用于语音情感分析的预训练模型表征及其抗噪声鲁棒性

Vikramjit Mit ra、*Vasudha Kowtha*、*Hsiang-Yun Sherry Chien*、*Erdrin Azemi*、*Carlos Avendan o*

苹果公司

arXiv:2303.03177v1 [eess.AS] 3 Mar 2023

摘要

预训练模型表征在语音识别、自然语言处理及其他应用中展现了最先进的性能。双向编码器表征（BERT）和隐藏单元 BERT（HuBERT）等语音模型，能够生成词汇和声学表征，从而助力语音识别应用。我们研究了利用预训练模型表征从语音中估计激活度、效价和支配度等维度情感。我们发现，效价可能主要依赖词汇表征，而激活度和支配度则主要依赖声学信息。在本研究中，我们使用预训练模型的多模态融合表征，实现了最先进的语音情感估计；与标准声学和词汇基线相比，在效价估计的 concordance 相关系数（CCC）上，相对提升分别达到 100% 和 30%。最后，我们研究了预训练模型表征对噪声和混响退化的鲁棒性，并注意到词汇表征和声学表征受到的影响各不相同。我们发现，与声学表征相比，词汇表征对失真具有更强的鲁棒性；此外，我们还证明，从多模态模型进行知识蒸馏有助于提升基于声学模型的抗噪鲁棒性。

索引词— 语音情感，鲁棒性，多模态

1. 引言

基于语音的情感识别模型旨在实时识别说话者的情绪状态。当前的人机交互系统能够识别说话者说出的词语，却无法识别其表达的情绪。可靠且鲁棒的基于语音的情感模型有助于改善人机交互和健康与福祉应用，例如语音助手 [1, 2], 临床心理健康诊断 [3, 4], 和/或治疗方案 [5]。

情感识别研究主要采用两种情感定义：（1）离散情感 [6, 7], 和（2）维度情感 [8]。离散情感分为恐惧、愤怒、喜悦、悲伤等类别，情感词典的类别数量从六类 [6]到二十七类 [9]不等。差异离散情感定义上的差异会导致标注复杂、难以实现一致的情感标签，以及无法涵盖罕见和／或复杂的情绪状态。维度情感使用由以下三项构成的三维模型来表示情感空间：（i）激活度（反映声音能量）、（ii）效价（表示负面性或正面性），以及（iii）支配度（说明一个人听起来可能有多强势或温和）。

先前研究表明，结合词汇和声学表征可以提升语音情感识别模型的性能 [10, 11, 12, 13]。近期研究显示，预训练模型表征能够实现最先进的情感识别性能 [13, 14], ，并显著提升效价估计性能 [12, 15]。对于实际应用而言，评估语音情感模型抵御背景失真的鲁棒性至关重要。文献 [16, 17], 研究了噪声对语音情感识别的影响，以及特征和建模方法的鲁棒性；然而，先前研究尚未探讨预训练模型表征的噪声鲁棒性。

在本研究中，我们展示了：

(1) 融合从预训练模型中获得的多种词汇和声学表征后，在一个大型公开真实世界数据集上，情绪估计的性能达到了最先进水平。

(2) 与低层声学特征相比，预训练模型表征（包括词汇和声学表征）表现出更强的抗噪声和信道退化鲁棒性。

(3) 与相应的声学嵌入相比，词汇嵌入对声学失真具有更强的鲁棒性。

2. 数据

所有实验均使用公开可用的 MSP-Podcast 数据集 1.6 [18] ，其中包含英语使用者的语音。语音片段是由单个说话者发出的语句，时长为 3 至 11 秒。数据由多位评分者人工标注了效价、激活度和支配度评分（7 点李克特评分）。数据划分如表1所示。为使结果能够与文献中的结果进行比较， [11, 12], 我们报告了 Eval1.3 和 Eval1.6 上的结果（见表 1 ）。部分数据带有人工转录文本，用于执行第 4 节所述的自动语音识别分析。

