arXiv:2602.06000v1 [cs.AI] 5 Feb 2026

利用 OpenAI 的 Whisper 表征和注意力池化方法进行语音情感识别

Ali Shendabadi[1], Parnia Izadirad[1], Mostafa Salehi[1*], Mahmoud Bijankhan[2]

[1]德黑兰大学智能系统工程学院，伊朗德黑兰 N.卡加尔。[2]德黑兰大学文学与人文学院，伊朗德黑兰恩格拉布。

*通讯作者。电子邮箱：mostafa salehi@ut.ac.ir；投稿作者：alishendabadi@ut.ac.ir；parniaizadirad@ut.ac.ir；mbjkhan@ut.ac.ir；

摘要

由于缺乏标准且规模足够大的数据集，语音情感识别（SER）研究一直受到限制。近期研究利用预训练模型提取特征，用于 SER 等下游任务。本研究通过提出两种基于注意力的池化方法——多头注意力平均池化和 QKV 池化——探索 Whisper（一种预训练的自动语音识别系统）在语音情感识别中的能力。这些方法旨在高效降低 Whisper表征的维度，同时保留情感特征。我们分别使用 IEMOCAP 和 ShEMO 数据集，在英语和波斯语上采用 Whisper Tiny 和 Small 进行实验。我们的多头 QKV 架构在 ShEMO 数据集上取得了最先进的结果，非加权准确率提升了 2.47%。我们还比较了不同 Whisper 编码器层的性能，发现中间层在波斯语数据集上的 SER 任务中往往表现更好，为 HuBERT X-Large 等大型模型提供了一种轻量且高效的替代方案。研究结果凸显了 Whisper 作为 SER 表征提取器的潜力，并证明了基于注意力的池化方法在降维方面的有效性。

关键词：语音情感识别、Whisper、注意力池化、注意力机制

1 引言

近年来，人机交互已进入新阶段。以ChatGPT为代表的AI助手已成为人类的日常工具，这也使机器更好地理解人类及其需求的必要性进一步提升。实现这一目标的一个重要因素，是这些系统正确检测人类情感的能力Nimmagadda等 （2022年）。在口语交互中，情绪通过话语的语调来传达；更具体地说，即词语的说话方式及其语义。语音情感识别（SER）的任务是利用从语音信号中提取的特征来检测话语中的情绪。能够更好理解人类情感的语音情感识别系统可以对用户做出更恰当的回应，并有助于预判其需求。

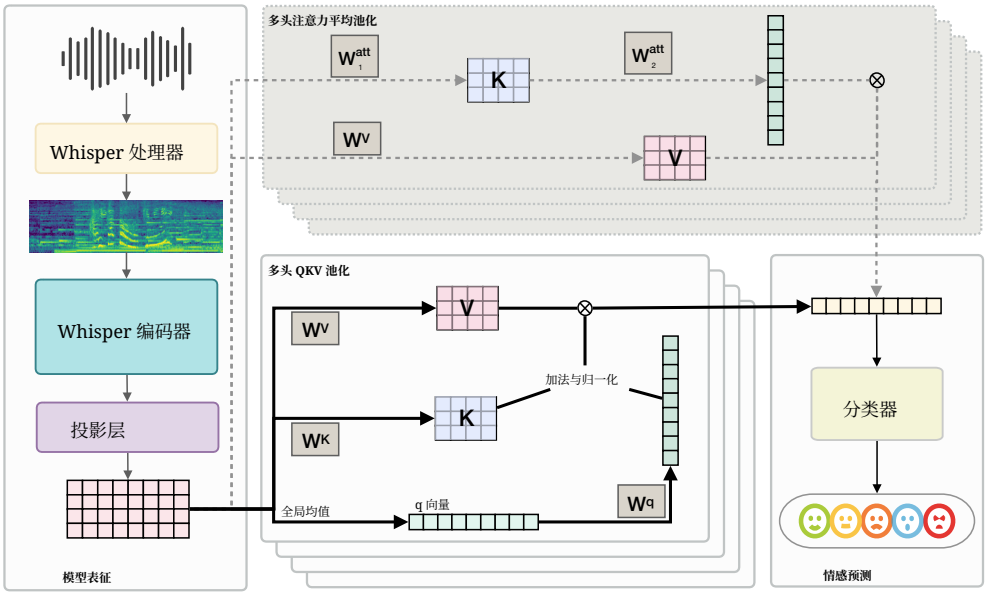


图 1：用于语音情感识别的多头注意力平均池化和多头 QKV 池化流程示意图。使用 Whisper 编码器提取语音表征后，应用多头池化方法降低表示矩阵的维度。堆叠的矩形表示注意力头。所有注意力头的输出经过拼接，随后通过权重矩阵 W^o 投影，生成一个最终的 256 维向量，作为分类器的输入。

近期研究中，研究人员使用预训练模型提取特征，而非采用梅尔频率倒谱系数等传统特征。这些研究通过在语音情绪识别语料库上微调预训练模型，取得了最先进的成果。另一些研究则将预训练模型作为上游模型来提取特征，方法是

冻结所有层，并构建一个下游模型，根据预训练模型生成的表征来预测情绪。

在本研究中，我们使用 WhisperRadford 等人(2022年)，这是一种预训练的自动语音识别模型。我们冻结编码器层，以此进行特征提取。随后，我们使用降维模块（DRM），从表示矩阵中池化出单个向量。我们提出了两种基于注意力的池化方法：注意力池化和 QKV 池化。在注意力池化中，我们首先沿用Nasersharif 和 Namvarpour (2024年的实验方法，为每一帧赋予权重，以评估其对于预测话语情感的重要性。我们进一步将这种方法扩展为多头形式。在 QKV 池化中，查询以预训练模型初始表示的全局平均池化结果为条件。结果表明，基于注意力的池化方法能够让较小的模型与较大模型的结果相媲美。

Th本研究的主要贡献如下：

- 提出了 QKV 池化，这是一种基于注意力的池化方法，能够从高维表示矩阵中提取单个向量，同时最大限度地减少信息损失。此外，我们还将注意力池化应用于 Whisper。结果显示，在 IEMOCAP 和 ShEMO 上，采用 QKV 池化的 Whisper 小型模型均取得了与 Whisper Large V3 相当的性能，展现出成本与准确率之间稳定且有前景的权衡。
- 探索将 Whisper 编码器作为表征提取器，用于波斯语这一低资源语言的语音情感识别的有效性。
- 分析 Whisper 编码器的不同层，并比较它们生成有效语音表征的能力。

尽管其他模型取得了更高的准确率，如第4节所述，但我们的研究强调，有效的池化方法可以使较小的预训练模型取得与规模大数百倍的模型相当的结果。我们在以下地址发布了全部代码：https://github.com/alishendabadi/AttentivePoolingWhisperSER。

