arXiv:2302.10912v2 [cs.LG] 8 Jun 2023

用于不平衡分析的平衡视听数据集

Wenke Xia[1,†], Xu Zhao[2,†], Xincheng Pang[1],Changqing Zhang[3],Di Hu[1,4,*]

[1]中国人民大学高瓴人工智能学院，北京[2]东北大学计算机科学与工程学院，沈阳[3]天津大学智能与计算学部，天津[4]北京市大数据管理与分析方法重点实验室，北京。项目网站：
https://gewu-lab.github.io/Balanced-Audiovisual-Dataset/

1. 引言

不平衡问题在多个领域都受到了广泛关注，包括多任务学习中的损失不平衡、长尾数据集中的类别不平衡等。在多模态学习领域，不平衡问题同样普遍存在：当不同模态之间存在差异时，模型往往会依赖占主导地位的模态 [7]。近期研究 [7, 9]致力于解决模态不平衡问题。Wang等人 [9]发现，由于模态过拟合行为，多模态模型的性能可能不如单模态模型。Peng等人 [7]和Wu等人 [10] 从优化过程的角度提出了解决模态不平衡问题的方法，而Han等人 [5] 则通过实现可靠的多视图融合缓解了这一问题。然而，我们观察到，尽管现有的不平衡处理方法在整体测试集上表现出色，但如图1所示，在某些模态偏好子集上进行推理时，其表现甚至不如单模态模型。以往研究尚未充分解决这一现象，但它可能会影响多模态模型在某些场景下的可靠性，例如存在模态噪声的场景。

为了全面分析这一现象，我们首先提出一种指标，利用单模态置信度估计样本级别的模态差异，并将现有数据集划分为不同的模态偏好子集。具体而言，我们将模态偏好子集定义为一组样本，其中对应模态的置信度高于另一模态。进一步地，我们评估了现有的不平衡方法，发现：尽管现有的不平衡方法在整体测试集上的表现优于单模态模型，但如图 1(a-b)所示，在不同数据集的视觉偏好测试子集上，这些方法的表现始终不如单模态模型。

然而，鉴于深度学习以数据驱动为本质 [10]，现有数据集（如 Kinetics-Sound）中存在的严重模态偏差必然会导致多模态模型学习这种偏差 [1] 以及 VGG-Sound [2] 因此，为了进一步研究不平衡方法的有效性，我们构建了一个平衡视听数据集，其中包含具有各种模态差异的样本，并且这种模态差异在整个数据集中的分配保持均匀，如图 1（c）所示。

基于我们的平衡数据集，我们重新评估了现有的不平衡方法，发现结果与此前的现象一致：如图 1(c)中的直方图所示，目前仍没有一种不平衡方法能在每个模态偏好测试子集上都超越所有单模态模型。我们进一步选取具有不同模态差异的测试子集来评估现有不平衡方法的性能，并意外地发现，模态差异越大，多模态模型相较于单模态模型的表现就越差。这一现象表明，现有的不平衡方法难以处理模态差异较大的多模态数据。因此，要全面分析多模态模型的性能，也应考虑不同程度的模态差异。

总而言之，我们的贡献如下：

•我们提出了一种评估指标，用于估计模态差异，并指出现有视听数据集存在严重的模态偏差。

• 我们提供了一个模态差异均匀分布的平衡视听数据集，以支持对多模态模型的进一步分析。

• 我们发现，在模态差异严重的场景中，现有多模态模型的表现会逊于单模态模型。我们强调，未来研究应重视解决这一问题，以确保多模态模型在各种应用场景中的可靠性。

2. 评估方法

2.1. 模态差异估计

受一些针对单模态数据集的样本难度度量研究启发，我们 [3]，提出了一种用于

† 贡献相同。 *通讯作者。

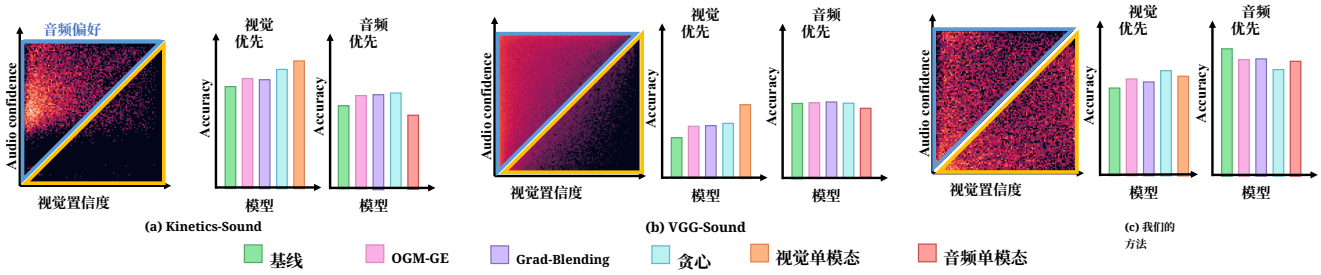


图 1：（**a-b**）展示了现有数据集上的模态差异分布热力图和分类结果。在热力图中，颜色表示样本所占比例，颜色越亮表示样本越多。（**c**）展示了我们的平衡热力图和分类结果。