为分析情感模型的鲁棒性，我们在 Eval1.6 集上添加了噪声和混响。我们添加了反映家用电器噪声的噪声，信噪比设为三个水平：20 至 30 分贝（平均 20 分贝）、10 至 20 分贝（平均 15 分贝）以及 0 至 10 分贝（平均 5 分贝）。每个文件的信噪比水平均从均匀分布中随机选取。因此，我们创建了三个额外的评估集，分别对应这三个信噪比水平（见表1）。我们对加噪数据进行了人工混响处理，以模拟真实的室内环境。为进行噪声感知训练，我们对表1中训练集“Train+noise”的 9% 文件添加了类似的失真。

表1。MSP-Podcast 数据划分及噪声退化集。

划分	小时	描述
Train1.6	85	训练集
Valid1.6	15	验证集
Eval1.3	22	Podcast1.3 评估集
Eval1.6	25	Podcast1.6 评估集
训练+噪声	93	含有 9% 噪声数据的训练集
$Eval1.6_{25dB}$	25	评估1.6 +噪声范围为 20-30 分贝
$Eval1.6_{15dB}$	25	评估1.6 +噪声范围为 10-20 分贝
$Eval1.6_{5dB}$	25	评估1.6 +噪声范围为 0-10 分贝

3. 表征与声学特征

3.1. 声学特征

作为基线声学特征，我们使用 40 维的梅尔滤波器组能量，并附加基频、基频差分和浊音特征，得到 43 维的 $MFBF_0$特征。

3.2. HuBERT 的声学表征

我们探究了由 HuBERT large 生成的嵌入。HuBERT large 是一种预训练声学模型 [19], , 在包含 60,000 小时语音的 Libri-light 数据集上进行了预训练，具有 24 个 Transformer 层和 1024 维嵌入。在本研究中，我们提取了以下嵌入：

(1) 来自预训练 HuBERT large 模型 $HUBERT_L$ 第 24^{th} 层的嵌入（未进行微调）。
(2) 来自 $HUBERT_L$模型（24^{th} 层）的 $HUBERT_A$嵌入，该模型在使用 100 小时 Librispeech 数据进行的自动语音识别（ ASR）任务上经过微调。

3.3. 来自 BERT 的词汇表征

我们研究了使用预训练的 BERT [20] 嵌入来训练情绪模型。由于 MSP Pod- cast 数据中只有部分语音片段提供了真实转录文本，因此我们依靠两个自动语音识别模型来估计语音的转录文本： $HUBERT_A$ 和一个内部自动语音识别系统。对于已转录的语音，我们从预训练 BERT 模型的 12^{th} 层提取了 768 维嵌入。

4. 基于自动语音识别模型文本数据的分析

我们在 MSP-Podcast 数据中包含人工转录文本的一部分上评估了自动语音识别的性能。我们使用了两种不同的自动语音识别系统（一种内部自动语音识别系统和$HUBERT_A$ 系统），图 1 展示了将每个维度视为相互独立时，效价、激活度和支配度在不同区间（低、中性和高）下词错误率（WER）的变化情况。图 1 显示，效价和激活度越高，词错误率越高；而支配度较低时也呈现出相同趋势。效价高、激活度高或支配度低时词错误率上升，这表明在这些条件下，自动语音识别系统会出现更多错误。图 1 表明，在不同情绪水平下，自动语音识别转录文本生成的 BERT 嵌入可能会受到词错误率变化的影响。

5. 情绪估计模型

5.1. 模型架构

此处介绍的所有模型均采用相同的基本架构（见图 2）进行有监督的维度情绪估计：先经过一个带跳跃连接的时序卷积（TC）层，再经过两层门控循环单元（GRU），其输出投影至一个 128 维嵌入层，最后经过输出层。

单模态输入：时间卷积门控循环单元模型以一种模态作为输入（第 3 节所述的任意一种特征）。输入特征的维度和时间分辨率可能因特征类型及其模态而异。

多模态输入：词汇-声学融合模型结合了 $HuBERT_L$ 和 $HuBERT_A$ 嵌入（2048 维），并将其投影至 128 维，如图 3 所示。随后，将所得的 128 维嵌入与分别使用来自 $HuBERT_A$ 的 BERT嵌入和内部自动语音识别转录文本训练的时间卷积门控循环单元模型所获得的嵌入进行拼接。