2 相关工作与背景

早期语音情感识别系统依赖于传统机器学习方法，如隐马尔可夫模型（HMM）、高斯混合模型（GMM）、支持向量机（SVM）和决策树Pepino等人（2021）。在这些方法中，研究人员采用了手工设计的声学特征，如梅尔频率倒谱系数、线性预测倒谱系数（LPCC）、能量、音高和语速。尽管这些方法具备一定能力，但在捕捉人类情感的复杂性与个体差异方面存在根本局限Jin等人（2025年）。

近期研究探索了深度学习方法，并取得了良好的实验结果。循环神经网络（RNN）和长短期记忆（LSTM）网络通过对语音信号进行序列处理，用于学习高层次的特征表征 Gupta 等人 (2022年）。在DialogueRNN Majumder 等人 (2019年）、AGHMN Jiao 等人 (2020）和 HiGRUJiao 等人 (2019年）中，研究人员

使用门控循环单元（GRU）捕捉情感语音中的时序依赖关系和上下文关系。在其他研究中 Zhao 等人 (2019年）；Zhang 等人(2019年）；Dutt 和 Gader (2023年），卷积神经网络（CNN）和循环神经网络（RNN）分别结合使用，以对情感语音中的频率特征和时序特征进行建模与提取。

尽管取得了显著成果，这些模型仍面临若干挑战。由于窗口大小固定，CNN 难以捕捉长期依赖关系。另一方面，循环神经网络在处理长序列时容易出现梯度消失或梯度爆炸问题，从而限制了其对情绪的整体理解。此外，结合 CNN 和循环神经网络结构往往会增加模型复杂度，导致计算成本更高、训练耗时更长，并增加过拟合风险，尤其是在小型数据集上。

针对上述问题，研究人员采用 Transformer 模型来建模语音情感识别任务。这类基于自注意力的模型更擅长处理较长的输入序列并捕捉其全局依赖关系。

语音情感识别研究长期面临的一个问题是缺乏标准数据集。现有数据集通常规模较小且分布不平衡。因此，研究人员在工作中采用迁移学习，利用预训练模型学习可应用于下游任务的有效表征。Wav2Vec、HuBERT 和 Whisper 等预训练模型提供了丰富的上下文语音表征，对包括语音情感识别在内的各种下游语音处理任务产生了重大影响。例如，Jiao 等人 Jiao 等人 (2024) 在 MFHCA 方法中使用 HuBERT 特征和对数梅尔频谱图来识别情感相关区域并融合特征。该语音情感识别方法基于多空间融合模块（MF）和层次协同注意力模块（HCA）。HuBERT 还被用于 DSTM 的特征表征中；DSTM Jin 等人 (2025年) 是一种基于 Transformer 的网络，采用动态静态特征融合，并使用局部自适应多头注意力模块（DTM）处理动态局部特征，使用全局静态注意力模块（STM）处理全局特征。

尽管 MFHCA 和 DSTM 利用 HuBERT x-large 表征取得了基准测试结果，但该模型的参数数量庞大，因而在处理过程中面临较高的时间和空间复杂度挑战。较新的预训练模型在语音识别方面更加稳健。借助这些模型，研究人员得以在不牺牲准确率的情况下降低计算成本。例如，Qu 等人Qu 等人（2024）开发了 SER-Whisper 框架，该框架将冻结的 Whisper 编码器与专用的基于 Transformer 的分类层相结合，在 RAVDESS Livingstone 和 Russo （2018年）数据集上取得了很高的准确率。同样，Sankaran 等人 Sankaran 等人 （2024）和 Ma 等人 Ma 等人 （2024）将基于 Whisper 的表征与其他预训练模型进行了比较，例如 Wav2vec 2.0、WavLM Chen 等人 （2022年）和 Data2vecBaevski 等人 （2022年），结果表明，Whisper 在音频滥用检测和语音情感识别任务中均持续优于其他方案。

后续研究致力于将 Whisper 应用于情感相关任务。Chou等人Chou (2024) 提出了 Whisper-SER，该方法利用 Whisper Tiny 和轻量级适配器实现多标签语音情感识别。他们采用了结合加权损失的两阶段训练策略，以缓解自动语音识别与语音情感识别目标之间的冲突。

然而，他们指出，与中性话语相比，Whisper Tiny 在 MSP-PODCAST 等情感语音数据集上的词错误率（WER）更高。Feng等人Feng 和 Narayanan (2024) 将 Whisper 的作用扩展至语音情感识别数据整理，通过生成转写文本并将编码器输出用作表征。他们还探索了使用包括 LLaMa 2、Falcon 和 Flan-T5 XXL 在内的大语言模型进行情感标注和数据增强。尽管大语言模型在零样本情感标注方面表现不佳，但结合多个模型并引入有限的人工反馈，显著提升了标注质量和语音情感识别性能。

为了高效处理预训练语音模型的高维输出，降维已成为降低计算成本和过拟合风险的关键技术。Nasersharif 等人 Nasersharif 和 Namvarpour (2024) 提出了一种降维模块（DRM），结合注意力平均池化、Maxout激活和线性层，以提取相关特征并降低维度。该方法能够动态选择信息丰富的模块，并通过注意力机制进行特征融合，从而提升语音情感识别性能。

另一类研究探讨了将 Wav2vec 2.0 用于语音情感识别，尤其关注如何利用中间 Transformer模块输出。研究表明，较早的模块编码声学特征，而较后的模块则捕捉语言信息。Pepino 等人 Pepino 等人 (2021) 提出通过可训练权重组合多个 Wav2vec 2.0 层，取得了更优的语音情感识别性能，并揭示了不同层的相对贡献。Nasersharif 等人 Nasersharif 和 Nam- varpour (2024) 在 IEMOCAP 和 ShEMO 上开展了研究Mohamad Nezami 等人 (2019) 并证明，中间层在情感识别方面优于初始层和末尾层。

3 方法论

本节介绍我们研究中的三个主要组成部分：1. 一个预训练模型，用于从音频样本中生成表征。2. 一种基于注意力机制的池化方法，可在降维过程中减少重要信息的损失。3. 一个用于预测最可能情绪的分类器。

3.1 模型表征

本研究使用 Whisper特征来表征音频样本。Whisper 在许多语音相关任务（包括语音情感识别）中展现了最先进性能。我们选择该模型的另一个原因是其多语言能力，因为我们的实验涉及英语和波斯语。下面我们将简要介绍 Whisper，然后详细说明特征提取过程。

3.1.1 Whisper

Whisper 是一个弱监督的多语言、多任务语音识别模型，使用 680,000 小时的标注音频数据进行预训练。Whisper 采用编码器-解码器Transformer架构。首先，将音频输入重采样至 16,000 Hz，并对音频

片段计算 80通道对数梅尔频谱图并进行全局归一化。随后，这些频谱图会经过两个卷积层。接着，编码器Transformer模块使用正弦位置嵌入处理卷积层的输出。最后，解码器部分使用可学习位置嵌入和绑定输入输出表示，将表征转换为词元序列。

