基于互信息最大化和对比损失的多模态对话情绪识别模型

黎倩尔, 黄沛杰*, 陈佳炜, 吴嘉林, 徐禹洪, 林丕源
华南农业大学, 数学与信息学院, 广东广州, 510642
li@stu.scau.edu.cn, pjhuang@scau.edu.cn,jw_chen@stu.scau.edu.cn,
galinwu@stu.scau.edu.cn, xuyuhong@scau.edu.cn, pyuanlin@scau.edu.cn

摘要

多模态的对话情绪识别（emotion recognition in conversation，ERC）是构建情感对话系统的关键。近年来基于图的融合方法在会话中动态聚合多模态上下文特征，提高了模型在多模态对话情绪识别方面的性能。然而，这些方法都没有充分保留和利用输入数据中的有价值的信息。具体地说，它们都没有保留从输入到融合结果的任务相关信息，并且忽略了标签本身蕴含的信息。本文提出了一种基于互信息最大化和对比损失的多模态对话情绪识别模型MMIC来解决上述的问题。模型通过在输入级和融合级上分级最大化模态之间的互信息（mutual information），使任务相关信息在融合过程中得以保存，从而生成更丰富的多模态表示。本文还在基于图的动态融合网络中引入了监督对比学习（supervised contrastive learning），通过充分利用标签蕴含的信息，使不同情绪相互排斥，增强了模型识别相似情绪的能力。在两个英文和一个中文的公共数据集上的大量实验证明了所提出模型的有效性和优越性。此外，在所提出模型上进行的案例探究有效地证实了模型可以有效保留任务相关信息，更好地区分出相似的情绪。消融实验和可视化结果证明了模型中每个模块的有效性。

关键词： 多模态对话情绪识别；图卷积网络；互信息；监督对比学习

Multimodal Emotion Recognition in Conversation with Mutual Information Maximization and Contrastive Loss

**Qianer Li, Peijie Huang*, Jiawei Chen, Jialin Wu,
Yuhong Xu, Piyuan Lin**
College of Mathematics and Informatics, South China Agricultural University, China
li@stu.scau.edu.cn, pjhuang@scau.edu.cn,jw_chen@stu.scau.edu.cn,
galinwu@stu.scau.edu.cn, xuyuhong@scau.edu.cn, pyuanlin@scau.edu.cn

Abstract

Emotion recognition in conversation (ERC) is a key component for building emotional dialogue systems. In recent years, graph-based fusion methods have been proposed to dynamically aggregate multimodal context features in conversation, which improve the performance of models on multimodal emotion recognition in conversation. However, these methods do not fully preserve and utilize valuable information in the input data. Specifically, they do not retain task-relevant information from input to fusion result, and ignore the information implied by labels themselves. In this paper, to overcome

*通讯作者

第二十二届中国计算语言学大会论文集，第264页-第276页，哈尔滨，中国，2023年8月3日至5日。
(c) 2023 中国中文信息学会计算语言学专业委员会

the above issues, we propose a new model based on Mutual Information and Contrast(MMIC), for multimodal emotion recognition in conversation. The model maximizes the mutual information between modalities at input level and fusion level hierarchically, which preserves task-relevant information during fusion process and generates richer multimodal representations. We also introduce supervised contrastive learning into graph-based dynamic fusion network, which leverages the information implied by labels to make different emotions repel each other and enhances the model's ability to recognize similar emotions. Extensive experiments on two public benchmark datasets and a new Chinese dataset demonstrate the effectiveness of our proposed model. In addition, case studies conducted on the proposed model effectively confirmed that the model can effectively retain task-related information and better distinguish similar emotions. Ablation experiments and visualization results demonstrated the effectiveness of each module in the model.

Keywords: multimodal emotion recognition in conversation , graph convolutional network , mutual information , supervised contrastive learning

1 引言

情绪是人类日常交流的重要组成部分。对话中的情绪识别（ERC）旨在对话过程中自动识别和跟踪说话者的情绪状态，如图1所示。近年来，它越来越引起了自然语言处理和多模态处理领域研究者的关注。ERC具有广泛的潜在应用范围，如心理学、人机交互和社交机器人等领域。具体来说，对话系统想要和人类进行有效的情感沟通，就必须具备一定的情感能力，因此对用户情绪进行正确的识别判断显得十分关键(Zhou et al., 2018)。在ERC的背景下，充分利用现有数据中的有价值的信息是至关重要的，因为对话是一个复杂而动态的过程，情绪可以通过各种方式表达，如言语，面部表情和肢体语言等。通过充分利用这些不同形式的多模态信息，我们可以对说话者的情绪状态有更为完整的理解。

图 1. M^{3}ED数据集中的一个对话示例

社交媒体的快速增长是促使人们研究对话情绪识别的原因之一。对话情绪识别和传统情绪识别的区别在于，对话情绪识别是一个分类任务，旨在对一段对话中的话语进行情绪分类。任务的输入是一段连续的对话，输出是这段对话中所有话语的情绪。传统情绪识别是一种识别单独话语中不同人情感类型的方法。对话情绪识别比独立情绪识别更困难更有意义的地方在于，对话层面的情绪分析有助于理解说话者在整个说话过程中所涵盖主题的情绪动态变化(Zadeh et al., 2016)。此外，对话情绪识别可广泛应用于各种对话场景中，如社交媒体中评论的情感分析、人工客服中客户的情绪分析等。对话中通常蕴含了丰富的语境线索(Hu et al., 2019)，能否捕捉到语境线索对情绪识别至关重要。在以往与语境相关的研究中，对话中的文本模态是研究的重点，而这些研究常常忽略了其他模态的有效组合。在考虑多模态上下文信息的工作中，对

话通常被建模为序列或图结构，而特征连接则是通过简单的早期或是晚期融合方法(Majumder et al., 2019; Ghosal et al., 2019)实现的，缺乏对多模态信息的有效利用探索。

