多模态混合专家中的路由坍塌：跨模态条件化恢复专家多样性，但未提升准确率

Syed Hussamuddin 和 Mubashra Fayyaz

arXiv:2611.28725v1 [cs.MM] 21 Nov 2026

摘要——**混合专家（MoE）分类器正越来越多地应用于多模态情感识别，但其路由分布的健康状况鲜少得到衡量。我们研究了IEMOCAP上的一个包含七个专家的混合专家头，在两种条件下进行比较；两种条件的唯一区别在于模态嵌入送入该头之前的融合方式：仅使用学习式标量门控，或在相同门控之前加入成对跨模态注意力。采用标量融合时，路由器发生坍塌，集中使用七个专家中的1.88个，归一化熵为0.241。跨模态融合将熵提高至0.537，将专家利用数提高至3.25，增加了** $2.23\times$ **（**$p = 0.035$**，Cohen's** $d = 2.55$**），并促成了标量融合情形中不存在的情感差异化路由。然而，这种多样性的恢复并未带来可检测的收益：每种条件使用六个随机种子时，加权准确率（**$p = 0.646$**）和不同运行间的方差（**$p = 0.768$**）在统计上没有显著差异。我们还观察到，在配置相同的情况下，准确率差异达到13个百分点。在这一条件下，路由器健康度与任务性能彼此脱钩；诊断出坍塌本身并不足以证明有必要修复它。**

索引词——**情感识别、混合专家、多模态融合、可复现性、路由坍塌。**

I. 引言

MIXTURE-of-Experts（混合专家）层会将每个输入路由至由并行子网络构成的某种学习得到的组合。近期，这类层已被用于多模态情绪识别（MER），其依据是不同专家会专门处理不同的情感模式 [2]–[4]。这一假设很少经过检验。已有研究将性能提升归因于专家专门化，却没有衡量路由器是否确实在各专家之间分配负载。

这一点很重要，因为已知混合专家路由器会发生坍塌：门控机制收敛到少数专家，导致有效容量远低于名义专家数量，而参数数量和报告的架构却保持不变。路由坍塌在大规模语言建模中已有充分记录，辅助负载均衡损失也已成为标准做法。但在小数据MER场景中，这种问题很少得到诊断；在该场景下，不到四千条训练话语必须同时支持路由和专家学习。

我们对这一问题进行了对照研究。在固定专家头的情况下，我们只改变其接收的融合表征是否先经过跨模态注意力增强，并直接测量路由行为，而不是从准确率推断。我们的贡献如下：

1）我们发现，在未进行跨模态条件化的情况下，七专家混合专家头在IEMOCAP上会坍缩至1.88个有效专家，个别运行结果低至1.53。2）我们发现，对该混合专家头的输入进行跨模态条件化可以修复这一问题（$2.23\times$ 熵，$p = 0.035$），并促成情绪差异化专家分配。3）我们发现，与每种条件下六个随机种子的结果相比，这种修复并未带来可检测到的准确率或稳定性提升。4）我们报告，在配置完全相同的情况下，六个随机种子之间的准确率差异达13个百分点，这表明在该基准上进行单次运行比较并不可靠。

作者来自巴基斯坦卡拉奇75030的FAST新兴计算机科学国立大学计算机科学系（电子邮件：syed.hussam.uddin1@gmail.com；mubashra.fayyaz@nu.edu.pk）。本研究已提交至IEEE，供其考虑发表。
版权可能在未另行通知的情况下转让，届时此版本可能不再 доступ。

II. 方法

A. 主干网络

我们在 IEMOCAP 的文本、语音和动作捕捉（MoCap）流上使用三模态编码器。文本使用冻结的 300 维 GloVe 向量 [5] 进行嵌入，并依次通过双向长短期记忆网络和 Transformer模块进行编码。语音由 34 个短时特征表示，包括梅尔频率倒谱系数、色度特征和频谱描述符，并通过两个 Conv1D层、一个双向长短期记忆网络以及一个 Transformer模块进行编码。MoCap 将头部、手部和面部标记流拼接成一个200×189 图，并通过三个带步幅的二维卷积层进行编码。每个分支输出一个 128 维嵌入 $\mathbf{t}$、$\mathbf{s}$、$\mathbf{m} \in \mathbb{R}^{128}$。在任何阶段均未使用预训练基础模型。

