用于识别针对社交机器人的主观自我表露的多模态神经网络

Henry Powell [1,2*] , Guy Laban [3,4,2*] , Emily S. Cross [2,5]
[1] Amazon, Edinburgh, UK.
[2] School of Psychology and Neuroscience, University of Glasgow, Glasgow, UK.
[3] Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer Sheva, Israel.
[4] Department of Computer Science and Technology, University of Cambridge, Cambridge, UK.
[5] Social Brain Sciences Group, Department of Humanities, Social and Political Sciences, ETH Zurich, Zurich, Switzerland.

***Abstract*— 主观的自我表露是人类社会互动的重要特征。尽管在社会和行为学文献中已经做了大量工作来描述主观自我表露的特征及其后果，但是在开发能够准确建模的计算系统方面，至今为止所做的工作却很少。尤其是关于人类与机器人伙伴互动时如何具体实现自我表露的建模研究更少。随着我们需要社会机器人在各种社会环境中与人类协同工作并建立关系，这一问题变得愈发紧迫。在本文中，我们的目标是开发一个定制的多模态注意力网络，基于情感识别文献中的模型，培训这个模型使用大型自我收集的自我表露视频语料库，并构建一个新的损失函数——比例保持交叉熵损失，提升这一问题的分类和回归版本。我们的结果表明，使用我们新颖的损失函数训练的表现最佳的模型，其 F1 得分达到 0.83，相比最好的基线模型提高了 0.48。此结果在实现社会机器人能够识别互动伙伴的自我表露这一目标上取得了显著进展，这种能力对于具有社会认知的社会机器人将是至关重要的。**

I. 介绍

自我揭露是在社交互动中分享自己的想法、感受或个人信息 [1] 。这是人类社交行为的重要方面，可以影响我们生活的许多方面。它可以影响我们形成彼此关系的程度——即我们认为与他人的关系有多么亲密和重要——同时对我们的心理和身体健康有显著贡献 [1]–[3] 。在本文中，我们特别关注主观的自我揭露，它强调个人在多大程度上认为自己正在与他人分享个人信息。例如，可能有人分享一些信息，这些信息一般来说可能不被认为特别个人或敏感。然而，这些信息可能对披露的人来说是有意义的。这里“主观”一词旨在澄清，在自我揭露行为中重要的是执行者认为自己正在分享有意义的想法、感受或个人信息。

考虑到主观自我披露在发展有意义的人际关系中的重要性，我们认为对这些自我披露的敏感性成为社交机器人在与人建立关系时的关键属性。社交机器人在社会环境中的研究越来越多，关注它们的情感角色。先前的研究强调了个人如何随着时间的推移与这些代理形成社会关系，为什么用户向其敞开心扉，以及它们可以支持人类用户的多种方式。因此，这些被设计为在人类中运作并与人类沟通的机器人代理必须具备适当处理此类披露的能力。然而，尽管具备潜力，社交机器人目前缺乏细微的认知能力，无法像人类一样轻松地推断这种社会信息并理解主观感知，这对于这些机器人承担社会和情感角色时至关重要。此外，在人机交互领域几乎没有研究试图建模检测和测量自我披露的能力，以期赋予一个具备社会导向的人工代理这种能力。我们的一个例外研究发现，一些标准的深度学习架构在对记录的互动中用户主观自我披露程度进行排序的任务上表现远高于平均水平。

在本研究中，我们旨在解决这个问题，并改进以前的结果。我们通过多种方式实现了这一目标。首先，开发了一个显著更大的数据集，该数据集不仅包括音频模式，还包括视觉模式，以捕捉面部特征随时间变化可能存在的主观自我披露标记。其次，开发了一种更为复杂的深度学习模型，该模型更适合于任务，即使用问题的领域知识和我们作为输入使用的数据表示来开发模型。第三，解决我们在该研究中遇到的实验框架问题，即如何建模既是分类又是比例的数据。

本文余下部分的结构如下：在第 II 节中，我们详细介绍了为形成用于执行深度学习实验的数据集而进行的实验的设计、数据收集和数据预处理。接下来，在第 III 节中，我们概述了我们从处理后的数据集中提取出的特征以及提取它们的方法。在第 IV 节中，我们详细描述了多模态注意网络的架构。然后，我们描述了用于生成基线的实验，以便与该模型的性能进行比较。此外，我们详细说明了消融实验的参数，以测试我们使用的损失函数、特征集和实验框架的效果，最后是训练实施的具体细节。接着，在第 V 节中，我们展示了消融研究的结果，最后在第 VI 节中，讨论了我们方法需要进一步改进的领域以及存在的一些问题。

A. 我们的贡献

我们对该领域的贡献如下：

1) 我们提出了到目前为止在人机交互中对主观自我披露进行建模的最广泛尝试，
2) 一种专门为从音频和视频数据进行自我披露建模而设计的多模态注意力架构
3) 一种新颖的损失函数，称为尺度保持交叉熵损失，能够有效处理介于回归和分类之间的问题，并在自我披露建模中优于平方误差和交叉熵方法。

