arXiv:2302.00765v1 [cs.CL] 1 Feb 2023

面向低资源语言的视觉接地关键词检测与定位

作者：Kayode
Kolawole Olaleye

提交于斯泰伦博斯大学工程学院，作为电子工程哲学博士学位的学位论文

导师：Herman Kamper 教授

2023年3月

致谢

我谨向导师 Herman Kamper 副教授致以最诚挚的谢意，感谢他在我的博士求学期间给予我无可比拟的支持与指导。他协助争取到的资助使我能够全心投入学业，而在困难时期，他的指导和沉稳的陪伴更是弥足珍贵。

我由衷感谢 Dan Oneat˘a 博士担任我的非正式联合导师，并为我的研究作出宝贵贡献。他的指导与支持对我的学位论文取得成功至关重要，我也真心感谢他为我的研究投入的时间和精力。

我要感谢德国学术交流中心、谷歌和 MIH Media Lab 为我的学业提供资助，也要特别感谢 Bruce Basset 教授、Aerobotics 的 Benji Meltzer、Olabode Koriko 教授、Babatope Omolofe 教授以及我的导师 Herman Kamper，感谢他们在我申请资助时给予的支持。我还要感谢 Marc Deisenroth 教授，让我有机会在博士学习之初参加于伦敦举办的机器学习暑期学校；那是一次令人难忘的学习经历。

我也感谢 LSL研究组的历届成员和现任成员 Benjamin van Niekerk、Leanne Nortje、Christiaan Jacobs、Lisa van Staden、Matthew Baas、Werner van der Merwe 和 Kevin Eloff，感谢他们在我研究的不同阶段给予的友谊、支持和宝贵贡献。特别感谢 Benjamin 为部分实验提供的有益意见，以及那些帮助我应对倦怠的（疗愈性）棋局；感谢 Leanne 的鼓励和美味蛋糕；感谢 Christiaan 始终如一的友谊与鼓励；也感谢 Matthew 帮我解决了硬盘问题。

我还要感谢我的考官 Willie Brink 教授、Murat Sara¸clar 教授和 Richard Klein 副教授，以及担任我口头答辩主持人的 Herman Engelbrecht 教授，感谢他们在答辩期间营造了轻松自在的氛围。我还要感谢 MIH Media Lab 的全体成员，感谢他们为我的研究营造了温馨的氛围。

我将永远感激我的父母 Olaleye 牧师和夫人，以及我的姐妹 Tolani Olaleye 和 Teniola Akinsanya 一直以来给予我的支持与鼓励。我还要感谢我的妻子 Olubunmi Olaleye 对我的爱与支持。

最后，我要感谢阿爸天父，在这段旅程中一直作我的来源和供养。

Stellenbosch UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

forward together
sonke siya phambili
saam vorentoe

剽窃声明 / 剽窃声明

1. 剽窃是指挪用并使用他人的思想、材料及其他知识产权，仿佛这些都是自己的作品。
剽窃是使用他人作品中的思想、材料及其他知识产权，并将其呈现为自己的内容。

2. 我承认，剽窃是一种可处罚的违法行为，因为它属于盗窃。

我同意，剽窃是一种可处罚的违法行为，因为它构成盗窃。

3. 我也明白，直接翻译属于剽窃。我也明白，直接翻译属于剽窃。

4. 因此，所有来自任何来源（包括互联网）的引文和贡献均已完整注明（致谢）。我承认，逐字引用文本却不加引号（即使已完整注明来源）属于剽窃。

因此，所有引文以及来自任何来源（包括互联网）的贡献均已完整注明出处。我明白，不使用引号复制文本（即使注明了来源）即属剽窃

5. 我声明，除非另有说明，本书面作业所载内容均为我自己的原创作品，而且我此前未曾将全部或部分内容提交，以供本课程模块/作业或其他课程模块/作业评分。

我声明，除非另有说明，本作业中包含的内容均为我的原创作品，且我此前未曾将其全部或部分提交用于本课程/作业或其他课程/作业的评分。

学号 / 学号 **21524319**	签名 / 签名 
姓名首字母和姓氏 / 姓名首字母和姓氏 **K.K. Olaleye**	日期 / 日期 **2023年2月3日**

摘要

英语

视觉接地语音（VGS）模型是在与未标注口语描述配对的图像上进行训练的。这类模型可用于在无法获取转录数据的环境中构建语音系统，例如记录无文字语言。我们在真实低资源设置下，研究使用经过训练的视觉接地语音模型进行语音中的关键词定位——即找出给定书面关键词在一段话语中的出现位置。现有的视觉接地语音研究在两个方面存在不足。首先，已有研究表明，视觉接地语音模型可用于跨模态检索、关键词检测和关键词检出等任务，但尚未探索关键词定位。其次，大多数以往的视觉接地语音研究使用的数据集都将图像与英语（或其他资源丰富语言）的语音配对。因此，英语常被用作低资源语言的替代语言，这使得在真实低资源设置下准确评估其性能变得困难。

基于此，我们探讨以下两个总体研究问题：(i) 能否使用视觉接地语音模型进行关键词定位？(ii) 在真实低资源设置中，我们能否跨语言实现视觉接地关键词定位？

为回答第一个问题，我们扩展并增强现有的视觉接地语音模型，使其不仅能够检测书面关键词，还能对其进行定位。针对这一研究问题，我们将研究范围限定在人工低资源环境中，使用英语视觉语音接地数据，以便与以往工作进行比较并直接扩展其成果。我们以一种现有的视觉接地语音模型训练方法为起点，该方法用于检测语音中的关键词：使用现有的离线图像标注器为训练图像添加软文本标签，再将这些标签作为目标来训练语音网络。也就是说，模型接收一个噪声目标，指示某些单词是否出现在某段话语中，但不指示它们出现的位置或顺序。我们采用四种定位方法扩展该模型。输入屏蔽会在不同位置屏蔽输入信号，并测量特定关键词对应输出单元的变化。注意力定位需要一个沿时间轴池化特征的注意力层；我们将注意力权重用作定位分数。Grad-CAM 是一种基于显著性的方法，可应用于任何卷积神经网络，以确定网络输入的哪些部分对特定输出决策的贡献最大。分数聚合方法采用一种特定类型的池化，使输出分数可视为局部分数的聚合；这些局部分数可用于选出查询关键词最可能出现的时间位置。在预言机定位测试中（模型已获知某个关键词出现在一段话语中，随后被要求指出其出现位置），基于掩蔽的定位

方法的准确率达到 57.0%，优于所有其他方法；基于注意力的方法以 46.0% 位居第二。

为了解决第二个研究问题（真实低资源设置下的跨语言关键词定位），我们首先收集并发布一个新的 VGS 数据集。Yor`ub´a Flickr 音频描述语料库（YFACC）数据集包含单一说话人用 Yor`ub´a 语录制的 6k 张 Flickr 图像的口语描述。Yor`ub´a 是尼日利亚使用的一种真实低资源语言。基于这些数据，我们研究跨语言关键词检测与定位问题：给定一个英语文本查询，我们检测该查询是否出现在 Yor`ub´a 语音中；若检测到，则定位该查询在话语中的具体位置。为构建这一 VGS 系统，我们自动为图像添加英语视觉标签，并将其作为基于注意力的模型的目标，该模型以 Yor`ub´a 语音为输入。随后，我们首次将基于注意力的定位方法应用于真实低资源设置下的跨语言关键词检测与定位。该跨语言模型在实际关键词定位任务中取得了 16.0% 的精确率；该任务需要先检测关键词是否出现，再进行定位。尽管单独来看这一结果并不突出，但该模型的训练未使用任何英语—Yor`ub´a 平行数据或任何转录文本。我们发现，若从基于英语图像—语音数据集预训练的模型初始化跨语言模型，性能便可提升至 22.8%。

为回答两个主要研究问题，我们做出了以下具体贡献：(1) 我们提出了一种新的VGS模型，利用注意力机制进行关键词检测和关键词检出，并与现有的基于VGS的方法进行了全面比较。(2) 我们为VGS模型扩展了四种定位方法。(3) 我们进行了详细的定量和定性分析，揭示上述模型的局限性，并展示其成功与失败模式。我们发现，在系统词汇表中的67个关键词里，有些词的定位结果良好（黑色、泳池、足球、树），而另一些词则会与语义相关词混淆：海洋 $\rightarrow$冲浪者；球 $\rightarrow$足球；游泳 $\rightarrow$泳池。(4) 我们发布了一个新的多模态多语言数据集，使得在真实低资源设置下进行VGS建模成为可能，并模拟了语言记录场景。该数据集在Flickr8k 图像-文本数据集的基础上，加入了约鲁巴语口语描述。 (5) 我们使用新建的约鲁巴语语音-图像数据集，提出了一个用于真实低资源设置下跨语言关键词检测和关键词定位的系统。(6) 我们对跨语言视觉接地语音模型进行了全面分析。我们发现，有些关键词的性能良好，例如棕色 ($b´ur´a`un$; 100。0% 精确率)、自行车 ($k`e.k´e.$；94。1%)和草($kor´ıko$ 90。9%)。但对于许多其他关键词，模型表现不佳，原因是视觉定位较差，且容易混淆语义相关的概念。

总而言之，我们展示了视觉接地语音模型可以在真实低资源设置下用于有限形式的关键词定位。我们希望，我们的新数据集和新发现能够推动更多关于将视觉接地语音模型用于真实低资源语言的研究。

约鲁巴语

A m´a ´ns.e `ıd´anil´e.k`o. ´o. f´un `awo. n `e.ro. `aw`os.e `o. r`o. `afe.nuso. t´ı a f`aw`or´an dar´ı n´ıpa l´ılo `awo. n`aw`or´an t´ı a so p`o. p`e.l´u `awo. n `ap`ej´uw`e´e `afe.nuso. wo. n t´ı k`o n´ı k´ ıko.. Ir´u `awo. n `e.ro. b´e.`e. j´e. k´ı ´o s.e´es.e l´ati k´o. `e.ro. `o. r`o. `afe.nuso. n´ı ibi t´ı k`o s´ı ` o. p`o.lo. p`o. d´at`a (o.ro. `afe.nuso. ati `afo. w´o. ko. wo.n), n´ı`ape.re., f´un `ıd´ab`ob`o `awo. n ` ed`e t´ı k`o n´ı k´ıko. . A s.e `ıw´ad`ı´ı b´ı a s.e l`e t´o. ka s´ı ibi t´ı `o. r`o. `afo. w´o.ko. kan w`a n´ın´u ` awo. n `o. r`o. `afe.nuso. n´ıpa l´ılo `awo. n `e.ro. `aw`os.e `o. r`o. `afe.nuso. t´ı a f`aw`or´an dar´ı t´ı a`k´o. n´ıbi t´ı k`o s´ı `o. p`o. lo. p`o. ohun `el`o `e.k´o. . Al´eb`u m´ej`ı w`a l´ara `awo. n `ıw´ad`ı´ı l´ ori `aw`os.e `awo. n `o. r`o.`afe.nuso. t´ı a f`aw`or´an dar´ı t´ı ´o w`a. N´ı `ak´o. k´o. , k`o s´ı n´ın´u ` awo. n `ıw´ad`ı´ı lori `aw`os.e `awo. n `o.r`o.`afe.nuso. t´ı a f`aw`or´an dar´ı t´ı ´o l´epa b´ı a s.e l`e l` o w´o. n l´ati t´o. ka s´ı ibi t´ı `o. r`o. `afo.w´o. ko. kan w`a n´ın´u `awo. n `o. r`o. `afe.nuso. ; b´ı ´o ti l`e. ti j´e. w´ıp´e a ti l`o w´o. n f´un `ıw´ad`ı´ı b´oy´a `o. r`o. `afo. w´o. ko. kan w`a n´ın´u `awo. n ` o. r`o. `afe.nuso. . N´ı `e.`e.kej`ı, `o. p`o. lo. p`o. n´ın´u `awo. n `ıw´ad`ı´ı l´or´ı `aw`os.e `awo. n `o. r`o. `afe.nuso. t´ı a f`aw`or´an dar´ı lo `awo. n `aw`or´an t´ı a so p`o. p`e.l´u `awo. n `o. r`o. ` afe.nuso. n´ı `ed`e g`e.´e.s`ı (`ab´ı `awo. n`ed`e t´ı ´o n´ı d´at`a p´up`o.). A m´a ´ns´ab`a lo `ed`e g`e.´ e.s`ı g´e.g´e. b´ı as.oj´u `ed`e t´ı k`o n´ı `o. p`o. lo. p`o. d´at`a, n´ıtor´ı `ey´ı, ´o nira l´ati l`e wo. n b´ı ` e.ro. `aw`os.e `o. r`o. `afe.nuso. t´ı a f`aw`or´an dar´ı s.e s.e d´aad´aa s´ı n´ı ibi t´ı k`o s´ı `o. p`o. lo. p`o. d´at`a.

El´ey`ı´ı l´o m´u wa koj´u `awo. n `ıb´e`er`e m´ej`ı w`o. ny´ı: (i) S. ´e ´o s.e´es.e l´ati lo `e.ro. ` aw`os.e `awo. n `o. r`o.`afe.nuso. t´ı a f`aw`or´an dar´ı f´un `ıt´o. kas´ı ibi t´ı `o. r`o. `afo. w´o. ko. w`a n´ın´u `awo. n `o. r`o. `afe.nuso. ? (ii) N´ı ibi t´ı k`o s´ı `o.p`o. lo. p`o. d´at`a, s.´e ´o s.e´es.´ee l´ ati lo `e.ro. `aw`os.e `awo. n `o. r`o. `afe.nuso. t´ı a f`aw`or´an dar´ı f´un `ıt´o.kas´ı `o. r`o. ` afo.w´o.ko. n´ıpa l´ılo `ed`e m´ej`ı pap`o. ?

L´ati koj´u `ıb´e`er`e `ak´o. k´o. , a r´o `e.ro. `aw`os.e `o. r`o. `afe.nuso. t´ı a f`aw`or´an dar´ı p`e.l´u agb´ara l´ati l`e se `ıt´okas´ı ibi t´ı `or`o `afow´oko w`a n´ın´u `awon `or`o `afenuso. A se ` ıw´ad`ı´ı wa n´ı ibi t´ı a ti lo `ed`e.t´ı ´o n´ı `o. p`o. lo. p`o. d´at`a g´e.g´e. b´ı as.oj´u f´un `ed`e t´ı k`o n´ı `o. p`o. lo. p`o. d´at`a. El´ey`ı´ı gb`a w´a l´ay`e l´ati s.e`afiw´e t`a`ar`a p`e.l´u is.´e t´ı ´o s.´a´aj´u is.´ e wa. L´ati b`e.r`e. `ıw´ad`ı´ı wa, a lo `ıl`an`a f´un k´ık´o. `e.ro. `aw`os.e`o. r`o. `afe.nuso. t´ı a f`a w`or´an dar´ı t´ı ´o w`a t´e.l`e.. Il`an`a y`ı´ı d´a l´or´ı k´ık´o. `e.ro. `aw`os.e `o. r`o. `afe.nuso. t´ı a f` aw`or´an dar´ı p`e.l´u `awo. n `aw`or´an `ati t´a`ag`ı t´ı k`o g´unr´eg´e wo. n, n´ıpa l´ılo `e.ro. t´ı a fi ´ n t´a`ag`ı `aw`or´an. T´a`ag`ı w`o. ny´ı ni w´o. n l`o l´ati fi k´o. `e.ro. `o. r`o. `afe.nuso. . Ir´u `e.ro. ` aw`os.e `awo. n `o. r`o.`afe.nuso. ti a f`aw`or´an dar´ı t´ı a s.e b´ay`ı´ı k´o. ni a r´o p`e.l´u agb´ara n´ ıpa l´ılo `ıl`an`a f´un `ıt´o. kas´ı `o. r`o.`afo. w´o. ko. n´ın´u `awo. n `o. r`o. `afe.nuso. . `Il`an`a `a k´o. k´o.（输入屏蔽） d´a l´or´ı b´ıbo `awo. n ib`ıkan`n´ın´u `awo. n `o. r`o. `afe.nuso. m´o. l`e. l´ati wo. n ipa t´ı ´o ma n´ı l´or´ı `ıd´am`o. `o. r`o. `afo. w´o. ko. kan. Il`an`a`e.`e.kej`ı（注意力） d´a l´or´ ı s.´ıs.e `ak´ıy`es´ı ibi t´ı ´o s.e p`at`ak`ı j`u l´ara `awo. n `o. r`o. `afe.nuso. l´ati s.e`ıd´am`o. `o. r`o. ` afo. w´o. ko. kan. `Il`an`a `e.e.ke.ta（Grad-CAM） d´a l´or´ı s.´ıs.e `ıw´ad`ı´ı ibo l´ara `o. r`o. ` afe.nuso.`ni ´o s.e p`at`ak`ı j`u f´un `o. r`o. `afo. w´o. ko. t´ı a f´e. d´am`o. . Il`an`a `e.`e.ke.rin（分数 聚合） d´a l´or´ı.`ık´ojo. p`o. `awo. n `es`ı l´ati l`e t´o. ka s´ı ibi t´ı `o. r`o. `afo. w´o. ko. w`a n´ın´u ` awo. n `o. r`o. `afe.nuso. . Il`an`a `ak´o. k´o. gb´egb´a or´ok`e p`e.l´u `ıs.ed´ed´e 57% l´or´ı is.´e t´ ıto. kas´ı ibi t´ı `o. r`o. `afo. w´o. ko. w`a l´a`ı b`ık´ıt`a f´un `ıd´am`o.`o. r`o. b´e.`e.. `Il`an`a `e.`e.kej`ı n´ı `ıs.ed´ed´e 46.0%。

L´ati koj´u `ıb´e`er`e `e.`e.kej`ı（n´ı ibi t´ı k`o s´ı `o. p`o. lo.p`o. d´at`a, s.´e ´o s.e´es.´ee l´ati lo `e.ro. ` aw`os.e `o. r`o.`afe.nuso. t´ı a f`aw`or´an dar´ı f´un `ıt´o. kas´ı `o. r`o. `afo. w´o. ko. n´ıpa l´ılo `ed`e m´ej`ı pap`o.? ）, a b`e.r`e. l´ati s.a d´at`a `aw`os.e `awo. n `o. r`o. `afe.nuso. t´ı a f`aw`or´an dar´ı tuntun jo.。 D´at`a y`ı´ı, t´ı a p`e n´ı Yor`ub´a

Flickr 音频描述语料库（YFACC）是一项包含来自 Flickr 的 6000 条英语图像描述的集合，这些描述由约鲁巴语母语者翻译。YFACC 数据集旨在支持低资源语言音频描述研究。该数据集提供了音频及其对应的文字描述，可用于训练和评估模型。它采用注意力机制，将音频与文本描述进行关联。数据集包含约鲁巴语图像描述，可用于研究低资源语言中的视觉接地关键词检测与定位。该数据集的基准结果显示，基于注意力的模型在约鲁巴语上的表现比基线模型高出 16.0%。不过，由于数据规模较小，模型性能仍有提升空间。我们预计，随着数据集规模扩大，性能将进一步提高。当前数据集中的描述数量为 22.8%，这为低资源语言的音频图像描述研究提供了有价值的资源。

N´ıpa k´ıkoj´u `ıb´e`er`e `ıw´ad`ı´ı wa, a s.e `awo. n `afik´un s´ın´u `ım`o. w`o. ny´ı: (1) A d´ab` a´a `ıl`an`a tuntun f´un b´ı a se l`e lo `ero `aw`ose `or`o `afenuso t´ı a f`aw`or´an dar´ı f´un `ıd´ am`o `ati `ıt´okas´ı `or`o.`afo. w´o. ko. ; a s`ı s.e `ıfiw´era p´ıp´e `aw`os.e n´a`a p`e.l´u `awo. n `ıl` an`a t´ı ´o ti w`a t´e.l`e.. (2) A r´o `e.ro. `aw`os.e`awo. n `o. r`o. `afe.nuso. t´ı a f`aw`or´an dar´ı l’ agb´ara p`e.l´u `ıl`an`a m´e.rin f´un `ıt´o. kas´ı `o. r`o. `afo. w´o. ko. . (3) A s.e `e.k´unr´e.r´e. `ıt´ upal`e. l´ati s.`afih`an `awo. n ibi t´ı `awo. n `e.ro. `aw`os.e n´a`a ti s.`as.eyo. r´ı `ati ibi t´ı w´o. n ti k`un`a. A s.`ak´ıy`es´ı p´e l´ara `awo. n `o. r`o. 67t´ı ´o w`a n´ın´u fokab´ul´ar`ı t´ı a l`o (黑色, 泳池, 足球, 树) ni `ıt´o. kas´ı wo. n s.ed´ed´e, n´ıgb`at´ı `ıt´o. kas´ı `awo. n `o. r`o. `ıy´ok`u k`un`a: 海洋 →冲浪者 ; 球 →足球; 游泳 →泳池. (4) A s.e `at´us´ıl`e d´at`a t´ı ´o w´ul`o f´un s.´ıs.e `ıd´anil´e.k`o. ´o. f´ un `awo. n `e.ro.`aw`os.e `awo. n `o. r`o. `afe.nuso. t´ı a f`aw`or´an dar´ı (5) A s.e `agb´ekal`e. ` e.ro. f´un d´ıd´am`o. `ati t´ıt´o. kas´ı`o. r`o. `afo. w´o. ko. n´ı `ed`e g`e. ´e.s`ı n´ın´u `o. r`o. `afe.nuso. n´ı `ed`e Yor`ub´a n´ı ibi t´ı k`o s´ı d´at`a p´up`o. n´ıpa l´ılo YFACC. (6) A s.e on´ır´uur´u `ıt´upal`e. ` e.ro. `aw`os.e `o. r`o. `afe.nuso. t´ı a f`aw`or´an dar´ı el´ed`e m´ej`ı. A s.`ak´ıy`es´ı p´e l´ara `awo. n ` o. r`o. `afo. w´o. ko. ni `ıs.ed´ed´e t´ı ´o d´ara, b´ı `ape.re., 棕色 (b´ur´a `un; `ıs.ed´ed´e al´a`ıt`as´e 100%), 自行车 (k`e. k´e. ; 94.1%) `ati 草 (kor´ıko 90.9%). S. `ugb´o. n `e.ro. n´a`a k`o s.e d´ad´aa l´or´ı `o. p`o. lo. p`o. `awo. n `o. r`o. m´ır`an.

N´ı `ak´ojo. p`o. , a fih`an p´e `awo. n `e.ro. `aw`os.e `o. r`o. `afe.nuso. t´ı a f`aw`or´an dar´ı s.e´e l`o f´un `ıt´o. kas´ı`o. r`o. `afo. w´o. ko. n´ı ibi t´ı k`o s´ı `o. p`o. lo. p`o. d´at`a. A n´ır`et´ı p´e d´at` a tuntun wa `ati `awo.n `aw´ar´ı tuntun t´ı a s.e y´o`o s.e `ıw´ur´ı f´un `ıw´ad`ı´ı s’i´ı ni l´ılo `awo. n `e.ro. `aw`os.e `o.r`o. `afe.nuso. t´ı a f`aw`or´an dar´ı f´un`awo. n `ed`e t´ı k`o n´ı `o. p`o. lo. p`o. d´at`a.

目录

声明 ion ii

摘要 iii

图目录 x

表格目录 xiv

1. 引言 1

1.1. 动机.1 1.1.1. 应用场景.2 1.1.2. 背景.2 1.2. 研究问题.6 1.3. 研究方法.7 1.4. 贡献.8 1.4.1. 出版物.9 1.4.2. 代码仓库.9 1.5. 学位论文大纲.10

2. 相关工作 12

2.1. 多模态数据集：文本、视觉和语音.13 2.1.1. 带语音的静态图像.13 2.1.2. 视频数据集.13 2.2. 多模态学习方法.15 2.2.1. 联合表示.15 2.2.2. 协同表示.18 2.2.3. 跨模态蒸馏.20 2.3. 视觉接地语音系统的可解释性与分析.23 2.4. 视觉接地的（粗粒度）语音识别.24 2.5. 与VGS建模密切相关的多模态范式.25 2.5.1. 音视频（细粒度）语音识别.25 2.5.2. 从图像和文本中学习.27 2.5.3. 手语关键词搜索.28 2.6. 关键词检出与定位.28 2.6.1. 不使用自动语音识别或视觉监督的关键词检出 29

2.6.2. 语音中的关键词定位.29 2.6.3. 跨语言学习.30 2.7. 相关工作总结.31

3. 视觉 grounding 的关键词检测与检索 32

3.1. 数据集.33

3.1.1. Flickr8k.33

3.1.2. Flickr 音频描述语料库（FACC）.33 3.1.3. FACC 中的视觉关键词.34 3.2. 评估用于关键词检索任务的视觉接地语音模型.35 3.2.1. 关键词检测.36 3.2.2. 关键词检出.36 3.3. 使用视觉定位进行关键词检测与检出.37 3.3.1. 网络结构.37 3.3.2. 训练.38

3.3.3. 测试视觉接地语音模型.39 3.3.4. 视觉接地语音架构与实现细节.40 3.4. 结果：关键词检测与关键词检出.42 3.5. 总结.43

4. 使用视觉接地语音模型进行关键词定位的方法 45

4.1. 评估视觉接地语音模型的关键词定位能力.46 4.1.1. 预言机关键词定位.47 4.1.2. 实际关键词定位.48 4.1.3. 关键词检出定位.49 4.1.4. 关于基于点的评估.50 4.2. 定位方法.50 4.2.1. Grad-CAM.50 4.2.2. 分数聚合.51 4.2.3. 注意力.53 4.2.4. 输入屏蔽.55 4.2.5. 架构与实现细节.57 4.3 结果：关键词定位.57 4.3.1. 定量分析.57 4.3.2. 定性结果.62 4.4. 总结.66

5. 真实低资源设置中的跨语言关键词定位 68

5.1. 与先前工作的关系.69 5.2. YFACC：约鲁巴语 Flickr 音频描述语料库.70

5.2.1. 数据收集.71 5.2.2. 跨语言差异.71 5.2.3. 与 Flickr8k 及其他约鲁巴语数据集的关系.73 5.3. 跨语言关键词定位评估.76 5.4. 使用视觉接地语音模型进行跨语言关键词定位.76 5.4.1. 训练：视觉监督.77 5.4.2. 测试：跨语言定位.78 5.5. 架构与实现细节.78 5.5.1. 数据集.78 5.5.2. 架构：注意力网络.78 5.5.3. 实现细节.79 5.6. 结果：跨语言关键词定位.79 5.6.1. 定量结果.79 5.6.2. 定性结果.83 5.7. 总结.84

6. 总结与结论 85

6.1. 研究问题与拟议答案.85 6.2. 研究结果.86 6.2.1. 人工低资源环境.86 6.2.2. 真实低资源环境.89 6.2.3. 研究贡献.90 6.3. 未来工作.91 6.4. 结论.92

参考文献 **94**

图目录

1.1. Flickr 音频描述语料库中的一个训练数据项 [1]，由一张静态图像（左）及其口语字幕（右）组成。图像和话语均未标注.3 1.2. 左：采用 Kamper 等人 [2]的方法训练用于关键词检测的视觉接地语音模型的框架示意图。右：采用 Harwath 等人 [3]的方法训练用于跨模态检索的视觉接地语音模型的框架示意图.4 1.3. 左：测试阶段，视觉接地语音网络可用于识别某个关键词是否出现在给定话语中（关键词检测）。右：本学位论文的主要目标之一是探究视觉接地语音模型是否也可用于定位输入话语中的关键词（关键词定位）.5 2.1. 联合表示被投射到同一嵌入空间中。本图展示了三种模态：文本、静态图像和语音。 . . . . . 16 2.2. 联合语音图像嵌入空间，其嵌入维度为 d，描述长度为 L。左侧分支处理图像，右侧分支处理语音输入。随后通过点积 [3]将两个分支连接起来.17 2.3. 协调表示存在于各自的空间中，但通过约束进行协调，以确保相似性.18 2.4. [4]的视听模型架构。图中以蓝色表示卷积神经网络层，以红色表示池化层，以黑色表示批归一化层。每个卷积神经网络层后均接一个 ReLU。图中展示了一个约 10 秒（1024 帧）的频谱图输入示例，以说明池化比例。右侧为输出示例，展示了图像特征与语音特征对之间的时空相似度.19 2.5. 离线跨模态蒸馏结构。教师模型（图像模型；底部）的训练与学生模型（音频模型；顶部）的训练相互独立.20 2.6. 用于知识蒸馏的教师-学生网络 [5].22 2.7. 用于生成概率目标的多标签图像分类器（左），以训练语音系统（右） [2].24

2.8. 一个训练样本示例，包含说话人的一系列嘴部区域（左）以及该说话人说出的相应话语（右） [6]。26 2.9. [7] 中用于手语关键词搜索的基于注意力的时空图卷积网络。我们提出的基于注意力的方法（第 4.2.3 节）深受这项工作的启发.28

3.1. Flickr 音频描述语料库（FACC）中的一个示例：一张图像及其对应的口语描述。训练期间，我们的任何模型都不会使用文本描述，此处仅为说明而提供。我们确实会在评估所提出的方法时使用文本转录.

34

3.2. 方法概述。左侧：我们训练一个音频神经网络，根据在对应图像 i上运行的固定预训练图像标注器提供的监督，预测关键词 w 是否出现在音频话语 a 中。右上：在测试阶段，该网络可用于识别某个关键词是否出现在给定话语中（关键词检测）。右下：它还可用于定位输入中的关键词（关键词定位），这将在下一章讨论.3

8 4.1. 在关键词定位任务中，测试阶段的任务是定位给定的关键词。在此示例中，网络用于在输入话语a中定位关键词球 .47 4.2. 图示说明如何评估我们的机制在四段话语中定位关键词“男性”的性能，这四段话语分别由曲线 (a)、(b)、(c) 和 (d) 表示。曲线 (a) 完整标注了曲线各部分所代表的内容.48 4.3. 分数聚合定位方法示意图。橙色区域表示编码器 (Enc) 的输出如何展开，为每一帧 t 和每个单词 w给出一个分数 $h_{w,t}$ 。在此示例中，对应索引 2 和 W 的关键词，其建议位置分别为帧 2 和 T，它们在图示橙色区域中取得最高分.52

4.4. 我们用于语音中关键词定位的基于注意力的架构。语音网络（蓝色部分）首先将音频输入$\mathbf{a}$ 传入卷积编码器。随后，使用基于查询关键词 w嵌入$\mathbf{q}_w$ 的注意力机制对生成的特征$\mathbf{h}_{1:\mathsf{T}}$ 进行池化。上一步得到一个向量，该向量最终经过多层感知机 (MLP)，输出关键词在音频中出现的概率 $p(w|\mathbf{a})$。视觉网络（灰色部分）使用 VGG-16 网络将相应图像$\mathbf{i}$ 嵌入为一个向量，并使用关键词专属的多层感知机计算关键词在图像中出现的概率$p(w|\mathbf{i})$。请注意，视觉网络仅在训练阶段使用（用于监督语音分支），其权重固定；在测试阶段，我们仅使用顶部的语音分支.54 4.5. 我们用于语音中关键词定位的掩蔽式方法。输入话语被拆分为时长不一的重叠片段。随后，将每个片段输入网络 (DNN_a)，以确定特定单词的输出检测概率。输出检测概率最高的单词所对应的片段可能就是该单词的位置。例如，包含索引为 $w = 1$ 的单词的概率最高的片段是片段 1，其 $\alpha_{1,1} = 0.7$.56 4.6. 按每个关键词报告的 F_1 分数衡量的实际定位性能。我们仅纳入至少检测到一段话语的关键词（67 个中有 45 个），也就是说，其检测分数高于阈值 θ 0.5。结果分别针对 CNN-Attend 架构和掩码输入定位方法报告.63 4.7. 对于图 4.6 中所列的 44 个关键词，语音网络的定位性能 (F_1) 与视觉网络的关键词定位性能（平均精度）对比。结果分别针对 CNN-Attend 架构和掩码输入定位方法报告.64 4.8. 使用 CNN-Attend 架构和掩码输入定位方法，对四个查询关键词（自行车、空气、汽车、男孩）进行预测的示例。每幅图中均展示输入音频信号（上方子图）和输出定位分数（下方子图）。我们还给出检测分数 $p(w|\mathbf{a})$、每个关键词是否被正确检测或定位，以及话语的转录文本。垂直灰线表示词边界，浅蓝色标出的单词是得分最高的单词。 67

5.1. 新 Yor`ub´a Flickr 音频描述语料库（YFACC）中的样本：Flickr8k 数据集中的图像（左）与 Yor`ub´a 音频描述（右）配对，这些音频描述基于原始英语描述。对于关键词的一个子集（绿色），我们提供了用于评估的时间对齐。70 5.2. YFACC 中 500 条测试话语里的英语–Yor`ub´a 关键词对的共现，使用标准化 Cohen's $\kappa$ score 计算。为便于阅读，我们从 67 词词汇表中选取关键词.74 5.3. 使用视觉扎根语音模型进行跨语言关键词定位。*左*： 在训练阶段，语音模型根据预训练且冻结的（英语）图像模型提供的视觉监督，训练其预测某个英语关键词是否出现在 Yor`ub´a 音频话语中。*右*： 在测试阶段，给定一段 Yor`ub´a 音频话语，模型输出某个英语关键词出现在该话语中的概率。如果该概率超过预先设定的阈值 $\theta$，模型还会根据网络中的注意力权重，在话语中定位该关键词。. . . . . 77 5.4. 视觉教师在英语和 Yor`ub´a Flickr 测试集上的平均精度均值.82 5.5. 使用 `yo-5k` VGS 模型进行跨语言定位的示例。英语查询关键词显示在右侧.83

表格目录

2.1. 视听数据集及其对应模态概览 . . 14 2.2. 视听数据集及其定量特征概览。. 14 3.1. Flickr 音频描述语料库（FACC）的定量特征。整个语料库的时长约为 36 小时. 33 3.2. 本学位论文所考察的 67 种独特关键词类型列表。“训练计数”和“验证计数”列分别给出了每种关键词类型在训练集和验证集中的出现次数。. . . . . 35 3.3. 四种架构及其组件（编码器、池化层、分类器）.42

3.4. 四种不同架构的关键词检测分数（%），任务是检测给定关键词是否出现在某段话语中（不考虑其位置）.43 3.5. 四种不同架构的关键词检出得分（%），任务是根据话语包含系统词汇表中各关键词的概率对话语进行排序（不考虑关键词的位置）。44 4.1. 四种架构的组件（编码器、池化层、分类器）和定位方法（Grad-CAM GC、输入屏蔽 M、分数聚合 SA 和注意力 Att）.50 4.2. 预知关键词定位任务的测试数据结果（假设检测完美）。我们以准确率报告结果（百分比）。我们考察视觉监督和理想化词袋监督。. . . . . 58 4.3. 实际关键词定位任务的测试数据结果。我们以 F_1报告结果（百分比）。我们考察视觉监督和理想化词袋监督。我们使用阈值 $\theta = 0.5$ 来判断关键词是否出现在话语中.58 4.4. 关键词检测定位任务的测试数据结果。我们以前10位精确率（P@10）报告结果（百分比）。我们考察视觉监督和理想化词袋监督.59 4.5. 视觉接地语音模型的实际关键词定位精确率、召回率和 F_1 分数。括号内为关键词检测分数；由于实际定位需要先进行检测，然后才能定位单词，因此检测分数构成定位性能的上界.60

4.6. 视觉接地语音模型的关键词检测定位分数，以 $P@10$ 和 $P@N$ 表示。括号内为相应的关键词检测得分（忽略定位）；这些得分是定位性能的上界。61 4.7. 十六个查询关键词子集中最常被定位到的单词。对于每个关键词，我们根据检测分数选取排名前20的话语，并在每段话语中根据定位分数找出预测词。我们报告最常见的5个定位词。我们采用 CNN-Attend 架构和掩码输入定位方法。“–”符号表示语音噪声，或强制对齐器的音标词典中没有的单词.65

5.1. 开源 Yor`ub´a 语音语料库的可用模态比较。新的 YFACC 数据集为 67 个关键词子集提供了对齐.75

5.2. 开源 Yor`ub´a 语音语料库的定量特征比较.75

5.3. 视觉扎根语音模型在跨语言任务（在 Yor`ub´a 音频中查询英语）上的结果，使用提出的 YFACC 数据集（前三行）；以及在单语任务（在英语音频中查询英语）上的结果，使用 FACC 数据集（最后两行）。评估模型在实际关键词定位这一主要任务上的表现。通过更改随机种子重复实验三次，并报告平均性能及其标准差.7