Whisper 与其他自动语音识别模型的不同之处在于其多语言能力。Whisper 使用海量数据进行训练，其中包含涵盖 96 种语言的 117,000 小时语音数据。此外，预训练所用语音数据的多样性，使 Whisper 成为适用于低资源语言的稳健模型。

Whisper 主要由三个部分组成：1. 两个卷积层，用于提取由 Whisper 处理器从原始音频生成的梅尔频谱图的高层表征。梅尔频谱图以 25毫秒窗口计算，步长为 10 毫秒。2. 编码器 Transformer 模块，在添加位置嵌入后处理频谱图表征。编码器模块使用自注意力生成一系列编码器隐藏状态。3. 解码器 Transformer 模块，利用可学习位置嵌入预测下一个词元。预测结果以编码器隐藏状态为条件，通过编码器模块中的交叉注意力进行处理；同时也以先前预测的词元为条件，通过自注意力进行处理。编码器和解码器模块具有相同的宽度和层数。

3.1.2 特征提取

对于特征提取，我们使用 Whisper 的编码器部分，且不更新其参数。话语首先通过 Whisper 处理器，该处理器将每个样本截断为 30 秒，并从原始音频生成对数梅尔频谱图。梅尔频谱图由两个采用 GELU激活函数 的卷积层处理。经过卷积层后，正弦位置嵌入被添加到输出中。这些嵌入可以为模型提供音频特征在序列中位置的信息。

$$X = \text{Processor}(\mathbf{x}(t)), \quad X \in \mathbb{R}^{80\times 3000} \tag{1}$$

最后，这些增强后的表征会通过编码器的 Transformer 模块进行处理，以得到最终输出。最终表征是一个包含 1500 个向量的矩阵，其维度取决于模型规模。为了使所有 Whisper 规模具有统一的维度，我们应用了一个全连接投影层。该层将编码器输出映射到固定的模型维度，然后对每个表征进行归一化。在我们的实验中，我们将 d_{model} 设为 256，因此在所有情况下都会得到用于分类的表示矩阵$\mathbb{R}^{1500\times 256}$ 。

$$H = \text{Projector}(\text{WhisperEncoder}(X)), \qquad H \in \mathbb{R}^{1500\times 256} \tag{2}$$

3.2 注意力池化

为了使 Whisper 表征适用于分类，我们需要将其转换为一维向量，这会导致一些信息损失。我们提出了两种

转换方法：多头注意力平均池化和多头 QKV 池化。在我们的实验中，QKV 池化的表现通常优于注意力平均池化，但在 IEMOCAP 上的加权准确率除外。为保持论文内容的连贯性和组织性，本节介绍这两种方法。

$$a = \text{Pooling}(H), \quad a \in \mathbb{R}^{256} \tag{3}$$

在公式2中， H 是为每个样本生成的最终模型表征，而 a 是经过池化后输入分类器的最终向量。

3.2.1 多头注意力平均池化

在平均池化中，话语的所有帧具有相同的权重，因此对情感预测产生相同的影响。然而，在许多情况下，话语中的某些帧比其他帧包含更多用于预测情感的信息特征。为了为每一帧赋予适当的权重，我们采用了注意力统计池化，这是一种由Okabe 等人于 (2018年)首次提出的基于注意力的池化方法。该机制通过结合统计池化与注意力机制，为话语中的不同帧分配重要性，从而生成更稳健的说话人嵌入。

先前的研究 Nasersharif 和 Namvarpour(2024)；Okabe 等人 (2018)；Nazari 等人 (2025年)通过将注意力统计池化应用于分类，取得了显著成果。在本研究中，我们通过引入多头注意力机制来改进缩放准确率，从而扩展了这一方法。具体而言，每个注意力头都与一个独立的小型神经网络相关联，该网络经过训练以计算每一帧的权重。

与 Okabe 等人 （2018年）以及 Nasersharif 和 Namvarpour （2024）的做法相同，我们将注意力网络描述为把一个值向量映射为一个输出，并将其所有元素分别乘以一个注意力权重。

$$e_t = f(\mathbf{h}_t W_1^{attn} + \mathbf{b})\mathbf{w}_2^{attn} + k \tag{4}$$

，其中 $W_1 \in \mathbb{R}^{d_{model} \times d_{hidden}}$ 和 $\mathbf{w}2 \in \mathbb{R}^{d_{hidden}}$。 f（。）是双曲正切激活函数，其输出会经过丢弃层，以防止过拟合。注意力权重通过 Softmax 函数计算。它并非对 d_{hidden} 维值向量及注意力权重执行单一注意力函数，而是对一组打包在一起的查询进行计算，这些查询构成一个 V 矩阵，而打包在一起的注意力权重则构成 $\mathbf{e}$ 向量。

$$\text{Average Pooling}(\mathbf{e}, V) = \text{Softmax}\left(\mathbf{e}^T\right) V \tag{5}$$

V 矩阵也是通过将表示矩阵中的每个向量映射为一个大小为 d_{hidden} 的向量而形成的。所有这些向量组合在一起，构成一个大小为 $\mathbb{R}^{1500 \times d_{hidden}}$ 的 V 矩阵。

$$V = HW^V, \quad W^V \in \mathbb{R}^{d_{model} \times d_{hidden}} \tag{6}$$

上述过程采用 4 个注意力头的方式执行 4 次，每次使用不同的 W^V。各注意力头的输出向量拼接在一起，再映射回 d_{model}，从而得到池化向量 a。

$$\text{Head}_i = \text{Average Pooling}(\mathbf{e}_i, V_i) \qquad \text{Concat}(\text{Head}_1, \ldots, \text{Head}_n)W^{out} \tag{7a}$$

$$\mathbf{a} = \text{MultiHead}(\text{Head}_1, \ldots, \text{Head}_n), \qquad \mathbf{a} \in \mathbb{R}^{d_{model}} \tag{7b}$$

，其中 $W^{out} \in \mathbb{R}^{(\#heads.d_{hidden})\times d_{model}}$ ，并将池化向量的维度映射到 d_{model}

3.2.2 多头 QKV 注意力池化

在该方法中，我们将多头 QKV 注意力机制融入池化方法，以减少降维过程中的信息损失。在 Vaswani 等人提出的 QKV 注意力机制中，Vaswani (2017)，通过对 Q和 K 进行缩放点积运算，为每个向量生成注意力权重，其中 Q 由先前生成的词元导出。然而，在我们的分类任务中，没有可用的先前输出。为此，我们采用了 CLIP 中使用的池化方法，Radford 等人(2021)，以音频样本的全局平均池化表示为查询提供条件。

$$\text{QKV Attention}(\mathbf{q}, K, V) = \qquad \text{Softmax}\left(\frac{\mathbf{q}K^T}{\sqrt{d_k}}\right)V \tag{8a}$$

$$\mu(H) = \frac{1}{1500}\sum_{i=1}^{1500}\mathbf{h}_i, \quad \mu \in \mathbb{R}^{256} \tag{8b}$$

$$\text{Head}_i = \text{QKV Att}(\mu(H)W_i^Q, HW_i^K, HW_i^V) \tag{8c}$$

3.2.3 分类器

降维后，得到的池化向量（a）为 256 维，并被传入分类层。分类器头是一个权重随机初始化的全连接层，它根据数据集中的情绪类别数量，将该向量映射为较低维度的输出。例如，在 IEMO-CAP 中，使用四种情绪类别：愤怒、快乐、悲伤和中性。通过 Softmax 激活函数，预测出最有可能的情绪。

$$\text{Emotion} = \text{Softmax}(\text{FC}(a)), \quad \text{Emotion} \in \mathbb{R}^4 \tag{9}$$