近年来，已经发表的著名研究是在图架构中模拟跨模态和单模态交互(Hu et al., 2021; Liu et al., 2021)，这为跟踪情绪模式的互补性提供了深刻的见解。研究(Hu et al., 2022)利用基于图的融合方法对来自不同语义空间的上下文信息进行动态建模，减少了连续聚合中的冗余信息。然而，这些方法不能充分利用数据中存在的有价值的信息，即从输入到融合结果的任务相关信息以及标签本身包含的信息。因此，这些方法无法捕获数据中底层关系和依赖关系的全部复杂性。我们认为，将关键任务相关信息从输入保留到融合结果中，可以提取和整合输入中的单模态原始数据，从而生成更丰富的多模态表示(Han et al., 2021)。此外，利用标签中蕴含的情绪信息，可以使相似的情绪之间相互排斥，从而更好地区分相似的情绪，如“沮丧”和“悲伤”，“快乐”和“兴奋” (Li et al., 2022)。

为了克服上述提及的问题，本文提出了一种基于互信息最大化与对比损失的多模态对话情绪识别模型——MMIC。为了更好地保存从输入到融合结果的任务相关信息，我们首先最大化了各模态表示之间的互信息。在通过图卷积运算得到融合结果后，我们进一步最大化它与低水平单模态表示之间的互信息。为了更充分地利用标签中隐含的信息，我们采用监督对比学习，使具有相同情绪的样本具有内聚性，不同情绪的样本相互排斥。我们在两个公共基准数据集和一个中文数据集上进行了大量实验，证明了所提出模型的有效性和优越性。此外，在所提出模型上进行的案例探究有效地证实了该模型可以有效保留任务相关信息，更好地区分出相似的情绪。消融实验和可视化结果进一步证明了该模型中每个模块的有效性。本文的主要贡献如下：

- 我们提出了一种基于互信息和对比损失的多模态对话情绪识别模型，实现了对输入数据的充分利用，提高模型在多模态对话情绪识别上的性能。
- 我们在输入级和融合级上分级最大化模态之间的互信息，更好地在融合过程中保存与任务相关的信息，从而生成丰富多模态表示。
- 我们将监督对比学习纳入基于图的融合网络，充分利用标签信息，使具有相同情绪的样本具有内聚性，不同情绪的样本相互排斥，从而增强对相似情绪的识别。
- 我们在两个英文公共基准数据集和一个新发布的中文数据集上进行了大量实验，验证了本文提出的模型取得了优于研究进展方法的效果。

2 相关工作

本文的研究通过充分保留和利用输入数据中有价值的信息，提高模型在多模态对话情绪识别上的性能。本节简要介绍相关技术方法，并阐述本文方案中融入这些技术方法的设计依据。

随着近年来社交媒体的快速进步，以及高质量拍摄设备的出现，我们见证了多模态数据的爆炸式增长，如电影、短视频等。在现实生活中，多模态数据通常由三个模态组成：视觉、声学和转录文本。一般来说，同一数据段中的不同模态往往是具有互补性的，为语义和情绪消歧提供额外的线索(Li et al., 2011)。多模态数据融合是情绪识别中至关重要的部分，模型要从所有输入数据中提取和整合信息，才能理解数据背后代表的情绪。因此，模型能否在过滤噪声的同时，生成具有带有任务相关信息的丰富多模态表示，对情绪识别十分关键。

早期多模态融合的典型代表是LSTM(Poria et al., 2017)和ICON(Hazarika et al., 2018)，它们通过拼接三种模态的特征来利用多模态信息，而不对模态之间的交互进行建模。Chen等人(Chen et al., 2017)在单词水平上进行多模态融合，对孤立的话语进行情绪识别。Zadeh等人(Zadeh et al., 2018)提出MFN来融合多视图的信息，从而很好地协调来自不同模式的特征。Hu和Liu等人(Hu et al., 2021; Liu et al., 2021)在图架构中模拟跨模态和单模态交互，很好利用了模态之间的互补性。方法(Hu et al., 2022)则利用基于图的融合方法对来自不同语义空间的上下文信息进行动态建模，减少了连续聚合中的冗余信息。然而，这些方法都缺乏对从原始输入到融合嵌入的信息流的控制，这可能会导致信息流在融合过程中丢失任务相关信息。

互信息（mutual information，MI）是信息论中的一个概念，指两个随机变量之间的关联程度。即给定一个随机变量后，另一个随机变量不确定性的削弱程度。互信息越大，说明两个随机变量之间的相关性越强；互信息为0，说明两个随机变量之间没有任何关系。互信息可以用来评价两个变量是否有关系，以及关系的强弱，其在机器学习、深度学习和数据挖掘等领域有

广泛的应用。Alemi等人(Alemi et al., 2016)首先将与MI与深度学习模型进行结合。从那时此，大量的著作(Bachman et al., 2019; He et al., 2020; Amjad and Geiger, 2019)研究并证明了MI最大化原则的好处。互信息最大化已经被证明了在减少与下游任务无关的冗余信息，保留任务相关信息方面起到有效作用(Poole et al., 2019; Han et al., 2021)。在本文的工作中，我们利用互信息最大化，分级对模态输入对以及多模态融合结果与单模态融合结果之间的互信息进行最大化，更好地过滤噪声，在融合过程中保存与任务相关的信息。

对比学习（contrastive learning，CL）(Hinton et al., 2006)是一种基于对比思想的判别式表示学习框架，已被用于图像、文本、语音等多种领域。监督对比学习（supervised contrastive learning, SCL）(Khosla et al., 2021)则是一种利用标签信息来提高对比学习效果的方法。对比学习是自监督学习，它通过构造正负样本对，让模型学习区分不同的数据表示。监督对比学习则是在一个批处理内，让同类别的样本表示彼此接近，而不同类别的样本表示彼此远离。这样可以增强模型的判别能力和泛化能力。利用好标签本身所携带的信息，我们可以对说话者的情绪状态有更为完整的理解。监督对比学习已经被证明，在充分利用标签信息的情况下，可以使具有相同情绪的样本具有内聚性，不同情绪的样本相互排斥(Li et al., 2022)。在本文中，我们将监督对比学习纳入基于图的融合网络，利用标签信息增强模型对相似情绪的识别能力。