9 5.4. 视觉扎根语音模型在跨语言任务（在约鲁巴语音频中查询英语）上的结果，使用所提出的 YFACC 数据集（前三行）；以及在单语任务（在英语音频中查询英语）上的结果，使用 FACC 数据集（最后两行）。模型在较简单的任务（预言机定位）上进行评估，该任务为实际定位性能提供了上界。通过更改随机种子重复实验三次，并报告平均性能及其标准差。80 5.5. 视觉扎根语音模型在跨语言任务（在约鲁巴语音频中查询英语）上的结果，使用所提出的 YFACC 数据集（前三行）；以及在单语任务（在英语音频中查询英语）上的结果，使用 FACC 数据集（最后两行）。模型在较简单的任务（关键词检测）上进行评估，该任务为实际定位性能提供了上界。通过更改随机种子重复实验三次，并报告平均性能及其标准差.80

第 1 章

引言

自动语音识别（ASR）通过转录音频媒体实现人机交互，并提升无障碍体验。然而，由于此类系统需要大量转写语音数据 [8]，目前只有世界上一小部分语言拥有准确的自动语音识别系统。因此，人们对语音处理系统的兴趣日益浓厚：这类系统不使用精确转录文本，而是能够从弱标注数据 [9–12] 中学习。

一种范式是视觉落地语音（VGS）建模。视觉接地语音模型是一种特定形式的多模态系统，通过两种模态进行学习：图像及其口语描述 [1, 3, 4, 13–19]。由于配对的语音和图像是早期语言习得过程中可获取的信号，因此，这种同时出现的视听输入非常适合用于模拟婴儿的语言学习 [20–26]。视觉接地语音模型也适用于教机器人学习新词 [27, 28]。视觉接地语音这一设置的吸引力还源于这样一个事实：图像及其描述性话语或许比语音和转录文本更容易收集，尤其是在为低资源语言 [29] 或没有书写形式的语言 [30–34] 开发系统时。

先前研究表明，视觉接地语音模型可用于跨模态检索、关键词检测和关键词检出等任务。本研究旨在探究如何使用在真实低资源语言环境中训练的视觉接地语音模型进行语音中的关键词定位，即找出话语中关键词出现的位置。先前研究尚未探索过这一任务及其问题设置。

本章余下部分将介绍本学位论文。首先，我们将阐述本研究的动机（第1.1节），然后介绍本研究的主要研究问题（第1.2节）、研究方法（第1.3节）和研究贡献（第1.4节）。最后，我们将概述本学位论文的结构（第1.5节）。

1.1. 动机

本节首先介绍两个场景，以说明本学位论文的研究发现可能适用的情形。随后，我们将介绍 VGS建模的背景，重点关注

以往研究中使用的方法和数据集。

1.1.1. 应用场景

为了说明基于视觉接地语音的建模如何应用于真实低资源设置，请考虑以下两个示例。

语言记录。 设想一位语言学家负责记录一种低资源语言（全世界约 7,000 种语言中，大多数都被视为低资源语言）。她从以该语言为母语的说话人那里收集话语数据集，并希望了解如何用这种新语言表达一般概念（例如“儿童”“水”“红色”）。她考虑使用机器学习工具来辅助这一过程。然而，开发自动化模型需要标注数据，例如音频和转录文本。在某些环境中，这些资源十分匮乏，甚至完全无法获得；没有书写形式的语言就是如此。在这种情况下，可以合理地假设她能够收集与口头描述配对的图像，用于训练机器学习系统。例如，这样的系统可以帮助她在收集的话语数据集中检测并定位特定关键词。

提供人道主义援助。设想一位外国医生正在为自然灾害受害者提供医疗援助。她需要能够与受害者交流，以了解应该采取何种治疗措施。不幸的是，这些受害者只会说自己的语言，而她不懂这种语言，并且人工翻译人员十分稀缺。在这种情况下，能够检出并定位关键词的视觉语音技术有助于快速扫描受害者的话语，查找感兴趣的关键词是否出现。理想情况下，她应能用自己的母语提供关键词，系统随后在外语语音中检出这些关键词的出现（这项任务称为跨语言关键词检索 [35]）。这样的系统可能远非完美。然而，在紧急情况下，面对未知语言却没有翻译系统可用这一替代情形时，即使不完美的系统也可能大有用处 [8]。

1.1.2. 背景

上述示例展示了视觉接地语音模型可以应用的一些实际场景。下面我们将简要回顾现有的VGS建模主要工作。第2章将对相关文献进行全面回顾。此处的目标则是提供背景信息，为我们的研究问题和贡献（将在后续章节中介绍）铺垫语境。

如图 1.1 所示，视觉接地语音模型从图像及其配对的未标注话语中学习。在这种视觉监督设置中，话语和与之配对的图像

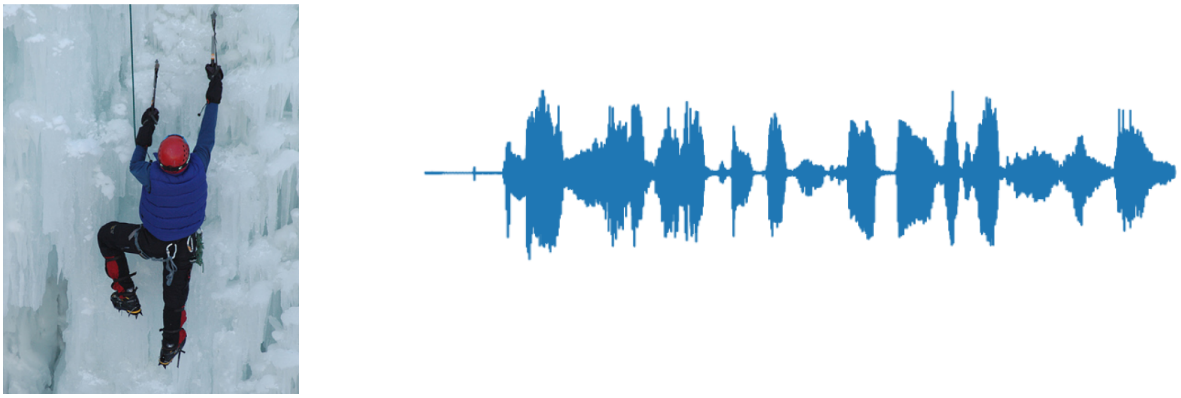


图1.1： Flickr 音频描述语料库 [1]数据集中的一条训练数据项，由一幅静态图像（左）及其口语字幕（右）组成。图像和话语均未标注。

都没有标注。因此，训练视觉接地语音系统需要某种形式的自监督。一种方法是训练模型，将图像和语音投射到联合嵌入空间 [1, 3, 14, 36, 37]中。这类模型通常通过优化基于排序的准则 [38] 进行训练，使得相互匹配的图像和图注在嵌入空间中的相似度高于不匹配的图像—图注对。如图 1.2 右侧所示。这类系统可用于跨模态检索：给定大量图像，任务是检索与给定口语字幕相匹配的图像（反之亦然，即给定图像查询，从大型音频库中找出正确的口语话语）。

虽然能够定义联合图像—语音空间（或能够在不同空间之间转换）的模型本身很有用，因为它们支持跨模态检索，但它们无法将语音映射到文本标签。Kamper等人 [2,39], 解决了这一问题；在本学位论文中，我们以他们的方法为视觉接地语音模型的起点（完整细节见第2.2.3节）。他们的方法如图1.2左侧所示，包括使用离线图像标注器获取软文本标签，并将其作为噪声标签来训练卷积神经网络。在测试阶段，再使用训练好的语音模型将输入话语映射到关键词（无需知道关键词的顺序或数量）。作者表明，即使没有任何转录文本，这类视觉接地语音模型也能以相对较高的精确率进行关键词检测 [2]。如图1.3左侧所示，关键词检测是指预测查询关键词是否出现在话语中的任何位置。在测试阶段，训练好的模型仅接收一段话语（不含图像）和一个查询关键词，并输出该关键词在话语中出现的概率。

这些基于关键词的视觉接地语音模型（可视为执行粗略形式语音识别的模型）也用于执行关键词定位任务。关键词定位与关键词检测相关，但并不相同。在关键词定位中，目标是根据一组话语包含某个

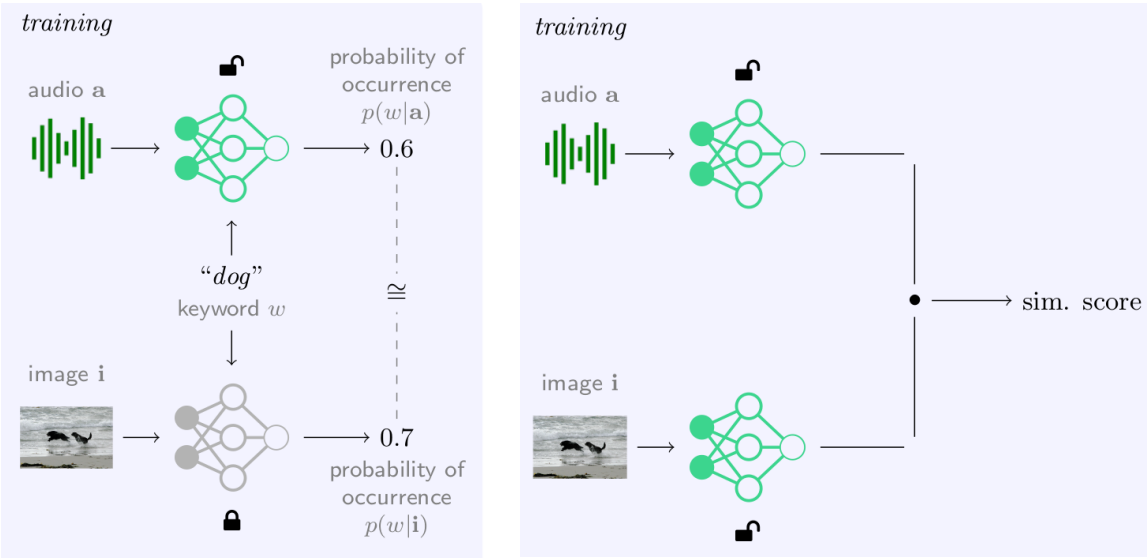


图 1.2： 左：使用 Kamper等人 [2]的方法训练用于关键词检测的视觉接地语音模型的框架示意图。右：使用 Harwath 等人 [3]的方法训练用于跨模态检索的视觉接地语音模型的框架示意图。

给定书面关键词查询的可能性，对这组话语进行排序。这不同于关键词检测，后者的任务是确定某个查询是否出现在单条话语中。同样，Kamper 等人 [2]表明，在不使用任何转录文本进行训练的情况下，视觉接地语音模型也能在一定程度上执行关键词定位。他们还表明，这些模型不仅能找到完全匹配项，还能找到语义匹配项 [39, 40]。例如，给定查询“儿童”，系统会检索出诸如“年轻男孩在公园踢足球”这样的语句。

除了检索、关键词检测和关键词检出任务外，近期研究还将视觉接地语音方法用于其他多种任务和建模方法。例如，一些研究提出了多语言变体，即使用两种或更多口语训练视觉接地语音系统 [16, 41–43], ，使系统能够执行跨语言检索，也就是说，给定一种语言的查询时，检索另一种语言中的内容。其他近期研究则考虑使用视频而非静态图像 [6, 44–47]。如前文所述，本学位论文仅讨论静态图像及其口语描述的情形。

尽管已有研究采用 VGS 建模来应对上述任务，但此前的研究都没有考虑关键词定位任务。关键词定位是指在语音话语中找出给定查询关键词出现的位置，如图1.3右侧所示。我们的工作主要目标之一就是弥补这一不足。

现在，我们转而讨论 VGS 建模工作的另一项主要不足，这也是我们在本研究中力图解决的问题。上述研究几乎全都以英语为基础。为说明这一点，我们简要回顾与此最相关的 VGS 数据集。其中值得一提的是 Flickr 音频描述语料库（FACC） [1]。FACC 是一个多模态

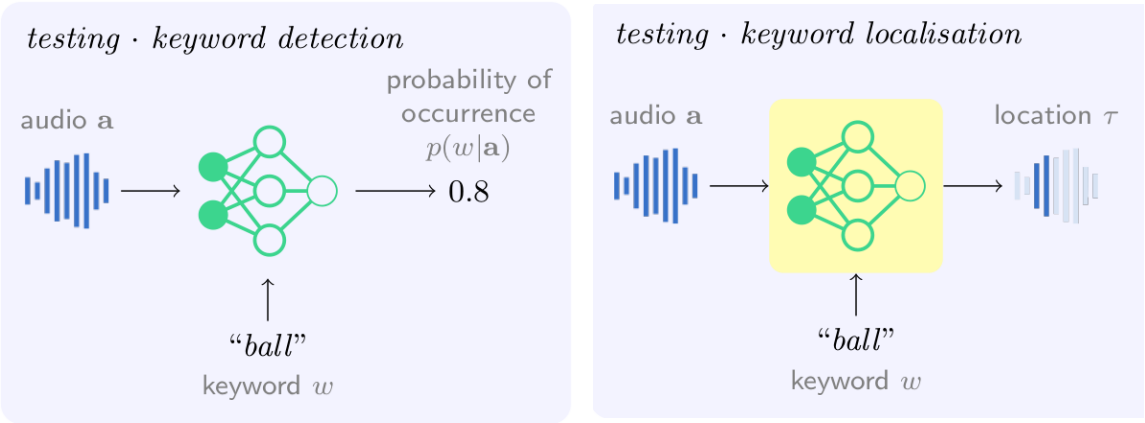


图1.3： 左：在测试阶段，可以使用视觉接地语音网络来识别任意给定的话语中是否出现某个关键词（关键词检测）。右：本学位论文的主要目标之一是探讨视觉接地语音模型是否也能用于定位输入语句中的关键词（关键词定位）。

数据集，包含与 Flickr8k 文本-图像数据集 [48, 49]相同的图像。众包工作者通过亚马逊土耳其机器人，根据 Flickr8k 数据集所含 8,000 张图像各自提供的五条文本描述，录制了 4 万条图像口语描述。研究者还使用强制对齐系统，为口语描述中说出的单词获取了时间对齐。FACC 数据集的发布随后推动了多项 VGS 任务的发展，例如图像-语音对齐 [1], 、语义语音检索 [39],、图像到语音合成 [50] 以及语音到图像生成 [51]。其他规模更大的语音-图像数据集则基于 MIT Places 图像数据集 [52], ，该数据集包含约 700 万张带有场景标签的图片。研究者为这一图像数据集补充了 40 万条英语口语描述 [3], 100、k 条印地语口语描述 [16], 以及 10 万条日语口语描述 [42]。然而，所有这些语言都可以视为资源充足的语言。总而言之，目前还没有能够支持在真正低资源环境下开展研究的 VGS 数据集。

尽管视觉接地语音研究存在这一不足，许多视觉接地语音研究仍将视觉接地语音模型在低资源环境中的应用列为主要动机。因此，大多数研究都以英语作为低资源语言的替代语言。例如，一项非常相关的研究 [35] 模拟了低资源环境，将英语视为低资源语言，将德语视为高资源语言。具体而言，作者使用视觉标注器生成德语文本标签，作为语音网络的目标；也就是说，采用了与图 1.2（左）相同的研究方法，但现在生成的不是图像的英语标签，而是德语标签（图 1.2 左下）。由此得到的语音模型（图 1.2 左上）接收英语语音，并生成德语文本标签。更具体地说，在跨语言关键词检索任务中，可以向该模型提供一个书写的德语关键词（例如“Hunde”，即德语中表示“狗”的单词）

并要求它检索包含该关键词的英语语音话语（例如，“两只狗在户外水边玩耍”）。以往关于跨语言关键词检索的研究采用了直接的方法，即使用文本转录训练系统，而不是使用视觉定位 [53]。因此，这是一项里程碑研究，因为它表明可以使用视觉接地语音模型进行跨语言关键词检索（以及检测）——这种方法可用于低资源环境。然而，研究仍使用英语来模拟低资源语言，因此尚不清楚这种方法能否适用于真实的低资源场景。（顺带一提， [35] 也只考察了关键词检测和关键词检出，并未考察关键词定位。）

显然，即使是那些试图明确考虑低资源环境的研究，也采用了人工方式。为此，我们收集了一个新的图像—语音数据集，使用一种真实的低资源语言，并首次展示如何利用在真实低资源语言上训练的视觉接地语音模型，进行跨语言关键词检测与定位。

1.2. 研究问题

根据上述背景，我们提出以下具体研究问题，并将在本学位论文中加以探讨。

研究问题1。 视觉接地语音模型能否实现关键词定位？ 如第1.1.2节所述，尽管视觉接地语音模型此前已用于关键词检出和检测，但尚未用于关键词定位。提醒一下，关键词定位的任务是找出给定查询关键词在话语中出现的位置，如图1.3右侧所示。由于此前大多数关键词检出和检测研究都是在人工低资源环境下使用英语视觉语音接地数据开展的，我们也将首先在这一人工环境中探讨该研究问题。这样便于我们与先前研究进行比较，并直接拓展其工作。不过，在下一个研究问题中，我们将探讨真实低资源场景。

研究问题 2。在真实低资源环境中，我们能否跨语言进行视觉接地关键词定位？在人工设定中将英语视为低资源语言，对视觉接地语音系统开展的分析有助于回答视觉接地语音模型能否实现关键词定位这一问题。但这无法回答视觉接地语音系统在真实低资源设置中的表现如何。对于一种低资源口语，问题在于我们能否构建一个可以跨语言定位单词的视觉接地语音系统。提醒一下，跨语言关键词检索任务涉及检索一种语言中包含另一种语言的指定关键词的话语 [35]。因此，首先要回答的问题是，能否在一种真实的低资源语言中使用视觉接地语音

模型来完成这项任务；其次，模型是否也可用于跨语言关键词定位（这是此前任何研究都未曾考虑过的任务）。

1.3. 研究方法

在本学位论文中，我们以 Kamper 等人 [2] 在第1.1.2节中提到的方法为起点。提醒一下，他们的方法是使用离线图像标注器从图像中提取软标签，并据此训练视觉接地语音模型，以进行关键词检测和定位。具体而言，我们在他们的方法基础上进行了扩展，使其也能够定位关键词。

由于训练阶段没有可用的位置信息，我们必须利用视觉接地语音模型自身定位 [54] 的能力。具体而言，我们提出了四种不同的定位方法，以使这些模型适用于关键词定位。其中一种方法采用了广受欢迎的可解释性方法$Grad\text{-}CAM$ [55]，该方法最初是在图像领域中开发的。这种方法可用于任何卷积神经网络架构。第二种和第三种方法利用了这样一个事实：对于某些架构，内部激活可以解释为定位分数。具体来说，第二种方法（分数聚合）假设最后一个池化层是一种简单的变换（例如平均池化或最大池化），因此输出分数可视为局部分数的聚合。这些分数可用于选出查询关键词最有可能出现的时间位置。第三种方法（注意力）假设网络包含一个沿时间轴池化特征的注意力层；我们将注意力权重用作定位分数。我们的注意力框架深受 [7]启发。第四种方法（输入屏蔽）通过在不同位置掩蔽输入信号，并测量训练好的模型对部分输入所预测的响应分数来实现；输出的大幅变化表明关键词可能存在。输入屏蔽与第一种方法类似，因为训练后它可以应用于任何分类模型，而不受架构限制。

截至目前所述的研究方法有助于解决第一个研究问题。为解决第二个研究问题，我们以 Kamper 和 Roth 的方法为起点 [35], ，该方法已在第1.1.2节中简要介绍，并将在第5.1节中详细说明。提醒一下，他们的方法是先为一种语言（资源丰富语言）中的图像获取软标签，再将这些软标签与另一种语言（低资源语言）的语音配对（他们通过使用两种资源丰富语言，人为模拟了这种设置）。为了在真实低资源设置中切实评估视觉接地语音系统的跨语言关键词定位性能，我们从一种真实的低资源语言约鲁巴语中采集了一个新数据集。约鲁巴语是尼日利亚的三种官方语言之一，尽管其母语说话人数超过 4400 万，但这种语言可用的高质量数据集却寥寥无几 [56–59]。随后，我们使用离线图像标注器，为与相应约鲁巴语音相关的图像添加英语标签，从而能够训练视觉接地语音模型

使其使用英语关键词和约鲁ˋ巴ˊ语音执行跨语言关键词检测。然后，我们首次在真实低资源设置中应用为第一个研究问题开发的定位方法，以执行跨语言关键词检测和定位。

1.4. 贡献

本学位论文为 VGS 建模领域的知识体系作出了以下具体贡献。

研究贡献 1. 提出了一种利用注意力机制进行关键词检测和关键词检出的新型 VGS 模型，并将该方法与现有的基于视觉接地语音的方法进行了全面比较。

研究贡献 2. 通过提出四种定位方法，扩展了现有 VGS 模型在关键词定位方面的应用。

研究贡献 3. 通过详细的定量和定性分析，揭示了上述模型的局限性，并展示了其成功之处和失败模式。该分析是在将英语视觉语音接地数据人为视为低资源语言数据的情境下进行的。

研究贡献 4。一个新的多模态、多语言数据集，使得在类似语言记录场景的真实低资源环境中进行 VGS 建模成为可能。该数据集将 Flickr8k 图像-文本数据集 [48, 49] 扩展至约鲁ˋ巴ˊ语，包含三种模态：其 6,000 条图注的约鲁ˋ巴ˊ语译文；由一位说话人录制的对应译文语音；以及针对其中 500 条图注的子集，对 67 个约鲁ˋ巴ˊ语关键词进行的时间对齐。

研究贡献 5。一个用于真实低资源环境下跨语言关键词检测和关键词定位的基线系统。该基线系统使用 5,000 条与英语软文本标签配对的约鲁ˋ巴ˊ语话语进行训练，生成一个能够在约鲁ˋ巴ˊ语语音中检测并定位英语关键词的语音网络。我们还表明，先使用规模更大的英语语音-图像数据集对该系统进行预训练，可以显著提升其性能。

研究贡献 6. 对上述跨语言视觉接地语音模型进行多方面分析。这些分析有助于更好地理解关键词定位任务的性能如何在系统词汇表中的 67 个关键词之间变化，并进一步帮助解读跨语言模型的失败模式。

1.4.1. 出版物

以下列出的是本学位论文所述研究成果带来的已发表贡献。

研究论文 *1*：（研讨会）K. Olaleye、B. van Niekerk 和 H. Kamper，“利用弱监督实现语音中的关键词定位”，*NeurIPS* 语音与音频处理自监督学习研讨会，2020年。

研究论文 *2*：（会议）K. Olaleye 和 H. Kamper，“利用视觉定位实现基于注意力的语音关键词定位”，*Interspeech*，2021年。

研究论文 *3*：（期刊）K. Olaleye、D. Oneat˘a 和 H. Kamper，“利用视觉接地语音模型定位未转写语音中的关键词”，《信号处理精选主题期刊》，2022年。

研究论文 *4*：（会议）K. Olaleye、D. Oneat˘a 和 H. Kamper，“YFACC：一个用于通过视觉定位实现跨语言关键词定位的 Yor`ub´a 语音-图像数据集”，*IEEE*口语技术（*SLT*），2023年。

1.4.2. 代码仓库

研究论文 *1*：基于弱监督的语音关键词定位代码网址

研究论文 *2*：基于视觉定位的注意力语音关键词定位代码网址

研究论文 *3*：使用视觉接地语音模型进行未转写语音中的关键词定位代码网址

研究论文 *4*：*YFACC*：约鲁 `巴 ´通过视觉定位实现跨语言关键词定位的语音-图像数据集 代码网址1，代码网址2 数据网址

1.5. 论文大纲

本节旨在概述本学位论文的结构，帮助读者了解各章的安排。我们简要总结每一章的目的和内容。

第2章：相关工作。 本章将深入回顾与本学位论文主题相关的既有研究。首先，我们回顾为支持多模态学习而开发和整理的各类数据集，重点关注三种模态：文本、视觉和音频。接着，我们回顾利用多种模态中的互补信息的不同方法。随后，我们回顾分析和解释一些多模态系统（尤其是视觉接地语音系统）的研究。此外，我们还综述了早于视觉接地语音的其他类型多模态学习（包括语音领域内外的研究），即视听语音识别——通过利用语音产生过程中可获取的视觉信息来训练自动语音识别系统；以及从图像和文本中进行的多模态学习。最后，我们回顾关于关键词检出与定位（有无视觉监督均包括在内）的研究，这些正是本学位论文所要完成的核心任务。

第3章：视觉接地的关键词检测与检索。 本章将介绍用于关键词检测与检索的视觉接地语音模型。本章旨在部分回答我们的第一个研究问题。具体而言，我们将阐释关键词检测模型，并在本学位论文的其余部分将其扩展用于关键词定位任务。我们将介绍本学位论文中使用的一个视觉接地语音数据集：第1.1.2节提到的英语 FACC 语料库。我们还将详细介绍评估流程和所采用的方法，包括视觉接地语音架构及其实现细节。此外，我们还将呈现视觉接地语音模型在关键词检测和关键词检出任务上的基线结果与上界结果。

第4章：基于视觉接地语音模型的关键词定位方法。 本章将全面探讨我们的第一个研究问题。我们将上一章（第3章）中的模型扩展至关键词定位任务。我们提出了四种不同的定位方法，以使模型适用于关键词定位（见第1.3节）。我们介绍视觉接地关键词定位的评估方案，并展示所提出的关键词定位方法的性能，对这四种不同的关键词定位方法进行系统且公平的比较。此外，我们还提供详细的定性分析，展示模型在关键词定位任务中的成功案例和失败模式。

第5章：跨语言关键词定位。 本章将探讨我们的第二个研究问题。我们通过引入一个多语言多模态数据集，朝着更贴近现实的学位论文任务环境迈进；该数据集支持在真实低资源设置中训练和评估视觉接地语音模型。我们为 Flickr8k 中的 6k 张图像收集并发布了一个新的单说话人音频描述数据集，语种为 Yor`ub´a——这是一种在尼日利亚使用的真实低资源语言。我们训练了一个基于注意力机制的视觉接地语音模型，其中图像会自动标注英语视觉标签，并与 Yor`ub´a 话语配对。我们介绍用于评估模型跨语言关键词定位性能的评估标准，并详细说明视觉接地语音系统的架构与实现。最后，我们展示视觉接地跨语言系统的定量与定性分析。

第6章：总结与结论。 本章总结并突出前几章的主要发现，针对第1.2节提出的研究问题给出直接回答，最后介绍未来可能的研究方向。

第2章

相关工作

本章回顾了与本学位论文主题相关的先前研究。我们首先回顾为支持多模态学习而开发和整理的各类数据集（第2.1节），重点关注三种模态：文本、视觉信号（图像或视频）和音频。然而，本学位论文开展的所有实验都仅使用音频和静态图像（训练阶段），以及仅使用音频（测试阶段）。我们在第2.2节回顾利用多种模态中的互补信息的不同方法。随后，我们在第2.3节回顾分析和解释多模态系统的研究，尤其关注视觉接地语音（VGS）模型。提醒一下，视觉接地语音模型通过图像及其口语描述进行学习，而这正是本学位论文始终关注的模型类型。

我们在第2.4节中进一步深入考察视觉接地语音系统，回顾与本学位论文最相关的研究——具体而言，是探讨第2.2.3节所述跨模态蒸馏方法的视觉接地语音模型研究。这些视觉接地语音系统专门针对粗粒度（非逐字）语音识别进行训练。所谓粗粒度语音识别，是指识别语音音频中说出的孤立关键词（语义关键词或精确关键词）。例如，给定转录文本为田野里有三只狗在看草丛中的某样东西的话语，粗粒度语音识别会使用视觉接地语音模型来预测话语中是否出现了单词草 。这不同于传统的自动语音识别（ASR）系统；对后者的评估依据是其能否生成整个话语的精确转录文本。此外，我们还回顾了VGS建模出现之前的其他多模态学习类型（包括语音领域内外的研究）。其中包括视听语音识别（AVSR），即利用语音产生过程中可获得的视觉信息来训练自动语音识别系统（第2.5.1节）；以及从图像和文本中学习（第2.5.2节）。最后，我们回顾了有视觉监督和无视觉监督条件下的关键词检出与定位研究（第2.6节）。这些是本学位论文试图解决的核心任务，具体而言，我们将使用视觉接地语音模型来完成这些任务。

2.1. 多模态数据集：文本、视觉和语音

可以组合多种不同类型的模态来构建多模态数据集。因此，为了让本节更加聚焦于本学位论文的核心主题，我们只介绍涉及三种模态的数据集：文本（包括书面文本和口语）、视觉（包括静态图像和视频）以及音频。多模态数据集由其中两种或更多模态组合而成。

2.1.1. 带语音的静态图像

视觉接地语音模型发展进步的关键因素之一，是视觉接地语音数据集的可用性 [60]。人们致力于收集描绘日常物体和情境的图像的口语描述，以促进建模方法的发展，并在缺乏传统监督的情况下，为下游任务提供语音技术。有关现有视觉接地语音数据集的最新综述，请参见 [60] 。

表 2.1 展示了其中一些数据集所包含的模态，表 2.2 则展示了它们的定量特征。在所有这些公开可用的视觉接地语音数据集中，与本学位论文最相关的是由 Harwath 和 Glass [1] 提出的 Flickr 音频描述语料库（FACC）（更多详情请参见第 3.1.2 节）。该数据集直接基于 Flickr8k [48, 49]。原始的 Flickr8k 是一个双模态（图像和文本）数据集，由 8k 张图像和 40k 条英语文本描述组成。FACC 使用亚马逊土耳其机器人，通过众包方式为每条文本描述收集英语音频描述，从而扩展了 Flickr8k。FACC 数据集的发布随后推动了多项视觉接地任务的发展，例如图像-语音对齐 [1],、语义语音检索 [39], 、图像到语音合成 [50] 以及语音到图像生成 [51]。在本学位论文中，我们进一步使用一种真正的低资源语言（Yor`ub´a）扩展 Flickr8k 数据集：为选定的关键词收集口语描述，并进行带有起止时间戳的人工对齐，以支持资源匮乏语言的跨语言关键词检索和定位（详见第5章）。

其他已开发的大型语音-图像数据集基于 MIT Places 图像数据集 [52]，该数据集包含约 700 万张带有场景标签的图片。该图像数据集新增了 40 万条英语口语描述 [3], 100k 条印地语口语描述 [16], 以及 10 万条日语口语描述 [42]。所有这些语言都可以视为资源丰富的语言。

2.1.2. 视频数据集

除了上述图像—语音数据集外，还有视频数据集，其中包含与视频所描绘的活动相关的书面或口述（或两者兼有）的描述

表 2.1： 视听数据集及其对应模态概览 s

数据集	语言	模态			
		音频	文本图像		对齐
Flickr 音频描述语料库（2015年） [1]	en	✓	✓	✓	✓
English Places 音频描述（2016年） [3]	en	✓	✓	✓	
合成语音 COCO（2017年） [36]	en	✓	✓	✓	
合成语音 STAIR（2017年） [61]	ja	✓	✓	✓	
Speech-COCO（2017年） [62]	en	✓	✓	✓	✓
印地语地点音频描述（2018年） [16]	hi	✓	✓	✓	
概念口语描述（2019年） [63]	en	✓	✓	✓	
SpokenCOCO（2020年） [64]	en	✓	✓	✓	
日语地点音频描述（2020年） [42]	ja	✓	✓	✓	

表 2.2： 视听数据集及其定量特征概览 s.

数据集	说话人 数量	时长 小时	话语 数量	采样 kHz
Flickr 音频描述语料库（2015年） [1]	183	36	40000	16
English Places 音频描述（2016年） [3]	2683	936	400000	16
合成语音 COCO（2017年） [36]	1	601	616767	24
合成语音 STAIR（2017年） [61]	1	793	616767	24
Speech-COCO（2017年） [62]	8	601	616767	16
印地语地点音频描述（2018年） [16]	139	316	100000	16
概念口语描述（2019年） [63]	6		3.3M	16
SpokenCOCO（2020年） [64]	2352	742	605000	48
日语地点音频描述（2020年） [42]	303	540	98555	16

内容。这些数据集来自多种场景，包括烹饪 [65–71],体育运动 [72–74]，电影 [75, 76] 或一般活动 [77]。这些数据集面向与视频理解相关的任务，例如视频描述 [67, 75, 76, 78–80] 和事件检测 [68, 69, 71, 81]。

Regneri 等人 [65] 提出了一个多模态数据集，将视频与对视频所描绘动作的多条自然语言描述对应起来，并标注动作描述与其对应视频片段的相似程度。该数据集旨在展示视频场景对基于文本的动作描述模型的影响。Das 等人 [66] 从 YouTube 收集了一个视频数据集，用于训练模型执行视频描述任务。该数据集包含 88 段视频，并收录了视频中所描绘物体的文本标注。Zhou等人 [70] 推出了一个多模态数据集，其中包含从 YouTube 下载的 2,000 段教学视频。该数据集包含每段视频中程序片段的时间位置信息和文本描述。该数据集用于促进将（仅视觉）视频划分为类别无关程序片段的任务。Spoken Moments in Time

（S-MiT）数据集 [82] 包含 50 万段独一无二的三秒口语音频描述，每段都来自 Moments in Time 数据集中的不同源视频。S-MiT 收录的是对视频中所描绘元素的口语描述，而非文本描述，因此可以联合训练语音-视频模型。MP II 电影描述语料库 [75] 包含来自 94 部好莱坞电影的转录音频描述，分为 6.8 万个片段；每个片段都配有一句电影剧本中的句子，以及对该片段视觉内容的音频描述。类似地，大规模电影描述挑战数据集 [76]包含 200 部电影和 12 万条句子描述。VideoStory [80]包含 2 万段社交媒体视频，每段视频都有一段篇幅为一段的描述。ActivityNet Captions 数据集 [83] 包含 2 万段视频和 10 万条文本描述。Microsoft Video Description 数据集 [78] 包含 2,000 段 YouTube 视频片段，时长为 10 至 25 秒，每段平均有 41 条单句描述。MSR-Video to Text [79] 包含 7,000 段视频，划分为 1 万个片段，每段视频配有 20 条图注。HowTo100M [77] 包含 1.36 亿个片段，来源于 122 万段教学视频，并根据与每段视频相关的字幕生成口述旁白。VaTeX [84]包含 4.1 万段视频，每段均标注有 10 条英文字幕和 10 条中文图注，用于多语言字幕生成。

所有这些视频数据集都使得以多模态方式学习下游任务模型成为可能。本学位论文中讨论的模型均使用口语描述和静态图像进行训练。将本学位论文中提出的方法扩展至视频领域，或许是未来工作中一个有趣的方向。

2.2. 多模态学习方法

多模态学习的成功很大程度上取决于各模态的表征如何相互作用，从而利用它们之间的互补信息 [85]。本节回顾了提出不同多模态学习方法的研究。其中一些方法涉及先学习各个模态的表征，再将这些表征融合到联合空间（第2.2.1节）或协调空间（第2.2.2节）中。另一种多模态学习方法是跨模态蒸馏（第2.2.3节），即将知识从一种模态（教师模型）迁移到另一种模态（学生模型）。

2.2.1. 联合表示

在联合表示研究方法中，每种模态的表征都会被投射到同一个多模态空间 [85]。图2.1展示了联合表示方法的示意图，其中静态图像及其文本和口语描述被投射到共同的嵌入空间。已有多项研究探讨如何从图像和文本构建联合表示。由此得到的表征使得诸如

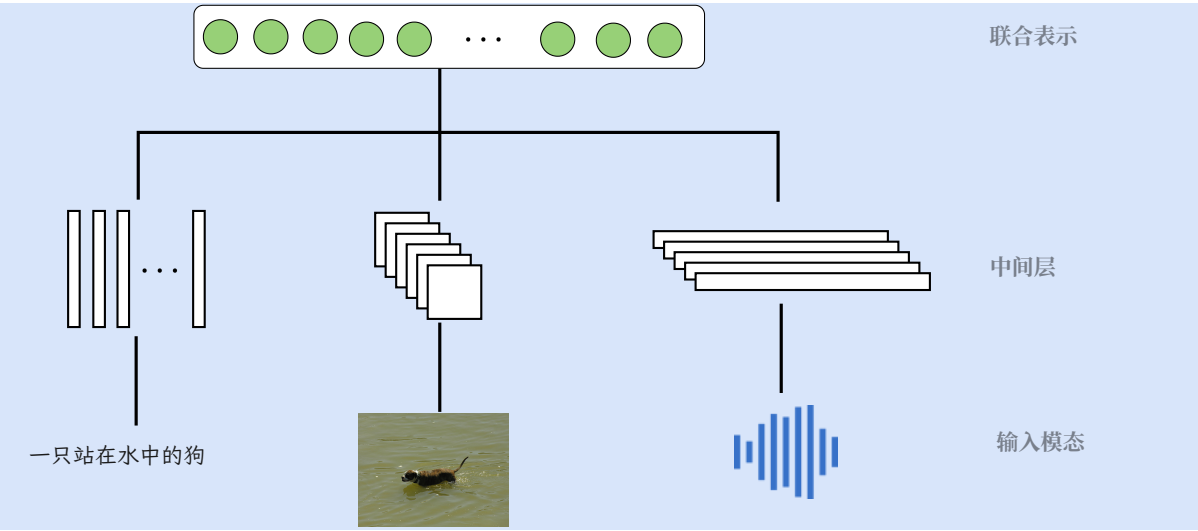


图2.1： 联合表示被投影到同一嵌入空间中。图中给出了三种模态：文本、静态图像和语音。

视觉问答 [86], 和图像描述生成 [87]等任务成为可能。联合表示方法也已从图像和文本模态扩展到包含音频 [46]。这种方法已用于视听语音识别（AVSR，见第2.5.1节综述）和视频分类 [88]。本小节将重点回顾静态图像及其口语描述之间联合表示构建方面的研究。