测量样本级模态差异的方法，该方法利用训练过程中单模态置信度对真实标签的赋值差异。

具体而言，我们考虑一个大小为 N,$S = \{(X_i,\ Y_i)\}_{i=1}^{N}$ 的训练数据集，其中第 i 个样本包含观测值 X_i 及其真实标签 Y_i。对于多模态数据集，我们认为每个观测值 X_i 由来自不同模态的 M 个输入构成，记为 $X_i = (x_i^{m_1}, .., x_i^{m_k}, .., x_i^{m_M})$，其中 m_k 分别表示第 k 个模态。我们首先针对每个模态，使用单一模态训练各个单模态模型，共训练 E_{m_k} 轮，并利用单模态模型的平均置信度表示第 i 个样本的学习难度：

$$\hat{u}_i^{m_k} = \frac{1}{E_{m_k}} \sum_{e=1}^{E_{m_k}} p_\theta^{m_k}(e)(y_i|x_i^{m_k}),$$

其中，$p_\theta^{m_k}(e)$ 表示参数为 θ 的单模态模型在第 e 轮的置信度。

我们将此类平均单模态置信度分数划分为间隔为 0.01 的区间，然后统计对应区间内的样本数量并计算其占比，这可视为模态差异的分布。我们在图1中以热力图展示所得分布，其中坐标对应平均单模态置信度分数，颜色表示相应区间内样本的占比。因此，我们的方法为直观呈现多模态数据集中的模态差异提供了一种方式。

2.2. 数据集划分策略

ˆ 我们使用平均单模态置信度 $\hat{u}_i^a$ 和 u_i^v来表示音频和视觉模态上的单模态置信度。为研究视听数据集中的模态偏差，我们根据单模态置信度分数将现有视听数据集划分为两个子集。我们将这两个子集分别命名为音频偏好子集和视觉偏好子集。如果 $u_i^a > u_i^v$，则将该样本划分到音频偏好子集中；而如果 $u_i^v > u_i^a$，则将该样本划分到视觉偏好子集中。

通过我们的划分策略，我们将现有数据集的测试集划分为不同的模态偏好子集。

如图1（a-b）所示，我们的估算结果证明，这些视听数据集中存在严重的模态偏差，可能不足以全面分析现有的模态不平衡方法。

3. 现有数据集上的实验

3.1. 实验设置

我们估计 Kinetics-Sound [6] 和 VGG-Sound [2] 数据集中的模态差异，并进一步评估这些数据集上现有的不平衡方法。我们使用 ResNet-18 作为两种模态的骨干网络。遵循 Peng等人 [7], 对音频输入的处理方式，我们以 $48kHz$的采样率提取频谱图；对于视觉输入，我们以 1fps 提取帧，并随机选择帧。

为估计模态差异，我们分别训练每种模态，以评估单模态置信度。鉴于不同数据集中各模态的覆盖速度有所不同，我们在训练过程中采用早停策略，以获得可靠的单模态置信度。

3.2. 实验结果

本节中，我们使用不同子集对现有的不平衡方法进行全面评估，并以朴素的后期融合方法作为基线。具体而言，我们首先评估这些方法在整体测试集上的表现，然后根据划分策略比较其在模态偏好子集上的结果。

如表 1 所示，我们发现一些结果与近期研究 [7]一致：所有不平衡方法在整体集合上的表现都优于单模态模型，这证明了多模态学习的有效性。此外，与基线相比，所有不平衡方法在不同数据集上均取得了显著提升，这表明现有方法在一定程度上缓解了模态不平衡问题。