B. 专家头

两种条件共享一个结构完全相同的七专家头。给定融合表示 $\mathbf{f} \in \mathbb{R}^{256}$后，路由器分配权重，预测结果是各专家输出的加权组合：

$$\begin{aligned} \mathbf{w} &= \mathrm{softmax}(\mathbf{W}_r \mathbf{f} + \mathbf{b}_r), \qquad \mathbf{w} \in \mathbb{R}^7, \\ \hat{\mathbf{y}} &= \mathrm{softmax}\Big(\textstyle\sum_{k=1}^{7} w_k \, E_k(\mathbf{f}) \Big), \end{aligned} \tag{1}$$

其中每个专家 E_k 都是一个将输入映射为四个类别对数几率的双层网络。不使用辅助负载均衡损失。这是有意为之：我们希望观察该架构自身产生的路由行为，因为采用混合专家头的多模态情绪识别研究通常会省略此类损失。

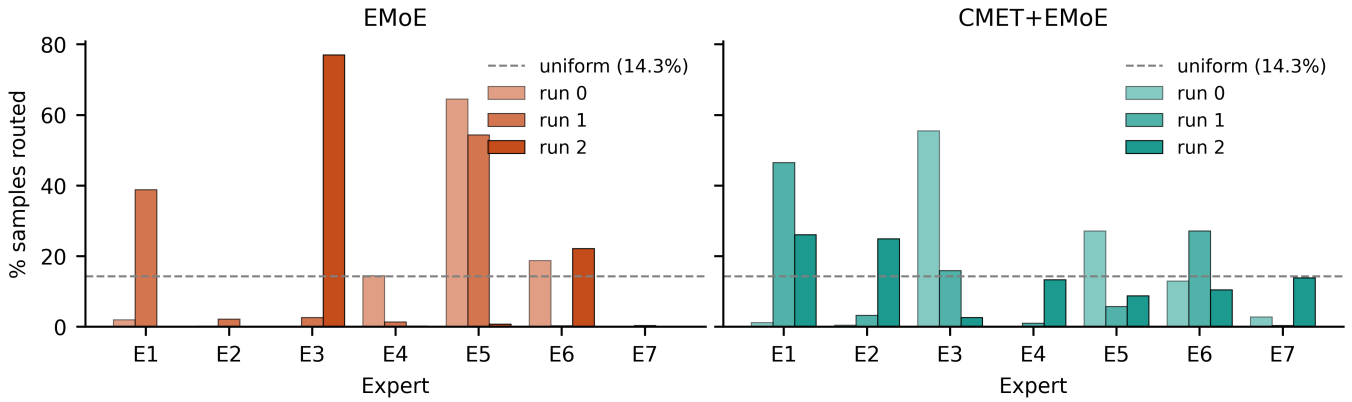


图 1. 在每种条件下，三个随机种子对应的留出样本中，各专家成为最大值索引的比例（运行 0–2 对应随机种子 42–44）。在 EMoE（左图）中，路由集中于一个或两个专家，且有若干专家完全没有接收到样本；在 CMET+EMoE（右图）中，负载分布更均匀。虚线表示均匀路由（14.3%）。

C. 两种条件

EMoE。三个分支嵌入由学习式标量门控组合，$\mathbf{f}_{\text{gate}} = g_t\mathbf{t} + g_s\mathbf{s} + g_m\mathbf{m}$ 其中 $\sum g = 1$，并投影至 $\mathbb{R}^{256}$。

CMET+EMoE。在融合之前，每种模态通过多头注意力（4 个头，$d_k = 32$）沿所有六个有向配对关注另外两种模态；这种设计与先前用于情感识别的门控跨模态融合方法 [8], [9]相关，但计算的是完整的成对图，而非单一交互。将模态 x 接收到的两个跨注意力输出之和记为 cross_x，则通过一个可学习的逐维门控结合自身内容与跨模态内容：

$$\begin{aligned} \boldsymbol{\alpha}_x &= \text{GELU}\big(\mathbf{W}_x[\mathbf{x}; \text{cross}_x] + \mathbf{b}_x\big), \\ \mathbf{x}_{\text{out}} &= \boldsymbol{\alpha}_x \odot \mathbf{x} + (1 - \boldsymbol{\alpha}_x) \odot \text{cross}_x, \end{aligned} \tag{2}$$