*Henry Powell and Guy Laban have contributed equally to this work and share first authorship. Corresponding author: guy.laban@cl.cam.ac.uk

The authors gratefully acknowledge funding from the European Research Council (ERC) under the European Union’ s Horizon 2020 research and innovation programme (Grant agreement 677270 to E.S.C.), the Leverhulme Trust (PLP-2018-152 to E.S.C.), and the European Union’ s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie to ENTWINE, the European Training Network on Informal Care (grant agreement no. 814072 to E.S.C.). Our thanks also go to Hubert Ramsauer for helpful discussions on the use of Hopfield networks.

II. 数据集和数据收集

为了给模型生成数据，按照 [4] 中的报道，进行了一项长期的介导线上实验。为了保持一致性，我们在此逐字重复该协议：采用 2（讨论主题：COVID-19 相关 vs. 普通）x 10（跨时间的聊天会话）的组间重复测量的实验设计。39 名参与者被随机分配到两个讨论主题组之一，根据分配，他们通过 Zoom 视频聊天与机器人 Pepper（软银机器人）就一般日常话题（如社会关系、工作生活平衡、健康和福祉）进行交流。其中一组的对话主题在 COVID-19 疫情的背景下进行（例如，疫情期间的社会关系），而另一组的对话主题相似，但从未明确提及 COVID-19 疫情。参与者被安排在预定时间与机器人进行为期五周的两次互动，总共进行 10 次互动。每次互动由机器人询问参与者 3 个问题（每个问题重复 3 次）组成，首先是建立融洽关系的常规问题（例如，你的周/周末过得怎么样），然后是两个与 10 个随机排序的主题之一相对应的附加问题（有关主题、问题和示例，请参见 [4] ）。每次互动的主题和问题顺序均为随机分配。通过 Zoom 聊天与 Pepper 交谈后，参与者填写了一份问卷，用以报告他们对其主观披露的感知，该问卷基于 Jourad 自我披露问卷的改编版本 [1] 。Zoom 聊天记录专用于分析目的。每次与机器人的互动持续 5 到 10 分钟，填写问卷则需要另外 10 到 20 分钟。该研究遵循严格的伦理标准，所有研究程序均获得英国格拉斯哥大学心理与神经科学学院研究伦理委员会的批准。

这导致了 $39 \times 10 = 390$ 个交互，每个交互至少包括三个对话段落，我们能够用来训练我们的模型。一旦数据集被收集，视频就会被手动分割以隔离仅包含参与者讲话的部分。大多数视频包含三个讲话段落，由参与者对 Pepper 问题的回答组成。然而，一些参与者对 Pepper 的回答进行了后续交流，从而产生了额外的讲话段落，我们能够将这些段落加入语料库中。然后，每个段落由实验者依据参与者对各自交互实例分配的自我揭露评分进行标记。这导致了用于我们深度学习实验的总计 1,248 个讲话和视频段落。

III. 特征提取

A. 视觉特征

我们使用先进的特征提取模型组合提取了多种视觉特征类型。首先，我们使用 OpenFace 2.2 库逐帧提取了注视和动作单元特征 [5] （视觉示例请见图 2 ）。为了处理由于 OpenFace 模型未能检测到人脸而导致的每个时间序列中的缺失帧，我们使用样条插值法将记录数据插值到缺失帧中。然后，我们使用一个 Savitsky-Golay 滤波器（采用 11 帧的滑动窗口和 3 阶多项式）对所得的多变量时间序列进行滤波和平滑处理。为了测试平滑和滤波对结果的影响，我们在初步实验中将平滑/滤波处理过的和未处理过的特征集分别对待。

接下来，我们使用预训练在 VGGFace2 数据集上的 InceptionV1 ResNet [6] [7] 架构提取面部特征。VGGFace2 由 331 万张名人面部图像组成，分为 9131 个主体类别，其姿势、年龄、光照和种族差异较大。我们使用的 InceptionV1 ResNet 在该数据集上获得了 99.6 % 的准确率。在这种情况下，视频帧的预处理包括提取每个帧的 160x160 像素子区域，该区域包含主体面部的像素和特征规范化（可以在图 1 中看到 MTCNN 输出的示例）。