尽管不平衡方法在整体子集上取得了出色的表现，我们观察到，现有方法在弱模态偏好子集上的表现始终不如单模态模型。具体来说，考虑——

	Kinetics-Sound			VGG-Sound		
方法	总体	音频优先	视觉优先	总体	音频优先	视觉优先
基线	61.73	60.96	69.09	48.65	52.38	27.50
OGM-GE	64.57	63.20	74.44	50.89	53.43	35.63
Grad-Blending	66.00	64.34	73.82	**51.47**	**53.73**	35.87
贪心	**66.35**	**64.51**	80.75	50.32	52.68	37.37
音频单模态	51.52	54.32	31.23	41.25	45.39	10.86
视觉单模态	47.67	42.28	**86.43**	32.44	28.65	**51.67**

表 1：现有视听数据集上的评估结果。

VGG-Sound [2] 数据集是一个经过筛选、以声音为主的数据集，其中音频偏好子集的样本数量显著多于视觉偏好子集。与音频模态相比，视觉模态较弱。然而，表 1 显示，尽管视觉单模态模型在整体集合上的表现不如现有不平衡方法，但在视觉偏好子集上，视觉单模态模型的表现大幅超过其他多模态模型，这表明现有不平衡方法仍未能充分利用弱模态。

然而，考虑到深度学习的贪心特性 [10]，多模态模型不可避免地会学到这些有偏视听数据集固有的模态偏差，这可能会影响对这些不平衡方法的评估。

4. 平衡视听数据集

为减轻模态偏差的影响，并进一步研究不平衡方法的有效性，我们充分考虑模态差异，构建了一个平衡视听数据集。如图 1(c) 所示，我们的平衡数据集包含具有各种模态差异的样本，并且这些模态差异均匀分布在整个数据集中。

为了自动收集带有相应标签的视频，我们使用现成的预训练单模态模型筛选爬取的视频（用于视觉模态的 SlowFast [4]在 Kinetics-400 [6] 上预训练；用于音频模态的 VGGish [8] 在 VGG-Sound [2] 上预训练）。受预训练数据集的限制，我们选取两个数据集中都存在且语义概念相同的若干类别，将它们整合为 30 个类别，并从 YouTube 爬取对应主题的视频。

随后，我们将完整视频切分为 10 秒的片段，并使用预训练模型评估每个片段的单模态置信度。根据单模态置信度，我们将片段分为三类，分别称为高对应度片段、音频正确片段和视觉正确片段，并据此构建数据集。其中，高对应度片段指两个单模态模型都能预测出真实标签的片段；音频正确片段指只有音频模态模型能预测出真实标签的片段，而视觉正确片段则相反。采用这种片段筛选策略，我们可以初步构建一个近乎均衡的视听数据集，其中三类视听片段的占比相近。

然而，由于从网站爬取数据存在局限，初步数据集呈现长尾分布，并且存在轻微的模态偏差。例如，乐器视频很容易收集，而某些人类动作类别则难以获取。因此，为解决这一问题，我们从 VGG-Sound [2] 和 Kinetics-400 [6] 中筛选额外样本，并将其纳入数据集。最后，我们估算数据集的模态差异，并根据估算结果通过移除噪声以及部分占主导地位的模态数据来重新校准数据集。最终，整个数据集包含来自 Kinetics-400 数据集的 10,000 个样本、来自 VGG-Sound 数据集的 8,000 个样本，以及从 YouTube 收集的 16,000 个样本。如图 1(c) 所示，我们的数据集中的模态差异仍呈均匀分布。

5. 我们数据集上的实验

5.1. 在模态偏好子集上的测试

如表 2所示，在我们构建的平衡音频-视觉数据集上进行训练时，所有不平衡方法在整体集合上的表现均优于基线和单模态模型，这证明了现有不平衡方法的有效性。

然而，目前仍没有一种方法能在每个模态偏好子集上都超过所有单模态模型。基线、OGM-GE [7]和 Grad-Blending [9]方法在音频偏好子集上取得了显著提升，但在视觉偏好子集上的表现仍不如视觉单模态模型。相比之下，贪心 [10]方法在音频偏好子集上的表现优于视觉单模态模型，但不及音频单模态模型。

表2 的结果表明，尽管现有的不平衡处理方法取得了显著进展，但在处理存在模态差异的样本时仍有困难。为了分析模态差异对多模态模型的影响，我们提出使用不同的模态差异子集来评估这些方法。

方法	总体	音频优先	视觉优先
基线	70.47	78.62	59.37
OGM-GE	**72.27**	78.92	63.59
Grad-Blending	72.04	**79.49**	61.89
贪心	72.01	72.46	**71.41**
音频单模态	62.96	77.76	44.79
视觉单模态	49.82	36.79	67.58

