有对对齐数据进行显式建模，因此未对齐数据集上的结果比对齐数据集上的结果稍差。

总的来说，我们提出的层次图对比学习可以充分学习样本的局部信息和全局共现特征，可以显着提高MSA任务的精度。

消融研究

为了验证分层图对比学习对性能的影响，我们对两个数据集进行了消融实验，结果如表4所示。从表4中我们可以看到，HGCL-LG中删除任何模块都会导致性能下降在模型性能方面。对于对比学习（CL），结果表明我们设计的L和L可以很好地探索多模态实例的全局信息和局部信息，并使模型能够学习局部信息和全局信息之间的复杂关系。对于边缘类型，“V→T←A”是最有效的边缘构建方法，这表明其他两种方法在消息聚合中产生负噪声特性。然后，对于信息类型，局部特征和全局特征在 MSA 任务中发挥着重要作用。最后，我们评估全局贡献特征的有效性，“CMU-MOSI”表示使用CMU-MOSI码本构建CMU-MOSI的全局图，“CMU-MOSEI”表示使用CMU-MOSEI码本构建CMU的全局图-MOSI，结果表明提取的全局共现特征能够有效地表征情感信息。

表示可视化

图 5 显示了 HGCL-LG 计算的融合多模态表示 O 的可视化，并对比了学习损失与否。在没有对比学习的情况下，正负样本的表示具有很强的可区分性，但中性样本是离散分布的，这意味着模型没有学习到样本的局部信息与全局共现特征之间的关系。引入设计对比学习后，正负样本有了更清晰的分界线，中性样本沿着分界线分布。这表明对比学习可以有效提高模型对不同样本的区分度，这也证明了所设计的对比学习任务在表示学习上的有效性。

案例研究

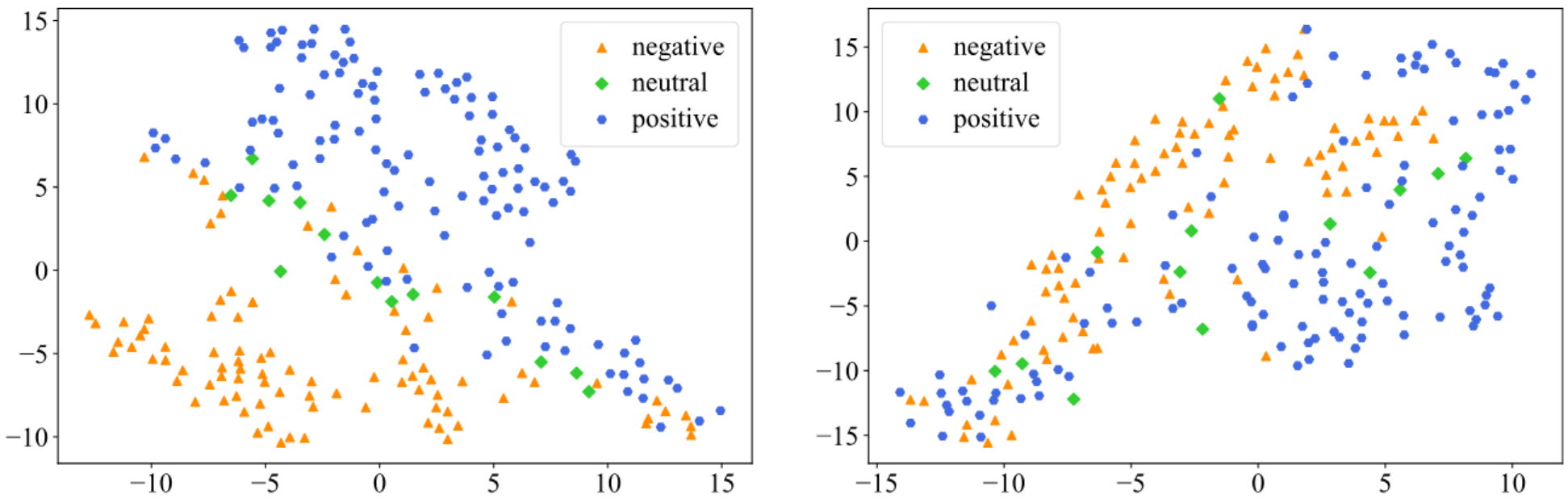
我们在图 6 中展示了图神经网络在多模态情感分析中应用的案例研究（该图像来自 CMU-MOSI。该数据集可公开下载，其中包含所有提取的特征）。