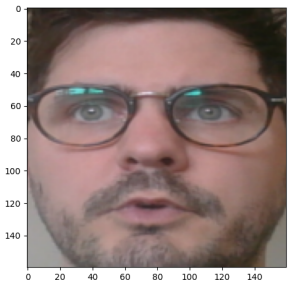

Fig. 1: MTCNN 用于面部特征提取的示例输出：一个 160x160 像素的标准化面部图像。

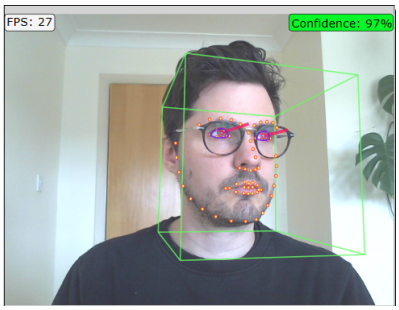

Fig. 2: OpenFace 2.2 处理面部动作单元和输入视频中的注视。

这是通过在每个视频帧上使用预训练的多任务级联卷积神经网络 [8] 来完成的。预训练的 ResNet 为每个视频帧生成 512 个面部特征。与我们使用 OpenFace 特征的方法类似，我们插值并过滤生成的时间序列，以实验这对模型得分的影响。

对于音频特征，我们首先为每个视频的音频模态生成了梅尔滤波倒谱系数（MFCC）矩阵。使用 PyTorch 的音频特征提取库和 256 个梅尔滤波器进行这项工作。选择这组特征的原因是因为 MFCC 被广泛认为能够有效捕捉人类语音识别任务中的重要音频特征。这是我们之前利用对数梅尔谱图识别主观自我披露工作的一个扩展，在该工作中我们发现谱图特征更能有效捕捉讲话者语音中与自我披露相关的重要特征。在本研究的情况下，我们发现 MFCC 特征在初步测试中产生了更好的结果，因此我们选择 MFCC 特征而不是对数梅尔谱图替代方案。我们还尝试了 MFCC 特征的倒谱均值和方差归一化对我们基线模型性能的影响（详见第 IV-A 章，因为这是在训练深度学习模型时也必须考虑的因素）。

其次，我们使用 Facebook AI 的 wav2vec2.0 架构直接从每个声音文件的振幅数组中提取音频特征。Wav2vec2.0 使用基于卷积神经网络的特征编码器堆栈，并通过 transformer 模型生成情境化的音频表示。我们使用了一个预训练在 LibriSpeech 数据集中 960 小时未标记音频数据上的 wav2vec2.0 模型。为了获取每个 wav 文件的特征集，我们提取了模型的 12 层 transformer 层的输出，结果是产生了 12 个 t x 768 的特征矩阵，其中值 t 取决于音频文件中的帧数。

IV. 深度学习实验

A. 支持向量机基线

由于我们处理的是专门为深度学习实验设计的新数据集，我们需要一种方法来建立一个基准，以便能够将我们

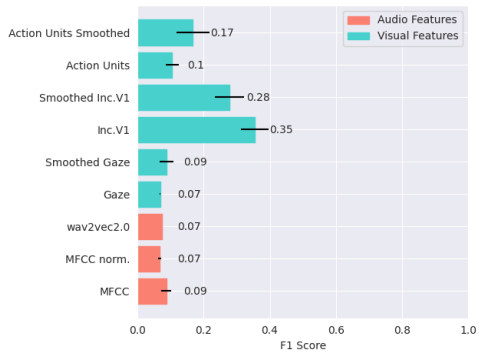

Fig. 3: 高斯 SVM 基线 F1 分数用于单独的平滑/过滤和未平滑/未过滤的音频和视觉特征集。标准差由黑色误差线表示。

的结果与之进行比较。根据 [9] ，我们分别使用训练在我们提取的音频和视觉特征上的高斯核支持向量机（SVM）来建立这样的基准。对于每种特征类型，计算一个表示每个样本中所有帧均值的向量，并让 SVM 负责对每个交互的自我披露分数进行分类。每个模型使用 3 折交叉验证进行训练，并使用平均 f1 得分作为衡量每个模型整体性能的标准。

这些基线实验的结果(见图 3)表明,使用在 VGGFace2 上预训练的 InceptionV1 提取的面部特征在任务中信息量最大，而对于音频特征，MFCC 表示的效果最好。整体来看，视频特征是在区分自我披露评分类别时最有用的特征集。结果还显示，由于测得的最佳 f1 分数仅为 0.36，因此此问题是一个困难的问题。一个令人惊讶的结果是，word2vec2.0 特征表现得如此糟糕。我们假设，由于 InceptionV1 特征的相对强劲表现，word2vec2.0 特征在大型相关任务数据上预训练，应该也会表现得相对较好。word2vec2.0 特征在基线测试中表现不佳的一个可能解释是与每帧均值向量的计算方式有关。每帧的 word2vec2.0 特征的维度远高于 MFCC 特征($12\times t^W \times 178$ vs. $256\times t^M$) 及最高维度的视觉特征集 (InceptionV1 特征为 $t^I \times 512$) [1] 。因此，将 word2vec2.0 特征在时间和注意力头维度上压缩为单一 178 维向量可能意味着丢失了过多信息，导致特征在任务中的辨别能力显著下降。