最早提出使用神经网络为图像和语音构建联合表示的研究之一是 [14]。作者使用多层感知机（带有 ReLU 激活）将语音和图像特征映射到同一嵌入空间。模型使用余弦平方余弦损失，学习关联共同出现的实词和图像片段：

$$\mathrm{L}_{\cos\cos^2} = \begin{cases} (1-\cos(\mathbf{v},\mathbf{s})) & \text{if } \mathbf{v} \text{ and } \mathbf{s} \text{ match} \\ \cos^2(\mathbf{v},\mathbf{s}) & \text{if different} \end{cases} \tag{2.1}$$

with

$$\cos(\mathbf{e_1},\mathbf{e_2}) = \frac{\mathbf{e_1}^\top \mathbf{e_2}}{\|\mathbf{e_1}\|\,\|\mathbf{e_2}\|}$$

其中 $\mathbf{v}$ 和 $\mathbf{s}$ 分别是输入图像片段和输入口语单词词元的输出表征。训练得到的视觉接地语音模型用于根据图像检索口语词，反之亦然。

在开启了大量 VGS 建模研究的开创性研究中，Harwath等人 [1,3]使用规模大得多的数据集（FACC，见第 2.1.1 节综述）扩展了这一方法。在 [3], 中，作者使用卷积神经网络为每种模态构建表征。图 2.2 展示了作者所用架构的示意图。

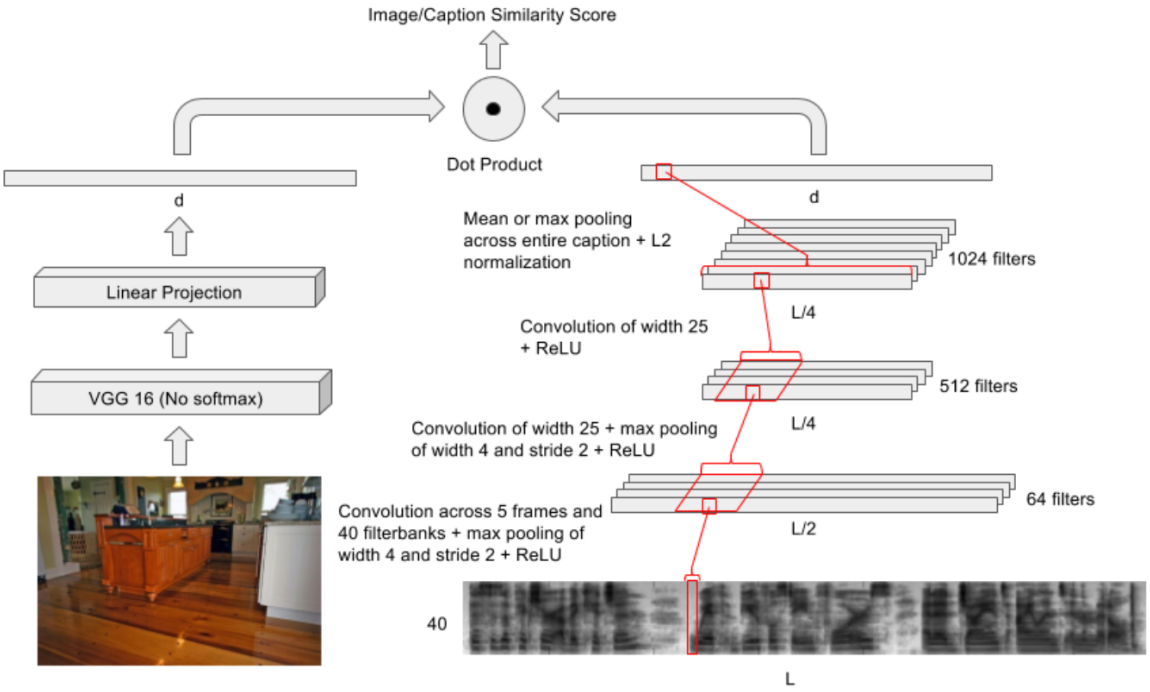


图 2.2：嵌入维度为 d 、描述长度为 L的联合语音图像嵌入空间。左侧分支处理图像，右侧分支处理语音输入。随后，两个分支通过点积 [3]联系起来。

该模型由一个图像分支和一个语音频谱图分支（长度为 L）组成。每个分支都会生成一个 d 维表征。模型对这些表征进行点积运算，并将运算结果作为图像与语音之间的相似度。作者还引入了对比边际损失，使匹配的图像-语音对比不匹配的图像-语音对更为接近；此后，这种损失的变体已成为大多数 VGS 模型的标准做法。在 [1], 中，作者还使用一对卷积神经网络，在词级构建图像对象和语音表征；这两个网络通过嵌入与对齐模型相连，该模型能够学习涵盖两种模态的联合语义空间。研究人员通过图像检索任务和图像标注任务评估了最终模型：前者要求模型找出与给定图注对应的图像，后者要求模型搜索所有图注，找出与给定图像对应的图注。

用于构建联合表示的其他网络架构也得到了探索：Chrupa la 等人 [36] 在语音分支中使用了带注意力机制的门控循环神经网络，而最近 Peng 和 Harwath [37]提出了一种基于 Transformer 的模型。其他研究提出了多语言变体，其中视觉接地语音系统使用两种或更多口语进行训练，并配备相同或不同的视觉组件。Harwath等人 [16] 证明，同时使用英语口语和印地语图注以及共享的视觉组件训练双语模型，可以生成能够

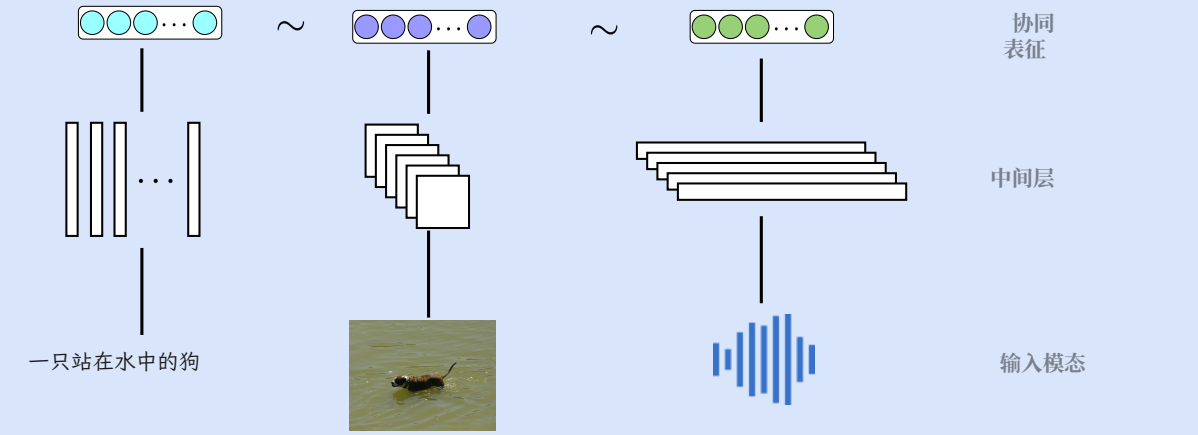


图2.3： 协调表示存在于各自的空间中，但通过约束进行协调，以确保相似度。

.

执行语义跨语言语音到语音检索的视觉接地语音系统。Azuh 等人 [41] 提出了一种方法，用于从跨两种语言的视觉接地语音模型中发现类词单元及其近似译文。Ohishi 等人 [42] 探索了用于跨模态和跨语言检索的三语视觉接地语音模型（使用英语口语、印地语和日语字幕进行训练），而 Ohishi 等人 [43]提出了一种数据扩增方法，用于学习多语言语义嵌入模型。该方法使用包含图像及其多语言音频描述的不相交数据集，且不同语言之间没有共享图像。

在本学位论文中，我们不采用联合表示方法来捕捉不同模态之间的互补信息。相反，我们以另一种方法为起点，并将在第2.2.3节中回顾该方法。

2.2.2. 协同表示

与第2.2.1节回顾的方法将各模态投影到联合空间不同，协同表示为每种模态分别学习独立的表征，并通过约束协调这些表征，以确保它们之间的相似度 [85]。图2.3展示了协同表示方法的示意图：静态图像、其书面字幕以及口语字幕各自都有独立的表征，但这些表征通过相似度函数进行协调。最早采用协同表示的研究基于图像和书面文本模态。Weston 等人 [89, 90] 通过简单的线性映射构建特征空间，并使用加权近似秩成对损失来协调该空间，使匹配的图像与文本图像描述表征之间的内积高于不匹配表征之间的内积，这是最早基于此类表征开展的研究之一。

与线性映射不同，近期的研究采用了神经网络。Frome等人 [91] 使用深度神经网络学习视觉语义嵌入空间，并协调

该空间与铰链排序损失；在这个协调空间中，书面单词汽车 与汽车 图像之间的距离，比书面单词汽车 与昆虫图像之间的距离更近。Kiros等人 [92] 通过使用长短期记忆网络模型构建特征空间，并使用成对排序损失协调该空间，将这一方法扩展到句子与图像的协同表示。Socher等人 [93] 研究了同一任务，但使用决策树循环神经网络模型将组合语义融入句子嵌入，以检索由书面句子描述的图像。

类似的方法如今已扩展到静态图像和书面图注之外。Pan 等人 [94] 提出了一种与 Socher 等人的方法类似的模型，使用视频而非图像。Xu 等人 [95] 也构建了视频与句子之间的协调空间；随后将这种表征用于跨模态检索和视频描述任务。最近，Harwath 等人 [4]利用 VGS 数据集扩展了这一方法，该数据集包含静态图像及其口语描述（见图 2.4）。作者构建了一个协调空间，在其中可以突出显示图像中与语音描述内容相关的区域。该模型主要通过训练来执行整幅图像与完整图注层面的语义检索。但它也展现出无需额外训练即可检测和定位视觉对象及口语词语的能力。Harwath 等人的这一模型与第2.2.1节中介绍的方法的区别在于：他们不是将整幅图像和口语话语映射到嵌入空间中的固定点，而是学习在空间和时间上分布的表征，使模型能够直接在两种模态中进行共同定位。在本学位论文中，我们不使用

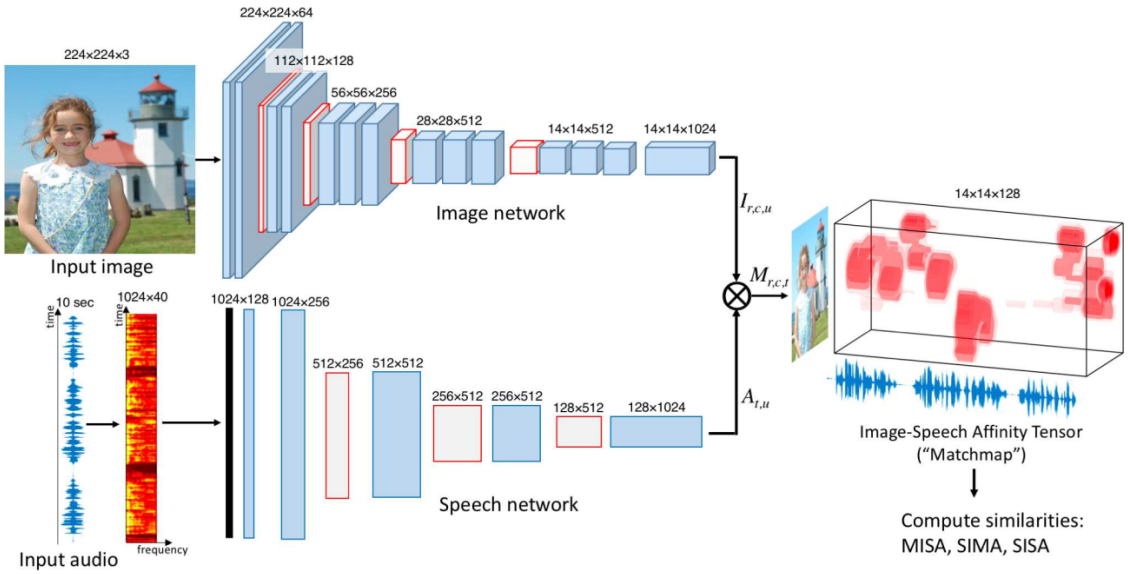


图 2.4： [4]的视听模型架构。蓝色表示卷积神经网络层，红色表示池化层，黑色表示批归一化层。每个卷积神经网络层后都接一个 ReLU。图中展示了一个约 10 秒（1024 帧）的频谱图输入示例，以说明池化比例。右侧展示了一个输出示例，呈现图像和语音特征对之间的时空相似度。

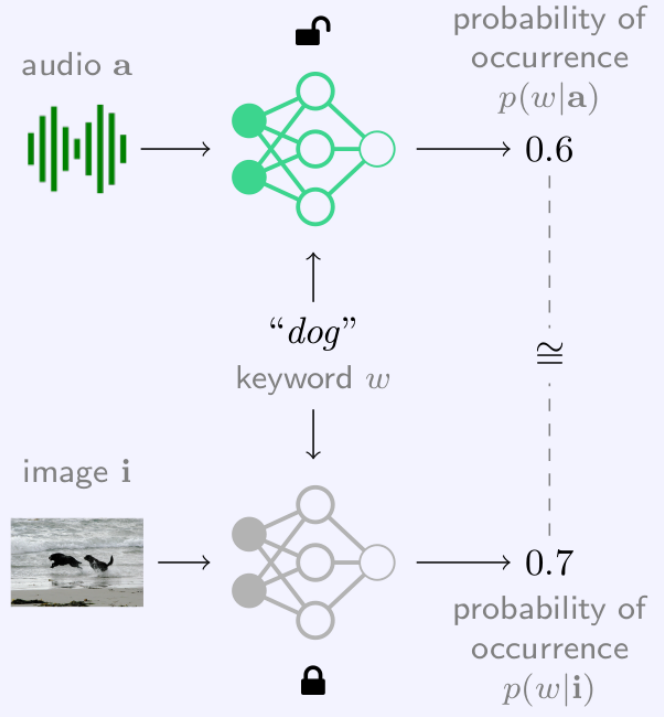


图2.5： 离线跨模态蒸馏结构。教师模型（图像模型；底部）的训练与学生模型（音频模型；顶部）的训练相互分离。

这种协同表示来捕捉不同模态之间的互补信息——我们采用另一种方法，并将在下文介绍。此外，尽管 Harwath 等人的方法能够在提供相应图像时定位某些单词，但我们研究的是在向模型提供书面关键词时进行关键词定位，这是一个不同的任务。

2.2.3. 跨模态蒸馏

在第2.2.1节和第2.2.2节中，我们回顾了多模态学习的两种不同方法。本节将讨论并综述与本学位论文最相关的方法。上述方法定义了联合多模态空间或协调空间，并在其中以端到端方式训练多模态系统；而本学位论文采用的多模态学习方法，则是利用一个系统（理想情况下是一个成熟的“教师”模型），从一种模态（理想情况下是资源充足的模态）中提取互补知识，并利用这些知识指导在另一种模态（理想情况下是资源匮乏的模态）上训练的“学生”模型的学习。图2.5展示了本学位论文所采用的具体方法，该方法最初由 Kamper等人 [2]在图像-语音场景中提出（我们将在第2.4节中对 [2] 作进一步介绍）。在这一方法中，图像模型（底部中间位置，带有 🔒 符号，表示离线训练）充当教师，音频模型（顶部中间位置，带有 b 符号）充当学生。

通过由教师网络监督学生网络，这种研究方法可以被视为一种知识蒸馏技术：一种将知识从较大的深度神经网络迁移到小型网络的方法 [96, 97]。知识蒸馏 [5] 最初旨在应对这样一项挑战：如何将深度学习模型部署到资源有限的设备上，例如手机和嵌入式系统。图2.6展示了用于知识蒸馏的教师-学生网络示意图——图左侧的教师模型描绘了一个层数很深的神经网络，它是被蒸馏到右侧浅层学生模型中的知识来源。最初的方法使用大型网络在单模态设置下训练较小的网络，即这些方法用于模型压缩。本学位论文不涵盖对这些方法的全面综述，但 [5]中提供了近期的综述。

与早期的知识蒸馏方法相比，我们在知识传递方式（通过教师模型生成的逻辑值）以及教师模型固定（离线）这一点上有相似之处，但不同之处在于我们的方法跨模态运行。这一思路属于跨模态蒸馏：通过利用另一种（资源丰富的）模态中的知识，帮助对一种（资源匮乏的）模态进行建模 [85]。

跨模态蒸馏此前已应用于多种任务。Gupta 等人 [98] 提出了一种技术：利用在标注充分的（RGB）图像模态上训练的教师模型获取监督信号，并将这些监督信号用作未标注深度图像和光流图像的标签。具体而言，作者在这些未标注的深度图像和光流图像上训练卷积神经网络模型，使其重现从标注充分的（RGB）图像模态中学习到的中层语义表征。两种模态是并行的。Zhao 等人 [99] 采用视觉模型为一个解析无线电信号以估计二维人体姿态的深度神经网络提供跨模态监督。具体而言，在训练期间，系统使用同步的无线和视觉输入，从视觉模态中提取姿态信息，并以此指导训练过程。训练完成后，网络仅使用无线信号进行姿态估计。Aytar 等人 [100]利用未标注视频中视觉与声音之间的自然同步，提出了一种学生-教师训练流程，将成熟视觉识别模型中的判别性视觉知识迁移到声音模态，以学习自然声音表示。与此类似， [100], Albanie 等人 [101] 利用未标注视频作为桥梁，通过蒸馏预训练面部情感识别网络的知识，训练了一个用于语音情感识别的卷积神经网络。Thoker和Gall [102] 研究了如何将基于RGB视频训练的动作识别模型适配为基于骨架的人体动作识别模型。具体而言，作者提取了一个在RGB视频上训练的教师网络中的知识，并将其迁移到一个由小型学生网络组成的集成中，该集成以骨架序列作为输入。每个RGB视频都与一个骨架序列配对。

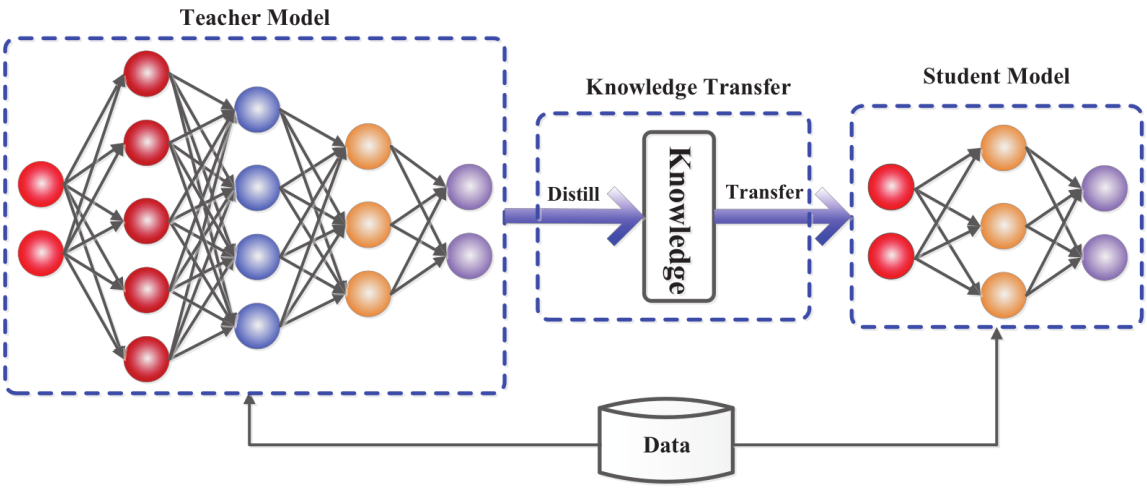


图 2.6： 用于知识蒸馏的教师-学生网络 [5]。

为提升人体动作识别模型的性能，Garcia 等人 [103]引入了深度图像这一额外模态，以进行跨模态蒸馏。在测试阶段，训练好的模型接收RGB视频流并生成幻觉流。为了帮助学生模型捕捉教师网络中的重要信息，Tian 等人 [104] 引入对比损失，以迁移不同模态之间的结构关系。作者证明，新的对比目标损失在多种知识蒸馏设置下的知识迁移任务中取得了具有竞争力的性能，包括单模型压缩、集成蒸馏和跨模态迁移。

与本学位论文采用的跨模态蒸馏方法类似，所有这些研究都利用视觉网络进行监督，且不需要任何标注数据。相反，它们都使用两种模态的序列对进行监督。此外，辅助模态（教师）仅在模型训练期间使用，在测试阶段不会使用。其他研究也表明，教师–学生角色可以互换。Owens 等 [105, 106] 使用声音信息监督视觉网络；更一般地说，Alwassel 等 [107], 通过对来自另一模态的特征进行聚类所得的伪标签，协同训练两个网络（音频和视觉）。这些方法学习到的表征已成功迁移至声音分类 [100, 107], 、视频分类 [107], 以及物体或场景识别 [106]等任务。

在本学位论文中，我们的下游任务与语音有关，具体而言是关键词检测和关键词定位。作为起点，我们采用 Kamper 等人 [2, 39] 提出的模型；他们的跨模态蒸馏方法与本学位论文开展的研究最为相关。虽然我们在此介绍了 Kamper 等人 [2] 工作的若干方面，以说明其如何被视为一种知识蒸馏方法，但有关这一研究方法的完整概述及详细内容，我们将在第2.4节中介绍。

2.3. 视觉接地语音系统的可解释性与分析

除了探索如何以最佳方式结合多种模态的表征之外，也有若干研究分析并解读这些表征。本节将回顾此类研究，尤其是与视觉接地语音模型相关的研究。

Harwath 和 Glass [18] 研究了在一种基于卷积的视觉接地语音系统中，如何产生可解释的子词语音单元。该系统经过训练，能够将原始语音波形映射到语义相关图像。Gelderloos 和 Chrupa la [25] 表明，他们的视觉接地语音模型以层级方式捕捉音素序列中的语言信息，其中低层对形式更敏感，高层对意义更敏感。Drexler 和 Glass [108] 表明，视觉接地语音模型生成的特征包含更多具有判别力的语言信息，且比传统语音特征更少依赖说话人。Havard 等人 [109]表明，视觉接地语音模型并不依赖全局声学模式来激活准确的词表示，只需获取目标词的第一个音素。他们还发现，只有一部分语音帧在特定单词的表征中发挥关键作用。Scholten等人 [110]研究了能否通过嵌入孤立词，并用它们检索其视觉指称物的图像，来识别单词。Havard等人 [111] 表明，在英语和日语图注上训练的基于注意力的视觉接地语音模型，会在图像的引导下学习定位话语中的名词。Havard等人 [109]和 Scholten等人 [110] 研究了视觉接地语音模型学习词语到其视觉指称物映射的效果。

Kamper等 [39] 研究了一个从图像及其（未经转录的）口语描述中学习的模型，如何捕捉语音中的语义特征。在第4章中，我们通过定性分析表明，视觉接地语音模型的一些错误关键词定位源于语义混淆，例如在查询关键词为游泳时，将单词仰泳 定位出来。我们还在第5章中表明，即使在跨语言环境下，这类语义混淆依然存在；在这种环境中，视觉接地语音模型经过训练，能够在 Yor`ub´a 语音中定位英语关键词。

在第4章中，我们对视觉接地语音模型进行了更多基于任务的分析：通过详细实验揭示视觉接地语音定位检测步骤的局限性、视觉接地语音架构对各项任务的影响，以及模型在单个关键词上的性能。在第5章中，我们探究使用更多数据会对视觉接地语音模型产生什么影响，以及从预训练模型初始化跨语言视觉接地语音模型的训练会如何影响模型性能。

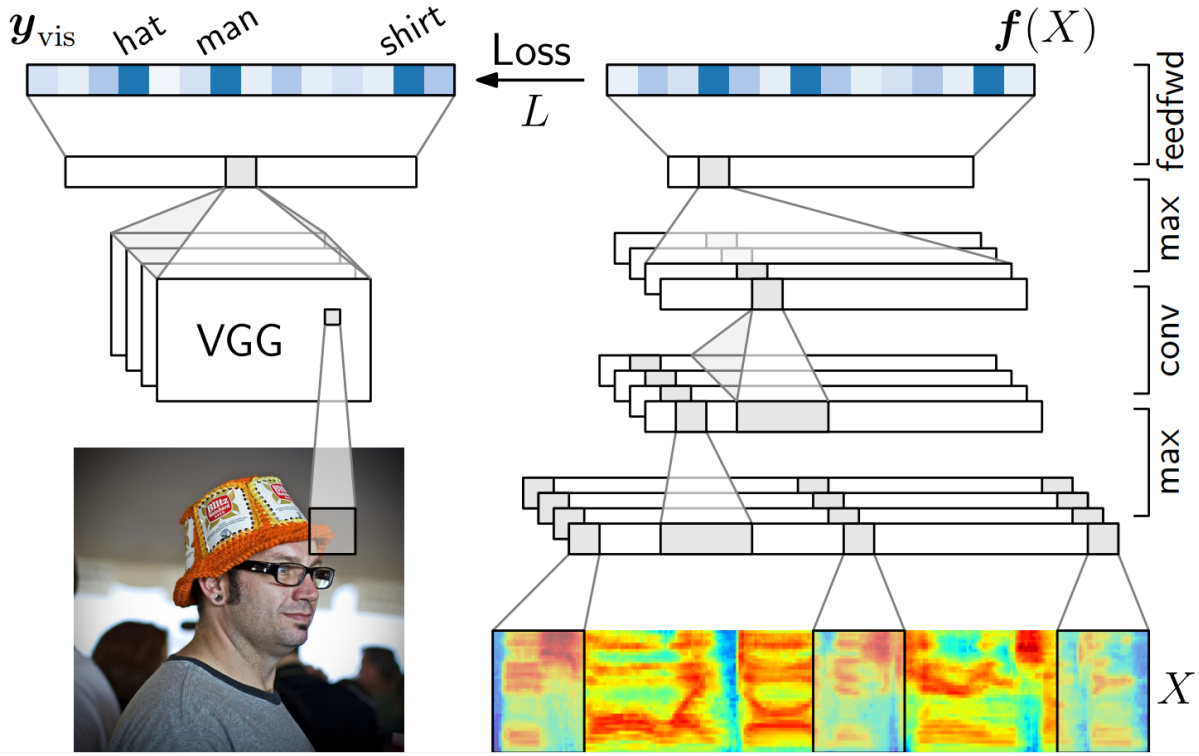


图2.7： 用于为语音系统（右侧）生成训练概率目标的多标签图像分类器（左侧） [2]。

2.4. 视觉接地（粗粒度）语音识别

在第2.2节中，我们从更广泛的角度回顾了与多模态学习相关的研究。在此，我们具体回顾学位论文中所采用的视觉接地语音方法，该方法由 Kamper 等人提出 [2]。由于该模型采用知识蒸馏方法，这一方法已在第2.2.3节中作了简要介绍。采用这一方法，视觉接地语音模型通过静态图像及其口语描述进行训练，以执行一种粗略形式的语音识别：给定一个关键词，模型能够检测该关键词是否出现在某段口语话语中。

第2.2.1节中回顾的研究（如图2.1所示）以及第2.2.2节中介绍的研究（如图2.3所示），分别定义了联合图像—语音空间，以及能够在不同空间之间转换的模型。尽管这些研究本身很有价值，但并未提供从语音到文本标签的映射。Kamper 等人 [2,39]解决了这一问题。在他们开创性的工作中，他们首次在 VGS背景下应用了离线跨模态蒸馏方案（见第2.2.3节）。Kamper 等人 [2] 采用了两阶段训练过程：(i) 首先在一组标注完善的静态图像上训练大型教师模型（视觉模型）；(ii) 然后利用教师模型以逻辑值的形式提取知识，并在蒸馏过程中用这些知识指导学生模型（音频模型）的训练。他们的架构如图2.7所示，其中包含一个图像分支，由成熟的 VGG-16 网络担任教师模型；

另有一个语音分支，由结构不同的卷积神经网络担任学生模型。（离线蒸馏与在线蒸馏的关键区别在于：在线蒸馏中，教师模型和学生模型均采用端到端方式训练，并同时更新；而在离线蒸馏中，教师模型的训练与学生模型的训练相互分离，如图2.5所示。）

在整篇学位论文中，我们都以这一特定模型 [2] 为出发点。不过，在我们的案例中，语音输入是梅尔频率倒谱系数（MFCCs），而非图2.7所示的频谱图。具体而言，为了检测书面关键词是否出现在某段话语中，作者使用了一个预训练视觉标注器（教师模型），为训练图像生成软标签；这些软标签随后作为卷积神经网络（学生模型）的标签，该网络将语音映射为无序的软词语目标。在测试阶段，语音模型便能够将输入语句映射为一组可能出现在该话语中的文本词语。关键在于有一个预训练图像标注器可用，它被用于监督语音模型。更确切地说，给定一幅图像，视觉标注器会为固定词汇表中的视觉类别生成软标签，而语音模型则尝试在相应的口语话语上模仿这些标签。大量大规模标注图像数据集 [83,112–115] 以及性能出色的视觉模型 [116–119] 确保了预训练视觉组件的可用性。Pasad 等人 [120] 指出，即使在有（有限）文本标签可用的情况下，这种方法也能发挥互补作用。

尽管这些研究已经证明可以检测关键词，但这些视觉接地语音模型是否也能定位查询关键词，仍然是一个悬而未决的问题。在本学位论文中，我们将这一研究方法加以扩展，构建同样能够定位语音中查询关键词的视觉接地语音模型。我们在人工低资源语言环境和真实低资源环境中均应用了这些模型。

2.5. 与VGS建模密切相关的多模态范式

本节简要回顾三种多模态学习范式，它们具体启发了近期一系列视觉接地语音研究。其中一些范式早于上述许多关于VGS建模的研究。

2.5.1. 音频-视觉（细粒度）语音识别

第2.4节回顾的研究着重于粗粒度语音识别，我们将其定义为从语音中稀疏地识别关键词的任务；而本节回顾的研究则涉及语音的细粒度识别。音视频语音识别（AVSR）的目标与自动语音识别（ASR）相似，即训练语音识别器自动将话语映射为转录文本；但不同之处在于，模型训练过程中还使用了另一种

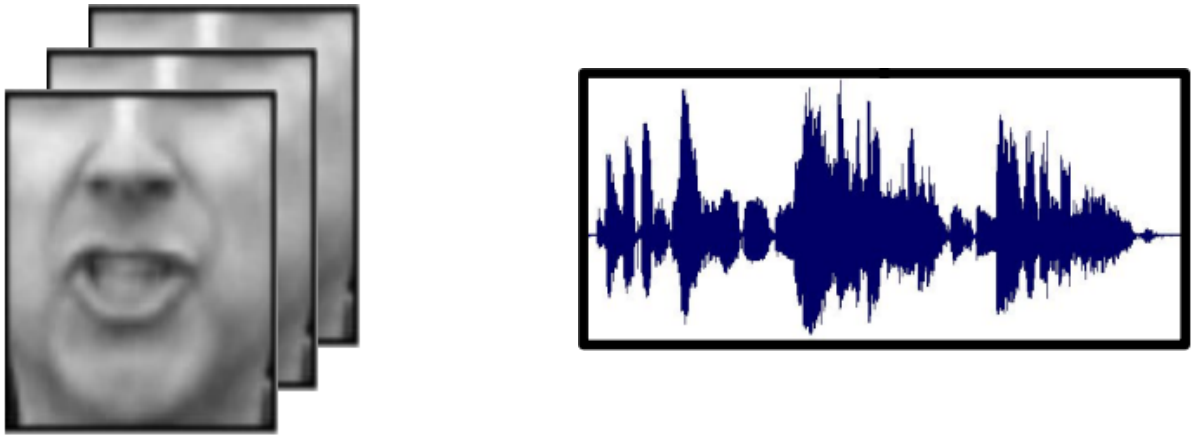


图2.8： 一个训练样本，其中包含一名说话人嘴部区域的序列（左），以及该说话人说出的对应话语（右） [6]。

模态（视觉信号），也就是说，语音是在视觉图像（通常是说话者的脸）的语境中出现的，如图2.8所示。音视频语音识别系统专门从语音音频、文本转录以及与语音相关的视觉上下文中学习。这种训练语音识别器的方法源于这样一种观察：大多数言语交流发生在听者既能看见又能听见说话者的情境中 [121]。人类经常通过唇读可见的面部言语发音动作来增强语音识别能力，尤其是在声学信号受到噪声或听力损伤影响而退化时 [122]。早在 1954 [123]，人们就已注意到在噪声环境下利用视觉信息处理语音的益处。随后， [121], 开展了一项关于言语感知过程中听觉与视觉交互作用的研究，展示了音频和视觉刺激在语音识别中的整合。此后的一系列后续研究 [122, 124, 125] 致力于将视觉信息与语音信息相结合，以改进自动语音识别。大多数此类研究试图回答的核心问题是：（i）能否自动提取视觉语音信号所传达的语音信息？（ii）如何将这些信息与声学信号中的信息结合起来，以改进自动语音识别？ [122]。与本学位论文更相关的是后一个问题，即如何将视觉信息与语音信息相结合，以便在缺乏充分转写语音数据的语言环境中促进下游任务。

许多近期研究采用神经网络技术，学习并融合文本、音频和视觉模态的表征。Ngiam 等人 [44] 对语音音频及其配套的唇部视频进行音视频语音分类，并证明在学习时存在多种模态，能够学到更适合单一模态的特征。Huang 和 Kingsbury [45] 利用深度信念网络，从音频和视觉特征中学习噪声鲁棒语音识别。在连续语音数字识别任务中，他们表明其模型在词错误率方面表现更佳。Mroueh 等人 [46] 提出了一种基于深度神经网络架构的多模态学习方法

用于融合语音和视觉信息，以实现视听语音识别。作者表明，在音素分类任务中，该视听语音识别系统优于未使用视觉组件训练的音频网络。Afouras 等人 [6] 研究了音频信号存在噪声时，唇读如何补充音频语音识别。Petridis 等人 [47] 提出了一种基于残差网络和双向门控循环单元的端到端音视频模型，用于在大型公开数据集上进行上下文内词语识别。该模型由两条流组成，分别对应一种模态，直接从嘴部区域和原始音频波形中提取特征。Ma 等人 [126] 提出了一种混合连接主义时序分类与注意力模型，该模型基于 ResNet-18 和卷积增强型Transformer（Conformer），能够学习直接从原始像素和音频波形中提取特征。在所有这些音视频语音识别系统中，视觉和音频模态的标签均可用于指导系统训练。

我们在整篇学位论文中使用的数据集，与视听语音识别所用数据集之间有一项重要区别：前者由静态图像及其口语描述组成，而后者通常由说话者的脸部图像或视频，以及他们所说的话组成。此外，我们的目标不是自动语音识别，而是在话语中检测、定位并确定预先指定的关键词类型。Prajwal等人 [127] 研究了静音视频中的口语关键词定位，而我们研究的是音频话语中的书面关键词定位。

2.5.2. 从图像和文本中学习

另一种早于VGS出现的多模态学习范式涉及从图像和文本中学习。鉴于人类婴儿的语言习得很大程度上植根于周围的视觉世界这一观点 [128–130], ，许多早期研究考虑在配对图像和人工撰写的文字描述上训练多模态系统。这方面的研究表明，此类系统能够通过一串关键词描述视觉场景来生成图像描述 [131, 132], ，甚至能将图注中的特定关键词与特定图像区域关联起来 [133–136]。

将图像描述中的关键词与图像区域相链接时，通常会考虑两种范式。一种是稀疏连接关键词与图像分割区域，即多模态系统优先将特定词性（例如名词）的关键词与特定图像区域相链接 [137]。另一种是将图像描述中的关键词与特定图像区域进行密集连接，而不考虑每个关键词的词性。例如，在“一位女性拿着气球”中，“拿着”应与该动作发生的图像区域相链接 [138]。然而，这种方法需要多模态图像标注，并且在训练期间能够获取图像描述中每个关键词与其出现的特定图像区域之间的对齐信息。