3 多模态对话情绪识别模型

形式上，一个包含N句话语的对话可以被描述为一系列话语$\{u_1, u_2, \ldots, u_N\}$。每个话语都包含三种话语层面的特征，包括声学(a)、视觉(v)和文本(t)，它们可以表示为$u_i = \{u_i^a, u_i^v, u_i^t\}$。多模态ERC的目标是预测每个话语$u_i$的情绪标签$y_i$。

所提出的MMIC如图1所示，其包括四个基本模块：模态编码器、基于图融合的有监督对比损失、互信息最大化和情绪分类器。不同模态的编码器首先捕获对应模态的上下文和说话者特征。然后我们最大化编码后的多模态输入之间的互信息。对于每个对话，我们为每个模态构造一个全连通图，在不同模态之间连接着对应于同一话语的节点。为了结合多模态上下文特征，基于图的融合网络以动态和顺序的方式堆叠。我们将经过网络后的特征与编码后的多模态输入进行拼接，得到最终的融合特征，接着最大化融合特征与编码后的多模态输入之间的互信息。最后，利用得到的融合特征和预测的情绪标签计算有监督对比损失和交叉熵损失。

3.1 模态编码器

使用相应的情态编码器，我们为每个情态创建上下文感知的话语特征编码。对于文本模态，我们使用双向长短期记忆网络(LSTM)对顺序文本上下文数据进行编码(Hochreiter et al., 1997)。对于声音或视觉模态，我们应用了一个全连接网络:

$$\mathbf{c}_i^t, \mathbf{h}_i^c = \overleftrightarrow{LSTM}_c\left(\mathbf{u}_i^t, \mathbf{h}_{i-1}^c\right) \tag{1}$$

$$\mathbf{c}_i^\varsigma = \mathbf{W}_c^\varsigma \mathbf{u}_i^\varsigma + \mathbf{b}_c^\varsigma, \varsigma \in \{a, v\} \tag{2}$$

考虑到说话者的个人信息也会影响对话，因此，我们还采用双向GRU(Chung et al., 2014)网络来捕获同一对话相邻话语之间的自依赖性。说话者嵌入可计算为:

$$\mathbf{s}_i^\delta, \mathbf{h}_{\lambda,j}^s = \overleftrightarrow{GRU}_s\left(\mathbf{u}_i^\delta, \mathbf{h}_{\lambda,j-1}^s\right), j \in [1, |U_\lambda|] \tag{3}$$

其中$\delta \in \{a, v, t\}$. $\mathbf{h}_{\lambda,j}^s$ 是第j个p_λ，$\lambda = \phi(u_i)$的隐藏状态。在一段对话中，所有的p_λ都被表述为U_λ。

3.2 基于图融合的有监督对比损失

在前面工作(Hu et al., 2021)的基础上，我们构建了一个无向图来模拟对话，表示为$\mathcal{G} = (\mathcal{V}, \mathcal{E})$，其中$\mathcal{V}(|\mathcal{V}| = 3N)$表示三种形式的话语节点，$\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$是连接的集合。节点通过上下文和说话者嵌入进行初始化:

$$\mathbf{x}_i^\delta = \mathbf{c}_i^\delta + \gamma^\delta \mathbf{s}_i^\delta, \delta \in \{a, v, t\} \tag{4}$$

其中$\gamma^a, \gamma^v, \gamma^t$ 是权衡超参数。边的权值$\mathcal{A}_{ij}$被计算为$\mathcal{A}_{ij} = 1 - \frac{\arccos(\text{sim}(x_i, x_j))}{\pi}$。

第二十二届中国计算语言学大会论文集，第264页-第276页，哈尔滨，中国，2023年8月3日至5日。

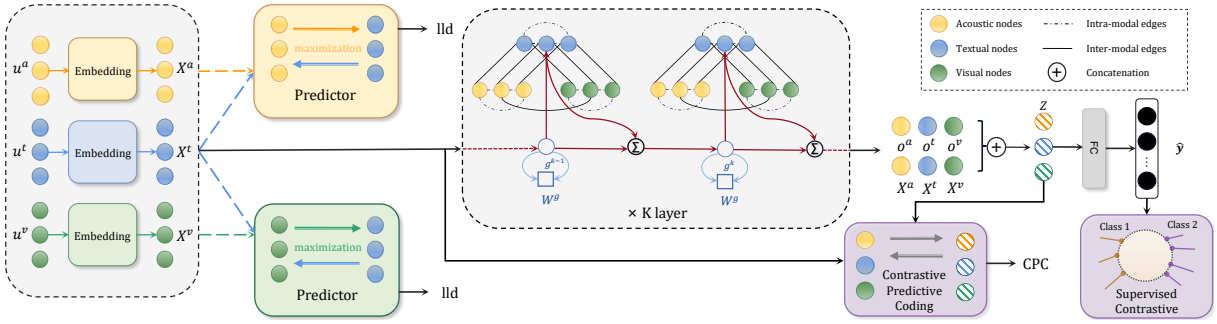

图 2. MMIC模型结构图

为了将模态间和模态内上下文信息在网络特定的语义区域内结合，我们在每一层使用图卷积操作。在(Hu et al., 2022)之后，门控方法用于动态融合会话中的多模态上下文信息:

$$\mathbf{g}^{(k)}, \mathbf{C}^{(k)} = \overrightarrow{LSTM}_c\left(\mathbf{g}^{(k-1)}, \mathbf{H}'^{(k-1)}\right) \tag{5}$$