其中，$\odot$ 表示逐元素相乘。请注意，$\boldsymbol{\alpha}_x$ 经过 GELU 激活而非 sigmoid 激活，因此不受 $[0, 1]$ 的约束：(2) 是一种学习得到的仿射组合，而非凸组合。三个输出经过层归一化、拼接后，再通过同一个 $\mathbb{R}^{256}$ 投影。

其他条件均相同，仅融合表示$\mathbf{f}$有所不同。由于$\mathbf{f}$ 在式（1）中同时输入路由器和专家，因此该比较考察的是跨模态融合对整个混合专家头的影响，而非仅对路由器输入的影响。CMET 还增加了注意力层和门控层的参数。

III. 实验设置

A. 数据

我们使用 IEMOCAP [1] 的四类别子集：愤怒（1,103）、悲伤（1,084）、中性（1,708）和兴奋（1,041），共包含 4,936 条话语。较小的快乐类别在以往一些研究中会与兴奋合并，但本研究将其排除。第 1–4 个会话构成训练集（3,838 个样本），第 5 个会话中的说话人在训练阶段未出现，因此将其留作测试集。我们采用这一单一的说话人无关划分，而不是完整的五折交叉验证，从而使每种条件保留六个随机种子仍在可承受范围内，因为随机种子方差本身就是这里关注的一个量。

B. 训练

两种条件均使用 Adam 优化器，采用余弦衰减学习率（$2 \times 10^{-3} \to 10^{-6}$）、带 0.1 标签平滑的分类交叉熵、平衡类别权重、批量大小 64，并设置耐心值为 12 的提前停止。提前停止和检查点选择均以留出会话为监测对象，因此绝对准确率应视为偏乐观；由于两种条件采用相同流程，该流程不会偏向任何一方。除其他配置相同外，每种条件均使用**六个随机种子**（42–47）进行训练。准确率根据全部六次运行报告。路由统计数据需要训练后的权重，因此根据保留了检查点的三**次运行**（随机种子 42–44）报告。均值采用双侧 Student's t-检验进行比较，方差则采用双侧 F-检验进行比较。

C. 路由指标

对于每次运行，我们都会记录加权准确率（WA），并在留出会话中记录每个测试样本的路由器输出 $\mathbf{w}_i$ for every testsample$i$。对于每个样本，我们计算归一化路由熵

$$H_i = -\frac{1}{\log K} \sum_{k=1}^{K} w_{ik} \log w_{ik}, \qquad K = 7, \tag{3}$$

其中，$H_i = 0$ 表示完全坍缩到单个专家，$H_i = 1$ 表示在所有 K 上进行均匀路由。这两项路由统计数据均以 N 个测试样本的平均值报告：平均熵 $\bar{H} = N^{-1} \sum_i H_i$ 和平均有效专家数量

$$\bar{E} = \frac{1}{N} \sum_{i=1}^{N} \exp\big(H_i \log K\big), \tag{4}$$

每个样本路由分布的平均困惑度。$\bar{}$ $\bar{}$我们注意到，根据詹森不等式，$E \geq \exp(H \log K)$，且差距会随着 H_i的离散程度增大；我们对每个样本的困惑度取平均，而不是对平均熵进行指数运算，因此$\bar{}$ E 反映了典型情况下有多少个专家被调动