尽管我们的工作与这些研究相关，但我们关注的是这样一种情形：训练期间无法获得书面

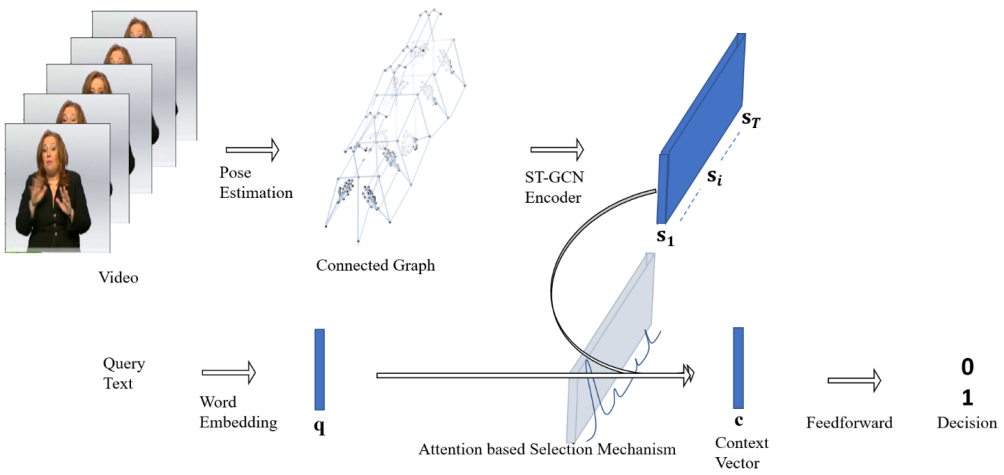


图 2.9： [7]中用于手语关键词搜索的基于注意力的时空图卷积网络。我们的基于注意力的方法（第 4.2.3 节）深受这项工作的启发。

图注，只能获得图像及其口语描述。此外，与 [138] 中的情况不同，我们在训练期间没有对齐信息。

2.5.3. 手语中的关键词搜索

另一种与 VGS 建模相关的多模态学习范式，是构建一个能够在手语中搜索书面查询的系统 [7, 139]。为了将这项技术扩展到手语，Tamer 和 Sara˙clar [7]提出了一种手语关键词搜索系统。如图 2.9 所示，作者首先从手语句子的各帧中提取身体和手部关节，并用骨架关节构成的统一时空图来表示。随后，他们以弱监督端到端方式共同训练图卷积网络编码器、查询嵌入和选择机制。这项研究与本学位论文有两方面关联。首先，手语可视为一种低资源语言（带有手语标签（即手语词 gloss）的可用数据有限 [7]）。本学位论文同样考虑低资源场景。其次，我们采用了与 Tamer 和 Sara˙clar 提出的方法类似的注意力机制来进行关键词定位。不过，我们将注意力融入卷积神经网络，以定位语音中的关键词；训练模型时使用视觉监督，而不是使用显式标签标明某个关键词是否出现在话语中。关于该方法的完整细节见第 4.2.3 节。

2.6. 关键词检出与定位

本节回顾与我们尝试通过视觉接地语音（VGS）方法完成的主要任务相关的研究。首先，我们回顾不涉及自动语音识别（ASR）的关键词检出研究，以及相关模型

的训练过程未使用自动语音识别或视觉监督（第2.6.1节）。接着，我们回顾与语音中的关键词定位相关的研究（第2.6.2节）。最后，我们回顾与跨语言学习相关的研究（第2.6.3节），这是第5章的主要焦点。

2.6.1. 无需自动语音识别或视觉监督的关键词检出

传统上，关键词检出通过自动语音识别系统实现，而这需要大量转录语音。对于资源匮乏语言而言，收集这些资源的成本很高。因此，已有多项研究探索其他途径，以便在资源受限的环境中实现语音技术，而不依赖传统的自动语音识别。本文将重点回顾这类研究，具体来说，是那些在没有视觉监督的情况下学习语音表征并构建关键词检出系统的研究。这与本学位论文的重点有所不同：本论文采用第2.4节回顾的跨模态蒸馏方法，在多模态设置下利用图像及其口语描述训练关键词检出系统。

关键词检出与关键词检测和口语术语检测密切相关。正如第1章所述，关键词检出的目标是找到包含关键词精确匹配 [140–142] 或语义匹配 [143–146]的话语。Bird [32] 指出，在假设已有少量孤立口语词语可用于启动流程的前提下，口语术语检测能够实现濒危口语的稀疏转写。Ferrand 等 [147] 进一步推进了这一工作，提出一种口语术语检测系统，借助人工专业知识，在几乎零资源的设置下支持语音转录。Van der Westhuizen 等 [148] 研究了特征学习，旨在实现一种计算高效的关键词检出方法，可应用于资源极度匮乏的环境。其目标是在几乎没有语言资源的地区实现语音技术。

所有这些研究都使用了有限的转写语音资源：它们都依赖一小组易于收集的孤立关键词。在本学位论文中，我们提出了用于关键词检出的视觉接地语音模型，这些模型无需任何文本转录或任何孤立关键词集即可训练。我们更进一步，为这些模型增添了关键词定位能力，使其在训练期间无需关键词对齐信息。人们也曾尝试在零资源设置下进行语音识别 [149–151] ，但本学位论文不涉及这种情况——在我们的研究中，我们依赖于某种监督信号，这些信号是通过使用成熟的现成视觉标注器处理资源丰富的图像模态而获得的。

2.6.2. 语音中的关键词定位

关键词定位是指找出给定关键词在话语中出现的位置。这不同于关键词检测，后者只需判断给定关键词是否出现在

话语中，而不考虑关键词的位置。使用视觉接地语音模型进行语音中的关键词定位是一项极具挑战性的任务，因为训练阶段既没有提供文本转录，也没有提供对齐信息 [152]。

视觉接地关键词定位可应用于濒危语言的记录工作 [147]。语言学家可以利用此类工具，在话语集合中快速定位包含查询关键词的语音片段；唯一的前提是拥有来自目标语言的图像-语音对集合，用于训练模型。近期研究开始构建语音模型，使其不仅关注生成语音的转录文本，还能对语音中的关键词进行对齐。目前用于对齐语音中关键词的常见方法采用自视觉领域。下面将讨论其中一些方法。

弱监督定位

通常，定位任务可以表述为一个多示例学习问题[153, 154]：样本是由多个实例组成的包；如果至少有一个实例为正，则该包被标记为正，否则标记为负。其目标是仅根据包级标注来推断实例的标签。在计算机视觉领域，相关的弱监督目标定位任务近年来受到了广泛关注，综述请参见 [155, 156]。

我们从这些方法中汲取灵感，但重点关注能够利用训练好的模型中已有线索的定位方法。除其他方法外，我们采用了 [54, 157, 158]，该方法利用网络激活来进行定位；还采用了 [159, 160]，，通过遮挡输入的部分内容来识别与分类相关的区域。我们的工作对这些方法进行了调整，使其能够作用于听觉信号，尤其适用于关键词定位任务。

2.6.3. 跨语言学习

在第5章中，我们考察了一种比第2.4节所述语言环境更为极端的语言环境。在这种极端环境下，濒危语言中只有数量有限的未转录话语，而书面关键词仅以另一种资源丰富语言提供，在濒危语言中则没有。许多研究通过跨语言学习，探索了在此类语言环境下处理语音数据的方法。

跨语言学习的核心思想是利用资源丰富语言的数据来支持资源匮乏语言的处理。早期的跨语言学习研究采用了一种直接方法，即使用文本转录来训练系统，而不是进行视觉定位。例如，Sheridan 等人 [53] 提出将自动语音识别与基于文本的跨语言信息检索级联起来 [161]。然而，只有在目标

语言中有已转录语音可用于构建识别器时，这种方法才行得通。近期研究提出了一些模型，可以将一种语言的语音直接转换为另一种语言的文本 [9, 12, 162, 163],，但这些方法同样需要转录数据：带有译文的平行语音。Kamper 和 Roth [35] 研究了无需转录文本的跨语言关键词检索，并利用视觉信息作为语言之间的纽带（详见第 5.1 节）。我们的方法也采用了这一核心思想，但所用语言和解决的任务有所不同。

跨语言学习面临的挑战

如上所述，跨语言学习涉及利用一种来源语言（通常资源丰富）来处理一种目标语言（理想情况下资源较少）。正如 [164] 中详尽讨论的那样，不同语言之间的语言和文化差异可能会给跨语言模型带来异常。Gao [165] 和 Larina [166] 强调了一种语言与另一种语言在交际风格上的文化差异，并指出，在一种语言中被认为礼貌的表达，在另一种语言中可能被认为具有冒犯性。例如，德语母语者在德语交流中往往会表现出较高的直接程度，而正如 House 和 Kasper [167]所讨论的，这在英语中会被视为冒犯。正如 [168–171]所讨论的，情感表达也因文化而异。我们将在第5章重新讨论这一问题，届时将说明语言和文化变异如何影响我们的跨语言数据集以及跨语言语音关键词检索任务。

2.7. 相关工作总结

本章回顾了与现有数据集相关的主题，这些数据集经过开发和整理，旨在支持多模态学习。随后，我们回顾了利用多种模态中的互补信息的不同方法。此外，我们还回顾了分析多模态系统的研究，重点关注视觉接地语音模型（VGS）——这是本学位论文始终讨论的多模态系统类型。我们进一步聚焦于 VGS，回顾了与本学位论文最相关的研究。

本章还回顾了其他类型的多模态系统，包括视听语音识别系统以及从图像和文本中学习的系统。随后，我们回顾了与本学位论文所开展的主要任务相关的研究：首先回顾了不使用自动语音识别和视觉监督训练的模型所涉及的关键词检出研究；接着回顾了语音中的关键词定位研究；最后回顾了跨语言学习研究。

第3章

视觉 grounding 的关键词检测与检索

本章介绍用于关键词检测和检出的视觉接地语音（VGS）模型。这些模型此前已有提出 [2, 39]，但在本学位论文中，我们对其进行了更新，使其适用于关键词定位（这项任务仅在后续章节中讨论）。本章还提出了利用注意力机制进行关键词检测和检出的新方法，此前尚未有研究探讨过这类方法。提醒一下，我们的关键词检测系统使用从与口语描述配对的图像中提取的软标签进行训练；在测试阶段，训练好的视觉接地语音系统接收一段话语和一个书面查询，以预测该话语中是否出现查询关键词。在关键词检出任务中，视觉接地语音系统接收一组话语和一个书面查询，并输出包含该查询实例的话语。

本章旨在介绍关键词检测模型。在本学位论文后续章节中，我们将扩展这些模型，以实现我们的主要目标——关键词定位。首先，我们将介绍本章多项实验所用的数据集（第 3.1 节）。随后，我们将在第 3.2 节详细说明评估流程。在第 3.3 节中，我们将介绍所采用的方法，包括视觉接地语音架构及其实现细节。在第 3.4 节中，我们将展示视觉接地语音模型在关键词检测和关键词检出任务上的基线结果。此外，我们还将展示上界结果，该结果是通过使用无序转录文本训练的模型进行估算的，而非使用自动获取的视觉标签训练的模型。

本章为回答研究问题1（第1.2节）奠定基础：视觉接地语音模型能否实现关键词定位？本章不对此问题作答，但会介绍我们将在后续章节中扩展的检测模型。

本章部分内容发表于：K. Olaleye、D. Oneat˘a 和 H. Kamper，“使用视觉接地语音模型在未转写语音中进行关键词定位”，《信号处理精选主题期刊》，2022年。

3.1. 数据集

本节介绍本章和第4章使用的主要数据集，以及第5章部分对比实验所使用的数据集。由于我们的任务涉及三种模态——评估时使用音频和文本（关键词），训练时使用音频和视觉——因此我们在文献中可能是最常见的多模态数据集上开展实验：Flickr8k 及其音频扩展数据集 Flickr 音频描述语料库（FACC）。

3.1.1. Flickr8k

Flickr8k 数据集 [48, 49] 是一个双模态（图像和文本）数据集，包含 8,000 张图像和 40,000 条英文图注。每张图像都由五位不同的人分别标注文本描述。这些图像（及其图注）通常描绘涉及人物或动物的关系和动作。Flickr8k 最初用于图像描述生成和跨模态检索，后来扩展加入了译文 [172, 173], 口语描述 [1],以及视觉概念的边界框标注 [137]。我们尤其关注该数据集的语音版本，下一节将对此进行介绍。

3.1.2. Flickr 音频描述语料库（FACC）

Flickr 音频描述语料库（FACC）由 Harwath 和 Glass [1]开发。他们使用亚马逊土耳其机器人，通过众包为原始 Flickr8k 数据集（见上文）中的每条文本描述收集了约 36 小时、采样率为 16 kHz 的非静音英语语音。随后，FACC 支持了多种视觉接地任务，例如图像—语音对齐 [1], 语义语音检索 [39], 图像到语音合成 [50] 以及语音到图像生成 [51]；这些研究将在第2章中回顾。

在本学位论文的其余部分中，我们在 FACC 上进行实验。FACC 为约 8k 张图像中的每张图像提供五条英语音频描述。训练、开发和测试分别使用 3万、5k 和 5k 条话语。我们的数据划分与其中使用的标准训练、验证和测试划分相对应

表 3.1： Flickr 音频描述语料库（FACC）的定量特征。整个语料库的时长约为 36 小时。

特征	总计	训练集	Val	Test
说话人	183	183	176	176
时长（小时）	36.0	27.0	4.5	4.5
图像	8k	6k	1k	1k
话语	40k	30k	5k	5k

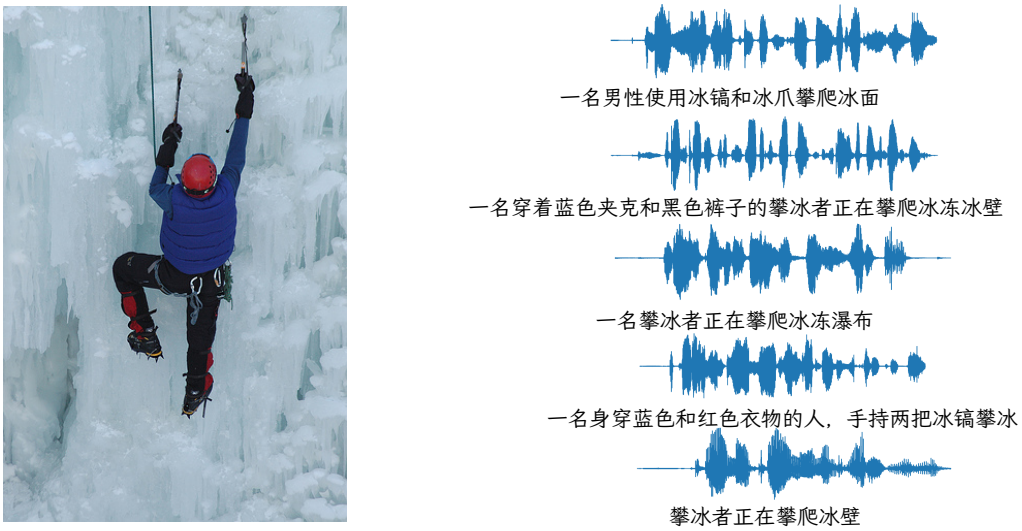


图 3.1： Flickr 音频描述语料库（FACC）中的一个示例，展示了一张图像及其对应的口语描述。训练任何模型时都不会使用文本描述，此处仅用于说明。评估我们的方法时，我们确实会使用文本转录。

。虽然各划分之间的图像或单条话语均不重叠，但不同集合之间存在说话人重叠。

表 3.1 列出了 FACC 各个数据划分（训练集、验证集和测试集）中的图像数量和话语数量。表中还显示了录制语音的说话人数量，以及整个数据集的时长。本章和第4章中的所有实验均基于 FACC，我们还在第5章中将该数据集用于部分比较模型。图 3.1 展示了一张图像及其对应的口语描述和文本描述。在训练视觉接地语音模型时，我们从不使用文本描述；但实际上，我们会用它们训练理想化模型，以估计视觉接地语音模型性能的上界。这些词袋（BoW）标签被视为一种弱监督形式，因为它们指出了单词是否出现，却没有提供其位置、顺序或出现次数（详见第 3.3.2 节）。评估我们的模型时也会使用文本转录。

3.1.3. FACC 中的视觉关键词

一些研究通过视觉定位，在 FACC 上进行了关键词检测和关键词检出。这些研究 [39, 120] 考虑了 67 种不同的关键词类型。表 3.2 展示了这些关键词类型及其在训练集和验证集中的出现次数。选择这些关键词的流程详见 [39]；该流程包含人工审阅者一致性步骤，将原始词集从 70 个缩减至 67 个。这些关键词也带有时间戳，这些时间戳是通过强制对齐自动获取的。在后续章节中，我们还将使用这组相同的关键词进行关键词定位。

表 3.2： 本学位论文中所考虑的 67 个不同关键词类型列表。“训练计数”和“验证计数”列分别给出了每种关键词类型在训练集和验证集中的出现次数。

	关键词	训练计数	验证计数		关键词	训练计数	验证计数
1	air	154	23	35	race	58	7
2	baby	58	10	36	red	378	66
3	ball	264	47	37	骑行	73	11
4	海滩	147	26	38	骑行	145	21
5	bike	122	25	39	road	67	9
6	黑色	566	86	40	rock	117	10
7	boy	536	93	41	跑步	336	57
8	棕色	393	64	42	sand	58	15
9	建筑物	70	9	43	衬衫	264	41
10	相机	108	15	44	sits	84	15
11	car	61	12	45	坐着	202	40
12	携带	66	8	46	滑板	73	10
13	儿童	168	33	47	小	187	24
14	攀爬	95	11	48	微笑	65	11
15	dirt	88	16	49	snow	236	32
16	dogs	312	47	50	下雪的	64	14
17	face	67	8	51	足球	84	12
18	田野	191	20	52	看台	126	25
19	足球	78	11	53	棍子	69	14
20	草	238	40	54	街道	163	22
21	hair	62	8	55	游泳	61	13
22	hat	94	17	56	网球	65	12
23	持有	206	40	57	三	188	34
24	夹克	112	14	58	top	74	9
25	跳跃	149	30	59	toy	90	12
26	大的	180	35	60	tree	58	7
27	小的	265	61	61	步行	85	11
28	山	77	11	62	水	390	75
29	嘴	160	29	63	穿着	468	69
30	海洋	73	16	64	白色	583	85
31	橙色	105	13	65	女性	101	17
32	park	74	15	66	黄色	194	20
33	pink	93	25	67	年轻	393	72
34	pool	104	20				

3.2. 评估用于关键词检索任务的视觉接地语音模型

在整篇学位论文中，我们将探讨关键词检测、关键词检出和关键词定位，具体而言，是使用视觉接地语音模型。但在本章中，我们将严格聚焦于前两项任务，因为它们为第4章和第5章提出的关键词定位方法奠定了基础。本节将介绍关键词检测和

关键词检出任务。这些任务通常使用有监督的纯语音模型完成（见第 2.6.1 节），但在本学位论文的其余部分，我们将考察现有的视觉接地语音模型，并提出用于解决这些任务的新模型。在评估用于关键词检索任务的视觉接地语音模型时，需要注意的是，本学位论文中完全不会使用图像。

3.2.1. 关键词检测

在关键词检测中，我们给定一个关键词和一段音频话语（模型此前从未见过），任务是预测该关键词是否出现在话语中。给定一段音频话语，视觉接地语音模型会为我们考虑的每个关键词 w 生成一个分数 $\hat{}y_w \in [0,1]$ 。要执行硬关键词检测，我们设定一个阈值 θ ，并为所有满足 $\hat{}y_w \geq \theta$ 的 w 输出标签。我们将预测标签与转录文本中真实的单词集合进行比较，并计算精确率（P）、召回率（R），以及所有纳入考虑的关键词类型的平均 F_1。

3.2.2. 关键词检出

在关键词检出任务中，我们会得到一个关键词 w 和一个话语集合，并希望根据各话语包含该关键词的概率对话语进行排序。为此，我们按 $\hat{}y_w$ 值从大到小对话语进行排序。然后，使用此排序列表，通过若干度量指标评估关键词检出：

- 前10名精确率（P@10）是得分最高的十个候选项的平均精度。

- 前 N名精确率（P@N）是排名前 N个候选项的平均精度，其中 N表示该关键词的真实出现次数。例如，对于 FACC 验证数据中的关键词比赛， N 为 7，也就是说，我们将报告检索结果中排名前 7 的话语里包含关键词比赛的比例。

- 等错误率（EER）是误接受率与误拒绝率相等时的平均错误率。误接受率是指被错误分类为包含关键词的话语数量与实际不包含关键词的话语总数之比；误拒绝率则是指被错误分类为不包含关键词的话语数量与实际包含关键词的话语总数之比。

3.3. 使用视觉定位进行关键词检测与定位

下面详细介绍我们用于关键词检测和关键词定位的视觉接地语音模型在训练和测试时的推理过程。这里采用的研究方法已在第2.4节中概述。现在，我们将这一方法形式化。简而言之，我们获得图像和口语描述（两者均未标注），并将图像输入视觉标注器，为口语描述生成软标签。随后利用这些标签训练语音网络，以预测留出话语中是否包含关键词。训练和测试的操作流程如图3.2所示。

我们工作的主要组成部分是一个系统：给定一段音频话语$\mathbf{a}$ 和一个关键词 w，估计该关键词在该话语中出现的概率。该系统通过音频深度神经网络 DNN_a 实现，输出一个预测结果向量 $\mathrm{DNN}_a(\mathbf{a}) \in \mathbb{R}^{\mathsf{W}}$，其中每个元素对应系统词汇表中 W 个单词之一的得分。随后将该向量输入 Sigmoid函数σ ，以确保预测结果位于 $[0,1]$ ，并可将其解释为概率：

$$\hat{\mathbf{y}} \triangleq p(w|\mathbf{a}) = \sigma(\mathrm{DNN}_a(\mathbf{a})). \tag{3.1}$$

要获得单个关键词 w 的概率，我们从 $\hat{\mathbf{y}} \in [0,1]^{\mathsf{W}}$ 中选取向量里对应第 w 个关键词的条目，即 $\hat{y}_w$。下文将说明如何获得真实目标$\mathbf{y}$。（简而言之，我们使用视觉监督或词袋监督。）

3.3.1. 网络结构

我们假设音频网络 DNN_a 由按顺序排列的三个组件组成——音频编码器 Enc、池化层 Pool 和分类器 Clf:

$$\mathrm{DNN}_a = \mathrm{Clf} \circ \mathrm{Pool} \circ \mathrm{Enc}. \tag{3.2}$$

音频编码器 Enc 会生成一个由 K 维嵌入组成的序列 $[\mathbf{h}_1, \ldots, \mathbf{h}_{\mathsf{T}}] = \mathbf{H} \in \mathbb{R}^{\mathsf{K} \times \mathsf{T}}$，随后由池化层在时间维度（共 T 步）上对其进行聚合。最后，结果会输入分类器 Clf，以获得一个 W 维向量，其中包含各关键词对应的分数。我们采用这种结构（编码器、池化层、分类器）的主要原因在于定位任务（将在第4章详细讨论），其中池化层发挥着核心作用——某些方法依赖池化前激活来进行定位。

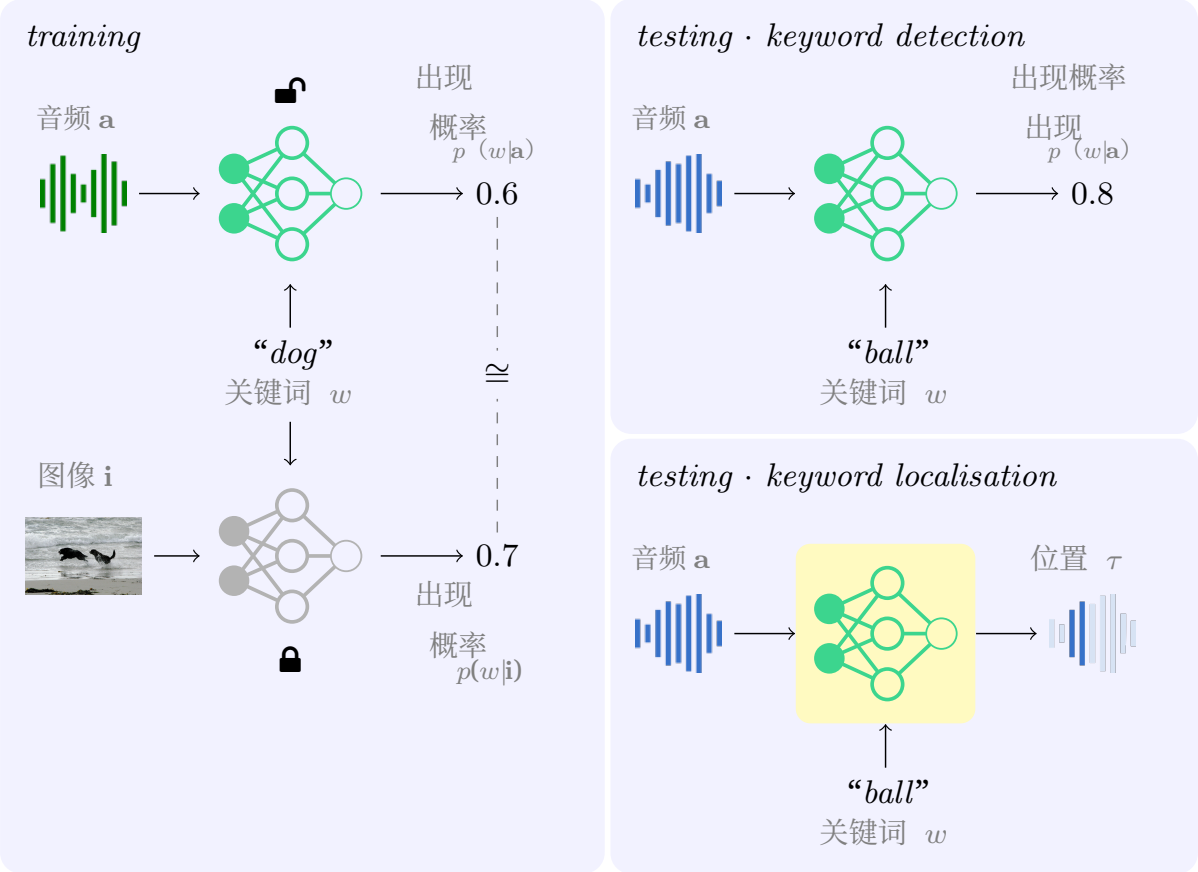


图3.2： 方法概览。左侧：我们训练一个音频神经网络，根据在对应图像 **i**上运行的固定预训练图像标注器所提供的监督，预测音频话语 **a** 中是否出现关键词 w 。右上：在测试阶段，该网络可用于判断给定话语中是否出现某个关键词（关键词检测）。右下：它还可用于定位输入中的关键词（关键词定位），这将在下一章讨论。

3.3.2. 训练

为了训练模型，我们使用视觉监督。具体而言，我们假设：(1) 训练数据中的每条话语都有一张与之匹配的图像；(2) 我们可以使用预训练图像标注器，根据图像生成覆盖同一 W 个单词词汇表的概率标签。音频网络的目标输出$\mathbf{y}^{\text{vis}} \in [0,1]^{\text{W}}$ 将是图像标注器生成的软标签：一个图像深度神经网络 DNN_i（下方紧接的小节将对此作更详细的说明）。因此，对于与话语**a**相关联的图像**i** ，我们得到：

$$\mathbf{y}^{\text{vis}} \triangleq p(w|\mathbf{i}) = \sigma(\text{DNN}_i(\mathbf{i})). \tag{3.3}$$

我们优化音频网络（以**a** 为输入）的预测结果与图像网络（以**i** 为输入）的预测结果之间的二元交叉熵：

$$L = y_w^{\text{vis}} \log \hat{y}_w + (1 - y_w^{\text{vis}}) \log(1 - \hat{y}_w). \tag{3.4}$$

L 对关键词和音频-图像对取平均，并针对音频网络的权重进行优化，同时冻结图像网络的权重。图3.2左侧展示了这一训练过程。

视觉标注器

我们的视觉接地语音模型通过将每张训练图像输入 [39] 的多标签视觉标注器，获得软文本标签，并据此训练模型进行关键词检测与定位。该标注器基于 VGG-16 [174] ，使用与本文所用数据互不重叠的图像进行训练，具体而言，使用的是 MSCOCO 中的图像 [114]。该标注器的输出包含 1000 个图像类别，但本文使用其中一个包含 67 种独特关键词类型的子集；其他几项研究也使用了同一组关键词 [39, 120]。选择这些关键词的流程详见 [39]；其中包含人工审阅者一致性步骤，将原来的 70 个单词筛减至 67 个。（有关这组关键词的更多详情，请参见第 3.1.3 节。）

词袋（BoW）监督

视觉标注器提供的真实标签在两方面较弱：(1) 没有定位信息——单词可以出现在话语中的任何位置；(2) 标签可能有误，因为音频描述可能与图像中呈现的视觉信息不同。为衡量 (2) 的影响，即使用视觉标注器提供的可能不正确标签所产生的影响，我们还进行了目标为词袋（BoW）的实验。词袋标注表示单词是否出现，但不提供其位置、顺序或出现次数。在这种情况下，我们根据音频样本的文本转录生成目标标签 y^{bow} ，即 $\mathbf{t}$ $y_w^{\text{bow}} = [\![w \in \mathbf{t}]\!]$，其中 $[\![-]\!]$ 表示示性函数。对于这一理想化设置，我们使用与公式 3.4 相同的二元交叉熵损失，但采用词袋标签。

3.3.3. 测试视觉接地语音模型

关键词检测

在测试阶段，音频神经网络的使用方式与训练时相同：给定一段音频话语和一个关键词，将它们输入网络，以获得指定关键词出现的概率。请注意，测试阶段不需要视觉信息。我们可以通过将概率分数 $\hat{y}_w$ 与值 θ_w 进行阈值化，来对关键词是否存在作出二元决策 w ，即 $[\![\hat{y}_w \geq \theta_w]\!]$。

图3.2右上方展示了这一过程。我们可以按照上文第 3.2.1 节所述评估这一预测的准确率。

关键词定位

输出 $\hat{}y_w$ 也可用于关键词定位。在这种情况下，视觉接地语音模型会获得一个关键词和一个话语集合，并必须按照 $\hat{}y_w$ 值从高到低对话语进行排序——$\hat{}y_w$ 值最高的话语排名第一，接着是次高的话语，依此类推。这意味着，话语是根据其包含该关键词的概率进行排序的。与关键词检测类似，此任务不需要视觉信息。视觉接地语音模型在关键词定位任务上的性能可按照第 3.2.2 节所述进行评估。

3.3.4. 视觉接地语音架构与实现细节

我们根据第 3.3.1 节提到的三个组件来定义具体的模型架构：编码器、池化层和分类器。

编码器（编码器）

我们使用两种类型的编码器：CNN 和 CNN-Pool。两者都基于卷积神经网络，并且已在以往的研究中得到探讨 $[2, 39]$。不过，在这里我们对它们稍作调整，将二维卷积神经网络改为一维卷积神经网络，并针对关键词定位进行更新（关键词定位仅在后续章节中讨论）。CNN 与 CNN-Pool 的主要区别在于，CNN-Pool 使用中间最大池化，而 CNN 不使用。我们考虑这两种方案的原因是，在早期研究中，CNN-Pool 在关键词检测方面取得了更好的分数，而 CNN 在下一章所述的某些设置下，关键词定位效果更好。

CNN-Pool 编码器由三个带有 ReLU 激活的单维卷积层组成。前两层采用对 3 个单元进行最大池化的中间操作。第一卷积层有 64 个宽度为 9 帧的滤波器。第二层有 256 个宽度为 11 个单元的滤波器，最后一层有 1024 个宽度为 11 的滤波器。CNN 编码器由 6 个带有 ReLU 激活的单维卷积层组成。第一层有 96 个核宽度为 9 帧的滤波器。接下来的四层各有 96 个宽度为 11 个单元的滤波器。最后一个卷积层使用 1000 个宽度为 11 个单元的滤波器。第一卷积层的填充大小为 4，其余各层为 5。

池化层（Pool）

我们的所有模型都以时长可变的语音数据作为输入，然后生成固定数量的输出预测。因此，我们需要将语音编码器的输出汇总为固定维度的向量。我们具体比较了三种池化层：沿时间维度进行最大池化、对数均值指数池化和注意力池化。

对数均值指数池化定义如下 [10]:

$$\mathrm{Pool}(\mathbf{H}) = \sigma\left(\frac{1}{r}\log \underset{\mathsf{T}}{\mathrm{mean}} \exp\left(r\mathbf{H}\right)\right) \tag{3.5}$$

它是一种软池化方法，可根据温度参数 r在最大池化和均值池化之间进行权衡。这种池化方法的主要动机是改善定位效果，但我们将在第 4.2.2 节中再讨论这一点。

对于注意力池化，我们使用以下公式：

$$\mathrm{Pool}(\mathbf{H})_w = \sum_{t=1}^{\mathsf{T}} \alpha_{w,t}\mathbf{h}_t. \tag{3.6}$$

这里$\alpha_{w,\ t}$ 是与特定时间点相关联的关键词专属权重；我们将在第4章详细介绍如何使用该权重预测关键词位置。现在，我们用它对每个时间步 t 的中间特征$\mathbf{h}_t$ 进行池化，以获得关键词检测分数。这种基于注意力的方法也深受 [7],的启发，不过我们同样会在第 4.2.3 节中详细说明。

请注意，我们所考虑的所有不同池化方法都已集成到现有的视觉接地语音模型中 [2, 39]。

分类器（分类器）

分类器由一个包含两层的多层感知机（MLP）组成，并使用 ReLU 激活。第一层的输出维度为 4096。第二层的输出维度有所不同（具体细节如下）：在某些情况下，分类器会为输出词汇表中的 W 个单词分别输出一个概率（例如，在针对多标签分类进行训练时）。在其他情况下，它会输出一个单一数值，表示是否存在某个特定关键词（例如，使用注意力池化进行训练时）。

架构

基于第3.3.4节所述的组件，我们组装了四种架构（PSC、CNN-Pool、CNN-Attend 和 CNN-PoolAttend），并在表 3.3 中进行了总结。我们将在第4章中展示，这些架构可以与四种不同的定位方法相结合。本章将评估每种架构在关键词检测和关键词检出任务上的表现。此处的 CNN 指的是其卷积层堆叠中不含

表 3.3：四种架构的组件（编码器、池化层、分类器）

架构	Enc	Pool	Clf
PSC	CNN	对数均值指数	—
CNN-Pool	CNN-Pool	max	多层感知机(1024, W)
CNN-Attend	CNN	注意力	多层感知机(1000, 1)
CNN-PoolAttend	CNN-Pool	注意力	多层感知机(1000, 1)

中间最大池化的编码器类型（见上文）。CNN-Pool 则包含中间最大池化。PSC 使用公式 3.5 所示的对数均值指数池化；这种架构名称的由来是，该池化方法最初由 Palaz、Synnaeve 和 Collobert 提出 [10]。

实现细节

所有架构均以梅尔频率倒谱系数作为输入。我们使用 SpecAugment [175], 对数据进行增强，这是一种增强策略，包括扭曲语音特征、掩蔽频率通道块以及掩蔽时间步块。该策略在端到端语音识别中实现了最先进的性能 [175]。所有模型均使用 PyTorch 实现，并采用 Adam 优化 [176]，学习率为 $1 \cdot 10^{-4}$。我们对每个模型训练 100 个训练轮次，并根据其在开发集上的性能选出最佳模型。所有模型均在配备 8 Gb 随机存取存储器的单块 GeForce RTX 2070 图形处理器上训练。

3.4. 结果：关键词检测与检出

我们在本学位论文中考虑关键词检测有两个原因。首先，为了完成我们的主要任务——关键词定位，我们通常需要先检测待定位的关键词是否出现在话语中，然后再对其进行定位。其次，模型在关键词检测上的性能也可以作为其在关键词定位任务上性能的上界。我们在此考虑的关键词检测模型是在先前工作中开发的 [2,10,39]，但我们对架构进行了多项改动，例如加入注意力，以实现基于注意力的关键词检测和关键词检出。

为了确定检测模型的基线性能，我们按照第3.3.2节所述，训练了两种弱监督模型——词袋和视觉接地语音。随后，我们评估它们在关键词检测和关键词定位任务上的性能。如前所述，我们在这里考虑词袋模型，是因为它可被视为一种完美的视觉接地语音模型，并且能让我们了解：如果拥有完美的视觉标注器，视觉接地语音模型在该任务上能够达到怎样的性能。

表 3.4： 四种不同架构的关键词检测分数（%），任务是检测给定关键词是否出现在某个话语中（无论其位置如何）。

	视觉			词袋		
架构	P	R	F_1	P	R	F_1
PSC	28.5	11.3	16.2	86.9	75.7	80.9
CNN-Pool	37.4	26.8	31.2	87.7	75.7	81.2
CNN-Attend	**38.9**	**28.2**	**32.7**	**89.6**	**79.6**	**84.3**
CNN-PoolAttend	35.1	22.6	27.5	81.5	65.0	72.3