B. 多模态注意力网络

我们设计了一个多模态注意力网络，该网络在独立的流中处理每个视频的音频和视觉特征，然后在通过线性神经网络层分类之前以晚期融合的方式结合这些表示。这种方法的动机来自于 [10] 中的先前观察得出的结论：“现成的”神经网络架构，即那些没有针对手头任务专门设计且没有使用预训练的架构，在仅音频版本的任务中产生了不理想的结果。我们的目标是通过以下方式改进我们之前的尝试（即 [10] ）：首先，考虑交互的视频录制。其次，设计一个定制的神经网络架构，该架构在将音频和视觉特征组合成一个潜在表示之前独立处理这些特征。第三，在每个特征处理流中使用预训练的神经网络骨架，最后，通过使用主成分分析进行特征融合来实验，以防止我们仅仅因为只能使用单一特征表示而限制结果。

[1]这里的 t^W 、t^M 和 t^I 分别指的是 word2vec2.0 特征、MFCC 特征和 InceptionV1 特征的时间维度。

该架构的设计灵感来源于其他深度学习方法，这些方法利用从深度卷积神经网络中利用的注意力机制来完成从人类主体获取的视觉和音频数据的情感识别及相关任务 [11] [12] 。我们的方法与 [12] 中的方法相似，因为我们为每种注意力机制使用他们的卷积架构，尽管在我们的案例中，我们仅在音频和视觉流中使用逐帧注意力。我们还使用在 VGGFace2 上训练的 InceptionV1 ResNet，而不是该研究中使用的 3DResnet，因为它更适合我们的问题，而我们的基线 SVM 实验明确表明这一特征表示对于此任务最具信息性。正如 [12] 中一样，我们以下述方式计算用于音频和视觉流的逐帧注意力（即在每种情况下沿着时间维度)。我们在这里调整他们的公式，以保证完整性及明确性，以说明我们对他们的方法所做的修改。完整架构展示在图 4 中

1) 音频时间注意子网络： 设 x_i^A 为 i^{th} 音频特征矩阵输入。首先，将 x_i^A 居中裁剪到固定长度 l ，使得 $\frac{l}{s} \in \mathbb{N}$ 对于某个正整数 s 给定 $x_i^{A'}$ 。如果 x_i^A 的时间维度小于 l ，则在输入的两侧用零填充，使其长度如今等于 l 。然后我们将 $x_i^{A'}$ 分成 s 段并将它们堆叠在一起，使得 $x_i^{A'} \in \mathbb{R}^{s\times\frac{l}{s}\times n}$ 。模型然后接收这些张量的大小为 b 的批次，随后通过模型的音频流进行处理。

音频流的第一步是通过一个在 ImageNet 数据集上预训练的 ResNet18 模型 [7] 来处理每个 $b \times s$ 特征段。虽然我们使用 ImageNet 训练的模型来处理 MFCC 音频表示（因为 ImageNet 中没有 MFCC 示例）可能听起来令人惊讶，但研究表明在 MFCC 特征矩阵上使用这样的 ResNet 确实可以可靠地提高模型分数 [13] ，而我们的实验也确实发现了这一点。然后，我们采集预训练 ResNet18 模型的第五卷积堆栈的输出 F_j^A ，并在特征图上执行空间平均池化，生成 $F_j^{A'}$ （其中 j 在特征矩阵段中索引)。这将 ResNet 的输出从 $F_j^A \in \mathbb{R}^{s\times h\times w\times c}$ 降采样到 $F_j^{A'} \in \mathbb{R}^{s\times c}$ ，其中 s 是段数，h 和 w 分别是特征图的高度和宽度，c 是通道数，为每个段创建一个 $1 \times c$ 长度的描述符。现在的目标是学习音频特征矩阵段的一个 $s \times 1$ 长度描述符，其中描述符的 k^{th} 元素根据其在分类输入样本中的重要性来加权 k^{th} 段。这个描述符是通过包含一个 1D 卷积层、一个全连接线性层和一个 ReLU 非线性层的卷积堆栈来学习的：

$$H^A = W_1^A(W_2^A(F_j^{A'})^T)^T \tag{1}$$

其中 W_1^A 和 W_2^A 分别是线性层和卷积层可学习的参数矩阵 $s \times s$ 和 $1 \times c$ 。然后，我们计算音频注意子网络 A^A 的激活，即 $s \times 1$ 长度段描述符，如下：