型号	卡内基梅隆大学莫西					卡内基梅隆大学莫塞伊分校					数据设定
	Acc7↑	Acc2↑	F1↑	MAE↓	Corr↑	Acc7↑	Acc2↑	F1↑	MAE↓	更正↑	
TFN*	33.7	78.3	78.2	0.925	0.662	52.2	81.0	81.1	0.570	0.716	u
LMF*	32.7	77.5	77.3	0.931	0.670	52.0	81.3	81.6	0.568	0.727	u
MuLT	35.5	80.6	79.3	0.972	0.681	49.0	81.4	81.7	0.630	0.664	a
MISA	43.5	81.8	81.7	0.752	0.784	52.2	81.6	82.0	0.550	0.758	a
自MM*	45.8	82.7	82.6	0.731	0.731	50.6	82.6	82.8	0.547	0.752	a
TCM-LSTM	35.4	81.7	81.8	0.903	0.672	50.6	81.4	81.6	0.606	0.673	a
MTAG	31.9	80.5	80.4	0.941	0.692	48.2	79.1	75.9	0.645	0.614	u
图笼	35.4	82.1	82.1	0.933	0.684	48.9	81.7	81.8	0.609	0.670	u
HGG-LG(我们的)	41.7	84.0	83.9	0.725	0.788	49.3	84.2	84.3	0.545	0.769	a
HGG-LG(我们的)	35.1	83.5	83.6	0.765	0.776	49.5	84.0	84.1	0.511	0.753	u

表示评价指标越高越好，↓表示评价指标越低越好。结果*代表我们在实验室取得的结果，其中Self-MM*是使用作者发布的源代码重现的。带 † 的结果表示来自的结果，带 ‡ 的表示来自的结果，对于数据设置，a和u分别表示对齐和未对齐。粗体表示最好的结果，斜体表示次好的结果。

消融	加速器2↑	F1↑	MAE↓	更正↑
对比学习(CL)				
二	84.0	83.9	0.725	0.788
L	82.5	82.6	0.736	0.778
L	83.1	83.0	0.730	0.780
边缘类型（图2）				
V → T ← A	84.0	83.9	0.725	0.788
T ← A ← V	83.1	83.0	0.736	0.742
T ← V ← A	82.3	82.1	0.749	0.727
信息类型（无 CL）				
仅限本地	82.1	82.2	0.712	0.720
仅限全球	81.5	81.5	0.722	0.717
本地、全球	84.0	83.9	0.725	0.788
密码本				
卡内基梅隆大学莫西	84.0	83.9	0.725	0.788
卡内基梅隆大学莫塞斯分校	83.5	83.6	0.730	0.780

表 4. 对齐的 CMU-MOSI 验证数据集的消融研究。最佳结果以粗体突出显示。 L 表示跨层图对比损失，L 表示 Land Lglobal 之和。



(a) HGCL-LG 与对比学习　　(b) 没有对比学习的HGCL-LG

图 5. CMUMOSI 有效集上嵌入空间中多模态表示的 T-SNE 可视化。

首先，将非对齐的多模态序列转换为具有异构节点和边的图，该图可以捕获不同模态之间随时间的交互。然后，使用多模式时间注意力有效地处理该图。情感分析结果是通过对流行模型进行检测获得的。该方法得到了相关工作者的认可，证明了图神经网络模型在现实世界中的适用性。

结论

本文提出了一种用于多模态情感分析（MSA）的新型分层图对比学习（HGCL-LG）框架，其中图对比学习在局部级、全局级和跨级进行。对于在局部级别和全局级别执行的图对比学习策略，我们设计了基于节点的对比损失和基于图的对比损失。基于节点的对比损失旨在通过捕获情感线索来改进情感线索的学习。本地/全局图的潜在情绪表示。跨级对比损失的设计是为了利用局部图和全局图内的情感关系。此外，为了探索更好的多模态图结构，我们引入了一种用于自动图增广的自适应图增广机制。两个基准数据集的实验结果表明，我们的方法优于 MSA 中最先进的基线。