表3.4展示了在词袋标签和视觉上下文监督下，不同架构的检测分数。在这两种监督形式中，词袋标签带来的检测性能始终优于视觉监督。这并不令人意外，因为不同说话人可能会用许多不同的方式描述同一张图像。此外，视觉标注器（在此提供训练信号）可能会为任何说话人都不会提及、但确实存在于图像中的概念赋予较高概率，也可能会标注语义相关的词语。与PSC结构相比，CNN-Pool似乎也能取得更高的分数。我们之所以考虑在编码器结构中使用中间最大池化的CNN-PoolAttend和CNN-Pool架构，是因为仅考察关键词检测而不考虑定位时，它们的分数优于另一种结构。但在这里可以看到，视觉CNN-Attend的检测性能略优于CNN-Pool F_1（32.7% 对比 31.2%）；我们将在第4章中展示，CNN-Attend还具有定位性能显著更好的优势。请注意，此前尚未有人针对这一任务考察过基于注意力的模型。

表3.5展示了与关键词检测使用相同模型的关键词检出性能。总体而言，我们发现，关键词检测任务中的类似结论同样适用于关键词检出：词袋模型的表现始终优于视觉接地语音模型；CNN-Pool 的 P@10（44.4%）、 $P@N$（28.2%）和等错误率（22.8%）均优于 PSC（分别为 19.7%、14.0% 和 35.5%）；CNN-Attend 在大多数度量指标上的分数略高于 CNN-Pool（分别为 43.7%、30.1% 和 22.7%）。

3.5. 总结

本章介绍了用于关键词检测和关键词检出的视觉接地模型。我们从 $[2, 39]$, 之前提出的模型入手，并采用此前尚未考虑过的注意力机制对其进行进一步修改和扩展。我们介绍了将在本学位论文后续部分扩展的关键词检测模型，以用于实现我们的主要目标——关键词定位。我们还介绍了之后将使用的那

表 3.5： 四种不同架构的关键词检测得分（%）。任务是根据话语包含系统词汇表中各个关键词的概率对话语进行排序（不考虑位置）。

	视觉			词袋		
架构	P@10	$P@N$	EER	P@10	$P@N$	EER
PSC	19.7	14.0	35.5	80.2	68.1	10.0
CNN-Pool	**44.4**	28.2	22.8	95.5	75.0	6.2
CNN-Attend	43.7	**30.1**	**22.7**	**95.7**	**80.2**	**5.9**
CNN-PoolAttend	31.6	24.4	27.0	92.8	72.6	10.2

个数据集。最后，我们展示了视觉接地语音模型的基线和上界结果（使用词袋（BoW）模型），并分析了视觉接地语音模型在关键词检测和关键词检出任务上的表现。具体而言，我们训练了四种不同的关键词检测架构，以建立关键词检测和关键词检出的基线。如前所述，词袋模型有助于在本研究中估算视觉接地语音模型性能的上界，并作为合理性检查。

我们发现，与视觉监督相比，词袋标签始终能带来更好的关键词检测和定位效果。这一结果并不意外，因为不同的说话人可能会用许多不同的方式描述同一幅图像。我们还发现，基于注意力机制的视觉接地语音模型在大多数情况下，其关键词检测和定位性能优于非注意力模型。在下一章，即第4章中，我们将扩展这些模型，使其能够用于关键词定位——这是首次考虑将这些模型用于此任务。

第4章

使用视觉接地语音模型进行关键词定位的方法

在第3章中，我们介绍了用于训练和评估视觉接地语音模型的关键词检测与定位方法框架。我们考察了现有的视觉接地语音模型 [2, 39]，这些模型通过将每张训练图像输入 [39] 的多标签视觉标注器，获得软文本标签，并利用这些标签进行关键词检测与定位的显式训练。此外，我们还介绍了一种基于注意力机制的新型视觉接地语音检测方法。本章将全面探讨我们的第一个研究问题（第1.2节）：视觉接地语音模型能否实现关键词定位？我们将第3章中的视觉接地语音模型扩展至关键词定位任务。提醒一下，视觉接地关键词定位是指使用在图像及其口语描述上训练的模型，找出给定书面关键词在话语中的出现位置。

本章提出四种不同的定位方法，以适配模型进行关键词定位。其中一种方法采用了广受欢迎的可解释性方法 *Grad-CAM* [55], ，该方法最初是在图像领域中开发的。这种方法可用于任何卷积神经网络架构。第二种和第三种方法利用了这样一个事实：对于某些架构，内部激活可以被解读为定位分数。具体而言，第二种方法（分数聚合）假设最后的池化层只是简单的变换（例如平均池化或最大池化），因此输出分数可视为局部分数的聚合结果。这些局部分数可用于选出查询关键词最可能出现的时间位置。第三种方法（注意力）假设网络包含一个沿时间轴池化特征的注意力层；我们将注意力权重用作定位分数。第四种方法（输入屏蔽）会在不同位置屏蔽输入信号，并测量训练好的模型对部分输入所预测的响应分数；输出变化较大，表明可能存在关键词。输入屏蔽与第一种方法类似，因为训练完成后，无论架构如何，都可以将其应用于任何分类模型。

所有为关键词定位改造的不同模型，都与上一章中的关键词检测模型采用完全相同的方式进行训练：训练过程中不进行对齐

数据（训练过程中，模型不会获知任何关键词的位置）。唯一的区别出现在预测时，此时我们会从模型中提取额外信息。

第一种和第二种定位方法及其结果发表于：K. Olaleye、B. van Niekerk 和 H. Kamper，《利用弱监督实现语音中的关键词定位》*NeurIPS* 语音与音频处理自监督学习研讨会，2020年。

第三种定位方法发表于：K. Olaleye 和 H. Kamper，《利用视觉定位实现语音中的基于注意力的关键词定位》，*Interspeech*，2021年。

第四种定位方法发表于：K. Olaleye、D. Oneat˘a 和 H. Kamper，《利用视觉接地语音模型实现未转写语音中的关键词定位》，《信号处理精选主题期刊》，2022年。

本章余下部分将介绍这些视觉接地语音模型在不同关键词定位任务中的评估（第4.1节）；并详细介绍所提出的定位方法，简要综述现有的相关研究（第4.2节）。最后，我们将展示所提出定位方法的性能，并对这些方法进行系统、公正的分析（第4.3节）。

4.1. 评估用于关键词定位的视觉接地语音模型

本章（以及整篇学位论文）的主要任务是关键词定位：给定一段音频话语和一个关键词，找出该关键词在输入中的出现位置。正如本章引言中所述，核心思路是使用在第3章中经过检测任务训练的同一音频网络来执行定位——也就是说，网络的训练过程不使用关键词位置标签。我们详细介绍了将关键词检测模型适配为定位模型的不同方法；每种方法都在第4.2节中进行说明。这里，我们仅介绍如何评估视觉接地语音模型在定位任务上的表现。

为计算位置，每种定位方法都会为输入音频序列中的每个关键词 w 和位置 t 分配一个分数$\alpha_{w,\,t}$。输入关键词 w 的预测位置 τ_w 是整个时间范围内分数最高的位置：

$$\tau_w = \operatorname{argmax} \{\alpha_{w,1}, \ldots, \alpha_{w,\mathsf{T}}\}. \tag{4.1}$$

这是 il　如图4.1所示（即图3.2中此前展示的右下部分 d

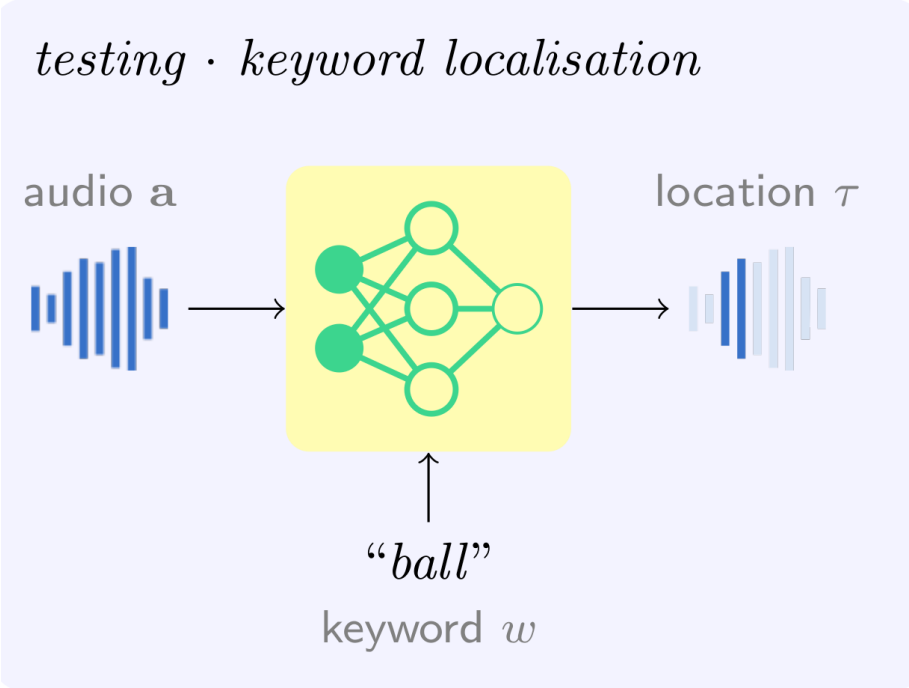


图4.1： 在关键词定位中，测试阶段的任务是定位给定的关键词。在此示例中，网络用于定位输入话语 **a** 中的关键词 球 。

）。我们用 T 表示输入序列的长度。

关键词定位可以根据最终使用场景的不同，以不同方式进行评估。在某些情况下，我们可能已知某个关键词出现在一段话语中，并需要预测该关键词出现的位置（下文称为预言机关键词定位）。在其他情况下，我们可能需要先检测某个关键词是否出现，然后再进行定位（实际关键词定位）。或者，我们可能需要对语音集合中的话语进行排序，仅对最有可能包含给定关键词的话语进行定位（关键词检出定位）。

在所有不同的定位场景中，如果预测的（点）位置落在根据强制对齐确定的关键词真实位置所对应的区间内，我们就认为该关键词已被定位。具体细节如下。

4.1.1. 预言机关键词定位

在此设置中，我们已知指定的关键词出现在某段话语中，需要找到该关键词的位置。因此，我们无需先进行检测，因为已知该关键词存在。这种情形称为预言机定位性能 in [10]，代表检测完美的情况，此时我们只考虑关键词定位性能。定位按照公式 4.1 进行，我们报告准确率：提出的单词位置落在真实词边界内的比例。

对于 cl 为说明这一点，我们在图 4.2 中给出一个示例。我们假设有四段话语 in

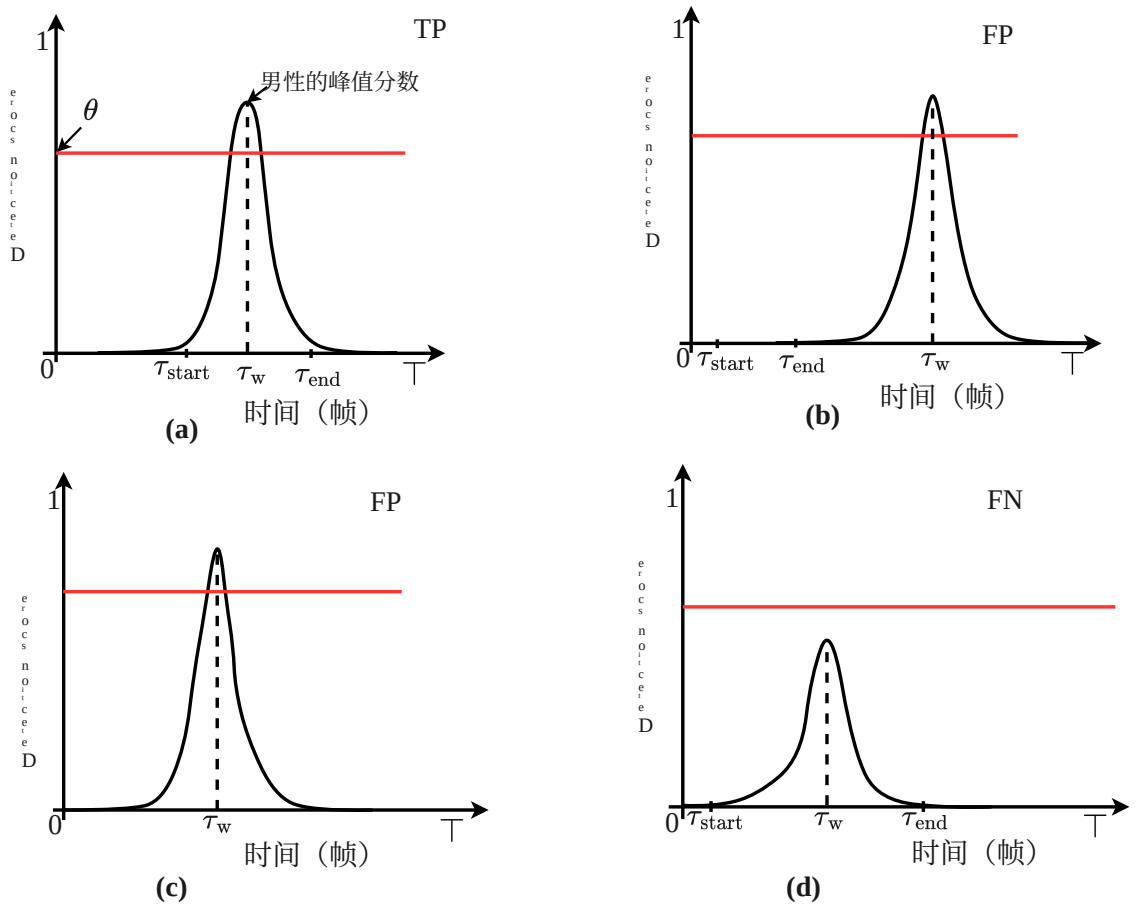


图 4.2： 示意图展示了如何评估我们的机制在四段由曲线 (a)、(b)、(c) 和 (d) 表示的话语中定位关键词“man”的性能。曲线 (a) 已完整标注，以说明曲线各部分所代表的内容。

测试语料库中只有一个需要定位的关键词 男性 。为计算预言机准确率，我们只考虑包含关键词 男性的话语。在所有情况下，根据数据的强制对齐结果，关键词 男性都从 (τ_{start} 开始，并在 τ_{end}) 结束。在这种情况下，预言机准确率为：

$$\begin{aligned}\text{Oracle accuracy} &= \frac{\#\text{ peaks on } man}{\#\ man \text{ in ground truth}} \\ &= \frac{\#[(a),(d)]}{3} = \frac{2}{3}.\end{aligned}$$

4.1.2. 实际关键词定位

在实际关键词定位设置中，我们不知道指定的关键词是否出现在某个话语中。因此，为了定位关键词，我们首先检查关键词 w 的检测分数 $\hat{}y_w$ （如第 3.2 节所述）是否高于指定的阈值 θ，然后选择注意力权重最高的位置$\alpha_{t,\ w}$ 作为

预测位置 τ_w，如公式 4.1 所示。如果位置 τ_w 落在与单词真实位置对应的区间内（该位置根据文本字幕与音频话语之间的强制对齐确定），则判定为正确。每当模型定位到一个检测分数小于或等于 θ 的单词时，都会受到惩罚；也就是说，如果某个单词未被检测到，即使$\alpha_{t,\ w}$ 在该单词范围内达到最大值，也会被计为定位失败。我们将 θ 的值设为 0。5，这一设置在开发数据上取得了最佳性能。我们报告精确率（P）、召回率（R）和 F_1。

为了更好地理解这一点，我们再次考察图 4.2 中的示例。在本例中，为了计算实际定位性能，我们考虑全部四段话语——图 4.2 中的曲线 (a)、(b)、(c) 和 (d)。精确率、召回率和 F_1 计算如下：

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} = \frac{1}{1+2} = \frac{1}{3}.$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{1}{1+1} = \frac{1}{2}.$$

$$F_1 = 2 \times \frac{P \times R}{P + R} = \frac{2}{5}.$$

4.1.3. 关键词检出定位

回顾第 3.3 节，在关键词检出任务中，我们会得到一个关键词和一个话语集合，并需要按照各段话语包含该关键词的概率对它们进行排序。关键词检出通过 P@10（得分最高的十个提议的平均精度）和 P@N（排名前 N 个提议的平均精度）进行评估，其中 N 表示该关键词的真实出现次数（参见第 3.2.2 节）。与实际关键词定位评估流程（见第 4.1.2 节）类似，该流程用于衡量模型定位已检出关键词的能力；我们将相关的关键词检出任务调整为定位场景。

对于关键词检出定位，我们还会考虑给定关键词的定位：只有当样本排名靠前（即检出该关键词），且预测位置落在真实词语范围内时，才视为正确。本章将关键词检出定位性能报告为 P@10，即位置预测正确的话语在排名前 10 的话语中所占的比例。如果某个话语被错误地排进前 10 名（也就是说，该话语不包含指定的关键词），则会对定位性能进行惩罚（即将该话语视为定位失败）。

表4.1： 四种架构的组件（编码器、池化层、分类器）及定位方法（Grad-CAM GC、输入屏蔽 M、分数聚合 SA 和注意力 Att）。

架构	Enc	Pool	Clf	GC	M	SA	Att
PSC	CNN	对数均值指数	—			•	
CNN-Pool	CNN-Pool	max	多层感知机(1024, W)	•			
CNN-Attend	CNN	注意力	多层感知机(1000, 1)		•		•
CNN-PoolAttend	CNN-Pool	注意力	MLP(1000, 1)		•		•

4.1.4. 关于基于点的评估

在本学位论文中，我们考虑的基于点的定位的一种替代方案，是预测并评估单词起始和结束时间构成的区间。然而，出于以下几个原因，我们选择采用基于点的评估方法：(i) 即便是使用转录文本训练的语音模型，也很难准确预测单词的时间区间 [177]。(ii) 我们的评估方法无需额外的后处理步骤，因此能够更直接、公平地比较不同定位方法。否则，对于提出的三种方法（Grad-CAM、分数聚合、注意力；将在第 4.2 节讨论），我们还需要设计一种方法，将最有希望的片段合并成一个更大的片段。(iii) 计算机视觉文献中也采用了类似的评估方法，尤其是在提出显著性模型的论文中 [178–180]。

4.2. 定位方法

在第3章中，我们详细介绍了如何训练和评估用于关键词检测和关键词检出的视觉接地语音模型。我们特别考察了四种不同的基于卷积的架构（汇总于表4.1）。这些架构的核心部分采用自现有的视觉接地语音模型 [2, 39]，本节将对其进行修改以用于关键词定位。具体而言，我们假设已获得一个为关键词检测任务（第3.3节）训练的音频网络，并介绍四类可为网络输出添加定位分数的方法。

4.2.1. Grad-CAM

基于梯度的类激活图 (Grad-CAM) [55] 是一种基于显著性的方法，可应用于任何卷积神经网络架构，以确定网络输入的哪些部分对特定输出决策的贡献最大。直观而言，它通过确定卷积层中的每个滤波器对特定输出预测（在我们的例子中，是指定的关键词）的重要程度来发挥作用。随后，通过根据滤波器对所考察关键词的重要性，对滤波器激活值进行加权，为输入语句计算定位分数

。我们将最后一个卷积层之后计算出的分数用作定位分数。

Grad-CAM 首先确定最后一个卷积层输出中的每个滤波器对单词 w 有多重要。更正式地说，将最后一个一维卷积层的输出记为 $\mathbf{h}_t$，其中每个 $\mathbf{h}_t \in \mathbb{R}^{\mathsf{K}}$ 都是一个 K 维向量，K 为滤波器的数量。我们首先确定滤波器 k^{th} 对单词 w 的“重要性”：

$$\gamma_{k,w} = \frac{1}{\mathsf{T}} \sum_{t=1}^{\mathsf{T}} \frac{\partial \hat{y}_w}{\partial h_{t,k}}. \tag{4.2}$$

该值表示根据训练后的模型权重，k^{th} 卷积滤波器对单词 w 的关注程度。随后，将时间步 t 上的输出激活 $\mathbf{h}_t$ 按其对单词 w 的重要性 γ_w 加权，从而得到定位分数。我们只关心能够提高 w 分数的变化，因此应用 ReLU，结果为：

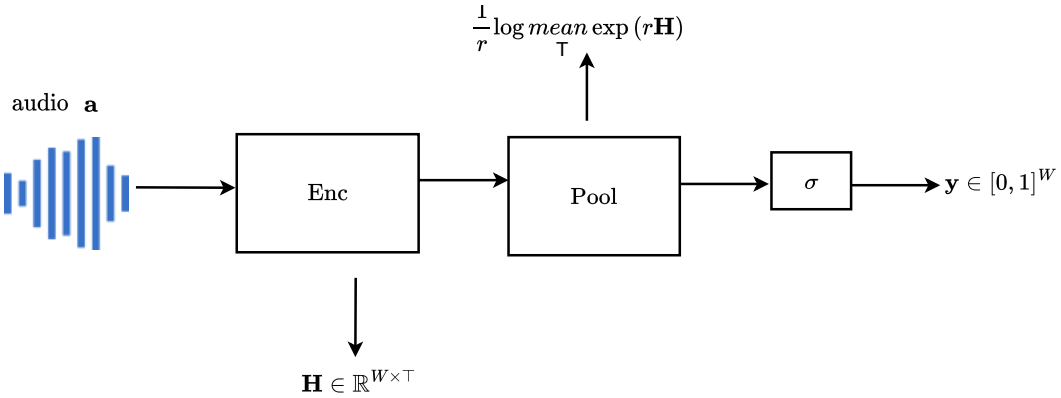
$$\alpha_{w,t} = \text{ReLU}\left[\sum_{k=1}^{\mathsf{K}} \gamma_{k,w} h_{t,k}\right]. \tag{4.3}$$

与先前研究的关系

为便于理解，我们简要介绍 Grad-CAM 的发展背景。类别激活图（CAM）最初在 [157] 中被提出，用于目标定位。其网络由卷积层、对特征图进行的全局平均池化，以及直接连接到（softmax）分类层的结构组成。随后，通过将输出层的权重投影回卷积特征图，识别对区分物体有重要作用的图像区域。最后一个卷积层的特征图加权求和后，即得到类别激活图。Grad-CAM [55] 旨在将 CAM 推广到全卷积模型之外。正如 Grad-CAM 论文所述，扩展的动机是避免像 CAM 那样为了更容易解释模型而牺牲模型复杂度和性能（这与移除全连接层有关）。他们使用 ReLU 作为阈值，只保留对目标类别有正向影响的特征图。

4.2.2. 分数聚合

分数聚合方法的核心在于使用我们在第3.3.4节中介绍的对数均值指数池化函数。这是一种软池化方法，可将语音编码器的输出汇总为固定维度的向量。本节将详细介绍如何使用该函数来实现关键词定位。分数聚合方法最初是在 [10] 中设计的，旨在使仅针对词分类进行训练的模型能够定位预测词。它通过明确地使架构



	1	2	3	⋯	T
1	0.2	0.01	0.025	⋯	0.4
2	0.1	0.5	0.03	⋯	0.2
3	0.05	0.009	0.02	⋯	0.09
⋯	⋯	⋯	⋯	⋯	⋯
W	0.05	0.3	0.07	⋯	0.6

图 4.3： 分数聚合定位方法示意图。橙色区域表示如何解包编码器（Enc）的输出，从而为每个帧 t 和每个单词 w给出一个分数 $h_{w,\ t}$。在此示例中，与索引 2 和 W 对应的关键词所提议的位置分别是帧 2 和 T，它们在图示的橙色区域中取得了最高分数。

具备位置感知能力来实现这一点。最后一个卷积层必须具有与模型最终输出相同数量的滤波器。通过以特定方式聚合最后一个卷积层的输出（下文将对此进行说明，并在图4.3中展示），我们可以从输出预测反向追溯，确定哪个时间位置上的哪个滤波器导致了该输出预测。

更正式地说，分数聚合方法将音频嵌入 $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_\mathsf{T}]$ 解释为定位分数 α，即 $\alpha_{w,\ t} = h_{w,\ t}$。该方法依赖于两个假设：

1. 嵌入维度 K 等于词汇表 W 中关键词的数量。

2. 池化层是保留激活值语义并将其作为分数的最后一层。因此，该方法不能应用于任意网络。具体来说，它需要与第3.3.4节中描述的 PSC架构搭配使用，该架构的池化层之后没有分类器层。PSC 这个名称源自原论文作者的姓名 [10]。

为了将帧级分数聚合为话语级检测分数（跨越 T 个步骤），我们使用对数均值指数函数，随后使用 Sigmoid函数 σ （这确保数值位于 $[0,1]$ 中，并可解释为概率）：

$$\mathrm{Pool}(\mathbf{H}) = \sigma\left(\frac{1}{r}\log\operatorname*{mean}_{\mathsf{T}}\exp\left(r\mathbf{H}\right)\right). \tag{4.4}$$

这里， r 是一个控制池化类型的正数：当 $r \to 0$ 时，我们得到平均池化；当 $r \to \infty$ 时，我们得到最大池化。因此，通过控制超参数，我们可以得到介于这两个极端之间的结果。根据 [10], ，这种中间的聚合操作会在训练过程中促使分数相近的帧具有彼此接近的权重，从而提升定位性能。

与先前研究的关系

Palaz 等人 [10] 提出了这种分数聚合方法，仅使用词袋监督即可联合定位和分类单词（见第 3.3.2 节）。在本学位论文中，我们改为使用视觉监督训练这种分数聚合方法。对于图像卷积神经网络，一种相关技术是 CAM [157] （见第 4.2.1 节），它将某一类别的分类分数分解为卷积神经网络最后一层特征图中各空间分数之和。该方法基于这样的假设：最后一层特征图会经过全局平均池化层和线性分类器。此方法与 Grad-CAM 相关，因为两者都通过沿通道轴池化卷积特征图来计算空间分数。主要区别在于，CAM 使用线性分类器的权重进行聚合，而 Grad-CAM 使用输出相对于特征图的梯度。

4.2.3. 注意力

注意力已成为一种标准机制，可让神经网络在生成某些（中间）输出时，自动为其输入的不同部分分配权重 [181, 182]。我们之前介绍过（参见第3.3.4节）如何使用注意力方法进行关键词检测。现在，我们将详细说明如何利用这一机制在语音中进行关键词定位。直观而言，给定的文本关键词会映射为一个学习得到的查询嵌入，随后用于衡量中间卷积滤波器的相关性。权重最高（即最受关注）的时间步，很可能最能指示特定关键词的位置。该方法如图4.4所示。

具体而言，我们首先使用基于索引的查找表，将给定的关键词 w 嵌入为 K 维嵌入 $\mathbf{q}_w$ （查询）。随后，注意力模块通过点积衡量卷积特征$\mathbf{h}_t$ 中每个时间步与查询嵌入之间的相似度，并据此为每个时间步 t 计算分数 $e_{w,\,t}$ ： $e_{w,\,t}$ =$\mathbf{q}_w\mathbf{h}_t$,

T

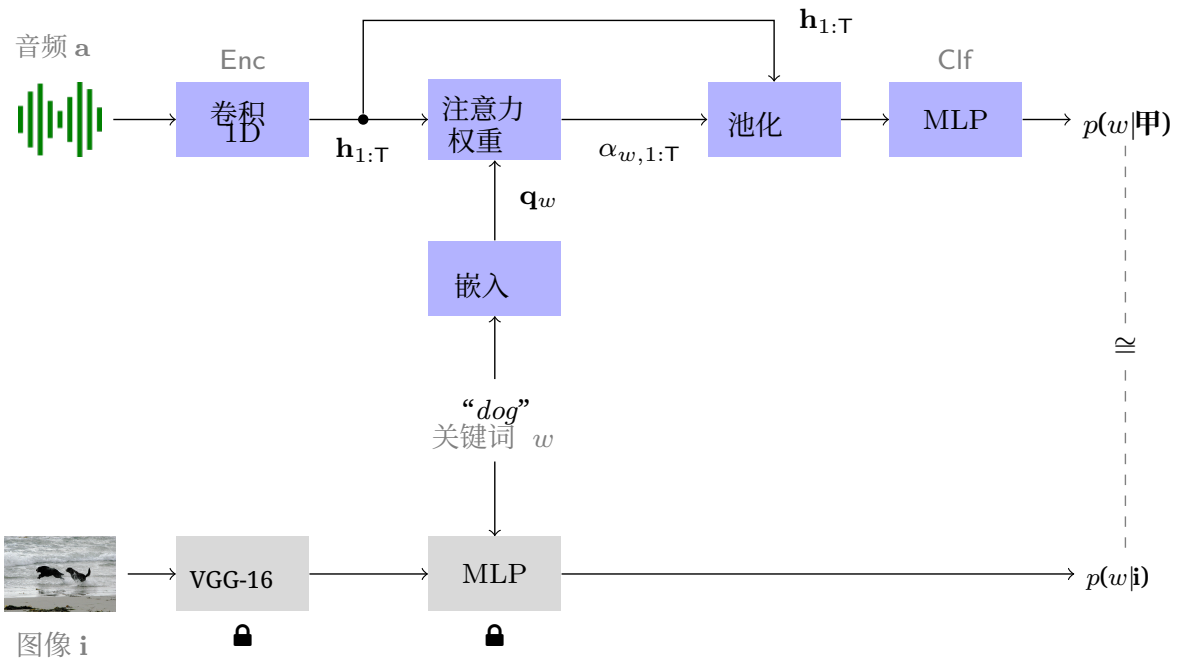


图4.4： 用于语音中关键词定位的基于注意力的架构。语音网络（蓝色）首先将音频输入$\mathbf{a}$ 传入卷积编码器。随后，利用基于查询关键词 w嵌入$\mathbf{q}_w$ 的注意力机制，对生成的特征$\mathbf{h}_{1:T}$ 进行池化。上一步会得到一个向量，最后将其输入多层感知机（MLP），输出关键词出现在音频中的概率 $p(w|\mathbf{a})$。视觉网络（灰色）使用 VGG-16 网络将相应的图像$\mathbf{i}$ 嵌入为一个向量，并使用关键词专属的 MLP 计算关键词出现在图像中的概率 $p(w|\mathbf{i})$。请注意，视觉网络仅在训练阶段使用（用于监督语音分支），且其权重固定；在测试阶段，我们只使用顶部的语音分支。

接着进行 Softmax 运算，将相似度分数 $e_{w,\ t}$ 转换为注意力权重：

$$\alpha_{w,t} = \frac{\exp\{e_{w,t}\}}{\sum_{t'=1}^{T} \exp\{e_{w,t'}\}}. \tag{4.5}$$

注意力权重用于关键词定位（作为定位分数）。如第3.3.4节所述，注意力权重还可用于关键词检测（作为中间特征）：将每个时间步 t 的特征$\mathbf{h}_t$ 池化为关键词专属上下文向量：

$$\text{Pool}(\mathbf{H})_w = \sum_{t=1}^{T} \alpha_{w,t}\mathbf{h}_t \tag{4.6}$$

并通过分类器 Clf 计算查询 w 的分数，该分类器通常由一系列全连接层（多层感知机）构成，并将上下文向量传入其中。

与先前研究的关系

我们的方法深受 Tamer等人 [7], 的启发，他们使用基于注意力的图卷积网络处理骨架关节，从而在手语中进行关键词搜索。我们调整了他们的架构，使其适用于语音中的关键词定位任务。注意力机制也被用于多实例学习的一般情形 [183]。该任务假设要对实例包进行分类：如果一个包中至少包含一个正实例，则该包为正。难点在于，标签仅在包这一层级给出，而实例标签未知。其思路是通过注意力对实例进行池化，并将相应的权重作为实例是否为正的指示。在一个实际应用案例中，Ilse 等人 [183] 运用该技术，在弱标注图像中突出显示癌变区域。

4.2.4. 输入屏蔽

本节介绍两种此前从未用于关键词定位的新技术。这些方法不再考察网络内部的中间激活，而是在特定位置修改网络输入，然后观察对输出检测概率的影响。例如，在下文将正式介绍的掩码输入 方法中，我们只将输入语句的一部分输入模型；如果这提高了某个特定单词的输出概率，就可能表明该单词出现在该片段中。在掩码输出 方法中，我们则将整段话语输入网络，但屏蔽掉部分输入；如果这降低了某个特定单词的输出概率，就可能表明该单词出现在被屏蔽的片段中。这两种方法的优点是都可应用于任何类型的检测模型。不过，具体如何对话语中的片段进行掩码输入或掩码输出，会影响计算开销；由于处理单段话语时可能需要多次运行模型，因此计算成本可能较高。

更形式化地说，如图 4.5 所示，掩码输入 方法将关键词定位任务视为关键词检测任务。具体而言，我们将一段测试话语切分成多个片段，每个片段的时长最短为 200 毫秒、最长为 600 毫秒，相邻片段之间重叠 30 毫秒。我们在开发数据上对最短和最长时长这两个超参数进行了粗略调优。

每个片段都会通过补零恢复到话语的实际长度。然后，我们将每个补零后的片段输入训练好的模型，计算某个关键词的检测分数。此过程会得到以下定位分数：

$$\alpha_{w,t} = \sigma\left(\mathrm{DNN}_a(\mathbf{a} \odot \mathbf{m}_t)_w\right), \tag{4.7}$$

其中 $\mathbf{m}_t$ 是一个二值掩码，其中与第 t 个片段对应的条目为 1，而 d

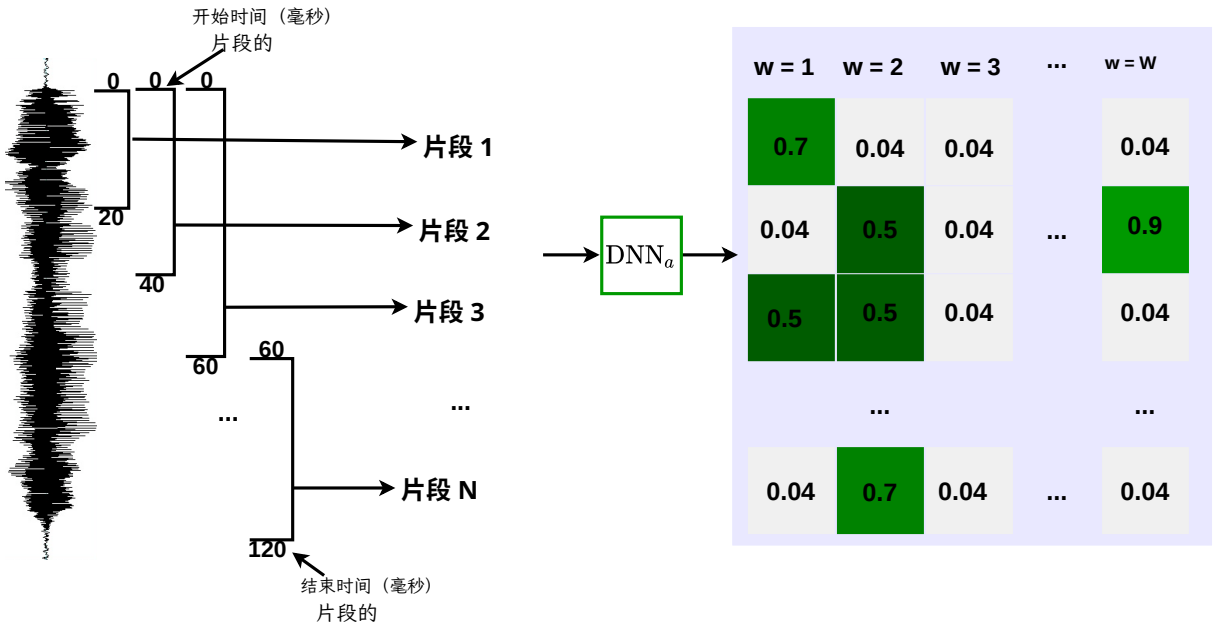


图 4.5： 我们用于语音中的关键词定位的掩蔽式方法。输入语句被切分为时长各异的重叠片段。随后，将每个片段输入网络（DNN$_a$），以确定特定单词的输出检测概率。与该单词输出检测概率最大值对应的片段，可能就是该单词的位置。例如，包含索引为 $w = 1$ 的单词的最可能片段是片段 1，其概率为 $\alpha_{1,\ 1} = 0.7$。

其余条目为 0，且 $\odot$ 表示逐元素乘法。将关键词检测分数最高的片段视为该关键词的位置。

掩码输出 方法与之恰好相反：我们将话语中的每个片段遮蔽（以零替换），并计算检测分数。如果掩码输出片段包含关键词，我们预期检测分数接近 0。为了根据最大值（公式 4.1）选择位置 τ，我们使用出现概率的补数作为定位分数：

$$\alpha_{w,t} = 1 - \sigma\left(\mathrm{DNN}_a(\mathbf{a} \odot (1 - \mathbf{m}_t))_w\right). \tag{4.8}$$

与先前研究的关系

掩码输入方法的运作方式与图像目标检测中常用的滑动窗口方法相似 [184]。一个重要区别是，在我们的情形中，反复应用于每个窗口的网络是在监督极弱的条件下训练的，使用了噪声监督标签，且没有任何位置信息。掩蔽输出方法与 Zeiler 和 Fergus [159] 的研究相关；他们探究了基于卷积神经网络的视觉模型是否真的在识别图像中物体的位置，还是只是在寻找周围上下文。我们改编了他们系统地

扰动输入图像不同部分的方法，但在这里将这种方法应用于语音。在语音领域，Harwath 和 Glass [15] 提出了一种方法，用于将话语中类似单词的声学单元与语义相关的图像区域联系起来。他们从话语中提取片段，然后利用多模态网络将这些片段与图像中相应的子区域关联起来。这与我们的设定不同：我们的方法不是将语音片段与相关图像区域匹配，而是将语音片段与文本标签匹配。Kamper等人 [40] 使用与我们的模型更为相似的模型，研究了检索与给定口语查询相关的话语这一任务。他们将查询嵌入与每个话语子片段的嵌入进行比较，计算余弦距离，并将最小余弦距离作为查询与话语相关性的分数。这里，我们采用了他们的方法，将每个话语划分为重叠片段，片段时长从某个最小时长到某个最大时长不等。不过，这里的查询是书面关键词（而非口语查询），更重要的是， [40] 未评估定位性能。

4.2.5. 架构与实现细节

如前所述，我们采用了第3章介绍的视觉接地语音架构作为基础。表4.1总结了这些架构及其对应的定位方法。某些定位方法与特定架构绑定，例如，分数聚合直接整合到PSC [10]架构和损失函数中，该架构和损失函数使用对数均值指数作为全局池化函数，并使用卷积神经网络作为音频编码器。注意力定位方法使用一个分类器，输出一个检测分数，用于判断给定关键词是否出现。其他定位方法则不受架构限制，例如，原则上，基于掩码的定位可用于任何架构。所有模型均按照第3.3.4节所述实现。

4.3 结果：关键词定位

4.3.1. 定量分析

主要结果见表4.2、4.3和4.4。表4.2展示了视觉接地语音模型和词袋监督模型在预言机定位任务上的性能。表4.3展示了实际关键词定位任务上的性能，表 4.4展示了关键词检出定位任务上的性能。

我们的主要研究目标是探究视觉接地语音模型在多大程度上能够实现关键词定位。因此，重点关注表4.2、4.3和4.4中“视觉”列的结果，可以看到，采用视觉监督的模型能够取得尚可的性能。在预言机设定（表4.2）中，系统已知关键词存在，

表 4.2： 预知关键词定位任务在测试数据上的结果（假设检测完美）。我们以准确率（百分比）报告结果。我们同时考察视觉监督和理想化词袋监督。

架构	定位	视觉	词袋
PSC	分数聚合	13.7	61.5
CNN-Pool	Grad-CAM	12.7	13.3
CNN-Attend	注意力	46.0	73.7
CNN-Attend	掩码输入	**57.3**	**87.5**
CNN-Attend	掩码输出	25.1	28.2
CNN-PoolAttend	注意力	16.5	33.6
CNN-PoolAttend	掩码输入	41.6	57.3
CNN-PoolAttend	掩码输出	22.3	28.7

表4.3： 实际关键词定位任务在测试数据上的结果。我们以 F_1 的形式报告结果（百分比）。我们同时考察视觉监督和理想化词袋监督。我们使用阈值 $\theta = 0.5$ 来判断某个关键词是否出现在话语中。

架构	定位	视觉	词袋
PSC	分数聚合	7.4	68.1
CNN-Pool	Grad-CAM	8.1	17.4
CNN-Attend	注意力	**25.2**	72.1
CNN-Attend	掩码输入	24.9	**79.8**
CNN-Attend	掩码输出	12.9	57.8
CNN-PoolAttend	注意力	8.2	39.0
CNN-PoolAttend	掩码输入	18.0	55.3
CNN-PoolAttend	掩码输出	10.4	48.1

然后再预测它出现的位置，最佳方法的准确率超过57%。检测后再进行定位时，最佳 F_1 分数为25.2%（表 4.3），而最佳关键词检出定位 P@10 为32.1%（表 4.4）。虽然这些分数并不高，但这些关键词定位结果是由未接受任何显式文本监督的模型取得的。

接下来，我们尝试量化首次检测或排序阶段造成的性能下降，考察架构选择的影响，并细分各个关键词的性能。

表 4.4： 关键词检测定位任务在测试数据上的结果。我们以排名前 10 位的精确率（P@10）报告结果（百分比）。我们考察了视觉监督和理想化词袋监督。

架构	定位	视觉	词袋
PSC	分数聚合	8.2	62.2
CNN-Pool	Grad-CAM	4.2	15.2
CNN-Attend	注意力	**32.1**	79.7
CNN-Attend	掩码输入	21.9	**86.6**
CNN-Attend	掩码输出	5.1	7.9
CNN-PoolAttend	注意力	8.1	42.7
CNN-PoolAttend	掩码输入	9.7	51.8
CNN-PoolAttend	掩码输出	5.5	11.9

作为视觉接地语音模型进行训练时，哪些架构和定位方法的表现最佳?