$$A^A = \text{ReLU}(H^A) \tag{2}$$

音频流的输出嵌入，即哪些音频片段对于将输入示例分类到特定自我暴露类别最为相关的表示，是通过以下方式计算的：

$$E^A = \sum_{j=1}^{S} F_j^{A'} A^A \tag{3}$$

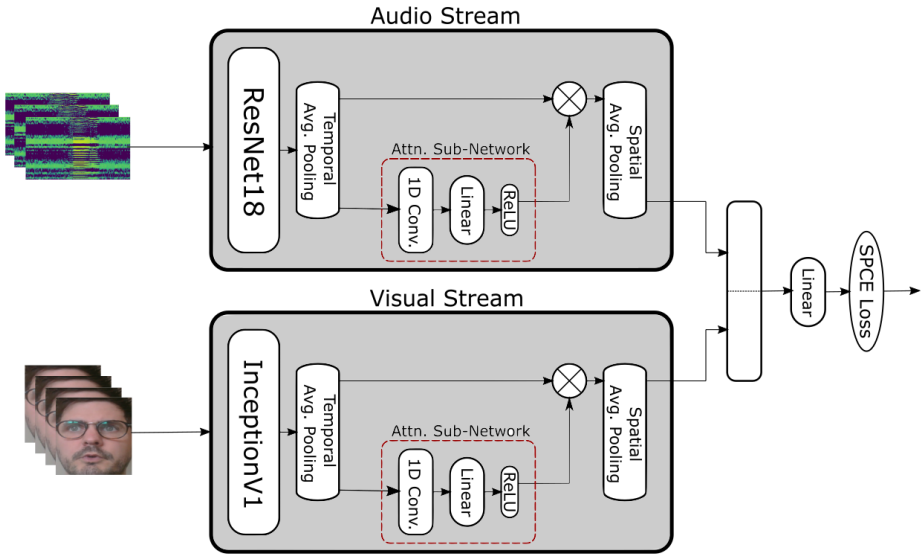

Fig. 4: 我们的多模态注意网络示例。MFCC 矩阵段（顶部）和裁剪面部的视频帧（底部）被输入到两个相似的流中。MFCC 段通过已在 ImageNet 上预训练的 2DResNet 骨干网络进行处理，然后进行平均池化，并复制。一个副本然后通过注意子网络处理，再与另一个 ResNet 输出副本相乘。这个表示随后再次进行平均池化，生成最终的音频嵌入。帧输入进行相同的过程，只是骨干网络是一个在 VGGFace2 上预训练的 InceptionV1 ResNet 架构。然后将所得的音频和视觉嵌入连接在一起，输入到一个线性分类层中。网络概率然后用于计算维护比例的交叉熵损失，通过该损失来优化网络参数。

*2) 视觉时间注意子网络：*用于从视频中提取视觉特征的音频嵌入 E^V 的方法完全遵循音频时间注意算法的相同步骤。实际上，主要的区别在于我们使用在 VGGFace2 上训练的 InceptionV1 ResNet 架构，而不是 ResNet18 模型，这个架构我们曾在 SVM 基线实验中使用过。

给定音频和视觉处理流的输出嵌入 E^A 和 E^V，然后我们通过计算每个嵌入向量在时间域（即跨段）的均值，使用平均池化来总结特征，得到 $E^{A'}$ 和 $E^{V'}$。然后在输入包含 7 个神经元的线性层之前将这些嵌入连接起来，以表示每个自我表露分数类。这产生输出 $\hat{y}$：

$$\hat{y} = \text{Softmax}(W^{AV}\text{concat}(E^A, E^V)) \tag{4}$$

其中 W^{AV} 是与线性输出层相关的可学习参数矩阵，concat() 是连接操作，而 Softmax() 是 softmax 函数，该函数返回在七个自我披露评分类别上的归一化概率。

在我们的实验中，我们测试了两种不同的视觉特征集、两种实验框架和四种不同的损失函数，以确定问题的最佳配置。首先，我们希望测试仅由 InceptionV1 ResNet 架构（在 VGGFace2 上预训练）输出的面部特征的效果，因为我们的 SVM 实验表明，这些可能是该任务中最具信息量的视觉特征。接下来，我们希望测试我们在 IV-A 节中详细提到的所有提取的视觉特征的组合。为了减少该特征空间的维数，我们将视觉输入通过视觉流中的 ResNetV1 骨干后获得的所有视觉特征连接在一起，并进行了主成分分析，参数设置为数据中 99 % 的方差由结果特征矩阵解释。这导致该特征空间的维数从每个视频帧的 555 维特征向量变为 67 维特征向量。

在 [10] 中，我们发现对自我披露分数进行分类的方法中存在一个细微差别。正如我们在那项研究中所述，参与者在他们的互动中使用 1 到 7 之间的李克特量表来评价自我披露的程度。这意味着每个分数都属于一个离散的类别，因此一种合理的方式是将该问题形式化为一个 n 类分类问题。然而，与 n 类分类问题相关的损失函数通常以相同的方式处理错误猜测，即没有一种方式可以表明一个猜测在数值上比其他可能的猜测更接近正确猜测。然而，自我披露的分数数据是有尺度的，意思是说，对于真实自我披露分数为 1 的情况下，模型猜测为 2 应该被视为比 6 或 7 更好的猜测。如此看来，可以认为这个问题更好地表现为回归问题。在 [10] 中，我们发现两种问题框架的形式化方法产生了类似的结果，因此不能明确得出哪种方法效果最好。因此，我们决定测试这两种方法对我们结果的影响。