5.2. 模型训练

对于单模态和多模态输入，我们训练了 2 层时间卷积门控循环单元模型，并使用留出验证集Valid1.6（见表1）上的性能进行模型选择。一致性相关系数（CCC） [21] 用作损失函数（L_{ccc}）（见式 (1)），其中 L_{ccc} 是由三种维度情绪各自获得的 CCC组合而成（$\alpha = 1/3$ 和 $\beta = 1/3$）。 CCC 由式 (2) 定义，其中 μ_x 和 μ_y 为均值， σ_x^2 和 σ_y^2 为估计变量和真实变量对应的方差， ρ 为这两个变量之间的相关系数。模型训练时的小批量大小为 32，学习率为 0.0005。

$$L_{ccc} = -(\alpha CCC_v + \beta CCC_a + (1-\alpha-\beta)CCC_d) \quad (1)$$

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}. \quad (2)$$

5.3. 蒸馏

为探究能否将知识从多模态模型蒸馏到较简单的单模态模型中，我们采用教师-学生框架，探索将前者学到的表征蒸馏到后者中。设

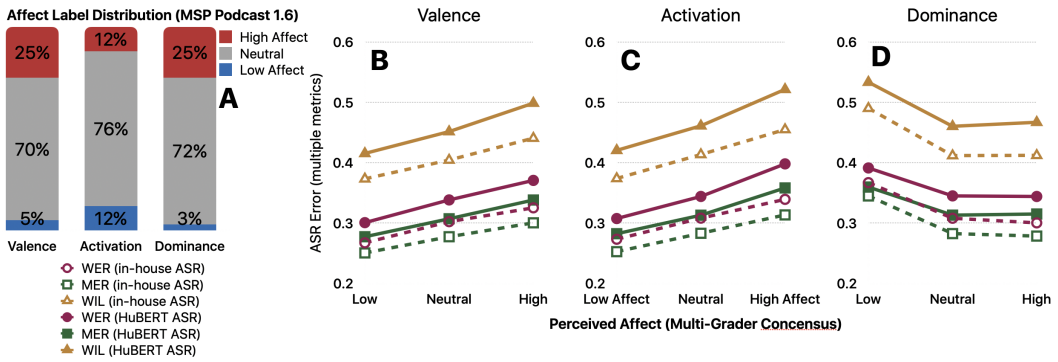


图 1。效价、激活度和支配度处于低、中性和高水平时的词错误率变化，数据取自自动语音识别模型生成的转录文本（WER：词错误率，WIL：词信息丢失率，MER：匹配错误率）

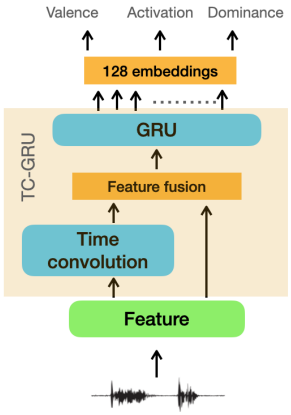


图 2。基本 TC-GRU 网络

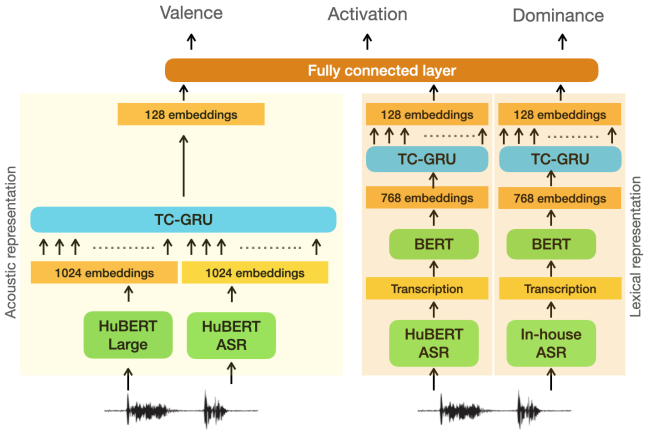


图 3。多模态情绪估计模型

多模态教师网络 T 针对语音样本 i 的嵌入为 $E_{T,\,i}$，学生模型 S 针对样本 i 学习到的嵌入为 $E_{S,\,i}$，真实标签为 $l_{k,\,i}$（其中 $k \in \{a, v, d\}$），且 T˜给出的估计标签为 $l_{k,\,i}$。我们提出的蒸馏损失如下：

$$L_{dis,k} = CCC_{E_{T,i},\hat{E}_{S,i}} \cdot \gamma_i, \tag{3}$$

$$\gamma_i = 1 - \frac{1}{M}\left(\frac{\sum_k |l_{k,i} - \tilde{l}_{k,i}|}{3}\right). \tag{4}$$