经过改进的卷积运算(Chen et al., 2020)表示为:

$$\begin{aligned}\mathbf{H}^{(k)} = \mathrm{ReLU}(&\left((1-\alpha)\tilde{\mathbf{P}}\mathbf{H}'^{(k-1)} + \alpha\mathbf{H}^{(0)}\right) \\ &\left((1-\beta_{k-1})\,\mathbf{I}_n + \beta_{k-1}\mathbf{W}^{(k-1)}\right))\end{aligned} \tag{6}$$

其中使用了正则化的图卷积矩阵是$\tilde{\mathbf{P}} = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}$。$\alpha$是一个超参数。$\rho$，$\beta_k$也是两个超参数。单位映射矩阵记为$\mathbf{I}_n$，$\mathbf{H}'^{(k)} = \mathbf{H}^{(k)} + \mathbf{g}^{(k)}$表示第$k$层的输出。

受到(Khosla et al., 2021)的启发，我们使用监督对比损失改进基于图的融合模块。有监督对比学习将批处理中所有具有相同标签的样本都视为正样本。在多模态对话情绪识别中，来自不同模态的特征在映射到共享嵌入空间之前被编码和融合。然而，不同模式之间的信息冗余会增加识别相似情绪的难度。通过有监督的对比学习，利用标记信息，可以增强模型对相似情绪类别的辨别能力，从而提高模型情绪识别的性能。由于数据集具有高度不平衡的特点，在批处理中有单个样本的情况下，损失计算可能会被掩盖(Poria et al., 2020)。为了解决ERC数据集高度不平衡的问题，我们创建话语隐藏状态$H_{d-\mathrm{win}}$的副本，并分离其梯度以确保稳定的参数优化。下式可表示批处理中所有样本的总监督对比损失:

$$X = \left[H_{d-\text{ win}}, \bar{H}_{d-\mathrm{win}}\right] \tag{7}$$

$$\mathrm{SIM}(p,i) = \log \frac{\exp\left(\left(X_i \cdot X_p\right)/\tau\right)}{\sum_{a\in A(i)} \exp\left(X_i \cdot X_a/\tau\right)} \tag{8}$$

$$\mathcal{L}_{\mathrm{SCL}} = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \mathrm{SIM}(p,i) \tag{9}$$

其中$P(i) = I_{j=i} - \{i\}$表示与i类别相同但不包括自身的样本，$A(i) = I - \{i, N+i\}$表示多视图批处理中除自身之外的样本。$X \in \mathbb{R}^{2N\times d}, i \in I = \{1, 2, \cdots, 2N\}$表示多视图批处理中样本的索引。$\tau \in R^+$表示用于控制实例之间距离的温度系数。

3.3 互信息最大化

在输入级和融合级上进行分级互信息最大化，有利于模型捕捉不同层次模态之间的依赖关系，保留任务相关信息，生成丰富的多模态表示(Barber et al., 2004; Han et al., 2021)。因此，我们采用了分级MI最大化框架(Han et al., 2021)。

输入级互信息最大化。对于最大化多模态输入对之间的MI，我们采用了(Barber et al., 2004)来计算精确的MI下界，用$q(y \mid x)$来近似$p(y \mid x)$。计算公式为：

$$\begin{aligned} I(X;Y) =& \mathbb{E}_{p(x,y)}\left[\log \frac{q(y \mid x)}{p(y)}\right] + \\ & \mathbb{E}_{p(y)}[KL(p(y \mid x) \| q(y \mid x))] \\ & \geq \mathbb{E}_{p(x,y)}[\log q(y \mid x)] + H(Y) \\ & \triangleq I_{BA} \end{aligned} \tag{10}$$

考虑到文本模态提供了更多的信息(Arandjelovic et al., 2017)，我们优化了两种模态对(文本，声学)和(文本，视觉)的边界。$H(Y)$为Y的熵，计算采用高斯混合模型(Nilsson et al., 2002)。假设$q_\theta(y \mid x) = \mathcal{N}\left(\boldsymbol{y} \mid \boldsymbol{\mu}_{\theta_1}(\boldsymbol{x}), \boldsymbol{\sigma}^2_{\theta_2}(\boldsymbol{x})\boldsymbol{I}\right)$(Cheng et al., 2020)。$x$决定了公式中的两个高斯分布参数$\boldsymbol{\mu}$和$\boldsymbol{\sigma}$，这些参数通常由回归多层感知器(MLP)来计算。似然最大化损失函数为：

$$\mathcal{L}_{lld} = -\frac{1}{N}\sum_{tv,ta}\sum_{i=1}^{N}\log q\left(y_i \mid x_i\right) \tag{11}$$

其中N是训练批大小。tv，ta是两个预测因子的概率之和。然后采用无参数估计$H(Y)$对不同类别(负和非负)的样本分别进行高斯混合模型建模。由$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c}\sum_{i=1}^{N_c} h_c^i$，$\hat{\boldsymbol{\Sigma}}_c = \frac{1}{N_c}\sum_{i=1}^{N_c} h_c^i \odot h_c^i - \hat{\boldsymbol{\mu}}_c^T\hat{\boldsymbol{\mu}}_c$，$H(Y)$ 被计算为(Huber et al., 2008)：

$$H(Y) = \frac{1}{4}\left[\log\left(\left(\det\left(\boldsymbol{\Sigma}_1\right)\det\left(\boldsymbol{\Sigma}_2\right)\right)\right] \tag{12}$$

其中$\boldsymbol{\mu}$是平均向量，$\boldsymbol{\Sigma}$是协方差矩阵，c表示负或非负。该层MI下界最大化的损失函数如下：

$$\mathcal{L}_{BA} = -I_{BA}^{t,v} - I_{BA}^{t,a} \tag{13}$$

融合级互信息最大化。对于融合结果与输入模态之间的MI最大化，优化的目标是融合网络生成的融合结果$Z = F\left(\mathbf{X}^t, \mathbf{X}^v, \mathbf{X}^a\right)$。我们使用一个评分函数来评估归一化预测和真实向量之间的相关性：

$$\overline{G_\phi(Z)} = \frac{G_\phi(Z)}{\|G_\phi(Z)\|_2}, \quad \overline{h_m} = \frac{h_m}{\|h_m\|_2} \tag{14}$$