我们希望研究损失函数对该问题的影响。一般来说，回归方法是通过最小化均方误差损失来优化给定模型的参数。由于我们没有充分的理由怀疑这个特定问题需要替代的基于回归的损失函数，所以我们选择仅根据均方误差损失来进行回归结果的研究。对于问题的分类版本，我们在实验中选择了一个分类交叉熵损失函数。对于这种损失，研究表明，标签平滑（一种通过一平滑参数 α 修改标准‘硬’标签的技术，公式为 $y_k^{LS} = y_k(1-\alpha) + \frac{\alpha}{K}$，其中 k 是总类别数的索引，在此研究中为七个类别）可以显著改善结果 [14]。因此，我们选择在消融研究中包含带标签平滑的交叉熵损失。最后，我们希望探索设计一种能在任务的分类和回归版本之间取得平衡的自定义损失函数的可能性，即一种利用数据是类别型的事实，同时也保留某些猜测相对于真实标签的正确性优于其他猜测的观念。受 [12] 的启发，我们设计了一种自定义交叉熵损失函数，对远离真实标签的猜测进行更大的惩罚。例如，对于一个标签自我披露分数为 7 的输入样本，猜测为 1 的损失会高于猜测为 2，猜测为 2 的损失会高于猜测为 3，以此类推。为此，我们修正了标准的交叉熵损失函数，其可表示为：

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{1}_{[c=y_i]}\log p_{i,c} \tag{5}$$

其中 N 是输入样本数，C 是类别数，$\mathbb{1}_{[c=y_i]}$ 是一个指示变量，当预测类别与真实类别相同时，该变量等于 1，$p_{i,c}$ 是 i^{th} 样本属于 c^{th} 类的概率。我们在 5 中加入了一个惩罚项，该惩罚项形式化了这样一种思想，即与真实值距离较远的预测应该受到更严厉的惩罚。这便是我们所称的尺度保留交叉熵损失：

$$\mathcal{L}_{SPCE} = -\frac{1}{N}\sum_{i=1}^{N}(1+\lambda(|y-\hat{y}|)^{\mu})\sum_{c=1}^{C}\mathbb{1}_{[c=y_i]}\log p_{i,c} \tag{6}$$

其中，λ 和 μ 是超参数，它们改变了根据不正确猜测与真实自我披露标签的距离来惩罚的程度。

综合来看，消融研究的参数导致多模态注意网络有八种不同的训练配置，其具体说明如图 5 所示。

C. 模型训练

回归和分类模型被训练了 100 个周期，而 SPCE 模型被训练了 150 个周期，因为我们发现它们需要更长时间才能收敛。所有网络版本我们都使用 Adam 优化器 [15]，初始学习率为 0.01，迷你批大小为 35 进行训练。音频特征输入被裁剪至长度 $l = 128$，并分为 $s = 4$ 段。视觉输入特征被裁剪至 $l = 210$ 帧，并分为 $s = 7$ 段。我们按照 [10] 准备训练数据，将训练和测试数据集按 80/20 划分，并使用加权随机采样来处理不平衡的类别。每个模型都被训练了五次，并计算所有五次训练实例的平均 F1 得分和标准偏差，以对模型的性能进行平衡评估。我们选择使用 F1 分数来验证模型，以便我们的结果可以直接与我们的 SVM 实验产生的结果进行比较。

V. 结果

我们研究的结果在 5 中展示。我们发现，多模态注意力网络的所有版本的得分都显著高于最佳的支持向量机基线。有趣的是，与 [10] 不同，在那里回归和分类模型的表现基本相当，我们发现分类框架（将自我披露得分视为离散类别）在建模问题上比回归框架（将得分视为源自连续数轴）更为有效。在所有情况下，我们发现，在每个实验框架中，从主成分分析中获得的特征表现优于仅训练在 InceptionV1 面部特征上的模型。这可能不足为奇，原因有二。首先，因为这一特征集包含了三倍于纯 InceptionV1 特征集的特征数量，然后被凝聚到其主成分。其次，因为显著减少特征数量（从纯 InceptionV1 情况下的 512 减少到主成分情况下的 67）意味着我们的模型不太容易受到维度灾难的影响，即需要更少的数据来有效建模这一较小的特征集。此外，我们发现，标签平滑在与非标签平滑版本的交叉熵损失相比时，对结果有所改善。最后，我们发现我们的规模保持交叉熵损失在所有模型版本中表现都优于所有，但与一个版本（使用标签平滑交叉熵损失的主成分特征）相比仅持平。