4 实验与结果

4.1 数据集

ShEMO，即 Sharif 情感语音数据库，包含时长 3 小时 25 分钟的在线广播剧，共有来自 87 名波斯语使用者的 3,000 条话语。

每条话语均被标注为六种情绪之一，包括愤怒、恐惧、快乐、悲伤、惊讶和中性，其中愤怒和中性占话语总数的 70%，悲伤占 15%，快乐和惊讶合计不足 15%。恐惧是规模最小的类别，因此按照以往研究的做法将其忽略。这些话语被划分为 10 组，每组均由不同的说话人组成。

IEMOCAP，即交互式情感双人动作捕捉数据集，是语音情感识别领域广泛使用的基准数据集。它包含约 12 小时的双人视听数据，来自 10 名演员（五对男女演员），录制于五个会话中。与以往研究一致，我们将分析范围限定为标注了四种情绪类别的即兴话语：愤怒、快乐（将兴奋和开心合并）、悲伤和中性。其余所有情绪标签均被排除，最终得到包含 2,793 条话语的数据集。评估时，我们采用五折交叉验证：每一折中留出一个会话用于测试，其余四个会话用于训练。这一成熟的方法确保我们的结果可与其他使用 IEMOCAP 数据集的研究直接比较。

4.2 实验设置

模型使用 PyTorch 框架实现，并从 Hugging Face 导入预训练的 Whisper 权重。在训练和测试阶段，IEMOCAP 和 ShEMO 数据集均被随机划分为批次大小为 16 的批次。根据数据集的划分方式，我们分别对 IEMOCAP 和 ShEMO 执行五折交叉验证和十折交叉验证。

超参数通过反复试验的方式确定。所有模型均训练 30 个训练轮次。对于 ShEMO，两种池化方法均使用 6 个注意力头，每个头的隐藏层大小为 4。对于 IEMOCAP，注意力头的数量设为 12，每个头的隐藏层大小为 4。采用 AdamW 优化器来优化可训练权重，因为其表现出色（Nasersharif 和 Namvarpour(2024）。使用余弦调度器调整学习率，学习率峰值为 10e−4，并设置 10% 的预热步数。

4.3 池化方法的结果

为确保与语音情感识别领域以往研究的可比性，选取两种准确率指标进行评估，即加权准确率和非加权准确率。

非加权准确率定义为各类别召回率的平均值，能够提供更均衡的性能衡量，尤其适用于 ShEMO 等测试集类别分布不均衡的数据集。

所提出的分类框架使用两种不同规模的 Whisper 模型作为表征提取器进行训练和测试。具体而言，我们采用了 Whisper Tiny 和 Whisper Small 来比较分类结果。

在第 3.2 节中，我们提出了两种基于注意力的池化方法。表 1 展示了它们与作为基线的全局平均池化相比的结果。

表 1：多头注意力平均池化（注意力）与多头 QKV 池化（QKV）的性能比较。均值平均池化（均值）作为基线模型。使用 Whisper Tiny 和 Whisper Small 编码器，在 ShEMO 和 IEMOCAP 数据集上评估了这三种池化方法。**WA** 表示加权准确率，**UA** 表示非加权准确率。

Size	池化	ShEMO		IEMOCAP	
		WA	UA	WA	UA
Tiny	Mean	84.12 ± 3.60	73.66 ± 3.87	68.22 ± 2.09	68.53 ± 2.35
	注意力	84.71 ± 4.59	74.71 ± 4.74	68.55 ± 1.79	68.67 ± 1.89
	QKV	**84.81** ± 3.64	**75.14** ± 4.24	**69.37** ± 1.84	**69.38** ± 1.98
小型	Mean	88.81 ± 3.26	82.41 ± 4.02	70.52 ± 1.73	70.61 ± 2.39
	注意力	88.94 ± 2.11	82.86 ± 4.88	**71.98** ± 1.75	72.64 ± 2.22
	QKV	**89.19** ± 2.65	**83。07** ± 4。99	71.82 ± 0.22	**72.96** ± 2.22

不出所料，基于注意力的池化方法优于简单的平均处理。这是因为音频样本中的情感特征分布并不均匀，有些帧承载的情感线索比其他帧更突出。与传统池化方法不同，利用注意力的方法能够应对特征提取中的这种差异。

QKV池化在 ShEMO 和 IEMOCAP 数据集上均取得了最高准确率，并且在不同模型规模下始终保持优异表现。然而，基于QKV的模型在较低层的准确率有所下降，这表明其性能提升主要来自高层表示（第 4.6节）。

4.4 模型规模因素

如表1所示，结果表明，使用从较大的 Whisper 模型中提取的表征时，该框架能够取得更优的分类性能。这是可以预料的，因为在 Whisper 的主要任务自动语音识别（ASR）中，Small模型的词错误率也低于Tiny模型，因此它能从输入语音中生成更好的表征。

ShEMO 上 Whisper Tiny 和 Whisper Small 编码器在多头注意力平均池化（AttW）和多头 QKV 池化（QKV）方法上的性能比较。雷达图显示，Whisper Small 的表现始终优于 Whisper Tiny，在样本较少的类别中提升尤为明显。

ShEMO的未加权准确率显著提升；与IEMOCAP相比，ShEMO是一个更不平衡的数据集，这表明更大的Whisper模型性能提升可归因于其更有效地对出现频率较低的类别进行分类的增强能力。如图2所示，开心和悲伤类别的准确率提升更为显著，相比