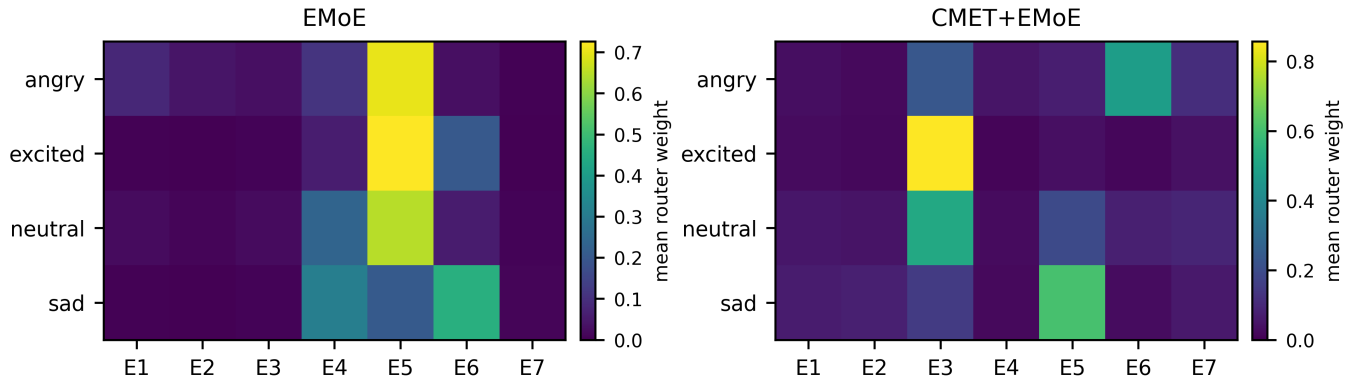


图 2. 在一个随机种子条件下的留出会话中，每种情绪类别的平均路由器权重（随机种子 42，图 1 中的运行 0）。专家索引是任意的，不同随机种子之间不匹配。在 EMoE（左）中，四种情绪中的三种由单个专家主导。在 CMET+EMoE（右）中，它们分布于三个不同的专家，表明存在情感差异化路由。

表 I 各条件下的结果（均值 ± 标准差）。路由指标基于三个随机种子，准确率基于六个随机种子。不同条件之间的路由存在显著差异；准确率及其方差则不存在显著差异。

	EMoE	CMET+EMoE	p
路由器熵 $\bar{H}$ [a]	0.241 ± 0.076	**0.537 ± 0.145**	**0.035**
有效专家 $\bar{E}$ [a]	1.88 ± 0.35	**3.25 ± 0.76**	**0.046**
加权准确率 [b]	66.51 ± 4.98	67.79 ± 4.33	0.646
准确率范围 [b]	58.20–71.58	59.29–70.86	—

[a] 三个随机种子 [b] 六个随机种子

的输入。我们还报告了每个专家作为最大值索引的测试样本所占比例，以及接收超过 1% 样本的专家数量。

IV. 结果

A. 无条件路由器发生坍塌

EMoE 达到 0.241 ± 0.076 的归一化熵，并使用七个专家中的 1.88 ± 0.35 个（表 I，图 1）。在单次运行中，这种行为尤为严重。在最极端的情况下，熵降至 0.155，有效专家数为 1.53；两个专家接收了 99.1% 的测试样本，而三个专家完全没有接收到样本。七个专家中有五个实际上是闲置参数，对决策没有任何贡献，却仍出现在报告的架构中。

B. 跨模态条件化修复路由

在使用完全相同的专家头时，CMET+EMoE 达到 0.537 ± 0.145 的熵和 3.25 ± 0.76 个有效专家。 $2.23\times$ 的熵增幅在 $p = 0.035$ 水平上具有显著性，且效应量较大（Cohen's $d = 2.55$）；有效专家数的差异在 $p = 0.046$ 水平上具有显著性。

这种现象在逐类别路由配置中也有定性体现，如图 2 所示，该图展示了一个具有代表性的随机种子。在 EMoE 下，四种情绪类别中有三种共享同一个主导专家。在 CMET+EMoE 下，这四种类别则分布在三个不同的专家上。因此，对头部输入进行跨模态条件化足以诱发这种架构旨在实现、但通常不会自行涌现的情绪差异化专门化。

C. 修复未带来可检测的收益

准确率并未随之提升。EMoE 的加权准确率为 66.51±4.98，CMET 为67.79 ± 4.33 +EMoE，差异为 1.28 个百分点（$p = 0.646$；95% 置信区间为 -4.7 至 $+7.3$）。运行间方差同样无法区分（$F = 1.32$， $p = 0.768$），这并不支持“坍缩路由器会导致训练不稳定”这一看法。在 6 个随机种子下，可以以 95% 的置信度排除准确率提升超过约 7 个百分点的可能性，但不能排除提升几个百分点的可能性。路由分布发生了显著变化；任务指标则没有可检测的变化。