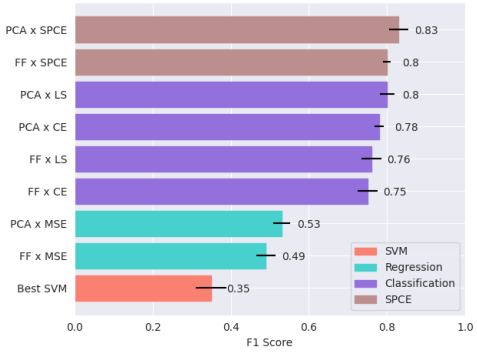

Fig. 5: F1 分数适用于我们的多模态注意力网络，该网络在不同的数据输入表示（主成分分析数据 (PCA)、仅面部特征 (FF)）和损失函数（类别交叉熵 (CE)、带标签平滑的交叉熵 (SE)、均方误差 (MSE)、以及我们的尺度保持交叉熵损失 (SPCE)）的组合下进行训练。我们还对深度学习实验中使用的不同实验框架进行了颜色编码。

VI. 讨论与结论

总体而言，与以往的尝试（例如，[10] ）相比，我们报告了在该任务中性能的显著提升。我们假设这得益于该工作中的一些重大进展。首先，我们收集了一个更大的数据集，使模型有更多的实例可以学习。其次，在这种情况下，所有的互动都是在相同的互动二人组之间记录的（即在人类和 Pepper 机器人的互动中）。在 [10] 中，我们收集了三种不同二人组之间的互动数据：人类–人类、人类–具象机器人和人类–语音代理。结果在这种情况下可能更差的一个原因是，由于参与者与不同代理进行互动，特定于每个自我揭露类别的声学特征可能受到调制。在我们的研究中，由于互动伙伴保持不变，变异性减少，这可能简化了学习任务。此外，我们使用了显著更先进的模型，利用了在大量数据集上训练的深层神经网络的表现力。此外，我们使用逐帧注意机制，利用了已证实在视频和语言建模任务中处于前沿的深度学习技术，可能为之前研究中使用的“现成”模型提供了直接的升级。最后，也许是最明显的，在本研究中，我们建模了两种不同的感官模式（音频和视觉），而不是 [10] 中考虑的单一感官模式。很可能音频领域与视觉领域相比，在自我揭示建模上具有更少的辨别信息，因此，之前的研究在仅考虑前者时就自然而然地处于劣势。

在我们的基线实验中，一个令人惊讶的观察是视觉特征在使 SVM 模型预测特定的主观自我揭示分数时最为有效。虽然关于自我揭示的文献在其定义上多种多样，但普遍认可的一点是，自我揭示主要是一种通过言语传递的社会现象 [16], [17] 。鉴于此，可以预期音频模态会产生最佳结果。这可能是因为音频特征的平均方式导致某些信息丢失。不幸的是，为什么视觉特征是最信息量丰富的一个彻底调查超出了本文的范围。

虽然本研究在使用神经网络对自我披露建模方面显示出较之前工作的显著提升，但我们在此详细介绍的改进是否足够显著以便使这些模型在社交机器人中实现，仍有待观察。在真实世界的环境中，错误的自我披露评分存在显著的风险 [10] 。假设一个人在自我披露很少的情况下实际上认为自己在分享大量的重要信息，这可能导致该人感到自己被忽视，或者他们所分享的敏感信息不值得听众的关注 [18] 。相反地，在一个互动对象不认为自己分享了大量个人信息的情况下，给予一个非常高的自

我披露评分可能会导致过度关注一个不重要的情况 [19] 。在这两种情况下描述的问题在情感健康干预的背景下会显得尤为复杂，其中未能察觉用户自我披露相关信号的风险可能会非常严重。因此，在考虑此类模型用于真实世界应用之前，需要做大量的工作。朝这个方向前进至少有两种方法。首先，应收集更多的数据以提高模型的性能。其次，应进行一项研究以评估模型性能与经过培训的专业人员在同一任务上的表现之间的差异。通常情况下，一个机器学习模型的质量及其作为真实世界应用的可行性是以其实现“类似人类”表现的能力来衡量的。一个有效识别个人信息披露程度的模型理应至少能够达到与训练有素的专业人员相同的水平（尤其是在该模型计划用于医疗干预背景时）。