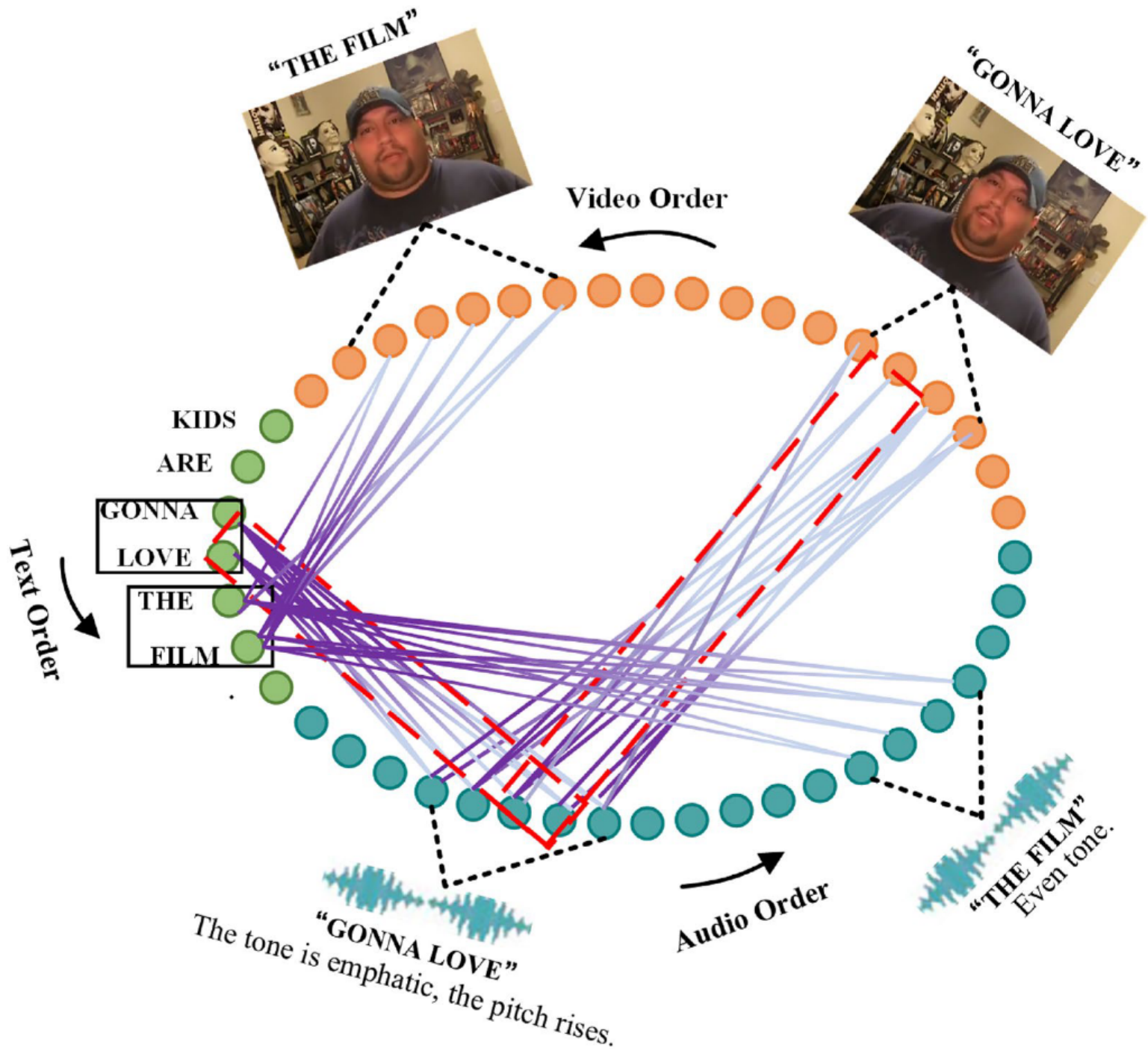


图6.图神经网络在多模态情感分析中的应用案例研究（图片来自CMU-MOSI，数据集可供公开下载）。

数据可用性

本研究期间生成或分析的所有数据均包含在这篇发表的文章中。声明：图6中的受试者图像来自CMU-MOSI，该数据集中的所有数据均可公开下载。所有受试者和/或其法定监护人均同意在充分知情后在科学报告中公布其身份信息或图像。