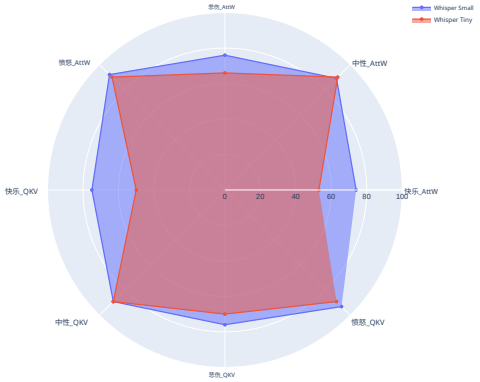
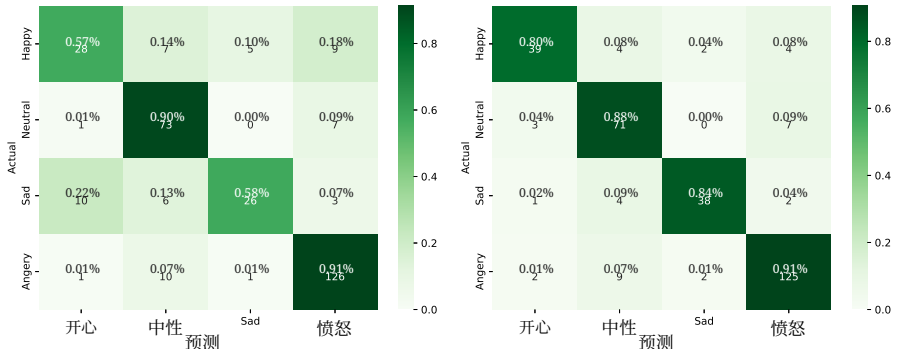


图 2：ShEMO 上 Whisper Tiny 和 Whisper Small 编码器采用多头注意力平均池化（AttW）和多头 QKV 池化（QKV）方法的性能比较。雷达图显示，Whisper Small 的表现始终优于 Whisper Tiny，在样本较少的类别中提升尤为显著。

中性和愤怒类别的表现尤其如此。查看图 3中的混淆矩阵时，也能观察到这一趋势。



(a) Whisper Tiny　　(b) Whisper Small

图 3：使用不同规模的 Whisper 进行表示提取时对 ShEMO 进行分类的混淆矩阵。

表 2：我们在 ShEMO 和 IEMOCAP 数据集上的主要结果与先前工作的比较。最佳结果以**粗体**显示。

数据集	基础模型	WA	UA
ShEMO	Wav2vec 2.0 大型版 Nasersharif 和 Namvarpour (2024)	86.80	80.60
	HuBERT Large Ma 等人 (2024)	83.35	64.29
	WavLM Large Ma 等人 (2024)	87.13	71.72
	Data2vec 2.0Large Ma 等人 (2024)	82.68	64.09
	Whisper Large V3 Ma 等人 (2024)	**89.55**	80.23
	Whisper 小型（我们的方法）	89.19	**83.07**
IEMOCAP	Whisper Large V3 Ma 等人 (2024)	72.86	73.54
	HuBERT Large Ma 等人 (2024)	63.10	63.87
	WavLM Large Ma 等人 (2024)	69.07	69.47
	Data2vec 2.0Large Ma 等人 (2024)	56.23	57.30
	HuBERT X-Large Jin 等人 (2025年)	70.65	71.46
	HuBERT X-Large Jiao 等人(2024)	**74.24**	**74.57**
	Whisper 小型（本文）	71.98	72.96

4.5 成本-准确率权衡

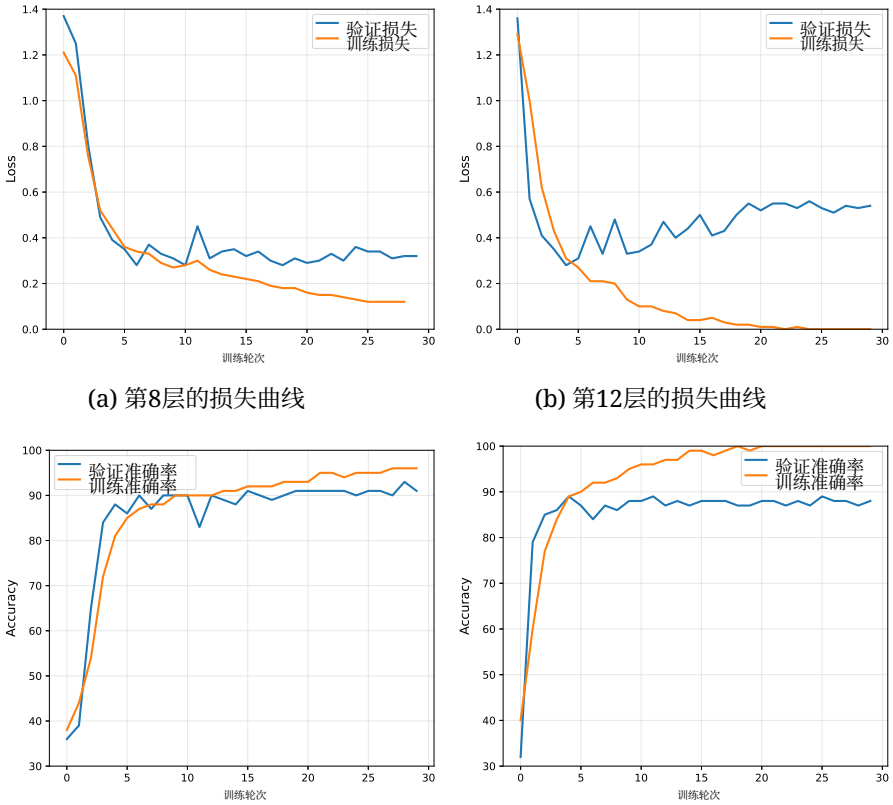
如表2所示，我们在两个数据集上取得的结果与最新的最先进水平（SOTA）研究相当。据我们所知，ShEMO 上的最高结果由 EmoBox 报告Ma 等人(2024)，而 Nasersharif 等人Nasersharif 和 Namvarpour(2024)报告的非加权准确率略高。然而，我们提出的多头 QKV 池化方法使用更小的模型取得了最佳非加权准确率，凸显了该方法在平衡性能与效率方面的有效性。

尽管与其他研究相比，我们在 IEMOCAP 上的结果未达到最先进水平，但我们认为该模型所需的计算成本显著更低。与最先进模型相比，准确率较低，但这可以由我们方法显著更高的效率和更轻量的架构来解释。需要指出的是，先前研究中用于提取表征的 HuBERT X-Large 模型，其参数量约为 Whisper Small 的十倍。因此，该模型在训练和推理过程中都需要显著更多的计算资源。

总体而言，我们的研究结果表明，QKV池化等高效的池化机制能够有效缩小较小型与较大型预训练模型之间的差距。不过，正如上一节对 Whisper Tiny 和 Whisper Small 的比较所示，模型大小会影响情感检测效果。因此，我们建议在较大型模型上采用基于注意力的池化方法，以进一步提升性能。

4.6 用于表征的不同 Whisper 编码器层

根据 Pepino 等人Pepino 等人（2021），预训练模型的不同层在训练过程中会捕捉不同类型的信息。虽然最后一个编码器层通常最适合为主要预训练任务生成表征，但中间层往往包含对语音情感识别等任务更有用的信息。