此外，在短期内有多种方法可以改进我们的方法。首先，由于我们发现当使用主成分分析结合视觉特征时，任务的表现有所提升，因此很可能通过结合音频特征也能实现性能提升。特别是，我们没有尝试将 wave2vec2.0 的变压器层输出与 MFCC 特征结合的方法。另外，还可以进行更多的实证研究，以确定在音频和视觉情况下结合特征集的最佳方法。例如，[9] 使用去噪自编码器来学习连接输入特征的压缩潜在表示。未来的研究应实证检验这样的假设，即这种潜在表示存在于比主成分分析产生的输入特征空间更有效的空间中。进一步的研究还可以着眼于尝试其他类型的注意力机制。在 [12] 中，作者在视觉流中使用了通道注意力和空间注意力，除了我们两种方法共有的逐帧注意力之外。沿着这些思路的发展可以实现一种注意力机制，该机制可以在特征（即输入矩阵的列）上产生描述符。这样，模型不仅可以表示输入中哪些帧对其分类很重要，还可以表示哪些特征很重要。最后，可以修改模型以利用 3D ResNets 在输入视频帧上产生更高维的特征。然而，这种方法需要一个针对人脸建模的大型视频数据集上训练的 3DResNet，据我们所知，目前没有这样的预训练模型是公开可用的。借鉴关于自我披露建模的文献 [20] ，表明在加入词汇特征后，可以在两个人类互动者之间的（非主观）自我披露建模任务上通过多模态取得非常好的结果。在该研究中，作者使用预训练的 BERT 语言模型 [21] 来提取与每次话语中使用的词语相关的特征。自我披露的重要部分（至少在人类之间的情况中）被认为是通过语言交流的 [16] [17] 。因此，将来可以研究在 HRI 版本的任务中加入这种模态。

我们相信，这项研究通过在该主题上显示出相对于之前任何研究的显著改进，在主观自我披露建模的新领域中取得了重大进展。

References

[1] S. M. Jourard, *Self-disclosure: An experimental analysis of the transparent self*. Oxford, England: John Wiley, 1971.

[2] H. Kreiner and Y. Levi-Belz, “Self-Disclosure Here and Now: Combining Retrospective Perceived Assessment With Dynamic Behavioral Measures,” *Frontiers in Psychology*, vol. 10, p. 558, 2019.

[3] S. M. Jourard and P. Lasakow, “Some factors in self-disclosure,” *The Journal of Abnormal and Social Psychology*, vol. 56, no. 1, pp. 91–98, 1958.

[4] G. Laban, A. Kappas, V. Morrison, and E. S. Cross, “Building Long-Term Human–Robot Relationships: Examining Disclosure, Perception and Well-Being Across Time,” *International Journal of Social Robotics*, vol. 16, no. 5, pp. 1–27, 2024.

[5] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, “Openface 2.0: Facial behavior analysis toolkit,” in *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 59–66, Ieee, 2018.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 06 2016.

[8] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[9] W. Lin, I. Orton, Q. Li, G. Pavarini, and M. Mahmoud, “Looking at the body: Automatic analysis of body gestures and self-adaptors in psychological distress,” *IEEE Transactions on Affective Computing*, 2021.

[10] H. Powell, G. Laban, J.-N. George, and E. S. Cross, “Is deep learning a valid approach for inferring subjective self-disclosure in human-robot interactions?,” in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, Hri ’22, p. 991–996, IEEE Press, 2022.

[11] Z. Zhao, Q. Li, Z. Zhang, N. Cummins, H. Wang, J. Tao, and B. W. Schuller, “Combining a parallel 2d cnn with a self-attention dilated residual network for ctc-based discrete speech emotion recognition,” *Neural Networks*, vol. 141, pp. 52–60, 2021.

[12] S. Zhao, Y. Ma, Y. Gu, J. Yang, T. Xing, P. Xu, R. Hu, H. Chai, and K. Keutzer, “An end-to-end visual-audio attention network for emotion recognition in user-generated videos,” *CoRR*, vol. abs/2003.00832, 2020.

[13] K. Palanisamy, D. Singhania, and A. Yao, “Rethinking cnn models for audio classification,” *arXiv preprint arXiv:2007.11154*, 2020.

[14] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, “Revisiting knowledge distillation via label smoothing regularization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3903–3911, 2020.

[15] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *International Conference on Learning Representations*, 12 2014.

[16] P. C. Cozby, “Self-disclosure: a literature review.,” *Psychological bulletin*, vol. 79, no. 2, p. 73, 1973.

[17] J. Omarzu, “A Disclosure Decision Model: Determining How and When Individuals Will Self-Disclose,” *Pers Soc Psychol Rev*, vol. 4, no. 2, pp. 174–185, 2000.

[18] N. L. Collins and L. C. Miller, “Self-disclosure and liking: a meta-analytic review,” *Psychological bulletin*, vol. 116, pp. 457–475, 1994.

[19] Z. A. Reese and K. Orrach, “Reciprocal self-disclosure: Although respondents are reluctant to steal the spotlight, self-disclosers feel validated, understood, and cared for when respondents share comparable experiences,” *Journal of Social and Personal Relationships*, vol. 40, pp. 3485–3514, 5 2023.

[20] M. Soleymani, K. Stefanov, S.-H. Kang, J. Ondras, and J. Gratch, “Multimodal analysis and estimation of intimate self-disclosure,” in *2019 International Conference on Multimodal Interaction*, pp. 59–68, 2019.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.