在（3）中，蒸馏损失通过教师（$E_{T,\,i}$）与学生（$E_{S,\,i}$）嵌入之间的 CCCˆ（见2）计算得出，并根据教师网络的目标标签（$l_{k,\,i}$）与估计标签（$l_{k,\,i}$）之间的残差误差（$|l_{k,\,i} - l_{k,\,i}|$）˜加权。我们探索了用于蒸馏的 L_2 和 $cosine$ 距离损失，发现所提出的基于 CCC 的损失表现更好。请注意，M 表示数据集中目标标签$l_{k,\,i}$的动态范围。γ_i 为每个样本提供置信度度量；对于教师网络残差误差较低的样本，该值更接近 1，而对于残差较高的样本则更接近 0。

残差误差较低时接近教师网络，残差较高时则更接近 0。最终损失定义为：

$$L_k = \kappa(L_{ccc,k} + L_{CE,k}) + \lambda L_{dis,k}, \tag{5}$$

其中，κ 和 λ 的取值经经验选择为 0.001 和 1，以更加侧重于蒸馏损失。

6. 结果

6.1. 情绪估计

我们使用以下输入训练了单模态TC-GRU情绪模型：(i) 基于 $MFBF_0$的声学特征，从(ii) $HuBERT_L$ 和(iii) $HuBERT_A$中提取的嵌入，以及(iv) 使用 $HuBERT_A$ 生成的转录文本，从BERT（$BERT_{HuB}$）获得的词汇嵌入（请注意，来自内部自动语音识别的BERT嵌入取得了与 $BERT_{HuB}$相似的性能）。我们使用语音嵌入（$HuBERT_L$和 $HuBERT_A$）作为输入，并结合 $BERT$ 嵌入，训练了多模态模型（图3）。单模态和多模态

模型在维度情感估计任务上的性能如表2所示。表2显示，相较于 $MFBF_0$，从任意一种预训练模型（无论是词汇模型还是声学模型）中提取的特征都能显著提升效价估计性能。基于$BERT$和 $HuBERT_L$ 的模型在效价估计方面的 CCC 相较于 $MFBF_0$分别相对提升了70%和60%。与所有单模态模型相比，多模态融合进一步提升了效价估计性能（见表2）；与基于 $MFBF_0$、 $BERT$ 和 $HuBERT_L$的模型相比，我们观察到其效价估计方面的 CCC 分别相对提升了100%、40%和44%。表2显示，词汇表征单独用于估计激活度和支配度时可能并不足够；然而，通过多模态融合，它们取得了整体最佳性能。

6.2 模型对声学失真的鲁棒性

我们研究了上述模型对训练期间未见过的声学失真的鲁棒性。表3显示，相较于预训练模型的表征，噪声退化对 $MFBF_0$的性能影响更为严重。我们观察到，对于效价而言，词汇嵌入比声学嵌入更能抵抗噪声。

此外，我们使用 Train+noise 数据集（表1）探索了噪声感知训练。我们观察到， $MFBF_0$特征在所有维度上均有所提升，而噪声感知训练对基于 BERT 的模型几乎没有影响或毫无影响。BERT嵌入在声学失真下表现出较强的鲁棒性。总体而言，多模态融合取得了最佳的最先进性能。这些发现表明，从预训练模型中学习到的表征能够抵抗声学失真。

接下来，我们研究了能否将多模态融合模型学习到的表征（采用第 5.3 节所述的流程）蒸馏到基于 $MFBF_0$和 $HuBERT_L$ 的模型中，以提升其鲁棒性。表4和表3显示，来自多模态模型的知识蒸馏有助于提升使用 $MFBF_0$和 $HuBERT_L$特征训练的两种模型的鲁棒性。

表2. 来自单模态模型和多模态融合模型的维度情感估计 CCC

系统	Eval1.3			Eval1.6		
	act	val	dom	act	val	dom
$MFBF_0$	0.73	0.33	0.63	0.71	0.33	0.64
$BERT_{HuB}$	0.33	0.56	0.29	0.30	0.55	0.27
$HuBERT_A$	0.73	0.53	0.64	0.70	0.52	0.62
$HuBERT_L$	0.76	0.54	0.68	0.74	0.53	0.66
多模态	**0.78**	**0.68**	**0.69**	**0.75**	**0.66**	**0.67**