(a) 第8层的损失曲线

(b) 第12层的损失曲线

(c) 第8层的准确率曲线

(d) 第12层的准确率曲线

图4：比较使用 Whisper 编码器不同层作为 ShEMO 表征时的学习速度

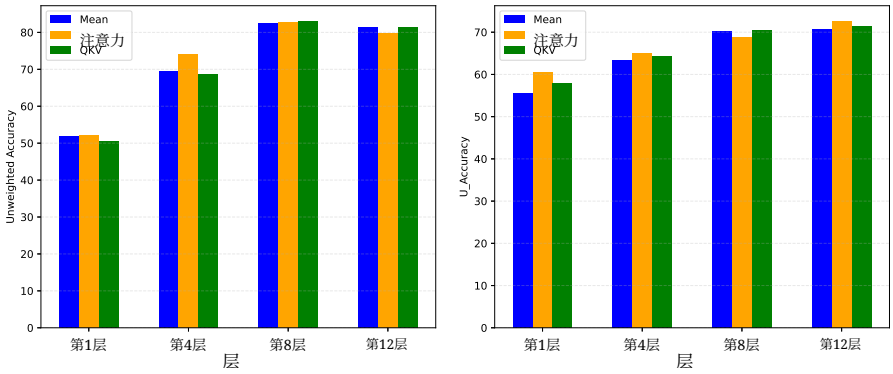
本研究使用的预训练模型 Whisper 经过多任务训练，可执行自动语音识别、语音活动检测（VAD）、语言检测和语音翻译等任务。该模型并未专门针对语音情感检测进行训练。此外，在多语言模型中，不同编码器层可能会根据输入语言习得不同类型的信息。

先前的研究已将 Whisper 用于语音分类和语音情感识别 Ma 等人(2024);Sankaran 等人(2025年)，但实验仅限于模型最后一层的表征。据我们所知，这是首项深入研究各层模型表征的工作。我们研究并报告了 Whisper Tiny 的全部四层以及 Whisper Small 的全部十二层的性能。

使用 Whisper 最后一层的模型能够更快地达到其最大学习潜力，但也往往会更早过拟合。相比之下，基于中间层和较低层的模型则需要更多训练轮次才能充分发挥潜力。

这一观察结果如图所示 4，并且在 IEMOCAP 和 ShEMO 数据集上均一致成立。

根据图5，对于 IEMOCAP 数据集，在大多数情况下，使用 Whisper 编码器最后一层的输出可获得最佳性能。然而，对于 ShEMO 数据集，使用中间层作为语音表征能够取得更好的结果。完整的实验结果见附录A。



(a) ShEMO (b) IEMOCAP

图 5：使用 Whisper Small 编码器的 4 个不同层时，均值平均池化（Mean）、多头注意力平均池化（Attentive）和多头 QKV 池化（QKV）的性能比较。

5 讨论与未来工作

在投影阶段，维度会大幅降低。例如，在小型模型中，通过一个权重随机初始化的全连接层，可将 768 维向量降至 256 维。随后，在分类阶段，再通过一个权重随机初始化的全连接层，将 256 维向量映射为 4 维输出。这会引入大量可训练参数，例如，在小型模型中，仅用于降维的权重就有 197,632 个。这些参数不仅需要随机初始化和训练，而且还需要比我们实验中可用数据集更大的数据集，同时也会增加模型开销。因此，采用更具创造性且高效的降维方法，可能有助于提高准确率并优化模型。

一个颇具前景的方向是集成自动语音识别系统，将语音转录为文本，并将文本作为额外模态用于情感识别。鉴于 Whisper 解码器本身能够将语音表征映射为文本，且受益于预训练权重，它或许可以免去使用额外嵌入模型（如 BERT）的需要。换言之，解码器会生成可直接用于情感分类的中间向量表示，从而避免

先生成文本，再借助独立语言模型将其嵌入的计算开销较大的两步流程。

另一个探索方向是多模态情感识别。例如，IEMOCAP 或 MELD 等数据集 Poria 等人 (2018年不仅包含语音数据，也包含视觉数据。利用面部手势作为辅助模态，可能有助于提升识别性能。

本研究仅使用 Whisper 各编码器层的输出作为独立的语音表征，并比较了它们的性能。由于部分层之间的性能差异很小，这表明每一层都编码了对情感分类有用但略有不同的信息。因此，设计一种基于注意力的机制，将所有编码器层的信息聚合到统一的表示矩阵中，或许能取得更好的结果。然而，这会增加模型复杂度，并需要更强大的硬件（例如显存更大的图形处理器）。

与依赖 HuBERT X-Large（10亿个参数）或 Wav2vec 2.0（3亿个参数）等模型，在 IEMOCAP 或 ShEMO 等数据集上进行语音情感识别的竞争方案相比，采用仅有 8800 万个参数的编码器的 Whisper-small 模型，是一种轻量但有效的替代方案。即使性能相当，使用 Whisper-small 也能提供更高效、更实用的解决方案。

此外，我们在 ShEMO 上的研究结果表明，Whisper 的中间编码器层比最后几层表现更好。这说明，中间层表征可作为有效的语音嵌入，同时降低计算成本，因为无需激活最后几层。这进一步提升了基于 Whisper 的语音情感识别系统的效率和实用性。

6 结论

在这项工作中，我们探索了如何弥补研究空白，找到一种能够避免语音表征中有价值信息丢失的池化方法。我们选择 Whisper，是因为它具有最先进性能和多语言能力，可从音频样本中生成表征。Whisper 的多语言能力也使我们能够为波斯语（这一低资源语言）和英语的语音情感识别作出贡献，从而既能将结果放在全球范围内比较，也能与更多研究进行对比。我们分别使用 ShEMO 和 IEMOCAP 作为波斯语和英语数据集。通过 Whisper 编码器获得音频表征后，我们应用了一种基于注意力机制的池化方法。第一种方法——多头注意力平均池化——在现有的注意力统计池化基础上，融入多头注意力机制，为话语中的不同帧分配重要性权重，因为有些帧所包含的情感预测相关特征比其他帧更丰富。第二种方法——多头 QKV 注意力池化——利用多头 QKV（查询、键、值）注意力机制进行池化；但由于分类任务缺少通常用于查询（Q）的预先生成词元，因此该方法以音频样本的全局平均池化表征为条件来确定查询。实验结果表明，在 ShEMO 数据集上，QKV 在非加权准确率方面取得了最先进水平；这对于 ShEMO 这类不平衡数据集尤为重要。此外，与

IEMOCAP 的最先进水平相比，我们的方法计算成本更低，在准确率与成本之间实现了良好的权衡。

参考文献

Baevski A, Hsu WN, Xu Q 等（2022）Data2vec：一种用于语音、视觉和语言自监督学习的通用框架。载于：国际机器学习会议，PMLR，第1298–1312页 Chen S, Wang C, Chen Z 等（2022）Wavlm：面向全栈语音处理的大规模自监督预训练。IEEE信号处理选题期刊 16(6):1505–1518 Chou HC（2024）A Tiny Whisper-Ser：统一自动语音识别与多标签语音情感识别任务。载于：2024 亚太信号与信息处理协会年度峰会暨会议（APSIPA ASC），IEEE，第1–6页