总体而言，CNN-Attend和CNN-PoolAttend架构优于PSC和CNN-Pool架构，而掩码输入和注意力定位方法优于其他定位方法。这是首次将基于掩码的定位纳入VGS建模。通过预言机定位可以看到，这种方法的准确率比最接近的竞争方法高出10多个百分点。与Grad-CAM和分数聚合定位方法不同，基于掩码的定位方法不受特定模型结构限制，能够应用于任何模型架构。当使用词袋监督时，采用掩码输入预测的CNN-Attend模型始终表现最佳；而使用视觉监督时，注意力定位在三项任务中的两项上表现最佳。我们推测，无论采用何种形式的监督，Grad-CAM在所有定位任务中的表现都较差（表4.2、4.3、4.4，第2行），这是因为它试图识别会导致特定输出单元发生较大变化的输入部分。Grad-CAM最初是为单标签多类别分类开发的，在这种分类中，某个输出的概率越高，其他输出的概率就越低。但对于本文采用的多标签分类，情况并非如此，多个单词的梯度都可能影响输出，而这些影响会体现在梯度中。训练期间有词袋监督时，分数聚合方法在所有定位任务中的表现相对较好；但我们观察到，改用视觉监督后，性能会下降（表4.2、4.3、4.4，第1行）。

表 4.5： 视觉接地语音模型实际关键词定位的精确率、召回率和 F_1 分数。括号内为关键词检测分数；由于实际定位需要先进行检测，再定位单词，因此检测分数是定位性能的上界。

	架构	定位	P	R	F_1
3	CNN-Attend	注意力	**28.8** (38.9)	**22.4** (28.2)	**25.2** (32.7)
4	CNN-Attend	掩码输入	28.3 (38.9)	22.2 (28.2)	24.9 (32.7)
7	CNN-Pool-Attend	掩码输入	21.9 (35.1)	15.3 (22.6)	18.0 (27.5)

我们付出什么代 **使用视觉监督而非词袋监督，需要付出什么代价?**

这是一个重要的问题，因为词袋监督可以被视为一种能够使用完美视觉标注器的情形：词袋标签仍然是弱标签（不包含任何位置信息），但词袋中的所有单词都与图像相关（依据人类提供的图像描述）。通过比较表4.2、4.3和4.4中的“视觉”和“词袋”两列，我们可以看到，从理想化监督转为视觉监督时，所有情况下的性能都大幅下降。例如，在 F_1 中，对于实际定位，最佳方法的性能从约 80% 降至约 25%。因此，改进视觉标注可能会显著提升视觉接地语音模型的关键词定位效果。不过需要指出的是，尽管词袋为视觉接地语音模型的性能提供了一个上界，但这个上界可能偏高：完美的视觉标注器或许能检测出人类不会描述的物体。在其他情况下，它可能无法为可能出现在话语中的某个关键词提供视觉支持（例如，相机 在“一个女孩正摆姿势站在相机前”中）。

作为一个子研究问题，我们指出，提升词袋监督模型的关键词定位能力本身就是一项有价值的工作。在许多情况下，词袋标签比转录文本更容易获得，例如参见 [185]。词袋监督下的定位也是原始 PSC 研究的主要目标 [10]。比较词袋模型后可以发现，在所有三项任务中，我们采用掩码输入定位的 CNN-Attend模型都显著优于 PSC 方法；在这三种情况下，准确率和 F_1分数的绝对提升均超过 10%。

首轮检测对定位性能的限制

进行实际关键词定位（表4.3）时，我们会受到第一轮检测的限制，因为这一轮只考虑高于某个阈值（$\theta = 0.5$）的关键词。由于无法定位未被正确检测出的单词，因此检测分数构成了可实现定位性能的上界。

表 4.6：视觉接地语音模型的关键词检测定位得分，以 P@10 和 $P@N$ 表示。括号中给出了相应的关键词检测得分（不考虑定位）；这些分数是定位性能的上界。

	架构	定位	P@10	$P@N$
3	CNN-Attend	注意力	**32.1 (43.7)**	**24.9 (30.1)**
4	CNN-Attend	掩码输入	21.9 (43.7)	19.5 (30.1)
7	CNN-Pool-Attend	掩码输入	9.7 (31.6)	10.4 (24.4)

为了量化部分视觉接地语音模型在定位性能方面的局限，表 4.5 列出了所达到的召回率、精确率和 F_1 分数，以及关键词检测分数（括号内所示）。（理解括号内分数的一种方式，是类比表 4.2 中的预言机结果：在预言机定位中，我们假设检测完美，只考虑定位；现在则假设定位完美，考察检测性能。）我们发现，在所有情况下，第一轮检测都会损害定位效果。就召回率而言，检测错误导致的结果与上界相差约 6%；就精确率而言，结果与上界相差约 10%。未来工作可以考虑采用一种方法，使用针对检测和定位这两项不同任务分别定制的模型来完成它们。

首轮关键词检出所能实现的定位上限

类似地，关键词检出定位结果（表 4.4）受限于标准关键词检出中执行的首次排序阶段。也就是说，如果首次排序阶段错误地将某个话语排在首位，而该话语实际上并不包含给定的关键词，就不可能正确定位该关键词。这里，表 4.6 展示了部分视觉接地语音模型所取得的 P@10 定位分数，以及上限关键词检出 P@10（括号内）。这里还列出了 $P@N$ 分数：实际取得的分数（及其上限）。我们再次看到，定位性能受到限制：即使是最佳方法（带注意力机制的 CNN-Attend），P@10 和$P@N$ 定位错误中，也只有约 10% 和 5% 并非由排序错误导致。因此，改进首次排序阶段的关键词检出，或许也能进一步提升后续定位效果。

中间最大池化对定位性能的影响

分析不同架构和定位方法的另一种方式，是考察它们对编码器结构的选择如何影响定位能力。每种架构都使用 CNN 或 CNN-Pool。在 CNN-Pool 编码器中，每个卷积层操作（最后一层除外）之后都会进行一次最大池化操作。CNN 编码器类型则不会执行这类中间池化操作。直观地说，最大池化操作会沿时间轴下采样，计算并返回输入中最显著的特征。尽管此类操作已被证明有助于

提升单词检测效果，但会降低视觉接地语音和词袋监督设置下的定位性能。我们通过比较两种编码器类型下表现最好的两种定位方法的定位分数来证明这一点。如表4.2、4.3和4.4所示，在三项任务中，使用注意力和掩码输入定位方法的 CNN 编码器架构，其表现均优于采用 CNN-Pool 编码器的架构。这一现象并不意外，因为 CNN-Pool 所采用的最大池化操作使得将每个输出预测追溯到与其最相关的滤波器变得更为困难。

各关键词的性能

我们考察词汇表中每个关键词的定位性能。我们报告实际关键词定位设置下的结果：对于每个关键词 w，我们首先选出检测分数高于阈值 θ 的话语，然后检查定位是否正确，方法是验证最大定位分数对应的位置 τ_w 是否落在与查询关键词 w相对应的片段内。结果见图 4.6。

词汇表中的 67 个关键词里，我们报告了其中 45 个的性能，因为其余 22 个关键词没有任何话语的检测分数超过设定阈值θ 0.5；如果没有检测到任何话语，就无法对其进行定位——此时精确率未定义，而召回率为 0。我们观察到，实际定位性能在词汇表中的不同关键词之间差异很大（从狗 的 1.4% 到 海滩的 65.7%），这表明这些方法高度依赖关键词。我们认为，关键词性能受到每个单词的视觉接地效果影响。

为了检验这一假设，我们考察定位性能与视觉网络性能之间的相关程度，见图 4.7。我们发现，这两个系统之间确实存在依赖关系：像相机、空气和穿着 这样的关键词显然难以进行视觉定位，这会影响语音定位能力；而海滩、足球和雪 这样的关键词很容易被视觉网络和基于语音的定位网络检测到。不过，也存在一些离群值：关键词足球、汽车和狗 虽然视觉定位效果良好，但定位效果却不佳。我们在第4.3.2节中研究了这些情况，并发现其原因是这些关键词与其他单词共现：足球–球员、汽车–比赛、狗–三 （见表 4.7）。在同一话语中经常共同出现的单词，对于任何无法获取额外信息的定位方法而言，都极难区分。

4.3.2. 定性结果

在本小节中，我们报告通过人工检查样本和网络的预测结果获得的具体示例。

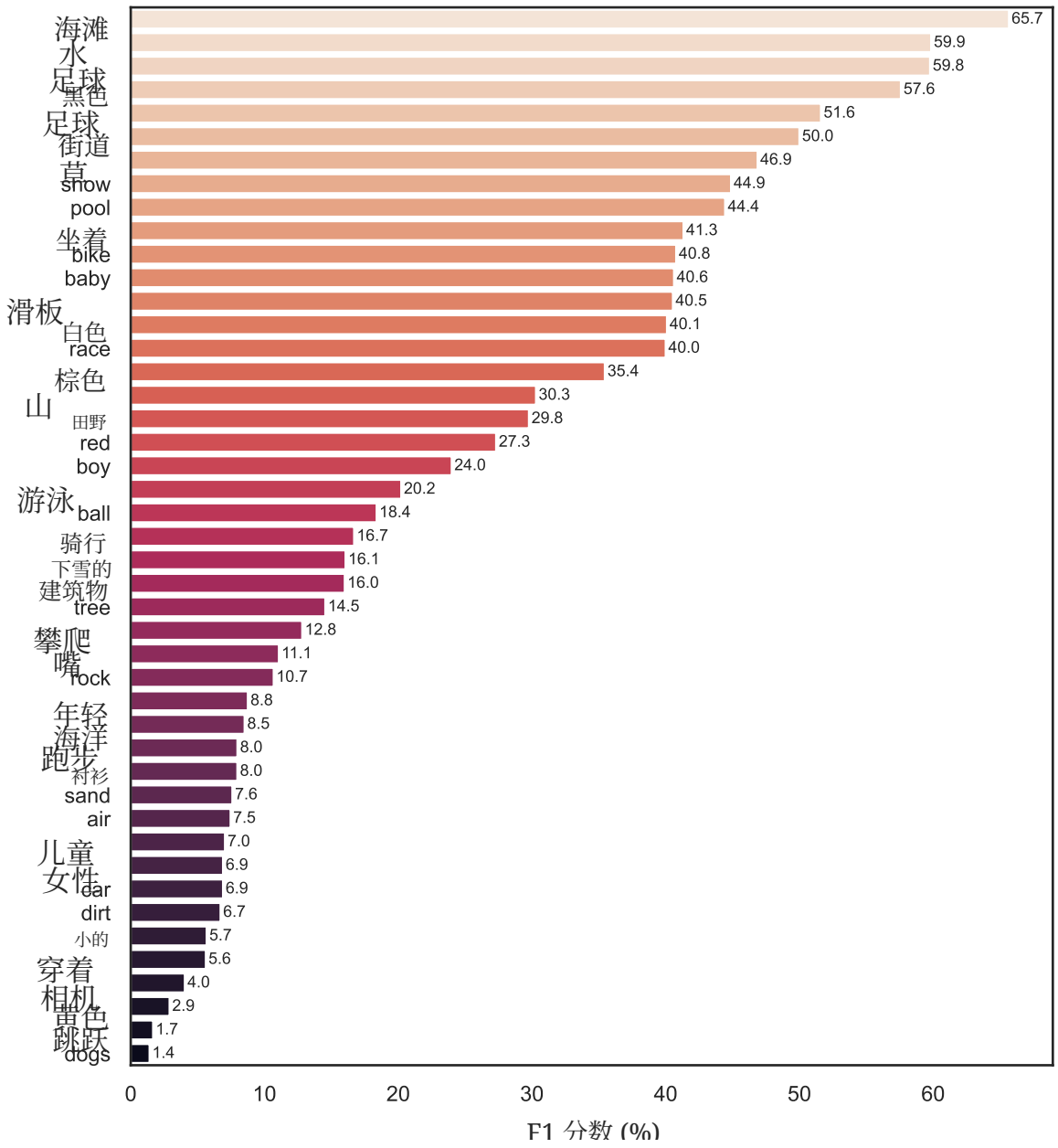


图 4.6： 按关键词报告的实际定位性能，以 F_1 分数衡量。我们仅纳入至少检测到一个话语的关键词（67 个中有 45 个），也就是说，检测分数大于阈值 θ 0.5）。结果针对 CNN-Attend 架构和掩码输入定位方法报告。

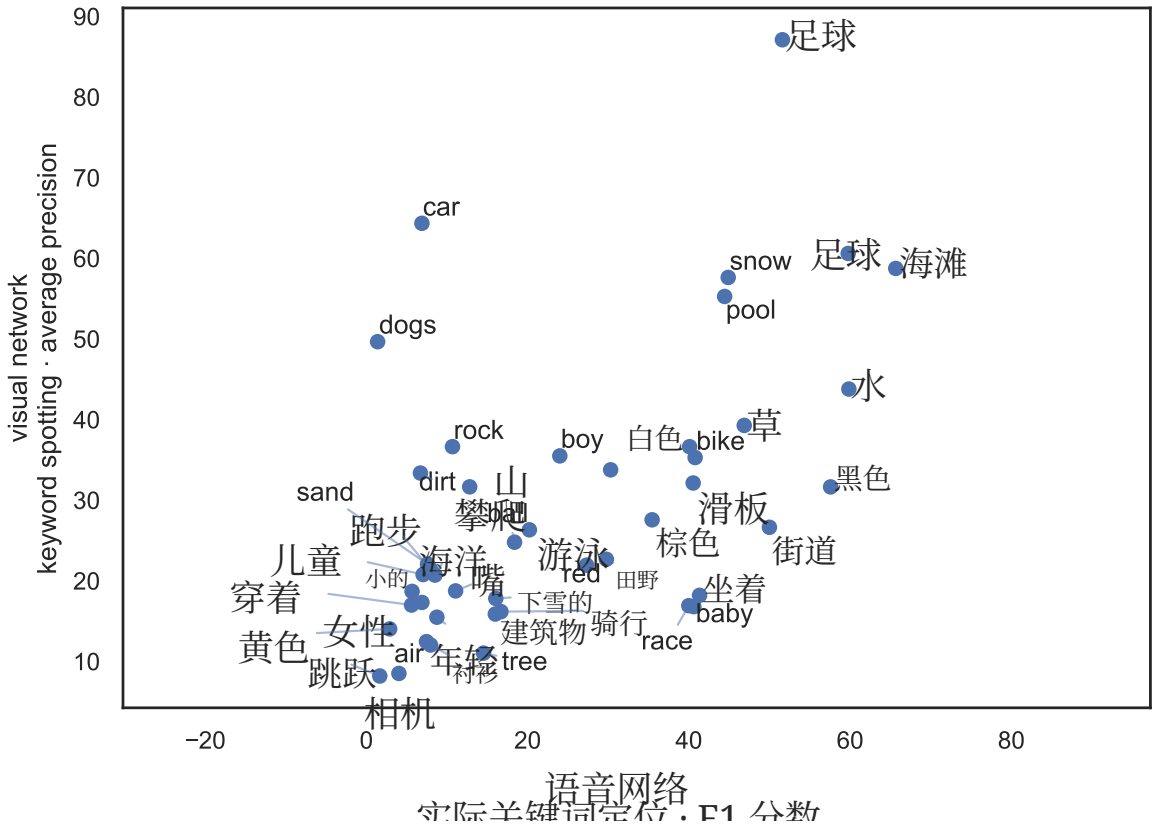


图 4.7： 针对图 4.6 中包含的 44 个关键词，语音网络的定位性能（F_1）与视觉网络的关键词定位性能（平均精度）之间的比较。结果基于 CNN-Attend 架构和掩码输入定位方法报告。

定性样本

图 4.8 展示了四个关键词的排名最高的话语（根据其检测分数）及其对应的定位分数：自行车、空气、汽车、男孩。对于自行车，可以看到检测和定位都正确，且分数都很高，接近最大值 1。在第二个例子中，查询关键词是空气，排名最高的话语确实包含该查询（因此检测正确），但单词单板滑雪者 的定位分数高于空气，导致定位错误。正如我们将在下一组实验中看到的，由于空气 和单板滑雪者 经常同时出现，两者常常被混淆。第三个例子突显了另一种常见的混淆——查询为汽车时，系统常常预测红色；在这种情况下，可能是红色围巾 与更常见的短语红色汽车在语音上的相似性加剧了这种情况。最后一个例子展示了定位差一点就正确的情况：虽然单词男孩的分数很高，但最大值出现在单词在的边界处。定位错误可能是强制对齐过程本身存在固有误差所致；该过程用于为转录文本中的单词分配位置信息。

表 4.7： 十六个查询关键词子集中最常被定位到的单词。对于每个关键词，我们根据检测分数选取排名前 20 的话语，并在每个话语中根据定位分数找出预测词。我们报告最常见的 5 个定位单词。我们使用 CNN-Attend 架构和掩码输入定位方法。“–”符号表示口语噪声，或强制对齐器的音标词典中没有收录的单词。

查询	排名前 5 的定位单词及其计数									
air	单板滑雪者	11	bike	2	–	2	滑雪板	1	滑雪	1
ball	足球	9	ball	5	球员	2	篮球	2	色彩鲜艳	1
黑色	黑色	18	dog	2						
car	race	7	red	4	赛车场	2	提高	2	car	2
dogs	三	8	二	6	dog	2	小狗	1	dogs	1
face	攀爬	7	rock	5	goes	1	岩石	1	金发	1
足球	足球	15	球员	4	运动员	1				
海洋	冲浪者	11	冲浪板	4	wave	3	冲浪	1	表面	1
橙色	red	9	穿着	1	a	1	寻回犬	1	吉他	1
pink	girl	8	年轻	2	女孩	2	穿着衣服	1	one	1
pool	pool	18	游泳	1	游泳者	1	的			
足球	足球	20								
游泳池		16	游泳	3	游泳	1				
tree	tree	9	树	7	训练	1	攀爬	1	–	1
黄色	黄色	17	自行车	1	bike	1	race	1		

最常见的混淆

对于词汇表中的每个关键词，我们找出音频话语中最常与其混淆的单词。为生成结果，我们采取以下步骤。(1) 给定一个关键词，按照话语级检测分数从高到低排列话语。(2) 对排名前二十的话语，选择定位分数 $\alpha_{w,\,1}$、……、$\alpha_{w,\,T}$ 中最大值所在位置对应的单词作为预测词。(3) 报告最常被预测的前五个单词及其计数。表 4.7 展示了部分关键词的结果。

我们观察到，部分关键词（黑色、泳池、足球、树）匹配良好，而其他关键词则会与语义相关的词混淆：空气 $\rightarrow$单板滑雪者；海洋 $\rightarrow$冲浪者；球 $\rightarrow$足球；游泳 $\rightarrow$泳池；脸 $\rightarrow$ 攀爬、岩石。汽车 的表现令人意外：考虑到汽车 具有视觉支持，我们原本预期其定位效果会更好，但该词常与比赛 和红色混淆。有趣的是，关键词狗 有时会与数词（*two* 和三）关联；我们认为这是因为我们区分单数词和复数词（单词狗 与单词狗不同）。对于一些表示颜色的关键词（黑色、白色、黄色），我们取得了较好的结果；而其他关键词则存在更多语义混淆：粉色 $\rightarrow$ 女孩；橙色 $\rightarrow$ 红色。

4.4. 总结

本章探讨了使用视觉接地语音模型进行关键词定位的不同方法，这些模型从图像和未标注口语描述中学习。具体而言，我们采用了一种视觉接地语音方法，用软文本标签为训练图像添加标注，作为语音网络的目标，使其能够检测话语中是否出现某个关键词。我们为这类视觉接地语音模型赋予了定位能力，并研究了四种定位方法。效果最佳的方法依赖于掩蔽：仅将语音输入的一部分送入模型，并通过观察输出概率的变化，判断话语的未掩蔽区域中是否出现某个关键词。通过使用宽度可变的滑动窗口，可以确定关键词的最佳位置。在预言机设定下，假设系统知道话语中出现了某个关键词，只需预测其位置，该方法的定位准确率达到 57%（而注意力方法达到 46%）；若先进行检测，则 F_1 为 25%。不过，与注意力方法相比，输入屏蔽的计算成本更高。

我们展示了，首次检测或排序阶段限制了系统的性能。通过改进视觉标注器，或许可以提升性能，因为每个关键词的性能与视觉标注器在该关键词上的性能高度相关。我们还表明，许多错误定位都是由预测语义相关词语造成的——探索这种做法能带来什么益处，是未来工作中另一个值得研究的方向（见第6.3节）。

到目前为止，我们考虑的是一组规模适中的语音-图像对，其中未标注口语描述使用英语。为了在真实低资源设置中应用所提出的方法，我们需要研究使用较少数据会如何影响性能。在一种极端低资源环境中，如果某种语言没有书写形式，还需要考虑如何使用书面关键词查询模型。我们将在下一章探讨这一问题，研究跨语言关键词定位：给定一种（资源充足的）语言中的关键词，然后在另一种低资源语言的语音中定位该关键词 [35]。我们在一种真实的低资源语言上开展这些实验，而不是人为地将英语视为低资源语言。

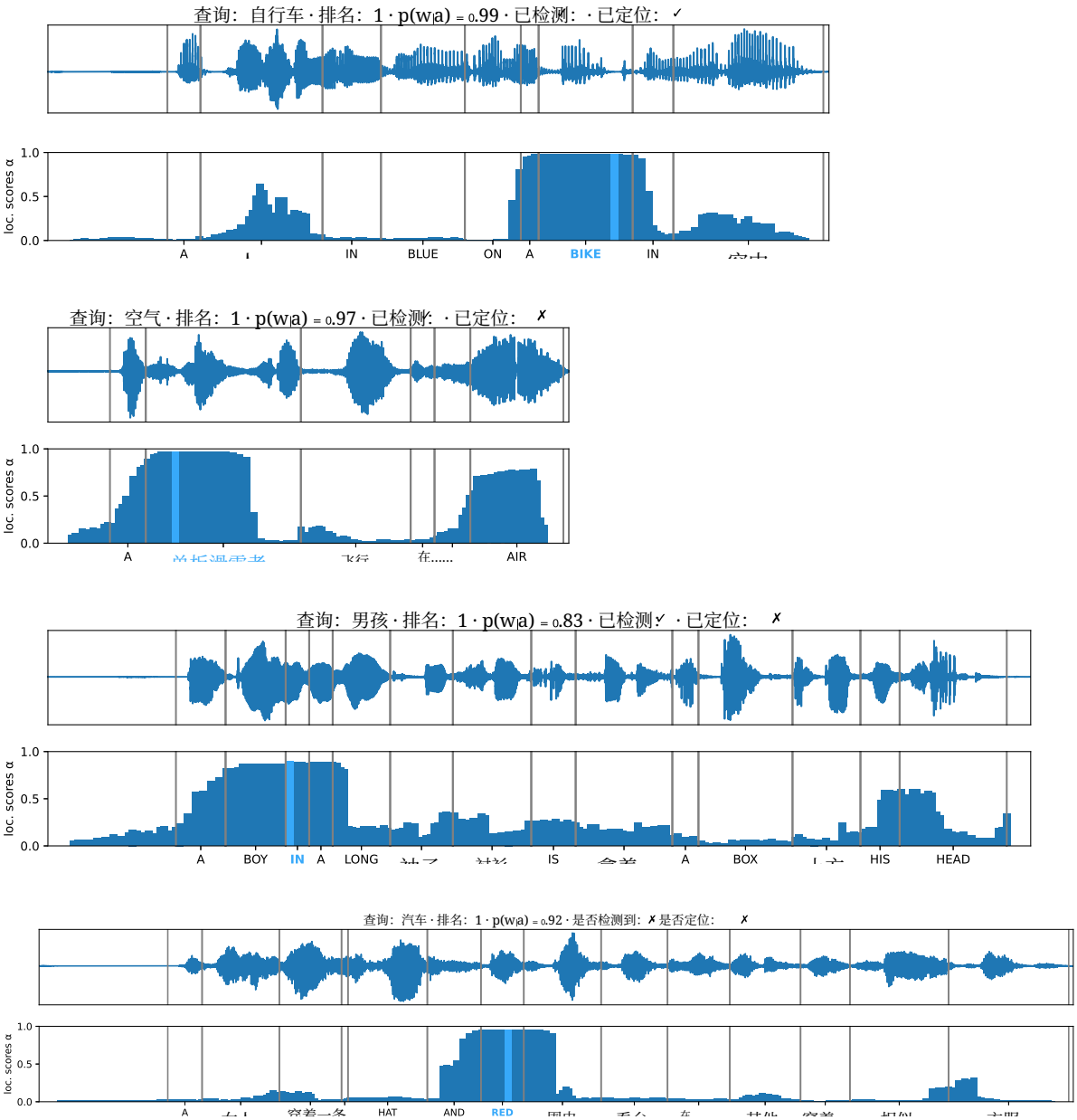


图 4.8： 使用 CNN-Attend 架构和掩码输入定位方法，对四个查询关键词（自行车、空气、汽车、男孩）进行预测的示例。每幅图中展示了输入音频信号（上方子图）和输出定位分数（下方子图）。此外，还给出了检测分数 $p(w|\mathbf{a})$、每个关键词是否被正确检测或定位，以及话语的转录文本。垂直灰线表示词边界，浅蓝色标出的单词是得分最高的单词。

第5章

真实低资源设置中的跨语言关键词定位

在前几章中，我们介绍了视觉接地关键词检测与定位的方法。这些方法的核心是现有模型 [2, 39]，我们为其扩展了四种不同的定位方法。正如第1章所述，此类方法可以帮助语言学家处理和研究仅以口语形式存在的语言。然而，由于缺乏真正的低资源 VGS 数据集，很难在这种情境下正确评估 VGS 系统的性能——截至目前对该系统所做的分析都处于人工设定中，其中英语被视为低资源语言。

本章将全面探讨我们的第二个研究问题（第1.2节）：我们能否在真实低资源设置中进行跨语言的视觉接地关键词定位？我们通过引入一个多语言多模态数据集，迈向更贴近现实的设置；该数据集支持在真实低资源设置中训练和评估视觉接地语音模型。具体而言，我们收集并发布了一个新的单说话人数据集，其中包含为 6,000 张 Flickr8k 图像录制的约鲁巴语音频描述（第5.2节）——约鲁巴语是一种在尼日利亚使用的真实低资源语言。我们训练了一个基于注意力机制的视觉接地语音模型，图像会自动标注英语视觉标签，并与约鲁巴语话语配对（第5.4节）。这使得跨语言关键词定位成为可能：系统可以在约鲁巴语语音中检测并定位书面英语查询。第5.3节介绍了用于评估模型跨语言关键词定位性能的评估方案。第5.5节介绍架构和实现细节。最后，在第5.6节中，我们展示了视觉接地跨语言系统的定量与定性分析。

本研究发表于 *IEEE SLT*：K. Olaleye、D. Oneat˘a 和 H. Kamper，《Yor`ub´a 语音-图像数据集 YFACC：通过视觉定位实现跨语言关键词定位》，*IEEE*口语技术（*SLT*），2023年。

5.1. 与先前工作的关系

本节旨在将本章开展的工作与文献中的先前研究联系起来。具体而言，新 Yor`ub´a 语音-图像数据集（在第 5.2 节介绍）所尝试并实现的任务——跨语言关键词定位——是在文献中探索的两个先前变体基础上迈出的新一步：**(i)** 单语关键词定位（第 4，[152] 章）；以及 **(ii)** 跨语言关键词检测 [35]。

如第 1 章所定义，从图像及其口语描述中学习的模型称为视觉接地语音模型 [1,3,4,13–19,28]。尽管先前关于视觉接地语音的研究表明，这类模型有望在低资源条件下执行多种下游任务，但只有极少数研究在真实环境中验证了这一方法：大多数论文使用未标注的英语语音或其他高资源语言的数据 [16,42] 来模拟低资源环境。

为弥补视觉接地语音研究中的这一不足，我们推出了一个新的低资源语言语音-图像数据集，语言为 Yor`ub´a。Yor`ub´a 是尼日利亚三种官方语言之一，尽管有超过 4400 万人以该语言为母语，但目前只有少数高质量数据集可用 [56–59]。我们的数据集通过为 Flickr8k 数据集 [48,49] 添加口语描述构建而成：我们录制了一位 Yor`ub´a 说话人的 6000 条 Yor`ub´a 话语，内容对应于原始英文图像说明的译文。我们还为 500 条测试话语中的每条话语所选出的 67 个关键词提供了带有起止时间戳的人工对齐。我们将最终得到的数据集称为 Yor`ub´a Flickr 音频描述语料库（YFACC）。图 5.1 展示了 YFACC 中的样本示例。扩展 Flickr8k 的一大优势是可以重复利用现有标注（英语录音 [1]，以及其他语言的译文 [172,173]），从而在未来工作中构建内容更丰富、也可能更有趣的任务。

先前研究探讨了跨语言关键词检索：在一种语言中检索包含另一种语言中特定关键词的话语。具体而言，Kamper 和 Roth [35] 研究了视觉定位是否可用于跨语言关键词检索。作者声称，这种跨语言关键词检索系统可以利用高资源语言的文本查询，在低资源语言的语音中进行搜索。为验证这一主张，他们使用英语语音（人为设定为低资源语言）和德语查询（高资源语言）。具体来说，他们使用德语视觉标注器为每张训练图像提供关键词标签，然后训练神经网络，将英语语音映射到德语关键词，而无需接触平行语音转录文本或翻译。作者还指出，大多数错误检索结果都包含含义相同或语义相关的关键词。本章在他们的工作基础上，直接纳入了关键词定位。此外，我们考虑的是真实低资源设置，而非人为将英语视为低资源语言。

致我们的 凭借我们的知识，我们是首个尝试跨语言关键词定位的团队。尽 le

管这项任务更具挑战性，但它也可能对开发面向语言学家的应用产生更大的影响。我们将第4.2.3节所述的定位方法应用于跨语言环境，并采用第3.3.4节所述的模型架构（该架构以现有架构 [2,39] 为基础）：自动为图像添加英语视觉标签，作为基于注意力的模型的目标，该模型以约鲁巴语`语´音为输入。这样一来，就可以使用英语文本查询约鲁巴语`语´音模型。

5.2. YFACC：约鲁巴语 Flickr 音频描述语料库

我们介绍全新的约鲁巴语 Flickr 音频描述语料库（YFACC）。YFACC 数据集将 Flickr8k 图像-文本数据集扩展至约鲁巴语，并包含三种模态：其 6,000 条图注的约鲁巴语译文；这些译文对应的口语录音；以及针对其中 500 条图注的子集所标注的 67 个约鲁巴语关键词的时间对齐。



图 5.1： 新建约鲁巴语 Flickr 音频描述语料库（YFACC）中的样例：Flickr8k 数据集中的图像（左）与约鲁巴语音频描述（右）配对，这些描述基于原始英语描述。我们为部分关键词（绿色）提供了时间对齐，以供评估。

5.2.1. 数据收集

我们从 Flickr8k 数据集中随机选取了 6,000 张图像，并为每张图像从其关联的五条图注中选取一条。随后，两位母语者将选出的英文字幕人工翻译成约鲁巴语。为便于录音，我们添加了部分变音符号。所有译文均由一位男性母语者录制。仅采集一位说话人的语音，主要是因为这符合真实的语言记录场景：语言学家通常只能接触到少数说话人。录音使用 BOYA BY-M1DM 双头全向领夹式麦克风，以 48 kHz 采样率和双声道进行。

这 6k 条话语被划分为训练集（5k）、开发集（500）和测试集（500）。对于这 500 条测试话语，一位母语说话人为一组 67 个约鲁巴语关键词生成了时间对齐，这些关键词分别对应表 3.2 中列出的 67 个英语关键词。这些关键词最初发表于 [39], ，并经过筛选，确保能够通过视觉进行定位且标注者间一致性较高，因此适合用于评估定位算法。我们使用 `Praat` [186] 生成对齐。对齐内容包括每个关键词的开始时间、结束时间和时长。表 5.2 展示了 YFACC 的定量特征详情，并将其与其他现有约鲁巴语数据集进行了比较（有关这些其他数据集的更多详情，请参见第 5.2.3 节）。图 5.1 展示了 YFACC 中的样例。

5.2.2. 跨语言差异

约鲁巴语在语言和文化方面都与英语有很大差异。接下来，我们将讨论其中一些与 YFACC 相关的具体方面。

声调符号

在 Yor`ub´a 中，发音和书写涉及使用三种声调符号 [187]：$d`o$，降调（重音符）；$re$，平调（无符号）；$m´i$，升调（尖音符）。这些符号标在七个 Yor`ub´a 元音（$a, e, e., i, o, o., u$）和四个鼻化元音（$e.n, in, o.n, un$）上方。声调符号是理解 Yor`ub´a 的关键，因为没有这些符号，许多单词都会产生歧义。例如，$okun$ 既可以表示 $`okun$（海滩）、$ok`un$（绳子），也可以表示$okun$（力量）；而$o.ko.$既可以表示$o.k`o.$（汽车）、$o.k´o.$（锄头），也可以表示$o.ko.$（丈夫）。实际上，书写时标注所有符号是一项耗时的活动。因此，对于我们的数据集，我们只为 67 个关键词标注变音符号，因为只有这些词会用于评估。不过，将来或许可以通过自动化方法 [188]为其余单词生成这些符号。

两种语言中的关键词

即使 YFACC 是通过翻译 Flickr8k 数据集构建的，英语图像描述中出现的关键词也不一定会出现在对应的 Yor`ub´a 译文中，反之亦然。例如，一位金发女性似乎在等车被译为 Yor`ub´a：*Ob`ınrin bilondi nd´ur´o de o.k`o.*（回译为一位金发女性站在汽车旁），因此关键词看台 和汽车 出现在 Yor`ub´a 图像描述中。我们通过计算科恩κ系数 ($\kappa$) 相关系数，量化两个语言中关键词共同出现的情况。测试集中的每个样本由两个句子表示：一个英语句子和一个对应的 Yor`ub´a 句子。给定一个英语单词和一个 Yor`ub´a 单词，我们会核实它们是否分别出现在对应的句子中；例如：

id 蓝色`
okun s011 0
s020 1 s031 1 ...
... ...