表2：在我们的数据集上的评估结果。

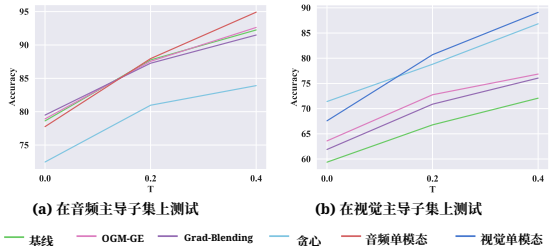


图2：在具有不同阈值 T 的模态主导子集上的评估结果。

5.2. 使用不同的模态差异进行测试

在获得每个样本的单模态平均置信度 $\hat{u}_i^m$ 后，我们提出一个阈值 T 来划分测试子集。具体而言，如果 $u_i^a - u_i^v > T$，则将该样本划分到音频主导子集；如果 $\hat{} \hat{} u_i^v - u_i^a > T$，则将该样本划分到视觉主导子集。随着阈值 T 增大，相应子集中的模态差异也会增大。在实验中，我们将阈值 T 设为 0、0.2 和 0.4，以选取模态差异程度不同的子集。

我们在图 T 2 中给出了不同阈值下音频主导子集和视觉主导子集的结果。仅考虑 $T = 0$ 情形时，某些多模态模型在相应子集上的表现优于单模态模型（例如，OGM-GE 在音频主导子集上的表现，以及贪心方法在视觉主导子集上的表现）。然而，随着阈值 T增大，模态主导子集中只会保留模态差异较大的样本，而在相应的模态主导子集上，单模态模型的表现优于其他多模态模型（当 $T = 0.4$时，音频模型在音频主导子集上的表现优于其他不平衡方法）。

人们预期多模态模型的表现会优于单模态模型，然而，额外的信息反而会妨碍模型在模态差异较大的多模态数据上的表现。考虑到模态差异较大的样本会包含更多模态噪声，这些结果表明，模态噪声会影响多模态模型的表现。

6. 结论

在这项工作中，我们首先提出了一种模态差异估计指标，并发现现有视听数据集存在严重的模态偏差。为了全面分析模态不平衡方法，我们构建了一个模态差异均匀分布的平衡视听数据集。此外，我们使用不同模态差异子集评估现有的不平衡方法，并发现多模态模型会……

在某些模态差异严重的场景中，其表现不如单模态模型。这种全面的评估方法应考虑不同程度的模态差异，以确保多模态模型在不同应用场景中的可靠性。

7. 致谢

本研究得到国家自然科学基金委员会（NO.62106272）、中国科学技术协会青年人才托举工程（2021QNRC001）以及中国人民大学公共计算云的支持。

参考文献

[1] R. Arandjelovic 和 A. Zisserman。Look, listen and learn。载于 ICCV，第 609–617 页，2017 年。1[2] H. Chen、W. Xie、A. Vedaldi 和 A. Zisserman。Vggsound：一个大规模视听数据集。载于 ICASSP，第 721–725 页。IEEE，2020 年。1, 2, 3[3] K. Ethayarajh、Y. Choi 和 S. Swayamdipta。数据集难度的信息论度量。arXiv 预印本 arXiv:2110.08420，2021 年。1[4] H. Fan、Y. Li、B. Xiong、W.-Y. Lo 和 C. Feichtenhofer。PySlowFast。https://github.com/facebookresearch/slowfast，2020 年。3[5] Z. Han、C. Zhang、H. Fu 和 J. T. Zhou。基于动态证据融合的可信多视图分类。IEEE 模式分析与机器智能汇刊（TPAMI），2022 年。1[6] W. Kay、J. Carreira、K. Simonyan、B. Zhang、C. Hillier、S. Vijayanarasimhan、F. Viola、T. Green、T. Back、P. Natsev 等。Kinetics 人体动作视频数据集。arXiv 预印本 arXiv:1705.06950，2017 年。2, 3[7] X. Peng、Y. Wei、A. Deng、D. Wang 和 D. Hu。通过即时梯度调制实现平衡多模态学习。载于 CVPR，第 8238–8247 页，2022 年。1, 2, 3[8] K. Simonyan 和 A. Zisserman。用于大规模图像识别的超深卷积网络。arXiv 预印本 arXiv:1409.1556，2014 年。3[9] W. Wang、D. Tran 和 M. Feiszli。是什么让多模态分类网络的训练变得困难？载于 CVPR，第 12695–12705 页，2020 年。1, 3[10] N. Wu、S. Jastrzebski、K. Cho 和 K. J. Geras。刻画并克服多模态深度神经网络学习的贪婪特性。载于 ICML，第 24043–24055 页。PMLR，2022 年。1, 3