7. 结论

在这项工作中，我们研究了预训练模型中的词汇和声学表征，并探讨了它们在不同声学条件下的鲁棒性。我们观察到，词汇这些表征有助于准确估计效价，但对激活度和支配度估计没有帮助。低层声学特征，例如 $MFBF_0$，被发现容易受到噪声退化的影响；然而，噪声感知训练有助于提升其鲁棒性。与 $MFBF_0$ 相比，使用 $HuBERT_L$ 嵌入训练的模型具有较强的鲁棒性，尤其是在不同信噪比水平下进行效价估计时；不过，在低信噪比水平下，激活度和支配度估计的表现有所下降。多模态融合有助于生成更优的表征，实现了最先进的性能，并且能够抵御各种信噪比条件下的噪声。通过多模态融合学习表征，并将这类信息蒸馏至仅使用声学信息的系统，能够提升这些系统在噪声条件下的性能。未来，我们计划研究领域失配会如何影响词汇特征和效价估计。

表3。在 Train1.6、Trian+噪声上训练的模型，并在噪声条件下进行评估的情绪估计 CCC 。

系统	Train1.6			Train+噪声		
	act	val	dom	act	val	dom
$MFBF_0$						
$Eval1.6_{25dB}$	0.67	0.31	0.58	0.69	0.33	0.61
$Eval1.6_{15dB}$	0.64	0.23	0.56	0.69	0.30	0.60
$Eval1.6_{5dB}$	0.45	0.11	0.41	0.60	0.23	0.53
$BERT_{Hub}$						
$Eval1.6_{25dB}$	0.31	0.55	0.27	0.29	0.55	0.26
$Eval1.6_{15dB}$	0.31	0.54	0.27	0.29	0.54	0.26
$Eval1.6_{5dB}$	0.29	0.51	0.26	0.28	0.51	0.26
$HuBERT_A$						
$Eval1.6_{25dB}$	0.68	0.52	0.58	0.69	0.54	0.59
$Eval1.6_{15dB}$	0.65	0.51	0.57	0.68	0.53	0.58
$Eval1.6_{5dB}$	0.55	0.48	0.47	0.63	0.50	0.51
$HuBERT_L$						
$Eval1.6_{25dB}$	0.72	0.51	0.61	0.71	0.51	0.61
$Eval1.6_{15dB}$	0.68	0.50	0.57	0.70	0.50	0.60
$Eval1.6_{5dB}$	0.52	0.44	0.38	0.64	0.45	0.55
多模态						
$Eval1.6_{25dB}$	0.72	0.66	0.62	0.72	0.66	0.62
$Eval1.6_{15dB}$	0.68	0.64	0.59	0.70	0.64	0.59
$Eval1.6_{5dB}$	0.53	0.59	0.51	0.57	0.57	0.43

表4. 多模态模型知识蒸馏后，单模态模型的情绪估计（CCC）

系统	$MFBF_0$			$HuBERT_L$		
	act	val	dom	act	val	dom
$Eval1.6_{25dB}$	0.70	0.35	0.60	0.73	0.55	0.63
$Eval1.6_{15dB}$	0.68	0.31	0.58	0.70	0.53	0.61
$Eval1.6_{5dB}$	0.54	0.22	0.46	0.54	0.47	0.43

8. 参考文献

[1] Vikramjit Mitra、Sue Booker、Erik Marchi、David SFar-rar、Ute D Peitz、Bridget Cheng、Ermine Teves、Anuj Mehta 和 Devang Naik，《利用声学线索和副语言嵌入从语音中检测表达》，*arXiv* 预印本 *arXiv:1907.00112*，2019年。

[2] Vasudha Kowtha、Vikramjit Mitra、Chris Bartels、Erik Marchi、Sue Booker、William Caruso、Sachin Kajarekar 和 Devang Naik，《从语音中检测情绪原语及其在辨别类别情绪中的应用》，载于*ICASSP*。IEEE，2020年，第7164–7168页。