（其中 1 表示存在，0 表示不存在）。

然后，我们取两个出现向量：一个对应英语单词 $[1, 0, 1, \ldots]$，另一个对应约鲁巴语单词 $[0, 1, 1, \ldots]$，并使用科恩卡帕分数的归一化变体 [189]（如下所述）计算二者的一致性。直观而言，该分数越高，这对单词就越有可能出现在图 5.2 所示的同一组样本中。

科恩卡帕系数（κ）是一种通过校正随机一致性来衡量两个二元向量一致性的指标。但该 κ 分数的上界取决于边缘概率之差（在我们的例子中，即某个单词在测试句子中的出现次数）：如果一个单词（属于一种语言）比另一个单词（属于另一种语言）出现得更频繁，那么可达到的最高分数就会小于 1。由于我们要比较不同语言中的单词（这意味着出现次数不同），因此我们将通常使用的系数除以可达到的最大值，得到我们所说的归一化科恩 κ 分数：

$$\kappa_{\text{norm}} = \frac{\kappa}{\kappa_{\max}} \tag{5.1}$$

其中

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{5.2}$$

and

$$\kappa_{\max} = \frac{p_{\max} - p_e}{1 - p_e} \tag{5.3}$$

with

- p_o 为相对观测一致率
- p_e 即偶然一致概率
- $p_{\text{最大值}}$ 即最大可达概率。

直观来说，如果一个英语单词和一个约鲁巴语单词倾向于共同出现（即出现在相应的英语—约鲁巴语翻译样本中，例如前一个例子中的$ride–o.k`o.$ 这一对），那么这对词就会得到较高的分数。

图 5.2 展示了 YFACC 中 500 条测试语句里英语—约鲁巴语词对之间相关性的图表。我们观察到图表对角线上存在很强的相关性，这表明英语关键词确实与其译文相关，但也有一些例外。有些约鲁巴语单词有多种意义，因此会与不止一个英语单词共同出现，例如 $`okun$，它既表示海滩，也表示海洋。反过来，有些英语单词也有多种意义：足球不仅指体育运动，也指球；小的 不仅指大小，也指年龄（小男孩；$o.mo.k`unrin\ k´eker´e$）。最后，有些相关性似乎仅仅是因为单词在两种语言中都共同出现：$´nwo.$（穿着）与衬衫、红色和足球相关。

文化差异

从英语数据集入手可能会造成文化方面的不匹配，正如 [164] 中所讨论的那样。Flickr8k 中选出的 67 个关键词（表 3.2）大多涉及普遍的跨文化概念（例如儿童、年轻、游泳），但有些英语单词在约鲁巴语中没有本土译法。例如，夹克和网球 等词通常会根据英语发音进行翻译，即在英语单词中添加额外的变音符号：$j´ak`e.t`ı$和$te.n´ ısa`ı$。或许，对于普通尼日利亚人来说，唯一不太有用的关键词就是英语单词滑板。

5.2.3. 与 Flickr8k 及其他约鲁巴语数据集的关系

在第 3.1 节中，我们详细介绍了 Flickr8k 数据集，这是一个双模态（图像和文本）数据集。我们还介绍了 Flickr 音频描述语料库（FACC）的详情，该语料库由 Harwath 和 Glass [1]，提出，是 Flickr8k 数据集的早期扩展版本，包含 36 小时的非静音英语语音，涵盖 Flickr8k 数据集中的每条文本描述。如本节引言所述，YFACC 扩展了 Flickr8k 数据集，加入了其中 6,000 条图注的约鲁巴语译文、约鲁巴语口语话语，以及 Flickr8k 中 500 条图注子集所对应的 67 个约鲁巴语关键词的时间对齐。

与以往的约鲁巴语数据集相比，我们的数据集有所不同，其模态更丰富，如表 5.1 所示。通过扩展 Flickr8k，我们能够重复利用其现有的

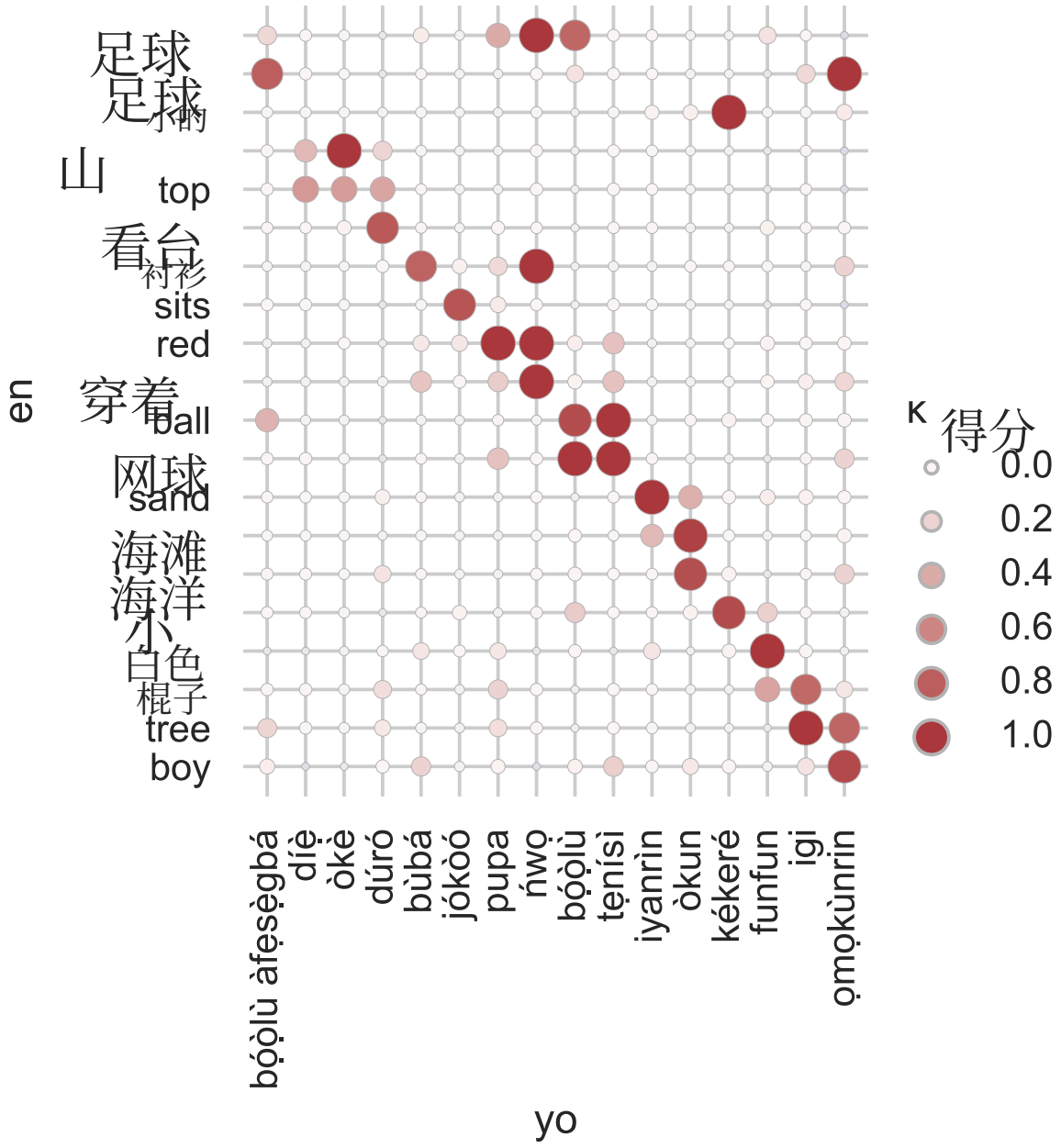


图 5.2： YFACC 中 500 条测试话语里的英语–约鲁巴语关键词对的共现，按归一化 Cohen's κ 得分计算。为便于阅读，我们从 67 词词汇表中选取关键词。

表 5.1： 按可用模态对约鲁巴语开源语音语料库进行比较。新的 YFACC 数据集为 67 个关键词中的一部分提供了对齐。

	模态				
数据集	语音	文本（yo）	文本（英语）	图像	对齐
van Niekerk 等人 [56]	✓	✓			
Gutkin 等人 [57]	✓	✓			
Meyer 等人 [58]	✓	✓			
Ogayo 等人 [59]	✓	✓			
YFACC（我们的）	✓	✓	✓	✓	部分

表 5.2： 开源 Yor`ub´a 语音语料库在定量特征方面的比较。

	说话人	时长	话语	采样
数据集	数量	小时	数量	kHz
van Niekerk 等人 [56]	33	2.75	4316	16
Gutkin 等人 [57]	36	4.00	3583	48
Meyer 等人 [58]		33.30	10228	48
Ogayo 等人 [59]		18.04	10978	
YFACC（我们的数据集）	1	13.3	6000	48

添加标注并构建多模态多语言数据集。与 FACC 类似，该数据集可用于定义各种各样的任务。本章特别探讨视觉基础跨语言关键词定位任务。未来可考虑的示例任务包括语音到语音机器翻译，或将视觉概念与尤鲁巴语语音中的图像关联起来。录音由单一说话人以高质量（48 kHz）完成，这也为未来开发尤鲁巴语文本转语音技术提供了条件。

约鲁`巴´语音语料库。直到最近，只有两个开源的约鲁巴语音数据集：拉各斯—西北大学约鲁巴语音语料库 [56]，由 17 名男性和 16 名女性说话人以 16 kHz 录制，总时长约 2.75 小时；以及 Gutkin等人 [57], 的语料库，由 36 名男性和女性志愿者录制的 48 kHz 音频组成，总时长约 4 小时。与我们同期开展的研究 [58, 59], 认识到为非洲语言创建数据集的重要性，并发布了规模更大的数据集，旨在构建文本转语音系统；表 5.2 列出了这些数据集的统计信息。

5.3. 跨语言关键词定位的评估

我们已经详细讨论了用于评估单语模型在关键词检测（见第 3.2.1 节）、预言机关键词定位（第 4.1.1 节）和实际关键词定位（第 4.1.2 节）方面性能的评估流程。本节将仅简要说明，我们如何针对本章所讨论的跨语言模型调整评估流程。

本章考虑的主要任务是在（约鲁巴语）语音中进行跨语言（英语）关键词定位。我们选择精确率作为主要指标，其计算方式为真正例样本数与检索样本数之比。当给定英语关键词的检测分数高于阈值 θ时，该样本即为检索到的 样本。若满足以下条件，该样本即为真正例：（1）该样本已被检索到；（2）给定的英语关键词在约鲁巴语话语中被正确定位，即预测位置落在相应约鲁巴语关键词的语音区间内。我们将这一指标称为（跨语言）实际关键词定位。在所有实验中，我们将 θ 设为 0。5，这一数值在 FACC 开发集上取得了最佳性能。

我们还通过考察另外两个较简单的任务来量化该任务的上界：（跨语言）关键词检测和（跨语言）预言机关键词定位。这两项任务及其评估流程与第 3.2.1 节和第 4.1.1 节中的描述完全相同，只是在这里我们考虑的是跨语言环境。我们在第 4.1.1 节中还补充说明，在预言机关键词定位环境下，我们旨在评估模型的定位能力，而不考虑检测。

5.4. 使用视觉接地语音模型进行跨语言关键词定位

本章采用的方法与第 3 章和第 4 章中已经介绍的方法相同——在此我们仅简要说明如何针对跨语言任务对其进行调整。

我们具体研究跨语言关键词定位任务：给定一个英语关键词和一段约鲁巴语口语话语，我们希望自动在音频中找到与给定英语关键词对应的约鲁巴语单词的位置。如果该关键词没有出现，模型应指出未检测到该关键词。关键的是，我们假设监督信号仅来自视觉通道。具体而言，在我们的案例中，我们假设可以获取一组与相应音频描述相关联的图像（YFACC）。为了利用这一弱监督信号，我们采用了与第 4.2 节所述单语基于 VGS 的定位方法几乎相同的方法，并将其调整至跨语言环境。

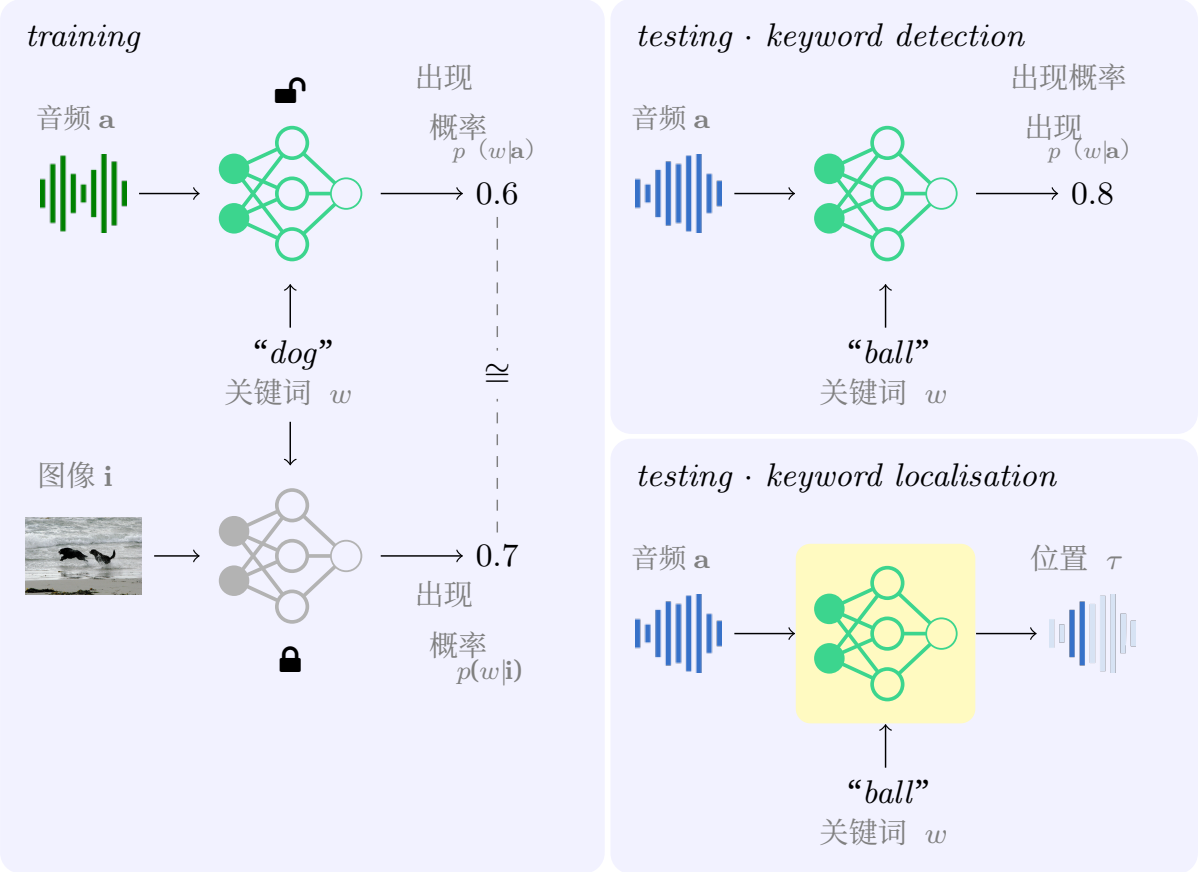


图5.3： 使用视觉基础语音模型进行跨语言关键词定位。左图： 在训练阶段，语音模型根据预训练且冻结的（英语）图像模型提供的视觉监督，学习预测约鲁巴语音频话语中是否出现英语关键词。右图： 在测试阶段，给定一段约鲁巴语音频话语，模型会输出英语关键词出现在该话语中的概率。如果该概率超过预设阈值 θ，模型还会根据网络中的注意力权重定位该关键词在话语中的位置。

5.4.1. 训练：视觉监督

我们的模型由一个音频网络组成，该网络处理约鲁巴语语音，并输出给定英语关键词在话语中（任意位置）出现的概率。该音频模型通过跨模态监督进行训练：对于一组预先定义的视觉类别，它学习复现预训练图像网络针对相关图像给出的预测结果。这一音频网络的权重通过图像预测结果与音频预测结果之间的二元交叉熵损失进行优化；该损失在词汇表中的关键词上取平均值。这些类别以英语关键词（例如 狗、儿童、球）定义，但也对应于音频中出现的约鲁巴语单词，视觉通道则将两种语言联系起来。训练过程示意图见图5.3（左）。

5.4.2. 测试：跨语言定位

在测试阶段，给定一段约鲁巴语话语——推理时不需要图像——我们的模型会输出视觉词汇表中某个英语关键词的出现概率。当该概率高于某个阈值 θ时，我们便认为该关键词出现在这段话语中。检测到关键词后，我们还希望定位相应的约鲁巴语单词。为了获得该约鲁巴语单词的时间位置，我们利用模型的额外结构，即模型在输入语句上的注意力权重。我们将权重最高的时间戳选作输入关键词最可能出现的位置。测试阶段的示意图见图5.3（右）。

5.5. 架构与实现细节

5.5.1. 数据集

在我们的实验中，除了新的 YFACC 数据集（第 5.2 节），我们还使用了先前的英语 FACC 数据集（第 3.1.2 节） [1]，用于我们拿来比较的单语英语视觉接地语音系统。

我们针对 FACC 的两种变体进行了实验，它们的训练数据量不同：`en-30k`包含全部原始的 3万个训练样本， `en-5k`包含 5千个训练样本的子集，以匹配 YFACC 的大小。对于在 FACC 上训练的模型，所用的测试数据是与 YFACC 测试集对应的 500 条英语话语。

5.5.2. 架构：注意力网络

我们采用第3.3.4节和第4.2.5节所述的 CNN-Attend 架构，但针对学习率、激活函数和最后一层的大小进行了超参数搜索。该搜索基于 FACC 开发集上的 F_1 关键词检测分数进行。与第3.3.4节中的原始架构相比，唯一的改动是将最后一层的大小从 4096 增加到 8192；这使开发集上的 F_1 从 14.4% 提高到了 19.4%。

音频网络由四个子组件组成：音频编码器、关键词编码器、注意力层和分类网络。音频编码器——在我们的案例中是一个卷积网络——将 Yor`ub´a 音频嵌入为一系列特征向量，而关键词编码器则使用可学习查找表为英语关键词分配一个嵌入向量。注意力层根据关键词嵌入对音频特征进行时间池化，将其汇总为一个向量。最后，再使用一个小型多层感知机网络将聚合后的特征向量投影为分类分数。如前一小节所述，我们将注意力分数最大值所在的位置作为查询关键词的位置预测。该架构受到 [7], an

表 5.3： 视觉基础语音模型在跨语言任务（在约鲁巴语音频中查询英语）中使用所提出的 YFACC 数据集（前三行），以及在单语任务（在英语音频中查询英语）中使用 FACC 数据集（最后两行）时的结果。模型针对实际关键词定位这一主要任务进行评估。通过更改随机种子重复实验三次，并报告平均性能及其标准差。

	关键词 语言	音频 语言	训练 话语数量	初始化	实际定位 精确率
`random`	随机基线				0.1 ± 0.0
`yo-5k`	英语	→ 约鲁`巴´	5k	随机	16.0 ± 1.6
`yo-5k-init`	英语	→ 约鲁`巴´	5k	`en-30k`	$\mathbf{22.8} \pm 4.3$
`en-5k`	英语	→ 英语	5k	随机	21.3 ± 4.4
`en-30k`	英语	→ 英语	30k	随机	$\mathbf{26.9} \pm 1.8$

用于手语视频视觉关键词检测的基于注意力的图卷积网络。

5.5.3. 实现细节

我们采用了与第3.3.4节和第4.2.5节所述实验类似的实现方式。具体而言，我们将 YFACC 下采样至 16 kHz，以匹配 FACC。所有话语均编码为梅尔频率倒谱系数，并在训练时使用 SpecAugment [175]进行增强。模型的视觉目标通过在 MSCOCO [114] 上训练的 VGG-16 网络 [174]计算得出，该网络针对 67 个关键词进行了训练，具体方法见 [39]。所有模型均按照第3.3.4节所述的方法实现。

5.6. 结果：跨语言关键词定位

5.6.1. 定量结果

我们的目标是探究，在极低资源环境下训练的视觉接地语音模型能在多大程度上实现跨语言关键词定位。我们具体考虑这样一种情况：我们只有 13.3 小时未转录的 Yor`ub´a 语音及其对应的图像。英语–Yor`ub´a 跨语言结果见表 5.3、5.4 和 5.5 的顶部。

表 5.3 展示了模型在我们的主要任务——实际关键词定位任务上的性能。跨语言模型 `yo-5k` 的性能约为 16.0%。尽管单独来看这一结果似乎并不突出，但它远优于随机模型的性能（0.1%）。更重要的是，该结果落在单语设置下取得的结果范围（21.3 ± 4.4%）之内；在该设置中，使用英语语音查询英语关键词，且模型是在相同数量的英语话语上训练的

表 5.4： 视觉基础语音模型在跨语言任务（在 Yor`ub´a 音频中查询英语）上的结果，数据来自所提出的 YFACC 数据集（前三行）；以及在单语任务（在英语音频中查询英语）上的结果，数据来自 FACC 数据集（后两行）。模型在难度较低的任务（预言机定位）上进行评估，该任务为实际定位性能提供了一个上界。实验通过更改随机种子重复三次，并报告平均性能及其标准差。

	关键词 语言	音频 语言	训练 话语数量	初始化	预言机 定位 准确率
random	随机基线				5.3 ± 0.9
yo-5k	英语	→ 约鲁`巴´	5k	随机	23.1 ± 3.0
yo-5k-init	英语	→ 约鲁`巴´	5k	en-30k	$\mathbf{35.0} \pm 2.5$
en-5k	英语	→ 英语	5k	随机	27.3 ± 2.8
en-30k	英语	→ 英语	30k	随机	$\mathbf{44.0} \pm 0.9$

表 5.5： 视觉扎根语音模型在所提出的 YFACC 数据集上的跨语言任务（在约鲁巴语音频中查询英语；前三行）以及在 FACC 数据集上的单语任务（在英语音频中查询英语；最后两行）中的结果。模型在较简单的任务（关键词检测）上进行评估，该任务可为实际定位性能提供上界。通过更改随机种子重复实验三次，并报告平均性能及其标准差。

	关键词 语言	音频 语言	训练 话语数量	初始化	关键词 检测 精确率
random	随机基线				2.4 ± 0.2
yo-5k	英语	→ 约鲁`巴´	5k	随机	26.7 ± 6.4
yo-5k-init	英语	→ 约鲁`巴´	5k	en-30k	$\mathbf{33.3} \pm 7.1$
en-5k	英语	→ 英语	5k	随机	29.5 ± 5.0
en-30k	英语	→ 英语	30k	随机	$\mathbf{39.2} \pm 1.7$

（en-5k）。有人认为，跨语言环境更具挑战性，因为视觉关键词更偏向英语，且未考虑约鲁巴语在文化和语言方面的差异（第 5.2.2 节）。另一方面，YFACC 相对容易，因为它是一个单说话人数据集，而 FACC 数据集涉及多个说话人。接下来，我们将探讨有助于提升性能的因素。

训练数据量

虽然在跨语言环境下，我们受限于 YFACC 中现有的 5k 条训练话语，但我们可以根据使用 FACC 的相应单语设置，推断增加数据量后的性能。在表 5.3 中可以看到，使用六倍于原来的数据进行训练，效果显著提升，这并不令人意外：从 21.3%（en-5k）提升到 26.9%（en-30k）。然而，尽管增加训练数据量是提升性能最直接的方法，但在实践中也可能是最具挑战性的方法。

语音表征

在低资源条件下从头训练模型颇具挑战。常见做法是从预训练模型出发，再针对下游任务对其进行微调。这里，我们使用在 FACC 中所有图像–英语音频对上训练的视觉接地语音模型，在 Yor`ub´a 上初始化训练。相比 Yor`ub´a 话语，这类英语数据对无疑更容易获取。我们观察到性能显著提升——从 16.0% (yo-5k) 提高到 22.8% (yo-5k-init)，如表 5.3 所示。

将任务解耦

关键词定位包含两个步骤：(1) 检测查询关键词是否出现在话语中的某处；若出现，则 (2) 在给定话语中定位该关键词。为了解我们的方法中哪些组件有待改进，分别评估这两个步骤很有帮助。为此，我们假设其中一个步骤能够完美完成，从而得到两个新任务：预言机定位（假设检测完美）和关键词检测（假设定位完美）。尽管这两个任务不能直接比较，但其结果（预言机关键词定位见表 5.4，关键词检测见表 5.5）表明，这两个组件都有进一步改进的空间。还需注意，先前关于不同模型相对性能的结论，在这些新任务上依然成立。

单个关键词表现

词汇表中的 67 个关键词，其性能差异很大。我们观察到，一些关键词表现良好（高于平均水平），例如 棕色(*b´ur´a`un*; 100%)、自行车 (*k`e.k´e.*; 94.1%) 或 草 (*kor´ıko* 90.9%)。但还有许多其他

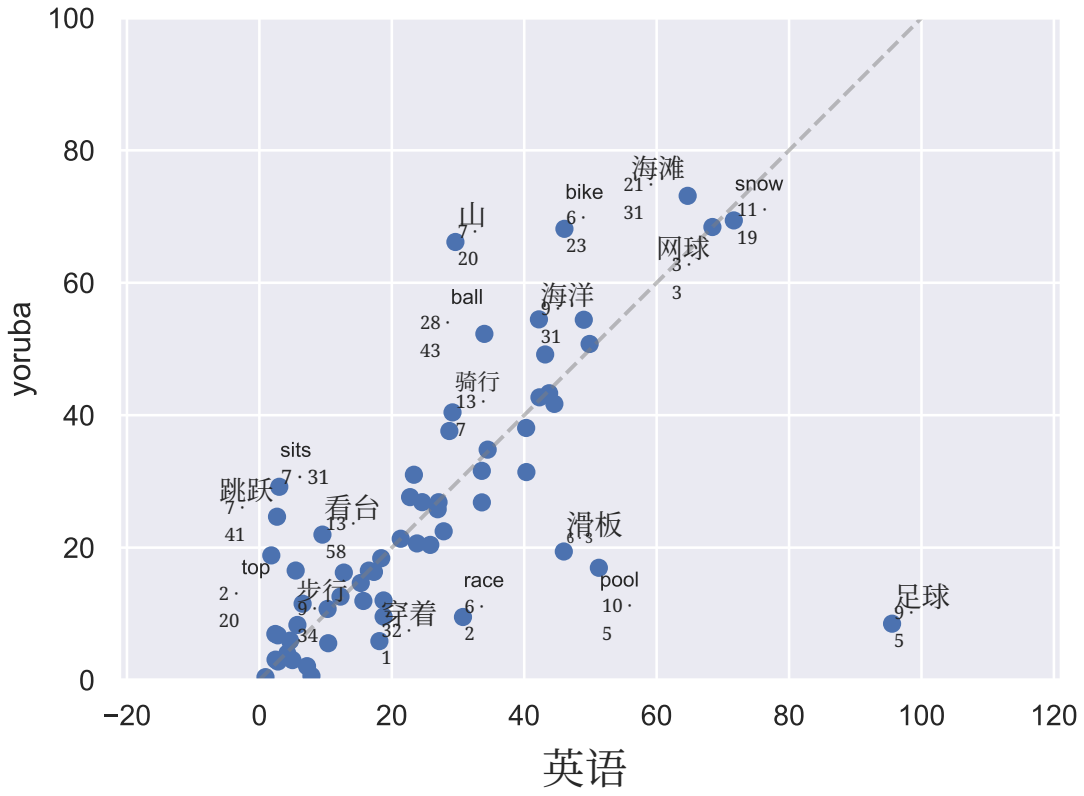


图5.4： 视觉教师在英语和 Yor`ub´a Flickr 测试集上的平均精度均值。

关键词会令模型难以应对。后者的原因包括视觉定位不佳（例如 相机、穿着、大型），以及对语义相关概念的混淆（骑行通常会检索到 $k`e.k´e.$，即 自行车；游泳 会检索到 $od`o\ ad´ag´un$，即 泳池）。密切关注单个关键词的具体特点，可以为未来的改进提供启发。

跨语言视觉教师性能

我们计算了视觉教师在英语和 Yor`ub´a Flickr 测试集上的性能。两种情况下的图像相同（因此视觉教师的预测结果也相同），但由于语言特征不同，图像描述可能会使用不同的关键词，因此性能自然会有所差异。我们报告了平均精度均值，并在图5.4中展示结果。（该图还显示了两个测试集中每个关键词的出现次数，分别对应原始英语测试集和 Yor`ub´a 测试集。）

值得注意的几点观察：

- 我们注意到，“足球”在英语中的性能非常好，但在约鲁巴语中的性能较差。这种现象或许可以用以下事实来解释：英语中的“足球”

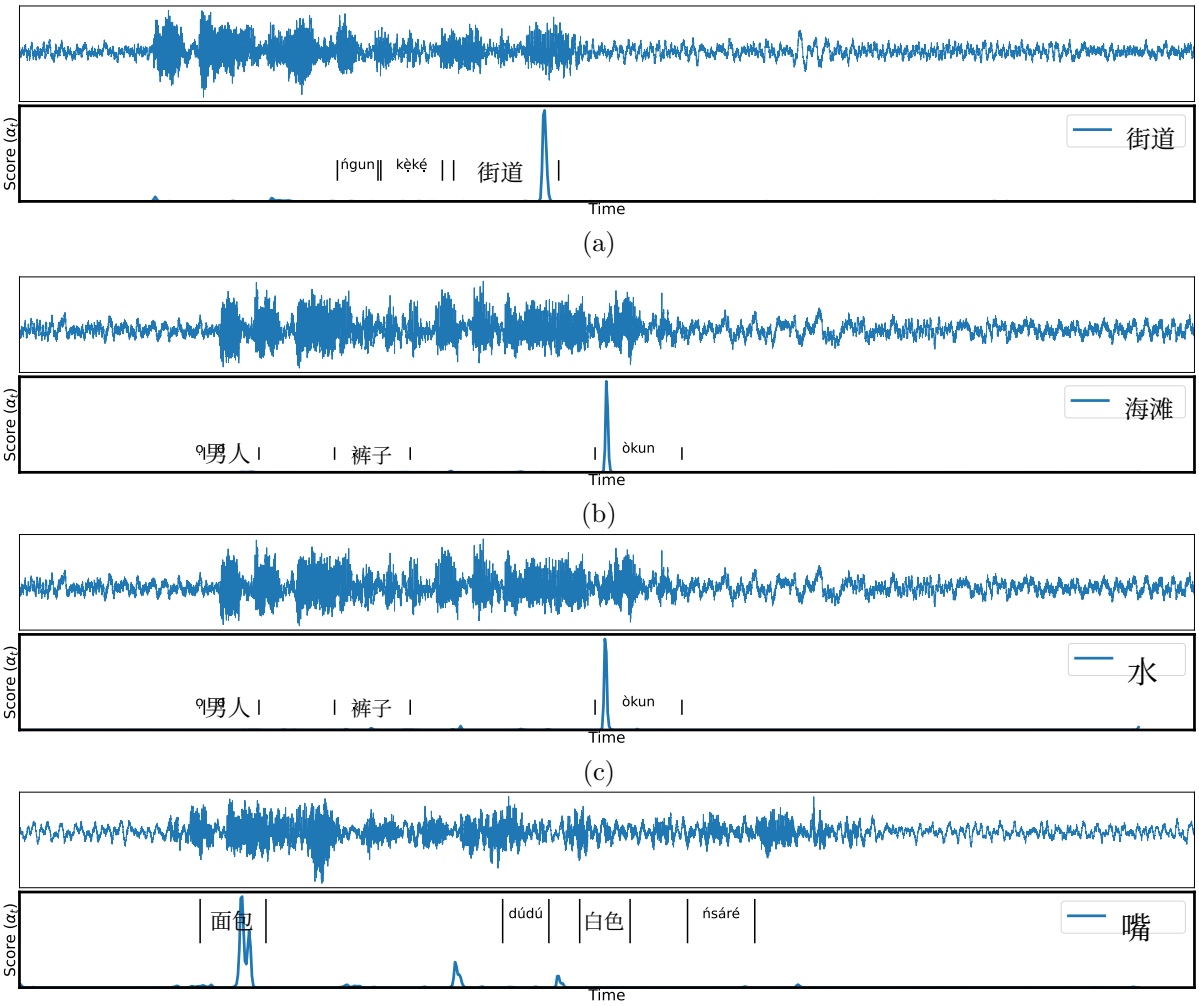


图5.5： 使用 `yo-5k` 视觉接地语音模型进行跨语言定位的示例。英语查询关键词显示在右侧。

和“足球”关键词被翻译成同一个约鲁巴语关键词，但它们在视觉上有所不同；视觉教师使用英语关键词的语义进行训练。

- 然而，虽然“海滩”和“海洋”在约鲁巴语中也各自有一个对应词，但两者的性能更加接近；这是因为这两个词在视觉上相关——它们通常会共同出现在图像中。

- “跳跃”“坐”和“看台”等词在约鲁巴语中的性能优于英语，因为它们在约鲁巴语图注中出现得更频繁（或许是因为其他一些英语单词被映射成了这些词？），所以它们的基线（随机）性能更高。

5.6.2 定性结果

为了更好地理解跨语言模型的失败模式，我们对结果进行了定性评估。图5.5(a)和(b)展示了对约鲁巴语中英语关键词的正确定位

话语：使用英语关键词 街道 进行提示时，模型会在包含口语单词 $\grave{}op\acute{}op\acute{}on\grave{}a$ 的帧中触发（5.5a）；使用关键词 海洋进行提示时，模型会在包含 $\grave{}okun$ 的帧中触发。图5.5(c)和(d)展示了失败案例。图5.5(c)表明，模型有时也会定位话语中语义相关的单词：当使用英语关键词 水进行提示时，口语约鲁巴语单词 $\grave{}okun$ （海洋）获得了最高的定位分数。图5.5(d)则是彻底的失败。

5.7. 总结

本章介绍了一个新的单说话人数据集，其中包含为 Flickr8k 图像录制的约鲁巴语口语描述。为了展示在真实低资源设置下，如何利用如此小规模的数据集（13.3 小时，6 千组语音–图像对）开展研究，我们使用这些数据，借助一个未经任何转录音频训练的视觉接地语音模型进行跨语言关键词定位。