收稿日期：2023 年 10 月 8 日；接受日期：2024 年 2 月 17 日
Published online: 04 March 2024

参考

1. Gandhi, A.、Adhvaryu, K.、Poria, S.、Cambria, E. 和 Hussain, A. 多模态情感分析：对历史、数据集、多模态融合方法、应用、挑战和未来方向的系统回顾。信息。融合 https://doi.org/10.1016/j.inffus。2022.09.025 (2023).
2. Yu, W. M., Xu, H., Yuan, Z. Q. & Wu, J. L. Learning modality-species with self-supervised multi-task Learning for multimodal情感分析，第三十五届 AAAI 人工智能会议 (AAAI)，10790–10797 。 https://ojs.aaai。org/index.php/AAAI/article/view/17289（2021）。
3. 张，D.等人。具有异构分层消息传递的多模式多标签情感识别，第三十五届 AAAI 人工智能会议 (AAAI)，14338–14346。 https://ojs.啊啊啊。组织/索引。 php/AAAI/文章/视图/17686 (2021)。
4. Cai, Y., Cai, H. & Wan, X. 使用分层融合模型在 Twitter 中进行多模式讽刺检测，计算语言学协会 (ACL) 第 57 届会议论文集，2506–2515。 https://doi.org/10.18653/v1/p19-1239（2019）。
5. Varshney, D.、Zafar, A.、Behera, N. K. 和 Ekbal, A. 使用增强图生成基于知识的医学对话。科学。报告 13(1), 3310 (2023)。
6. Truong, Q. T. 和 Hady W. L. VistaNet：用于多模态情感分析的视觉方面注意网络，载于第三十届 AAAI 人工智能会议 (AAAI)，305-312。 https://doi.org/10.1609/aaai.v33i01.3301305（2019）。
7. 吴云，刘红，陆鹏，张丽，袁芳。基于手势识别和服装转移算法的虚拟试衣系统设计与实现。科学。报告 12(1), 18356 (2022)。
8. 陈，Y.等人。用于神经血管分流器的微结构薄膜镍钛诺。科学。报告 6(1), 23698 (2016)。
9. 刘，Z.等人。与特定模态因素的高效低阶多模态融合，计算语言学协会 (ACL) 第 56 届年会论文集，2247–2256。 https://doi.org/10.18653/v1/P18-1209（2018）。

10. Chen, Q. P., Huang, G. M. & Wang, Y. B. 用于多模态情感分析的带有情感预测辅助任务的加权跨模态注意机制。 IEEE/ACM 传输。音频语音语言。过程。 30, 2689–2695。 https://doi.org/10.1109/TASLP。
2022.3192728 (2022).
11. 薛XJ，张CX，牛ZD和吴XD。用于多模态情感分析的多级注意力图网络。 IEEE 传输。
知道。数据工程https://doi.org/10.1109/TKDE.2022.3155290（2022）。
12. Tsai, Y. H. H.、Liang, P. P.、Zadeh, A.、Morency, L. P. 和 Salakhutdinov, R. 学习因式分解多模态表示，第七届国际学习表示会议 (ICLR)。 https://openreview.net/forum?id=rygqqsA9KX（2019）。
13. Hazarika, D.、Zimmermann, R. 和 Poria, S. MISA：多模态情感分析的模态不变和特定表示，第 28 届 ACM 国际多媒体会议记录 (MM ’ 20)，1122–1131。 https://doi.org/10.1145/3394171.3413678（2020）。

14. 杨，J.N.等人。 MTAG：未对齐的人类多模态语言序列的模态时间注意力图，计算语言学协会 (ACL) 第 59 届年会论文集，1009-1021。 https://doi.org/10.18653/v1/2021。
naacl-main.79 (2021)。
15. Mai, S. J., Xing, S. L., He, J. X., Zeng, Y. & Hu, H. F. 通过图卷积和图池进行未对齐多模态序列分析的多模态图。 ACM 翻译。多媒体计算。交流。应用。 https://doi.org/10.1145/3542927（2023）。
16. 林，Z.J.等人。模态内和模态间关系建模：用于多模态情感分析的分层图对比学习。第 29 届国际计算语言学会议论文集。 https://aclanthology.org/2022.coling-1。
622/ (2022).
17. Hu, X. & Yamamura, M. 用于多模态情感分析的全局局部融合神经网络。应用。科学。 12、8453。https://doi。org/10.3390/app12178453（2022）。
18. Caschera, M. C.、Grifoni, P. 和 Ferri, F. 使用多模态方法对视频中的语音和文本进行情感分类。多模式技术。相互影响。 6、28。 https://doi.org/10.3390/mti6040028 (2022)。
19. Oord, A. V. D.、Vinyals, O. 和 Kavukcuoglu, K. 神经离散表示学习， 《神经信息处理系统进展》30 (NIPS 2017)。 https://proceedings.neurips.cc/paper/2017/hash/7a98af17e63a0ac09ce2e96d03992fc-Abstract.html（2017）。