其中G_ϕ是一个参数为ϕ的神经网络，Z表示融合特征，h_m表示模态m。$\|\cdot\|_2$是欧几里得范数，我们除以它以得到单位长度向量。CPC分数表示为$s\left(h_m, Z\right)$，得分越高，该模态在融合模态中越显著。$\mathbb{E}_{\mathbf{H}}$为噪声对比估计框架，正样例表示存在融合关系的单模态表示和融合表示形成的对，负样例则是同一个批次内其他样本的单模态表示和融合表示形成的对。该层的损失函数为：

$$s\left(h_m, Z\right) = \exp\left(\overline{h_m}\left(\overline{G_\phi(Z)}\right)^T\right) \tag{15}$$

$$\mathcal{L}_{\mathbf{N}}\left(Z, \mathbf{H}_m\right) = -\mathbb{E}_{\mathbf{H}}\left[\log \frac{s\left(Z, h_m^i\right)}{\sum_{h_m^j \in \mathbf{H}_m} s\left(Z, h_m^j\right)}\right] \tag{16}$$

$$\mathcal{L}_{CPC} = \mathcal{L}_{\mathrm{N}}^{z,v} + \mathcal{L}_{\mathrm{N}}^{z,a} + \mathcal{L}_{\mathrm{N}}^{z,t} \tag{17}$$

3.4 情绪分类器

在经过网络K层的融合之后，每个语句i的三种形式的表示可以进一步细化为$\mathbf{o}_i^a; \mathbf{o}_i^v; \mathbf{o}_i^t$。最后，使用分类器来预测每个话语的情绪：

$$\hat{\mathbf{y}}_i = \mathrm{Softmax}\left(\mathbf{W}_z\left[\mathbf{x}_i^a; \mathbf{x}_i^v; \mathbf{x}_i^t; \mathbf{o}_i^a; \mathbf{o}_i^v; \mathbf{o}_i^t\right] + \mathbf{b}_z\right) \tag{18}$$

第二十二届中国计算语言学大会论文集，第264页-第276页，哈尔滨，中国，2023年8月3日至5日。

其中W_z和b_z是可训练的参数。为了得到任务损失，我们使用带有L2正则化的交叉熵损失:

$$\mathcal{L}_{\text{task}} = -\frac{1}{\sum_{l=1}^{L}\tau(l)}\sum_{i=1}^{L}\sum_{j=1}^{\tau(i)}\mathbf{y}_{i,j}^{l}\log\left(\hat{\mathbf{y}}_{i,j}^{l}\right)+\eta\|\Theta\|_2 \tag{19}$$

其中$c(i)$为对话中的话语数，N为对话数。$i, P_{i,j}$是对话中话语j预测情绪标签的概率分布，$i, y_{i,j}$是期望类标签。θ为l2正则化权值，三个变量都是可训练的参数。最终我们得到总加权损失来训练模型:

$$\mathcal{L}_{\text{main}} = \mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{SCL} + \beta\mathcal{L}_{CPC} + \zeta\mathcal{L}_{BA} \tag{20}$$

其中α代表监督对比损失的参数。超参数β和ζ是调节MI最大化的参数。

4 实验与分析

4.1 数据集

随着社交媒体的快速发展，交互数据越来越多，其中包括一些公开的对话数据集。我们在这次实验中选用的是两个英文的公共基准数据集和一个新发布的中文数据集，即IEMOCAP，MELD和M^{3}ED。对于IEMOCAP，MELD，我们选用了和研究(Hu et al., 2021)一致的话语级特征。对于M^{3}ED，我们使用从(Zhao et al., 2022)提供的帧级特征中获得的话语级特征。三个数据集的详细数据分布如表1所示。

	IEMOCAP			MELD			M3ED		
	train	val	test	train	val	test	train	val	test
说话者数量	8		2	307		100	421	87	118
对话数量	120		31	1153		280	685	126	179
话语数量	5810		1623	11098		2610	17427	2821	4201
平均话语长度	48.4		52.4	9.65		9.3	25.44	22.39	23.47

表 1. 实验数据集的详细信息

IEMOCAP(Busso et al., 2008)包含十个不同的说话人对之间的对谈视频。它包括7433个话语和151个对话。对话中的每句话都被标注了六个类别的情绪标签中的一个，标签包括快乐、悲伤、中性、愤怒、兴奋和沮丧。我们遵循之前的研究(Majumder et al., 2019)，使用前四次训练，使用最后一次测试，并随机抽取10个训练对话作为验证分割。

MELD(Poria et al., 2019)包含多方对话，这些对话都是从《老友记》系列电视剧中收集的视频得来的，每轮对话都有两个或两个以上的说话者参与。它包含1433个对话，13708个话语和304个不同的说话者。对话中的每句话都带有七种情绪标签中的一个，标签包括愤怒，厌恶，恐惧，喜悦，中性，悲伤和惊讶。为公平的比较，我们遵循(Hu et al., 2021)的设置进行实验。

M^{3}ED(Zhao et al., 2022)包含来自56部不同电视剧的990个二元情绪对话，是中文首个大规模多模态情绪对话数据集。它包含990个对话，24449个话语和626个不同的说话者。对话中的每句话都被标注了七个类别的情绪标签中的一个，标签包括中性、高兴、惊讶、悲伤、厌恶、愤怒和恐惧。我们使用M^{3}ED中预定义的训练/验证/测试分割进行实验。