Dutt A, Gader P（2023）基于小波多分辨率分析的语音情感识别系统，采用1D CNN LSTM网络。IEEE/ACM音频、语音与语言处理汇刊 31:2043–2054

Feng T, Narayanan S (2024) 基础模型辅助的自动语音情感识别：转写、标注与增强。见：ICASSP 2024-2024 IEEE国际声学、语音与信号处理会议（ICASSP），IEEE，第12116–12120页

Gupta V, Juyal S, Hu YC (2022年) 利用深度神经网络通过语音频谱图理解人类情绪。《超级计算杂志》78(5):6944–6973

Jiao W, Yang H, King I, 等 (2019年) HiGRU：用于话语级情感识别的分层门控循环单元。arXiv预印本 arXiv:190404446

Jiao W, Lyu M, King I (2020年) 基于注意力门控分层记忆网络的实时情感识别。见：AAAI人工智能会议论文集，第8002–8009页

Jiao X, Wang L, Yu Y (2024) MFHCA：通过多空间融合与分层协同注意力提升语音情感识别。arXiv 预印本 arXiv:240413509 Jin G, Xu Y, Kang H 等 (2025) DSTM：一种采用动态—静态特征融合的基于 Transformer 的语音情感识别模型。《计算机语音与语言》90:101733 Livingstone SR, Russo FA (2018) 瑞尔森情感语音与歌曲视听数据库（RAVDESS）：一个涵盖北美英语面部与声音表情的动态多模态数据集。《公共科学图书馆·综合》13(5):e0196391

Ma Z, Chen M, Zhang H 等 (2024) EmoBox：多语言多语料语音情感识别工具包与基准。arXiv 预印本 arXiv:240607162 Majumder N, Poria S, Hazarika D 等 (2019) DialogueRNN：一种用于对话情感检测的注意力循环神经网络。载于：AAAI人工智能会议论文集，第 6818–6825 页 Mohamad Nezami O, Jamshid Lou P, Karami M (2019) ShEMO：一个经验证的大规模波斯语语音情感检测数据库。《语言资源与评估》53(1):1–16 Nasersharif B, Namvarpour M (2024) 探索 Wav2vec 2.0 在通过分类器组合和基于注意力的特征融合进行语音情感识别中的潜力。《超级计算杂志》80(16):23667–23688 Nazari R, Salehi M, Shoeibi A (2025) 一种用于多模态自闭症谱系障碍分类的可解释连接组卷积 Transformer。《国际神经系统杂志》35(8):2550043–2550043 Nimmagadda R, Arora K, Martin MV (2022) 面向陪伴机器人的情感识别模型。《超级计算杂志》78(11):13710–13727 Okabe K, Koshinaka T, Shinoda K (2018) 用于深度说话人嵌入的注意力统计池化。arXiv 预印本 arXiv:180310963 Pepino L, Riera P, Ferrer L (2021) 使用 wav2vec 2.0 嵌入进行语音情感识别。arXiv 预印本 arXiv:210403502 Poria S, Hazarika D, Majumder N 等 (2018) MELD：一个用于对话情感识别的多模态多方数据集。arXiv 预印本 arXiv:181002508 Qu X, Sun Z, Feng S 等 (2024) 打破沉默：Whisper 驱动的人工智能心理支持模型中的情感识别。载于：2024 IEEE 人工智能大会（CAI），IEEE，第 290–291 页 Radford A, Kim JW, Hallacy C 等 (2021) 通过自然语言监督学习可迁移视觉模型。载于：国际机器学习会议，PMLR，第 8748–8763 页 Radford A, Kim JW, Xu T 等 (2022) 通过大规模弱监督实现鲁棒语音识别。网址https://arxiv.org/abs/2212.04356，arXiv:2212.04356Sankaran AN, Farahbakhsh R, Crespi N (2024) 面向低资源环境下基于少样本学习的跨语言音频滥用检测。arXiv 预印本 arXiv:241201408 Sankaran AN, Farahbakhsh R, Crespi N (2025) 面向低资源环境下基于少样本学习的跨语言音频滥用检测。载于：第 31 届国际

计算语言学会议

Vaswani A（2017年）Attention Is All You Need。神经信息处理系统进展

Zhang S、Zhao X、Tian Q（2019年）使用多尺度深度卷积长短期记忆网络进行自发语音情感识别。IEEE 情感计算汇刊 13(2)：680–688

Zhao J、Mao X、Chen L（2019年）使用深度 1D 和 2D CNN LSTM 网络进行语音情感识别。《生物医学信号处理与控制》47：312–323

附录A Whisper 编码器层的结果

在表 3 和 4中，我们列出了 Whisper 模型中每个 Transformer 层的加权准确率和非加权准确率。我们分别使用了 Whisper 小型版和 Whisper Large，二者分别包含 4 个和 12 个 Transformer 块。

表 3：Whisper Tiny模型配置在 IEMOCAP 和 ShEMO 数据集上，不同池化方法和层的性能比较。

	IEMOCAP			ShEMO		
层	WA	UA	F1	WA	UA	F1
Whisper Tiny +平均池化						
1	54.62 ± 2.27	55.40 ± 2.37	54.06 ± 2.23	70.94 ± 5.25	50.47 ± 2.87	67.08 ± 6.15
2	60.85 ± 1.30	61.76 ± 1.54	60.08 ± 1.65	80.88 ± 4.20	66.94 ± 4.82	79.54 ± 4.72
3	66.44 ± 1.93	66.92 ± 2.62	66.11 ± 2.07	83.26 ± 4.00	71.74 ± 3.78	82.48 ± 4.27
4	68.22 ± 2.09	68.53 ± 2.35	67.98 ± 2.13	84.12 ± 3.60	73.66 ± 3.87	83.47 ± 3.77
Whisper Tiny +注意力平均池化						
1	58.42 ± 2.46	58.88 ± 2.80	57.72 ± 2.89	78.35 ± 4.13	60.22 ± 6.33	76.42 ± 4.46
2	64.83 ± 1.84	64.84 ± 2.83	64.25 ± 2.11	80.33 ± 3.72	64.73 ± 6.21	78.98 ± 4.10
3	66.67 ± 1.44	67.47 ± 1.86	66.42 ± 1.66	84.71 ± 4.59	74.71 ± 4.74	84.29 ± 4.53
4	68.57 ± 2.17	68.69 ± 2.18	68.50 ± 2.42	83.09 ± 3.56	73.51 ± 3.48	82.78 ± 3.65
Whisper Tiny +QKV池化						
1	54.74 ± 2.22	56.43 ± 0.92	53.45 ± 2.65	73.95 ± 5.65	52.78 ± 4.72	70.24 ± 6.25
2	61.80 ± 1.60	62.51 ± 2.56	61.22 ± 1.97	79.89 ± 4.39	63.67 ± 6.01	78.06 ± 5.19
3	66.94 ± 1.41	67.88 ± 1.77	66.58 ± 1.79	84.81 ± 3.64	74.90 ± 4.48	84.26 ± 4.06
4	69.37 ± 1.84	69.38 ± 1.98	69.31 ± 1.91	83.32 ± 3.86	75.14 ± 4.24	82.97 ± 3.88