20. Veličković, P.、Cucurull, G.、Casanova, A.、Romero, A.、Liò, P. 和 Bengio, Y. 图注意力网络。 arXiv 预印本 arXiv：1707。10903。https://doi.org/10.48550/arXiv.1710.10903（2017）。
21. 蔡，Y.H.H.等人。用于未对齐的多模态语言序列的多模态转换器。过程。会议。副教授。计算。语言学家见面会。 https://doi.org/10.18653/2Fv1/2Fp19-1656（2019）。
22. Huang, K.、Xiao, C.、Glass, L. M.、Zitnik, M. 和 Sun, J. SkipGNN：通过跳图网络预测分子相互作用。科学。
报告 10(1), 21092 (2020)。
23. Huang, J., Lin, Z. H., Yang, Z J. & Liu, W. Y. 用于多模态情感分析的时间图卷积网络，载于 2021 年多模态交互国际会议记录 (ICMI '21)，239-247。 https://doi.org/10.1145/3462244.34799 39 (2021)。

24. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. 视觉表征对比学习的简单框架，第 37 届国际机器学习会议记录，1597-1607。 https://proceedings.mlr.press/v119/chen20j.html（2020）。

25. 刘，C.等人。 DialogueCSE：基于对话的句子嵌入对比学习，2021 年自然语言处理经验方法会议论文集，2396-2406。 https://doi.org/10.18653/v1/2021.emnlp-main.185（2021）。
26. Lin, R. H. & Hu, H. F. 通过单模态编码和跨模态预测进行多模态情感分析的多模态对比学习，计算语言学协会的调查结果：EMNLP 2022, 511–523。 https://aclanthology.org/2022。
调查结果-emnlp.36 (2022)。
27. 你，Y.N.等人。具有增强功能的图对比学习， 《神经信息处理系统进展》33 (NeurIPS 2020)，5812–5823。 https://proceedings.neurips.cc/paper/2020/hash/3fe230348e9a12c13120749e3f9fa4cd-Abstract.html（2020）。
28. 朱，Y.Q.等。深度图对比表示学习。 arXiv 预印本 arXiv：2006.04131。 https://doi.org/10.48550/arXiv。
2006.04131 (2020).
29. 尹玉华、王启智、黄思源、熊海Y. & 张，X. AutoGCL：通过可学习视图生成器进行自动图形对比学习，第三十六届 AAAI 人工智能会议 (AAAI)，8892–8900。 https://doi.org/10.1609/aaai.v36i8.20871。
(2022).
30. Xu, K. Y. L., Hu, W. H., Leskovec, J. & Jegelka, S. 图神经网络有多强大？ arXiv 预印本 arXiv：1810.00826。 https://doi.org/10.48550/arXiv.1810.00826（2018）。
31. 田，Y.L.等。对比学习的良好观点是什么？神经信息处理系统的进展 33 (NeurIPS 2020), 6827–6839。 https://proceedings.neurips.cc/paper_files/paper/2020/file/4c2e5eaae9152079b9e95845750bb9abPaper.pdf（2020）。

32. Zadeh, A.、Zellers, R.、Pincus, E. 和 Morency, L. P. Mosi：在线观点视频中情感强度和主观性分析的多模态语料库。 arXiv 预印本 arXiv:1606.06259 (2016)。
33. Zadeh, A. B.、Liang, P. P.、Poria, S.、Cambria, E. & Morency, L. P. 野外多模态语言分析：Cmu-mosei 数据集和可解释的动态融合图，计算语言学协会第 56 届年会论文集（第一卷：长论文），2236–2246。 https://doi.org/10.18653/v1/P18-1208（2018）。
34. Han, W., Chen, H. 和 Poria, S. 通过分层互信息最大化改进多模态融合以进行多模态情感分析，载于 2021 年自然语言处理经验方法会议论文集，9180-9192。在线和蓬塔卡纳，多米尼加共和国计算语言学协会（2021 年）。

35. Mai, S. J., Xing, S. L. & Hu, H. F. 使用通道感知时间卷积网络通过声学和视觉 LSTM 分析多模态情感。 IEEE/ACM 传输。音频语音语言。过程。 29, 1424–1437。 https://doi.org/10.1109/TASLP.2021.3068598（2021）。

36. Maaten, L. V. D. 和 Hinton, G. 使用 t-SNE 可视化数据。 J.马赫.学习。资源。 9, 2579–2605 (2008)。

作者贡献

DJ为本文提供了初始架构，并为模型构建提供了初步指导。 J.J.H.负责模型构建、数据选择和实验结果分析，作者审阅手稿。 Z.J.为本文提供了补充实验，并准备了图。 1、2、3、4、5，表 1、2、3、4。为本文提供了支持材料，并为本文的布局做出了贡献。

利益竞争

作者声明没有竞争利益。

附加信息

信件和材料请求应寄给 J.J.

卷：.(1234567890)