4.2 实验参数设置

在实验中，我们使用基于随机梯度下降的Adam(Kingma et al., 2017)优化器来训练我们的模型。为了避免过拟合，Dropout(Srivastav et al., 2014)被设置为0.1到0.5进行验证。具体超参数设置如下：三个数据集的GCN层数均为16层，批大小（Batchsize）设置为16，α in $\{0.8, 1.0\}$, β, ζ in $\{0.01, 0.05, 0.10\}$。特别地，由于IEMOCAP数据集中随机噪声较少，我们将ξ设为0。以下结果中的数据均为10次独立实验的平均值。所有实验均在GeForce RTX A5000 GPU上进行。

Model	IEMOCAP							MELD
	Happy	Sad	Neutral	Angry	Excited	Frustrated	Weight-Avg-F1	Weight-Avg-F1
BC-LSTM	33.82	78.76	56.75	64.35	60.25	60.75	60.42	57.29
MFN	48.19	73.41	56.28	63.04	64.11	61.82	61.60	57.80
ICON	32.80	74.40	60.60	68.20	68.40	66.20	63.50	-
DialogueRNN	32.20	80.26	57.89	62.82	**73.87**	59.76	62.89	57.11
DialogueGCN	47.10	80.88	58.71	66.08	70.97	61.21	65.04	57.34
MMGCN	43.44	77.80	60.03	67.25	75.20	61.79	65.20	57.87
MM-DFN	42.06	79.66	65.37	68.11	74.56	67.16	67.86	58.07
COGMEN	**54.89**	79.13	64.81	60.61	74.17	57.71	65.71	-
MMIC(Ours)	44.44	**82.83***	**66.27***	**69.72***	73.38	**67.27***	**68.74***	**58.57***

表 2. 在IEMOCAP和MELD数据集上的性能对比

Model	M^3ED
	Weight-Avg-F1
DialogueRNN	51.57
DialogueGCN	45.90
MMGCN	49.28
MM-DFN	53.27
COGMEN	52.06
MMIC(Ours)	**53.65***

表 3. 在M^3ED数据集上的性能对比

4.3 实验对比模型

为了评估我们提出的MMIC模型，我们与以下模型进行比较:

- **BC-LSTM** (Poria et al., 2017):该模型在话语级利用LSTM网络捕获多模态特征。
- **MFN** (Zadeh et al., 2022): 该模型采用多视图门控记忆单元同步多模态序列。
- **ICON** (Hazarika et al., 2018): 该模型通过多层跳跃存储器提供从模态到会话的特性。
- **DialogueRNN** (Majumder et al., 2019): 该模型引入循环网络来跟踪说话人的状态和对话中的上下文。
- **DialogueGCN** (Ghosal et al., 2019): 该模型利用图形结构来组合上下文依赖关系。
- **MMGCN** (Hu et al., 2021): 该模型使用基于图形的融合模块来捕获模态内和模态间的上下文特征。
- **MM-DFN** (Hu et al., 2022): 该模型用动态图形融合模块来融合对话中多模态的上下文特性。
- **COGMEN** (Joshi et al., 2022): 该模型使用基于图的架构在对话中建模复杂的关系(局部和全局信息)。

4.4 主实验

与已有的工作(Poria et al., 2017; Majumder et al., 2019; Joshi et al., 2022)一样，我们在多模态对话情绪识别的实验中使用加权平均F1分数作为评价指标。表2和表3比较了在多模态的设置下MMIC与其他模型在三个数据集上的性能。标有*的结果表示在配对t检验下，MMIC与最先进的分数相比有统计学意义($p < 0.05$)。由于一些模型的再现结果与原论文报告的结果存在差异，为了确保结果的可靠性和准确性，我们选择使用再现结果。

数据集	IEMOCAP	MELD	M^3ED
评估方法		Weight-Avg-F1	
MMIC	68.74	58.57	53.65
-MI	68.48(↓0.26)	58.37(↓0.20)	53.42(↓0.23)
-Contrast	68.34(↓0.40)	58.36(↓0.21)	53.46(↓0.19)
-MI, Contrast	67.86(↓0.88)	58.07(↓0.50)	53.27(↓0.38)

表 4. MMIC的消融实验结果

数据集	测试集样例	w/o MI,Contrast	w/o MI	w/o Contrast	MMIC	真实标签
M^3ED	我就生气怎么了!	Sad	Angcr	Sad	Angcr	Angcr
M^3ED	顺便躲避一下小蚯蚓的唉声叹气	Neutral	Neutral	Sad	Sad	Sad
MELD	I will!	Surprise	Anger	Joy	Joy	Joy
MELD	Yes!! Yes! Yes! Yes!! That's my Dad, that's Frank! Yeah!I'm sorry I'm getting all flingy.	Joy	Surprise	Joy	Surprise	Surprise

表 5. 案例探究

根据表2和表3可知，MMIC在三个数据集上的F1分数表现，超过了所有的基准模型和目前最先进的方法。具体来说，MMIC在IEMOCAP上的平均提高为0.88%，MELD上的平均提高为0.50%，M^3ED上的平均提高为0.38%。IEMOCAP数据集上，MMIC在7个评价指标中有5个指标均优于其他方法，并且在其他指标上也同样产生了具有竞争力的结果。实验结果表明了MMIC是有效的多模态对话情绪识别模型，在中英文和各种情境下均能获得优于其他所有模型的指标分数，取得优秀的性能表现。

4.5 进一步研究

通过比较以上结果可以看出，MMIC模型取得了良好的性能。为了进一步探究模型性能提升的原因，我们首先进行了消融实验，以分析本文模型建模的不同类型的关系对模型整体性能带来的影响。然后，我们在所提出模型上进行了案例探究和可视化操作，以证实该模型可以有效保留任务相关信息和利用标签信息，更好地区分出相似取消，提高在现实对话场景中识别不同情绪的准确度。

4.5.1 消融实验

表4显示了消融实验的结果，移除了所提出模型的关键模块。当最大化互信息模块或监督对比损失模块被移除时，数据集上的结果显著下降。当两个模块都被移除时，结果进一步下降。该实验证明了两个模块在三个数据集上的有效性，以及建模的不同关系对于模型的整体性能都是有价值的。详细分析如下：