表 4：Whisper Small 模型配置在 IEMOCAP 和 ShEMO 数据集上采用不同池化方法和层的性能对比。

	IEMOCAP			ShEMO		
层	WA	UA	F1	WA	UA	F1
Whisper Small + 平均池化						
1	55.58 ± 1.30	55.54 ± 2.06	55.29 ± 1.37	73.20 ± 6.78	51.97 ± 6.86	69.41 ± 7.98
2	58.42 ± 1.64	59.90 ± 2.58	57.79 ± 1.86	77.15 ± 5.99	58.08 ± 4.33	74.57 ± 6.66
3	59.88 ± 1.64	59.94 ± 1.65	59.27 ± 2.21	79.54 ± 5.51	62.41 ± 7.58	77.35 ± 6.51
4	62.43 ± 1.81	63.43 ± 1.76	61.93 ± 2.27	82.53 ± 4.55	69.48 ± 4.44	81.43 ± 5.13
5	66.68 ± 1.26	67.48 ± 2.49	66.42 ± 1.18	85.20 ± 3.96	76.03 ± 4.16	84.71 ± 4.23
6	67.77 ± 1.65	67.42 ± 2.52	67.78 ± 1.69	85.89 ± 3.37	77.30 ± 3.32	85.65 ± 3.44
7	68.27 ± 1.48	67.95 ± 1.83	67.99 ± 1.70	88.19 ± 2.46	79.84 ± 4.58	87.84 ± 2.53
8	69.99 ± 1.26	70.24 ± 1.84	69.88 ± 1.26	88.81 ± 3.26	82.41 ± 4.02	88.59 ± 3.25
9	70.05 ± 1.79	70.16 ± 2.37	69.94 ± 1.95	87.41 ± 3.76	78.40 ± 7.80	87.15 ± 3.86
10	70.52 ± 1.73	70.61 ± 2.39	70.43 ± 1.76	87.07 ± 3.52	77.92 ± 7.15	86.64 ± 3.73
11	70.65 ± 1.35	70.66 ± 2.35	70.56 ± 1.45	86.65 ± 3.35	77.55 ± 5.86	86.22 ± 3.47
12	70.69 ± 1.58	70.78 ± 2.37	70.62 ± 1.05	87.61 ± 3.15	81.65 ± 5.71	87.56 ± 3.09
Whisper Small + 注意力平均池化						
1	59.07 ± 2.73	60.47 ± 1.83	58.13 ± 3.44	73.15 ± 6.61	52.13 ± 5.49	69.62 ± 7.42
2	61.75 ± 1.49	62.54 ± 2.11	61.36 ± 1.64	77.55 ± 5.87	58.66 ± 5.28	74.71 ± 6.70
3	63.23 ± 1.51	64.32 ± 1.52	62.48 ± 2.26	80.69 ± 4.96	65.48 ± 4.18	78.87 ± 5.84
4	64.96 ± 1.48	65.03 ± 2.89	64.51 ± 1.57	83.18 ± 4.77	74.19 ± 3.58	82.80 ± 4.80
5	66.83 ± 0.87	67.70 ± 1.44	66.61 ± 0.82	85.19 ± 3.40	76.50 ± 4.61	84.84 ± 3.54
6	67.51 ± 2.22	67.46 ± 2.06	67.53 ± 2.53	85.98 ± 2.67	78.02 ± 5.47	85.75 ± 2.86
7	69.36 ± 2.09	70.81 ± 2.63	69.28 ± 2.14	88.31 ± 2.74	81.89 ± 3.18	88.12 ± 2.78
8	68.79 ± 1.97	68.70 ± 2.25	68.66 ± 1.94	88.94 ± 2.11	82.86 ± 4.88	88.79 ± 2.18
9	68.29 ± 2.26	68.79 ± 2.49	68.05 ± 2.23	88.63 ± 3.16	82.55 ± 5.21	88.44 ± 3.20
10	69.35 ± 2.43	69.68 ± 2.84	69.21 ± 2.45	87.02 ± 2.96	79.17 ± 5.44	86.59 ± 3.37
11	71.50 ± 3.09	71.11 ± 3.09	71.47 ± 3.14	88.19 ± 2.66	80.22 ± 6.32	88.02 ± 2.73
12	71.98 ± 1.75	72.64 ± 2.22	71.89 ± 1.64	86.71 ± 2.71	79.75 ± 3.60	86.44 ± 2.66
Whisper Small + QKV池化						
1	55.85 ± 1.77	57.82 ± 2.13	54.81 ± 2.05	73.06 ± 6.32	50.59 ± 5.55	68.90 ± 6.77
2	58.15 ± 2.00	59.89 ± 2.89	57.61 ± 2.23	76.91 ± 6.11	57.21 ± 4.31	73.91 ± 7.07
3	60.37 ± 1.75	62.17 ± 2.64	59.72 ± 2.21	78.80 ± 5.89	61.67 ± 5.86	76.58 ± 7.21
4	62.86 ± 1.58	64.40 ± 1.92	62.52 ± 1.89	82.43 ± 5.26	68.67 ± 6.71	81.07 ± 5.98
5	66.31 ± 0.91	67.87 ± 2.19	65.84 ± 1.01	85.24 ± 4.34	75.83 ± 5.98	84.95 ± 4.43
6	67.90 ± 2.29	68.20 ± 2.86	67.90 ± 2.37	85.84 ± 3.80	78.17 ± 4.53	85.61 ± 3.93
7	69.46 ± 1.46	69.23 ± 2.23	69.30 ± 1.47	88.16 ± 2.88	81.78 ± 3.04	88.01 ± 2.83
8	70.15 ± 1.42	70.55 ± 2.06	69.95 ± 1.45	89.19 ± 2.65	83.07 ± 4.99	89.04 ± 2.62
9	70.66 ± 1.34	70.70 ± 1.70	70.67 ± 1.36	88.45 ± 3.47	82.75 ± 4.70	88.35 ± 3.47
10	71.83 ± 1.44	72.04 ± 1.19	71.74 ± 1.45	88.55 ± 3.29	82.85 ± 3.07	88.40 ± 3.27
11	71.79 ± 2.20	72.96 ± 2.22	71.50 ± 2.25	88.02 ± 3.13	78.76 ± 6.90	87.63 ± 3.32
12	71.04 ± 1.61	71.51 ± 1.93	70.96 ± 1.46	87.34 ± 2.43	81.44 ± 3.42	87.17 ± 2.38