[3] DS Sahana 和 L Girish，《使用情感分析自动检测药物不良反应》，《国际计算机工程与技术高级研究期刊（*IJARCET*）》，第4卷，第5期，2015年。

[4]Brian Stasak、Julien Epps、Nicholas Cummins 和 Roland Goecke，《情绪语音在抑郁症分类中的研究》，载于*Interspeech*，2016年，第485–489页。

[5] Jared B Torre 和 Matthew D Lieberman，《将感受诉诸语言：情感标记作为隐式情绪调节》，《情绪评论》，第10卷，第116–124页，2018年。

[6] Paul Ekman，“关于基本情绪的论证”，《认知与情绪》，第6卷，第3–4期，第169–200页，1992年。

[7] Robert Plutchik，“情绪的本质：人类情绪具有深厚的进化根源，这一事实或可解释其复杂性，并为临床实践提供工具”，《美国科学家》，第344–350页，2001年。

[8] James A Russell 和 Albert Mehrabian，“情绪三因素理论的证据”，《人格研究期刊》，第11卷，第273–294页，1977年。

[9] Alan S Cowen 和 Dacher Keltner，“自我报告捕捉到27种不同的情绪类别，这些类别由连续梯度连接”，《美国国家科学院院刊》，第114卷，第38期，第E7900–E7909页，2017年。

[10] Saurabh Sahu、Vikramjit Mitra、Nadee Seneviratne 和 Carol YEspy-Wilson，《用于语音情感识别的多模态学习：自动语音识别输出与真实转录文本的分析和比较》，载于*Interspeech*，2019年，第 3302–3306 页。

[11]Ayoub Ghriss、Bo Yang、Viktor Rozgic、Elizabeth Shriberg 和 Chao Wang，《用于增强语音情感识别的情感感知自动语音识别预训练》，*arXiv* 预印本 *arXiv:2201.11826*，2022年。

[12] Sundararajan Srinivasan、Zhaocheng Huang 和 Ka- trin Kirchhoff，《通过跨模态条件教师-学生训练进行表征学习，以实现语音情感识别》，*arXiv* 预印本 *arXiv:2112.00158*，2021年。[13] Shamane Siriwardhana、Andrew Reis、Rivindu Weerasekera 和 Suranga Nanayakkara，《联合微调“类BERT”自监督模型以改进多模态语音情感识别》，*arXiv* 预印本 *arXiv:2008.06682*，2020年。

[14]Leonardo Pepino、Pablo Riera 和 Luciana Ferrer，《使用 wav2vec 2.0 嵌入进行语音情感识别》，*Interspeech* 论文集，第 3400–3404 页，2021年。

[15] Vikramjit Mitra、Sherry Chien Hsiang-Yun、Vasudha Kowtha、Joseph Yitan Cheng 和 Erdrin Azemi，《语音情感：研究模型表征、多任务学习与知识蒸馏》，载于 *Interspeech* 论文集，2022年。

[16] Seong-Gyun Leem、Daniel Fulford、Jukka-Pekka On-nela、David Gard 和 Carlos Busso，《并非所有特征都一样：噪声环境下语音情感识别的鲁棒特征选择》，载于*ICASSP 2022-2022 IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，2022年，第6447–6451 页。

[17] Andreas Triantafyllopoulos、Gil Keren、Johannes Wag- ner、Ingmar Steiner 和 Bj¨orn Schuller，《利用深度残差网络进行语音增强，实现稳健的语音情感识别》，*Interspeech 2019*年论文集，第1691–1695页，2019年。

[18] Soroosh Mariooryad、Reza Lotfian 和 Carlos Busso，《通过从现有语音语料库中检索表现性行为，构建自然情感语音语料库》，载于国际语音通信协会第十五届年会，2014年。

[19] Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdelrah- man Mohamed，《HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习》，*IEEE/ACM*音频、语音与语言处理汇刊，第29卷，第3451–3460页，2021年。

[20] JacobDevlinMing-Wei ChangKenton andLeeKristina Toutanova，《BERT：用于语言理解的深度双向变换器预训练》，载于 *NAACL-HLT* 论文集，2019年，第 4171–4186 页。

[21] I Lawrence 和 Kuei Lin，《用于评估可重复性的一致性相关系数》，*Biometrics*，第 255–268 页，1989。