- **消去最大化互信息**：即从模型中移除最大化互信息模块。从结果中我们可以看到，在IEMOCAP数据集中F1分数下降了0.26%，在MELD数据集中F1分数下降了0.20%，在M^3ED数据集中F1分数下降了0.23%。这表明，利用互信息最大化，分级对模态输入对以及多模态融合结果与单模态融合结果之间的互信息进行最大化，可以更好地过滤噪声，避免任务相关信息在融合过程中的丢失，提高模型在多模态对话情绪识别中的性能。

- **消去有监督对比损失**：即从模型中移除监督对比损失模块。从结果中我们可以看到，在IEMOCAP数据集中F1分数下降了0.40%，在MELD数据集中F1分数下降了0.21%，在M^3ED数据集中F1分数下降了0.19%。这表明，将监督对比学习纳入基于图的融合网络，有利于充分利用标签信息，增强模型对相似情绪的识别能力，提高情绪识别的准确度。

第二十二届中国计算语言学大会论文集，第264页-第276页，哈尔滨，中国，2023年8月3日至5日。

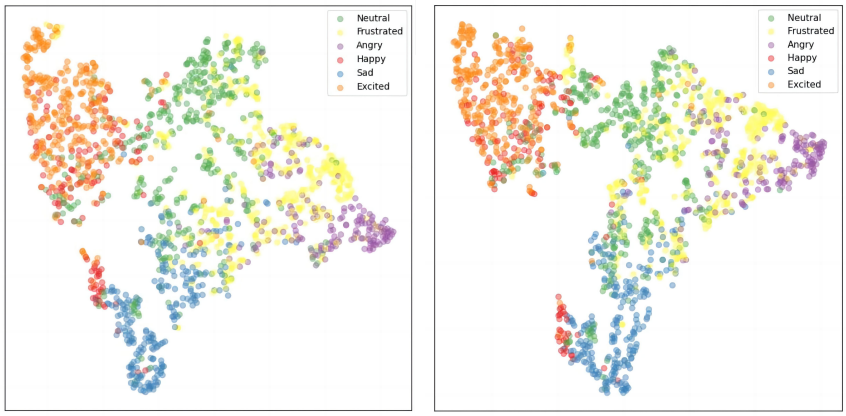

图 3. t-SNE可视化对比图

4.5.2 案例探究

我们在MELD和M^{3}ED数据集上进行了案例探究，如表5所示。可以看见，当去除了两个关键模块后，模型容易混淆两个相似的情绪，从而对说话者情绪进行错误的判别。当去除互信息模块后，模型可以准确识别案例2、3中的相似情绪，但是对案例1和4则进行了错误的识别。我们考虑这可能是由于案例1和4本身存在的信息较少或存在单独语义歧义，需要更多的上下文信息来识别情绪，而缺少了互信息模块，就导致融合信息中的任务相关信息受损，从而使模型进行了错误的判别。当去除对比损失模块后，模型可以准确识别案例1和4中的相似情绪，但是对案例2和3则进行了错误的识别。我们考虑这可能是由于案例2和3本身的信息较多或是存在两类不同的情感词语，导致语义相似度较高，使得不同模式之间的信息冗余。缺少了对比损失模块，模型无法利用标签信息来拉开相似情绪样本在表示空间的距离，从而进行错误的判别。而当两个关键模块都没有被移除时，模型可以准确判别出四个案例的情绪，这进一步证明了两个关键模块可以提高模型在多模态设置下各类情景中对话情绪识别上的性能。

4.5.3 t-SNE可视化

为了达到识别出不同相似情绪的目的，要尽量使不同类别的数据在表示空间中分开，而相同类别的数据在表示空间中靠近。这样可以有效地提高不同类别之间的耦合程度，增加相同类别之间的内聚性。耦合程度是指模块或类之间的关联和依赖的程度，内聚性是指模块或类内部元素之间的相关性。如图3所示，我们将模型的输出可视化在IEMOCAP数据集上，左边是我们模型在没有互信息最大化和监督对比损失的情况下的t-SNE可视化结果，右边是我们完整模型的结果。通过对比，我们发现我们所提出的模型使不同类别的数据在表示空间中更分离，相同类别的数据在表示空间中则更为靠近，即是模型可以更有效地提高不同类别之间的耦合程度，增加相同类别之间的内聚性，这证明了模型可以更好地区分相似的情绪。

5 总结

针对多模态对话情绪识别研究现有的问题，本文提出了一种基于互信息和对比损失的图融合模型——MMIC。为了更好地保存从输入到融合结果的任务相关信息，我们首先最大化了单模态表示之间的互信息。在通过图卷积运算得到融合结果后，我们进一步最大化它与低水平单模态表示之间的互信息。为了更充分地利用标签中隐含的信息，我们将监督对比学习纳入基于图的融合网络，使具有相同情绪的样本具有内聚性，不同情绪的样本相互排斥。我们在两个公共基准数据集和一个中文数据集上进行了大量实验，证明了所提出模型的有效性和优越性。所提出的模型经过案例探究，证实了它能有效保留任务相关信息，更好地区分相似情绪。消融实验和可视化结果证明了模型中每个模块的有效性。

致谢

本文受到广东省自然科学基金(2021A1515011864)、国家自然科学基金(71472068)、广州市智慧农业重点实验室(201902010081)、广东省普通高校特色创新项目(2020KTSCX016)、华南农业大学大学生创新训练计划项目(X202210564157)的资助。

参考文献

A. Zadeh, R. Zellers, E. Pincus, and L.P.e Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pp. 1128–1137.

X. Zhou and W. Y. Wang. 2018. MojiTalk: Generating Emotional Responses at Scale. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018*, pp. 1128–1137.

D. Hu, L. Wei, and X. Huai. 2019. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pp. 527–536.

N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. *Proceedings of the 33rd Association for the Advancement of Artificial Intelligence, AAAI 2018*, pp. 6818–6825.