D. 熵与准确率彼此脱钩

因此，这两种条件在路由上有所区别（p =0.035），但在准确率上没有区别（$p = 0.646$）。在 EMoE 内部也能观察到相同模式，但仅涉及 3 次运行：坍缩程度最高的一次运行，熵为 0.155，有效专家数为 1.53，其准确率在该条件下最高，达到 71.58%；而路由分布更分散的运行并未取得更好的表现。

我们认为，这表明这些专家并未学会互补功能。如果它们学会了，那么将负载分散到更多专家上理应有所帮助。分散负载并未带来帮助，这表明在这种数据条件下，将训练信号分散到更多专家会使其碎片化；路由器表面上的专门化并非功能上的专门化：路由分布能够区分输入，但专家无法区分输出。

E. 随机种子敏感性

在配置完全相同的情况下，EMoE 的准确率在六个随机种子下介于 58.20% 和 71.58% 之间，差距达到 13 个百分点（CMET+EMoE 为 11.6 个百分点），且有一个随机种子同时导致两种条件下的性能下降。从这一分布中抽取的单次运行结果，可能支持对底层架构得出范围更广的结论。一些经常被引用的 IEMOCAP

结果仅报告了一次运行，未提供方差估计 [6], [7]。如果这些设置下不同运行之间的变化幅度相近，那么此类数据之间相差一两个百分点的结果无法与随机种子效应区分开来。

V. 讨论与结论

在这种低资源环境中，若不加干预，混合专家路由器会发生坍塌。对其输入进行跨模态条件化可显著缓解这一问题，使熵提高 $2.23\times$，并促成情绪差异化专家分配，但未带来可检测到的准确率或稳定性提升。

这对如何呈现相关文献中的混合专家贡献有两点启示。首先，应直接报告路由统计数据。若不测量路由分布，就将性能提升归因于专家专门化，这并无我们的结果作为依据；一个名义上包含七个专家的头部，实际使用的专家可能不到两个。其次，应将路由器健康度视为诊断指标，而非优化目标。能够显著改善路由的干预措施未必能改善任务表现；而从大规模语言建模中借鉴的负载均衡目标，可能不适用于专家缺乏足够数据来实现差异化的场景。

这些发现基于单一基准、单一说话人无关划分（该划分也用于模型选择）、固定的专家数量，以及三个随机种子对应的路由统计数据。这种解耦是否会在更大规模的数据、完整交叉验证或显式负载均衡损失下依然存在，仍有待探究。可向作者索取各次运行的路由测量数据。

致谢

作者使用 Claude（Anthropic）协助起草和编辑本篇通讯的文本。研究本身，包括设计、实验、测量、统计分析和结果解读，均由作者完成。作者已核实所有报告的数值，并对内容承担全部责任。

参考文献

[1] C. Busso 等，“IEMOCAP：交互式情感二人动作捕捉数据库”，语言资源与评估，第 42 卷，第 4 期，第 335–359 页，2008 年。[2] A. G. Mengara Mengara 和 Y.-k. Moon，“CAG-MoE：采用交叉注意力门控混合专家的多模态情感识别”，数学，第 13 卷，第 12 期，文章编号 1907，2025 年。[3] S. Dutta、S. Balaji 和 S. Ganapathy，“一种用于对话中多模态情感识别的混合专家模型”，arXiv:2602 23300，2026 年。。[4] J. Li、R. Yu、H. Huang 和 H. Yan，“动态融合的多模态情感识别中的单模态驱动蒸馏”，arXiv:2503 23721，2025 年。。[5] J. Pennington、R. Socher 和 C. D. Manning，“GloVe：用于词表示的全局向量”，载于*EMNLP* 论文集，2014 年，第 1532–1543 页。[6] S. Tripathi 和 H. Beigi，“使用深度学习对 IEMOCAP 数据集进行多模态情感识别”，arXiv:1804.05788，2018 年。[7] M. Khan 等，“MemoCMT：采用基于跨模态Transformer的特征融合进行多模态情感识别”，科学报告，第 15 卷，文章编号 5473，2025 年。[8] S.-K. Lee 和 J.-H. Kim，“CAMER：贡献感知的多模态情感识别”，载于 *Interspeech* 论文集，2025 年，第 2690–2694 页。[9] J. He、J. Mi 和 T. Toda，“GIA-MIC：采用门控交互注意力和模态不变学习约束的多模态情感识别”，载于*Interspeech* 论文集，2025 年，第 2695–2699 页。