D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. *Proceedings of the 9th Int. Joint Conf. Natural Lang. Process, IJCNLP 2019*, pp. 154–164.

S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.P. Morency. 2017. Context-dependent sentiment analysis in user-generated videos. *Proceedings of the 55th Annu. Meeting Assoc. Comput. Linguist. ACL 2017*, pp. 873–883.

D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. *Proceedings of Conf. North Amer. Chapter Assoc. Comput. Linguist.: Human Lang. Technol, NAACL-HLT 2018*, pp. 2122–2132.

D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann. 2018. Icon: Interactive conversational memory network for multimodal emotion detection. *Proceedings of 2018 Conf. Empirical Methods Natural Lang. Process, EMNLP 2018*, pp. 2594–2604.

J. Hu, Y. Liu, J. Zhao, and Q. Jin. 2021. MMGCN: multimodal fusion via deep graph convolution network for emotion recognition in conversation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics, ACL 2021*, pp. 5666–5675

J. Liu, S. Chen, L. Wang, and Z. Liu. 2021. Multimodal emotion recognition with capsule graph convolutional based representation fusion. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Virtual and Singapore*, pp. 6339–6343.

D. Hu, X. Hou, L. Wei, L. Jiang and Y. Mo. 2022. MM-DFN: multimodal dynamic fusion network for emotion recognition in conversations. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore*, pp. 7037–7041.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Comput*, vol. 9, no. 8, pp. 1735–1780

J. Chung , C. Gulcehre, K.H. Cho and Y. Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *t arXiv:1412.3555*,

M. Chen, Z. Wei, Z. Huang, B. Ding, and Y. Li. 2020. Simple and deep graph convolutional networks. *Proceedings the 36th International Conference on Machine Learning, ICML 2020*, pp. 1725–1735.

D. Barber and F. Agakov. 2004. The IM algorithm: A variational approach to information maximization. *Proceedings of Advances in neural information processing systems*, 16:201.

W. Han, H. Chen, and S. Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. *arXiv:2109.00412.*

R. Arandjelovic and A. Zisserman. 2017. Look, listen and learn. *Proceedings of IEEE International Conference on Computer Vision , ICCV 2017*, pp. 609–617.

M. Nilsson, H. Gustaftson, S.V. Andersen, and W. B. Kleijn. 2002. Gaussian mixture model based mutual information estimation between frequency bands in speech. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. I–525.

M.F. Huber, T. Bailey, H. Durrant-Whyte, and U.D. Hanebeck. 2008. On entropy approximation for gaussian mixture random vectors. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 181–188.

P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin. 2020. Club:A contrastive log-ratio upper bound of mutual information. *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, pp. 1779–1788.

P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. 2021. Supervised contrastive learning. *arXiv:2004.11362.*

S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea. 2023. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 108–132.

C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee and S.S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, vol. 42, no. 4, pp. 335–359.

S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL 2019*, pp. 527–536.

C.J. Zhao, T. Zhang, J. Hu, Y. Liu, Q. Jin, X. Wang, and H. Li. 2022. M3ED: Multi-modal Multi-scene Multi-label Emotional Dialogue Database. *arXiv:2205.1023.*

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2020. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:2006.11477.*

A. Baevski, H. Zhou, A. Mohamed, and M. Auli. 2019. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv:1907.11692v1.*

G. Huang, Z. Liu, L.V.D. Maaten, and K.Q. Weinberger. 2017. Densely connected convolutional networks. *Proceedings of IEEE conference on computer vision and pattern recognition, CVPR 2017*, pp. 4700–4708.

A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.P. Morency. 2022. Memory fusion network for multi-view sequential learning. *arXiv:2205.1023.*

A. Joshi, A. Bhat, A. Jain, A.V. Singh and A. Modi. 2022. COGMEN: COntextualized GNN based Multimodal Emotion recognitioN. *Proceedings of the 20th Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL 2022*, pp. 4148–4164.

D.P. Kingma, and J.L. Ba. 2017. ADAM: A method for stochastic optimization. *arXiv:2205.1023.*

Shimin Li, Hang Yan, Xipeng Qiu. 2022. Contrast and Generation Make BART a Good Dialogue Emotion Recognizer. *arXiv:2112.11202v2.*

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y Ng. 2011. Multimodal deep learning. *Proceedings the 17th on Machine Learning, ICML 2011.* pp. 689-696

M. Chen, S. Wang, P.P. Liang, T. Baltrusaitis, A. Zadeh, and . Morency. 2017. Multimodal sentiment analysis with wordlevel fusion and reinforcement learning. *Proceedings of the 19th ACM International Conference on Multimodal Interaction.* pp, 163–171.

A. Zadeh, P.P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.P. Morency. 2018. Memory fusion network for multiview sequential learning. *Proceedings of the 32nd Association for the Advancement of Artificial Intelligence, AAAI 2018* vol 32.

B. Poole, S. Ozair, A.V.D. Oord, A. Alemi, and G. Tucker. 2019. On variational bounds of mutual information. *Proceedings of the 36th International Conference on Machine Learning, ICML 2019* pp. 5171–5180.

A.A. Alemi, I. Fischer, J.V. Dillon, and K. Murphy. 2016. On variational bounds of mutual information. *arXiv preprint arXiv:1612.00410.*

P. Bachman, R. Hjelm, and W. Buchwalter. 2019. Learning representations by maximizing mutual information across views. *arXiv preprint arXiv:1906.00910.*

K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* pp. 9729–9738.

R.A. Amjad and B.C. Geiger. 2020. Learning representations for neural network-based classification using the information bottleneck principle. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 9, pp. 2225-2239.

G.E. Hinton and R. R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. science. *Science.* vol. 313, issue 5786, pp. 504-507.

N. Srivastava, G.E. Hinton, A. Krizhevsky . 2014. Dropout:A simple way to prevent neural networks from overfitting. *Mach.Learn.Res.* vol. 15, no. 1, pp. 1929-1958.