除了跨语言检索任务本身固有的挑战外，英语–约鲁巴语这一组合也颇具挑战性，因为这两种语言属于不同的语系（印欧语系和尼日尔-刚果语系），且分别与迥异的文化相关。通过系统的实验，我们量化了这种语言不匹配的影响：分析所选关键词在不同语言间的匹配程度，将其与使用和 YFACC 相近规模数据训练的单语英语视觉接地语音模型进行比较，并将英语语音表征迁移至约鲁巴语（从而提升了检索性能）。

扩展 Flickr8k 的一大优势是可以重复利用现有标注（英语录音 [1], 其他语言的译文 [172, 173]），这有助于在未来工作中构建更丰富且可能更有趣的任务。

总而言之，本章的贡献包括：

- 发布了一个新的约鲁巴语语音-图像数据集`。
- 一个用于跨语言关键词检测和定位的基线系统。
- 对系统进行多层面的跨语言分析。

我们希望 YFACC 能推动相关研究，探索如何使用视觉接地语音模型解决真实低资源语言环境中的语音任务。

第6章

总结与结论

视觉基础语音建模（VGS）可用于在缺乏转写语音数据的低资源环境中实现语音应用。本学位论文探讨了如何利用在真实低资源语言环境中训练的视觉接地语音模型进行语音中的关键词定位，即找出给定的书面关键词在话语中的出现位置。下文首先直接回答本学位论文探讨的研究问题（第6.1节）。随后，我们概述并反思为回答这些问题所采取的方向和方法，并介绍我们对VGS建模的具体贡献（第6.2节）。我们还就本学位论文的主题提出未来工作的建议（第6.3节）。最后，我们简要总结本学位论文，重点阐述主要发现和重要启示（第6.4节）。

6.1. 研究问题与拟议答案

We 首先简要陈述我们两个主要研究问题的主要发现。

研究问题1。 视觉接地语音模型能否实现关键词定位？以往研究表明，视觉接地语音模型可用于跨模态检索、关键词检测和关键词检出等任务，但尚未探索关键词定位。视觉接地关键词定位的一项应用是濒危语言的记录工作 [147]。语言学家可以利用此类工具，快速定位话语集合中包含查询关键词的语音片段；唯一的前提是拥有目标语言的图像-语音对集合，用于训练模型。基于对人工低资源环境下视觉接地语音模型关键词定位性能的定量与定性分析，我们得出结论：在以英语模拟低资源语言环境的条件下训练的视觉接地语音模型，能够实现有限的关键词定位。此外，与许多其他视觉接地语音研究一样，我们发现许多错误的关键词定位都是由语义混淆造成的，例如，对于查询关键词“游泳”，却在某个话语中定位到了“仰泳”一词。

研究问题 2. 在真实的低资源环境中，我们能否跨语言进行视觉接地关键词定位？以往大多数视觉接地语音研究使用的都是图像与英语（或其他资源丰富语言）语音配对的

数据集。因此，研究中常以英语作为低资源语言的替代，这使得我们难以准确评估视觉接地语音系统在真实低资源环境中的性能。为解决这一局限，我们收集了一个新的约鲁巴语语音-图像数据集，随后在真实低资源环境中训练了视觉接地语音模型，以执行跨语言关键词检测与定位。根据视觉接地语音系统在跨语言关键词定位任务中的定量与定性分析，可以得出结论：该系统在其词汇表中数量有限的关键词上取得了尚可的性能。此外，还可以得出结论：在这一低资源环境中，通过从资源丰富语言上预训练的模型进行初始化，可以丰富跨语言视觉接地语音模型学到的表征。对跨语言模型的定性分析还显示，它们表现出与第一个研究问题相似的行为，即模型会出现语义错误。衡量模型性能时，是否应该因这类错误而惩罚模型？定位语义相关的关键词本身也可能是一种有用的应用。例如，在为洪灾受害者提供人道主义服务时，如果以英语单词“霍乱”为提示，在约鲁巴语话语中定位关键词“ẹ́ebì”（约鲁巴语中意为“呕吐”），仍可能让救援人员了解应采取何种治疗措施。

6.2. 研究结果

本节将总结并反思为解决上述研究问题所采取的方法，重点简要回顾我们作为起点所采用的视觉接地语音方法，总结我们在此视觉接地语音方法基础上提出的四种定位方法，最后反思研究结果。在整篇学位论文中，我们考察了两种不同的环境：用于解决第一个研究问题的人工低资源环境，以及用于解决第二个研究问题的真实低资源设置。

6.2.1. 人工低资源环境

在人工低资源语境中探究第一个研究问题——即使用英语模拟低资源语言环境——带来了诸多益处。首先，这使我们能够比较并直接扩展以往的研究。其次，这使我们能够先探究基于现有 VGS 数据集（具体而言，即现有的 Flickr 音频描述语料库（FACC））训练的视觉接地语音系统能否进行关键词定位。通过这项研究，我们还可以估算在真实低资源语言环境中开展研究所需的数据资源，而视觉接地语音系统正是在这种环境中真正发挥作用。第三，与……相比，模拟低资源环境，而不是

直接着手应对真实低资源环境，也可能有助于更快地进行原型开发，并探索更适合特定下游任务的方法。例如，试图构建系统以检测并定位某种濒危语言中关键词的计算语言学家，理想情况下应首先确定可能表现最佳、能够取得最佳结果的研究方法。

最适合回答我们的第一个研究问题的具体研究方法基于 Kamper等人 [2, 39], 的工作。他们证明，我们可以使用具有固定词汇表的外部图像标注器来获得软文本标签，并将其作为语音网络的目标；该语音网络将语音映射为关键词标签。我们以这一研究方法为起点，不仅用于解决第一个研究问题，也用于解决第二个研究问题。作为一项贡献，我们提出了四种定位方法。所提出的定位方法利用了视觉接地语音模型定位关键词的内在能力。在第一种方法中，我们改进了梯度加权类激活映射（Grad-CAM） [55]：这是一种在计算机视觉领域广泛使用的基于显著性的方法，用于确定输入图像的哪些部分对特定输出预测贡献最大。在第二种方法中，我们改进了分数聚合方法 [10] ，使其能够使用视觉监督。在第三种方法中，我们探究了注意力机制（其架构受 [7]启发）是否有利于视觉接地关键词定位。最后，我们提出了基于掩码的方法，通过在不同位置进行掩蔽，测量训练好的模型对部分输入所预测的响应分数；输出中的显著变化表明关键词的存在。与 Grad-CAM 一样，基于掩码的方法可以适用于任何架构。

尽管 Grad-CAM 适用于任何卷积神经网络架构，但它与用于训练视觉接地语音模型的特定多标签损失匹配不佳。结果表明，基于显著性的定位与所用的多标签损失并不匹配；检测性能更优的模型，其定位性能却较差。这一发现表明，如果采用一种与模型及多标签分类损失更好匹配的机制，定位性能应当能够提升。分数聚合的绝对性能较低。尽管如此，这种方法在关键词定位方面展现出了一些效果，且优于 Grad-CAM。我们发现，注意力方法显著优于现有的关键词定位方法。然而，其分数仍远低于采用无序词袋监督（通过转录文本获得）训练的模型；后者可视为理想化的视觉接地语音模型。此外，我们还发现，视觉训练的模型在定位语义相关词语时会受到惩罚，例如，对于查询关键词“游泳”，模型却定位到了单词“仰泳”。在预言机设定中，系统假定某个关键词确实存在，并直接预测其出现位置；在这种设定下，基于掩码的方法取得了 57.0% 的预言机性能，而基于注意力的方法为 46.0%。然而，基于掩码的方法在计算上的开销高于注意力定位方法。

We p对所采用方法的不同方面进行了详细分析 d

在人工低资源环境中。我们在下文总结了研究发现。

将视觉监督与词袋监督进行比较。 我们研究了词袋设定，并将其视为能够使用完美视觉标注器的设定。通过分析，我们得以估算使用离线预训练视觉系统提供的视觉监督所付出的代价。在所有情况下，从词袋监督转为视觉监督时，性能都显著下降。我们强调，尽管词袋为视觉接地语音性能提供了上界，但这个上界可能偏高，因为完美的视觉标注器可能会检测到人类不会描述的物体。我们得出结论，改进视觉标注可大幅提升视觉接地语音模型的关键词定位性能。

逐关键词性能。 我们研究了系统词汇表中每个关键词的定位性能。我们观察到，各种方法的性能高度依赖于关键词，并证实关键词的定位效果受其视觉接地程度的影响。此外，对于那些能够进行视觉接地、但定位效果仍然不佳的关键词，我们发现，它们与其他单词的共现导致了定位效果不佳。我们得出结论：对于经常在同一话语中共同出现的单词，任何无法获取额外信息的定位方法都难以将它们区分开来。

中间最大池化对定位性能的影响。我们还分析了不同的架构和定位方法，考察编码器结构的选择如何影响定位能力。我们发现，在每个卷积层操作之后（最后一层除外）都进行最大池化操作的架构中，定位性能有所下降。我们预料到会出现这种情况，因为最大池化操作会使人更难将每个输出预测追溯到相关滤波器。尽管先前的一项研究 [190] 表明，最大池化操作有助于提升关键词检测效果，但我们得出结论，关键词定位模型应避免使用该操作。

解耦关键词定位与检测。本学位论文研究的关键词定位任务包含两个步骤：(i) 检查查询关键词是否出现在话语中的某处；如果出现，则 (ii) 找出关键词出现的位置。为了量化这两个相互关联的步骤对模型造成的限制，我们假设其中一个步骤能够完美完成。由此形成了两种定位任务：预言机定位（假设检测完美）和关键词检测（假设定位完美）。我们观察到，在所有情况下，首次检测都会损害定位效果。因此我们得出结论，未来工作中解耦这两项任务可能会提升性能。

6.2.2. 真实低资源环境

我们在真实低资源设置下解决了第二个研究问题。我们首先在一种真实的低资源语言——Yor`ub´a——中收集了一个新的视觉语音接地数据集，随后展示了如何使用基于这些数据训练的视觉接地语音模型进行跨语言关键词检测与定位。我们收集的 Yor`ub´a Flickr 音频描述语料库（YFACC）数据集包含由一位说话人用 Yor`ub´a 语录制的、对应 6k 张 Flickr 图像的口语描述；Yor`ub´a 是尼日利亚使用的一种真实低资源语言。我们使用 YFACC 训练了一个用于跨语言关键词检测与定位的视觉接地语音系统：给定一条英语文本查询，检测该查询是否出现在 Yor`ub´a 语音中；如果检测到，则定位该查询在话语中的出现位置。

为了设计这一视觉接地语音系统，我们以 Kamper 和 Roth [35] 的方法为起点；他们通过使用两种资源丰富的语言来模拟低资源环境。具体而言，作者使用离线视觉标注器生成的德语软文本标签，训练一个接收英语语音并输出德语文本标签的语音网络。我们则使用离线（英语）图像标注器，为与相应 Yor`ub´a 语音关联的图像标注英语标签。这使我们能够训练一个基于注意力机制的视觉接地语音模型，利用英语关键词和 Yor`ub´a 语音执行跨语言关键词检测与定位。我们使用的注意力架构与人工设定中采用的架构相似。不过，我们嵌入的不是英语话语，而是 Yor`ub´a 话语，并将其表示为一系列特征向量。

在 5k 条 Yor`ub´a 话语上训练的跨语言模型，在实际关键词定位任务（检测关键词是否出现，然后对其进行定位）中取得了 16.0% 的精确率。单独来看，这一结果并不突出，但仍处于人工设定中使用 5k 条英语话语训练的单语模型所取得的结果范围（$21.3 \pm 4.4$%）之内。我们一方面认为，跨语言环境更具挑战性，因为视觉关键词更偏向英语，未能考虑 Yor`ub´a 语言中的文化和语言差异。另一方面，YFACC 相对容易，因为它是一个单说话人数据集；而我们使用的英语数据集——Flickr 音频描述语料库（FACC）——则包含多个说话人。

我们对跨语言关键词任务进行了详细分析，重点考察训练数据量、语音表征、单个关键词表现以及任务解耦的影响。此外，我们还对跨语言基于注意力的视觉语音模型进行了定性解读。下面总结我们的研究发现。

训练数据量。对于跨语言环境，我们只能使用 YFACC 中可用的 5k 条训练话语。为了了解模型在使用更多数据训练时的表现，我们根据使用 FACC 的人工设定，推算了数据量增加时的性能。我们发现，数据量增加到六倍后，训练效果显著提升。然而，

我们得出结论，在真实低资源设置中，扩大训练数据规模在实践中可能并不可行。

语音表征。 在真实低资源设置中，我们面临两项约束：(i) 在低资源条件下从头训练模型具有挑战性，部分原因是缺乏足够的数据；以及 (ii) 在这种环境下，收集更多数据可能是不可能的。因此，我们遵循一种常见做法，通过从预训练模型开始，然后在下游任务上对其进行微调，来缓解这些约束。具体而言，我们以一个在 FACC 数据集中所有图像–英语音频对上训练的视觉接地语音模型——一个更大的数据集——来初始化约鲁巴语的训练。这带来了显著的性能提升。

单个关键词表现。 跨语言模型在系统词汇表中的 67 个关键词上的性能差异很大。我们观察到，在实际关键词定位任务中，有些关键词具有良好的精确率（高于平均水平）。然而，模型在许多其他关键词上表现困难。后者的原因包括较差的视觉定位（例如 相机、穿着、大型），以及对语义相关概念的混淆（例如 骑行 经常检索到 $k\grave{}e.k\acute{}e.$，即 自行车；游泳 检索到 $od\grave{}o\ ad\acute{}ag\acute{}un$，即 泳池）。我们得出结论，密切关注单个关键词的特殊性，可以为未来的改进提供启发。

6.2.3. 研究贡献

本学位论文旨在回答两个研究问题：（i）视觉接地语音模型能否实现关键词定位？（ii）在真实低资源设置中，我们能否跨语言进行视觉接地关键词定位？为了解决这些问题，我们首先采用人工低资源环境，将被视为资源丰富语言的英语当作低资源语言。随后，我们转向真实低资源场景，在真实低资源语言环境中研究视觉接地语音模型。下面总结了我们为解决这些研究问题所取得的贡献。

研究贡献 1. 我们开发了一种新的视觉接地语音模型，利用注意力进行关键词检测和关键词检出，并将这种方法与现有的基于视觉接地语音的方法进行了全面比较。

研究贡献 2. 我们通过提出四种定位方法，扩展了视觉接地语音模型，使其能够进行关键词定位。

研究贡献 3。 我们进行了详细的定量和定性分析，揭示了扩展视觉接地语音模型的局限性，并展示了其成功案例和失败模式。

本研究是在将英语视觉语音接地数据人为视为低资源语言的设定下进行的。

研究贡献 4。 我们发布了一个新的多模态多语言数据集，可用于在类似语言记录的真实低资源环境中进行视觉接地语音建模。

研究贡献 5。 我们构建了一个基线系统，用于在真实低资源环境中进行跨语言关键词检测和关键词定位。该基线系统使用 5k 条与英语软文本标签配对的 Yor`ub´a 话语进行训练，生成了一个语音网络，能够检测并定位 Yor`ub´a 语音中的英语关键词。

研究贡献 6。 我们对上述跨语言视觉接地语音模型进行了多方面分析。这些分析有助于更好地理解系统词汇表中的 67 个关键词在关键词定位任务上的性能差异，并帮助解释跨语言模型的失败模式。

6.3. 未来工作

关于模态的选择。 本学位论文所考察的数据集不同于视频数据集，因为其中不包含视频模态，只包含静态图像。将本学位论文中开发的方法扩展到视频处理领域，或许是未来工作中一个有趣的方向。

关于视觉标注器的选择。 我们考虑使用一种成熟的现成视觉标注器，从图像模态中提取软标签。或许有更先进的视觉标注器，更适合本学位论文所探讨的关键词检出和定位任务。

迈向开放词汇表。 我们的视觉接地语音模型经过训练，可以检出并定位一组预定义的关键词。因此，目前尚不清楚这些模型在开放词汇表情境下会如何表现；在这种情境中，模型并未经过专门训练来识别预定义的关键词。最近已有一些朝这一方向开展的研究，例如 [191] 和 [192]。

超越基于点的定位。我们研究了基于点的定位任务：如果关键词预测的位置落在真实对齐标注范围内，就认为该关键词定位正确。这足以满足前文所述的应用场景，但研究如何预测话语中关键词的完整时长，即预测给定关键词的起点和终点，也很有意义。近期也有一些朝这一方向开展的研究 [193]。

评估视觉接地语音系统的语义关键词定位。我们通过定性分析表明，许多定位错误是由对语义相关词的定位造成的。未来工作可以探究这样一种评估方式：不仅评估模型能否提出目标关键词的确切位置，也评估其能否提出语义相关关键词的位置。若要在真实低资源设置中评估这一点，一种做法是为我们的 YFACC 数据集添加语义标注。

6.4. 结论

本学位论文在 VGS 建模的三个主要重点领域作出了贡献：(i) 研究方法、(ii) 数据集和 (iii) 分析。

在研究方法 (i) 方面，我们扩展了现有的基于 VGS 的关键词检测方法，使其也能够定位给定的书面关键词。具体而言，我们采用了一种 VGS 方法，将软文本标签作为目标标注到训练图像上，并将其用于语音网络，使该网络能够检测话语中是否出现关键词。我们为这类 VGS 模型赋予了定位能力，并研究了四种方法。表现最佳的方法依赖输入屏蔽：在不同位置屏蔽输入信号，并测量特定关键词对应输出单元的变化。在一种将英语视觉语音接地数据视为低资源数据的人工低资源环境中，该方法在预言机关键词定位任务上取得了 57.0% 的定位准确率；在这项任务中，我们假定系统已知话语中出现了关键词，只需预测其位置。注意力定位方法位居第二；该方法沿时间轴对特征进行池化，生成作为定位分数的注意力权重，其预言机定位准确率为 46.0%。在真实低资源设置中，我们使用了这种注意力定位方法，但嵌入的不是英语话语，而是 Yor`ub´a 语话语，并将其转换为特征向量序列。基于注意力的定位方法在实际关键词定位（检测关键词是否出现，然后定位该关键词）任务中，使用从头训练的模型时精确率为 16.0%，使用从在更大数据集上训练的模型初始化的模型时精确率为 22.8%。在 VGS 模型分析 (iii) 方面，我们系统地比较了四种不同的关键词定位方法，并通过详尽的实验揭示了架构选择的影响以及模型在单个关键词层面的性能。此外，我们还提供了定性分析，展示 VGS 模型在关键词检测和定位方面的成功之处与失败模式。

在数据集方面（ii），我们推出了一个新的单说话人数据集，其中包含约鲁巴语对 Flickr 图像的口语描述，名为 YFACC。为了展示在真实低资源设置下，如何利用这样一个小型数据集（13.3 小时、6k 个语音–图像对）开展研究，我们使用这些数据，借助一个未经任何转录音频训练的视觉接地语音模型进行跨语言关键词定位。

我们相信，我们的新数据集以及本文介绍的研究发现，将推动更多关于将视觉接地语音模型用于真实低资源语言的研究。

参考文献

[1] D. Harwath 和 J. Glass，“用于语音和图像的深度多模态语义嵌入”，发表于*Proc. ASRU*，2015年。[2] H. Kamper、S. Settle、G. Shakhnarovich 和 K. Livescu，“从未转写语音中进行视觉接地的关键词预测学习”，发表于*Proc. Interspeech*，2017年。[3] D. Harwath、A. Torralba 和 J. Glass，“结合视觉上下文的口语无监督学习”，发表于*Proc. NeurIPS*，2016年。[4] D. Harwath、A. Recasens、D. Sur´ıs、G. Chuang、A. Torralba 和 J. Glass，“从原始感官输入中联合发现视觉对象和口语词语”，发表于*Proc. ECCV*，2018年。[5] J. Gou、B. Yu、S. J. Maybank 和 D. Tao，“知识蒸馏：综述”，*IJCV*，2021年。[6] T. Afouras、J. S. Chung、A. Senior、O. Vinyals 和 A. Zisserman，“深度视听语音识别”，*IEEE PAMI*，2018年。[7] N. C. Tamer 和 M. Sara¸clar，“手语关键词搜索”，发表于*Proc. ICASSP*，2020年。[8] L. Besacier、E. Barnard、A. Karpov 和 T. Schultz，“资源匮乏语言的自动语音识别：综述”，*Speech Commun.*，2014年。[9] L. Duong、A. Anastasopoulos、D. Chiang、S. Bird 和 T. Cohn，“无需转录的语音翻译注意力模型”，发表于*Proc. NAACL-HLT*，2016年。[10] D. Palaz、G. Synnaeve 和 R. Collobert，“利用卷积网络联合学习词语定位与分类”，发表于*Proc. Interspeech*，2016年。[11] S. Settle、K. Levin、H. Kamper 和 K. Livescu，“利用判别式神经声学词嵌入进行基于示例查询的搜索”，发表于*Proc. Interspeech*，2017年。[12] R. J. Weiss、J. Chorowski、N. Jaitly、Y. Wu 和 Z. Chen，“序列到序列模型可直接翻译外语语音”，发表于*Proc. Interspeech*，2017年。[13] J. Driesen、H. Van hamme 和 W. B. Kleijn，“通过输入空间缩放增强的非负矩阵分解，从图像和语音中学习”，发表于*Proc.IEEE SLT*，2010年。

[14] G. Synnaeve、M. Versteegh 和 E. Dupoux，“从图像和语音中学习词语”，*NeurIPS Workshop Learn*，2014年。[15] D. Harwath 和 J. Glass，“通过联合音视分析学习类词单元”，发表于*Proc. ACL*，2017年。[16] D. Harwath、G. Chuang 和 J. Glass，“视觉作为中间语：学习未转写语音的多语言语义嵌入”，发表于*Proc. ICASSP*，2018年。[17] R. Eloff、H. A. Engelbrecht 和 H. Kamper，“语音和图像的多模态单样本学习”，发表于*Proc. ICASSP*，2019年。[18] D. Harwath 和 J. Glass，“迈向视觉接地的子词语音单元发现”，发表于*Proc. ICASSP*，2019年。[19] D. Harwath、W.-N. Hsu 和 J. Glass，“从视觉扎根语音中学习分层离散语言单元”，发表于*Proc. ICLR*，2019年。[20] P. C. Bomba 和 E. R. Siqueland，“婴儿形式类别的性质与结构”，*J. Exp. Child Psychol.*，1983年。[21] S. Pinker，《语言本能》，*Harper Perennial, New York*，1994年。[22] P. D. Eimas 和 P. C. Quinn，“低龄婴儿基于知觉的基本层次类别形成研究”，*Child Develop.*，1994年。[23] D. Roy，“扎根式口语习得：词汇习得实验”，*IEEE Trans. Multimedia*，2003年。[24] L. Boves、L. ten Bosch 和 R. Moore，“ACORNS——迈向交流与识别技能的计算建模”，发表于*Proc. ICCI*，2007年。[25] L. Gelderloos 和 G. Chrupa la，“从音素到图像：视觉扎根语言学习循环神经模型中的表征层级”，发表于*Proc. COLING*，2016年。[26] O. R¨as¨anen 和 K. Khorrami，“基于低龄婴儿视听经验的早期语言习得计算模型”，发表于*Proc. Interspeech*，2019年。[27] M. Sun 和 H. Van hamme，“非负Tucker分解与离散密度隐马尔可夫模型的联合训练”，*Comput. Speech Lang.*，2013年。[28] L. Nortje 和 H. Kamper，“语音和图像多模态单样本匹配的无监督学习与迁移学习”，发表于*Proc. Interspeech*，2020年。

[29] V. De Sa 和 D. Ballard，“通过多模态感知进行类别学习”，*NeuralComput.*，1998年。[30] O. Scharenborg、L. Besacier、A. Black、M. Hasegawa-Johnson、F. Metze、G. Neubig、S. St¨uker、P. Godard、M. M¨uller、L. Ondel、S. Palaskar、P. Arthur、F. Ciannella、M. Du、E. Larsen、D. Merkx、R. Riad、L. Wang 和 E. Dupoux，“无文字语言中从多模态输入发现语言单位：‘Speaking Rosetta’JSALT 2017 研讨会总结”，载于*Proc. ICASSP*，2018年。[31] F. L¨upke，“语言记录中的研究方法”，*Lang. Documentationand Description*，2010年。[32] S. Bird，“稀疏转写”，*Comput. Linguistics*，2020年。[33] O. Scharenborg、L. Besacier、A. Black、M. Hasegawa-Johnson、F. Metze、G. Neubig、S. St¨uker、P. Godard、M. M¨uller、L. Ondel、S. Palaskar、P. Arthur、F. Ciannella、M. Du、E. Larsen、D. Merkx、R. Riad、L. Wang 和 E. Dupoux，“面向无文字语言的语音技术”，*IEEE/ACM TASLP*，2020年。[34] X. Wang、T. Qiao、J. Zhu、A. Hanjalic 和 O. Scharenborg，“根据口头描述生成图像”，*IEEE/ACM TASLP*，2021年。[35] H. Kamper 和 M. Roth，“语音中视觉接地的跨语言关键词检索”，载于*Proc. SLTU*，2018年。[36] G. Chrupa la、L. Gelderloos 和 A. Alishahi，“视觉接地语音信号模型中的语言表征”，载于*Proc. ACL*，2017年。[37] P. Peng 和 D. Harwath，“用于语音视觉接地的快速-慢速变换器”，*Proc. ICASSP*，2021年。[38] A. Karpathy、A. Joulin 和 L. F. Fei-Fei，“用于双向图像-句子映射的深度片段嵌入”，载于*Proc. NeurIPS*，2014年。[39] H. Kamper、G. Shakhnarovich 和 K. Livescu，“利用未转写语音的视觉接地模型进行语义语音检索”，*IEEE/ACM Trans. Acoust.,Speech, Signal Process.*，2019年。[40] H. Kamper、A. Anastassiou 和 K. Livescu，“使用视觉定位进行语义示例查询语音搜索”，载于*Proc. ICASSP*，2019年。[41] E. Azuh、D. Harwath 和 J. R. Glass，“探索从视觉接地语音音频中发现双语词典”，载于*Proc. Interspeech*，2019年。

[42] Y. Ohishi、A. Kimura、T. Kawanishi、K. Kashino、D. Harwath 和 J. Glass，“利用自注意力机制实现视觉接地语音的三语语义嵌入”，载于*Proc. ICASSP*，2020年。[43] Y. Ohishi、A. Kimura、T. Kawanishi、K. Kashino、D. Harwath 和 J. R. Glass，“利用不相交的视觉接地语音音频数据集，通过配对扩展学习多语言语义嵌入”，载于*Proc. Interspeech*，2020年。[44] J. Ngiam、A. Khosla、M. Kim、J. Nam、H. Lee 和 A. Y. Ng，“多模态深度学习”，载于*Proc. ICML*，2011年。[45] J. Huang 和 B. Kingsbury，“用于噪声鲁棒语音识别的视听深度学习”，载于*Proc. ICASSP*，2013年。[46] Y. Mroueh、E. Marcheret 和 V. Goel，“用于视听语音识别的深度多模态学习”，载于*Proc. ICASSP*，2015年。[47] S. Petridis、T. Stafylakis、P. Ma、F. Cai、G. Tzimiropoulos 和 M. Pantic，“端到端视听语音识别”，载于*Proc. ICASSP*，2018年。[48] C. Rashtchian、P. Young、M. Hodosh 和 J. Hockenmaier，“使用亚马逊土耳其机器人收集图像标注”，*NAACL HLT Workshop*，2010年。[49] M. Hodosh、P. Young 和 J. Hockenmaier，“将图像描述构建为排序任务：数据、模型和评估指标（扩展摘要）”，载于***Proc.*** 国际人工智能联合会议，2015年。[50] M. Hasegawa-Johnson、A. Black、L. Ondel、O. Scharenborg 和 F. Ciannella，“Image2speech：自动生成图像的音频描述”，载于*Proc.ICNLSP*，2017年。[51] J. Li、X. Zhang、C. Jia、J. Xu、L. Zhang、Y. Wang、S. Ma 和 W. Gao，“直接语音到图像翻译”，*IEEE JSTSP*，2020年。[52] B. Zhou、A. Lapedriza、J. Xiao、A. Torralba 和 A. Oliva，“利用Places数据库学习用于场景识别的深层特征”，载于*Proc. NeurIPS*，2014年。[53] P. Sheridan、M. Wechsler 和 P. Sch¨auble，“跨语言语音检索：建立基线性能”，载于*Proc. SIGIR*，1997年。[54] M. Oquab、L. Bottou、I. Laptev 和 J. Sivic，“目标定位是否免费？利用卷积神经网络进行弱监督学习”，载于*Proc. CVPR*，2015年。

[55] R. Selvaraju、M. Cogswell、A. Das、R. Vedantam、D. Parikh 和 D. Batra，《Grad-CAM：通过基于梯度的定位解释深度网络的视觉结果》，载于国际计算机视觉会议论文集，2017年。[56] D. van Niekerk、E. Barnard、O. Giwa 和 A. Sosimi，《拉各斯-西北大学约鲁巴语语音语料库》，https://repo.sadilar.org/handle/20.500.12185/431，2015年。[57] A. Gutkin、I. Demirsahin、O. Kjartansson、C. Rivera 和 K. T´ubos´un，《构建约鲁巴语语音开源语料库》，载于*Interspeech*，2020年。.[58] J. Meyer、D. I. Adelani、E. Casanova、A. Oktem、D. W. J. Weber、S. Kabongo、E. Salesky、I. Orife、C. Leong、P. Ogayo等，《BibleTTS：大型、高保真、多语言且独具非洲特色的语音语料库》，载于*Interspeech* 论文集，2022年。[59] P. Ogayo、G. Neubig 和 A. W. Black，《构建非洲语音》，*arXiv* 预印本 *arXiv:2207.00688*，2022年。[60] G. Chrupa la，《视觉接地的口语模型：数据集、架构和评估技术综述》，*JAIR*，2022年。[61] Y. Yoshikawa、Y. Shigeto 和 A. Takeuchi，《STAIR 图像描述：构建大型日语图像描述数据集》，载于*ACL* 论文集，2017年。[62] W. N. Havard、L. Besacier 和 O. Rosec，《Speech-COCO：与 MSCOCO 数据集对齐的 60 万条视觉关联语音描述》，*GLU*（*Interspeech*）研讨会，2017年。[63] G. Ilharco、Y. Zhang 和 J. Baldridge，《从视觉接地的未转写语音中进行大规模表征学习》，*arXiv* 预印本 *arXiv:1909.08782*，2019年。[64] W.-N. Hsu、D. Harwath、C. Song 和 J. Glass，《利用学习到的分段单元进行无文本的图像到语音合成》，*arXiv* 预印本 *arXiv:2012.15454*，2020年。[65] M. Regneri、M. Rohrbach、D. Wetzel、S. Thater、B. Schiele 和 M. Pinkal，《将动作描述关联到视频》，计算语言学协会汇刊，2013年。[66] P. Das、C. Xu、R. F. Doell 和 J. J. Corso，《寥寥数语描述千帧：通过潜在主题和稀疏对象拼接对视频进行语言描述》，载于*CVPR* 论文集，2013年。[67] A. Rohrbach、M. Rohrbach、W. Qiu、A. Friedrich、M. Pinkal 和 B. Schiele，《以不同细节层次进行连贯的多句视频描述》，载于*GCPR* 论文集，2014年。[68] M. Rohrbach、A. Rohrbach、M. Regneri、S. Amin、M. Andriluka、M. Pinkal 和 B. Schiele，《利用以手部为中心的特征和脚本数据识别细粒度复合活动》，*IJCV*，2016年。

[69] D. Damen、H. Doughty、G. M. Farinella、S. Fidler、A. Furnari、E. Kazakos、D. Molti-santi、J. Munro、T. Perrett、W. Price等，《扩展自我中心视觉：EPIC-KITCHENS 数据集》，载于*ECCV* 论文集，2018年。[70] L. Zhou、C. Xu 和 J. J. Corso，《从网络教学视频中自动学习操作流程》，载于*AAAI* 论文集，2018年。[71] D. Damen、H. Doughty、G. M. Farinella、A. Furnari、E. Kazakos、J. Ma、D. Moltisanti、J. Munro、T. Perrett、W. Price等，《重新扩展自我中心视觉》，*arXiv* 预印本 *arXiv:2006.13256*，2020年。[72] K. Soomro、A. R. Zamir 和 M. Shah，《UCF101：包含 101 类野外视频中人类动作的数据集》，*arXiv* 预印本 *arXiv:1212.0402*，2012年。[73] H. Idrees、A. R. Zamir、Y.-G. Jiang、A. Gorban、I. Laptev、R. Sukthankar 和 M. Shah，《面向“野外视频”动作识别的 THUMOS 挑战赛》，计算机视觉与图像理解，2017年。[74] A. Karpathy、G. Toderici、S. Shetty、T. Leung、R. Sukthankar 和 L. Fei-Fei，《利用卷积神经网络进行大规模视频分类》，载于*CVPR* 论文集，2014年。[75] A. Rohrbach、M. Rohrbach、N. Tandon 和 B. Schiele，《电影描述数据集》，载于*CVPR* 论文集，2015年。[76] A. Rohrbach、A. Torabi、M. Rohrbach、N. Tandon、C. Pal、H. Larochelle、A. Courville 和 B. Schiele，《电影描述》，*IJCV*，2017年。[77] A. Miech、D. Zhukov、J. Alayrac、M. Tapaswi、I. Laptev 和 J. Sivic，《HowTo100M：通过观看一亿个带旁白的视频片段学习文本-视频嵌入》，载于国际计算机视觉会议论文集，2019年。[78] D. Chen 和 W. B. Dolan，《收集用于释义评估的高度平行数据》，载于计算语言学协会人类语言技术会议论文集，2011年。[79] J. Xu、T. Mei、T. Yao 和 Y. Rui，《MSR-VTT：用于连接视频与语言的大型视频描述数据集》，载于*CVPR* 论文集，2016年。[80] S. Gella、M. Lewis 和 M. Rohrbach，《讲述社交媒体视频故事的数据集》，载于*EMNLP* 论文集，2018年。[81] R. Krishna、K. Hata、F. Ren、L. Fei-Fei 和 J. Carlos Niebles，《视频中的密集事件描述》，载于国际计算机视觉会议论文集，2017年。

[82] M. Monfort, S. Jin, A. Liu, D. Harwath, R. Feris, J. Glass, 和 A. Oliva, 《口语片段：从视频描述中学习联合音视表示》，载于*CVPR* 会议论文集，2021年。[83] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma 等, 《视觉基因组：利用众包密集图像标注连接语言与视觉》，*IJCV*，2017年。[84] X. Wang, J. Wu, J. Chen, L. Li, Y.-F. Wang, 和 W. Y. Wang, 《VATEX：用于视频与语言研究的大规模、高质量多语言数据集》，载于国际计算机视觉会议论文集，2019年。[85] T. Baltrušaitis, C. Ahuja, 和 L.-P. Morency, 《多模态机器学习：综述与分类体系》，*IEEE PAMI*，2018年。[86] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, 和 D. Parikh, 《VQA：视觉问答》，载于国际计算机视觉会议论文集，2015年。[87] A. Karpathy 和 L. Fei-Fei, 《用于生成图像描述的深度视觉语义对齐》，载于*CVPR* 会议论文集，2015年。[88] Z. Wu, Y.-G. Jiang, J. Wang, J. Pu, 和 X. Xue, 《利用深度神经网络探索特征间和类别间关系以进行视频分类》，载于国际多媒体会议论文集，2014年。[89] J. Weston, S. Bengio, 和 N. Usunier, 《网络规模图像标注：利用联合词图嵌入学习排序》，载于欧洲机器学习会议论文集，2010年。[90] ——, 《WSABIE：扩展至大词汇量图像标注》，载于国际人工智能联合会议论文集，2011年。[91] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, 和 T. Mikolov, 《DeViSE：一种深度视觉语义嵌入模型》，载于*NeurIPS* 会议论文集，2013年。[92] R. Kiros, R. Salakhutdinov, 和 R. S. Zemel, 《通过多模态神经语言模型统一视觉语义嵌入》，*arXiv* 预印本 *arXiv:1411.2539*，2014年。[93] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, 和 A. Y. Manning, 《用于通过句子查找和描述图像的基于视觉的组合语义》，计算语言学协会汇刊，2013年。[94] Y. Pan, T. Mei, T. Yao, H. Li, 和 Y. Rui, 《联合建模嵌入与翻译以衔接视频和语言》，载于*CVPR* 会议论文集，2016年。

[95] R. Xu, C. Xiong, W. Chen, 和 J. J. Corso, 《在统一框架中联合建模深度视频与组合文本，以衔接视觉和语言》，载于*AAAI* 会议论文集，2015年。[96] G. Hinton, O. Vinyals, 和 J. Dean, 《蒸馏神经网络中的知识》，*arXiv* 预印本 *arXiv:1503.02531*，2015年。[97] L. J. Ba 和 R. Caruana, 《深度网络真的需要很深吗？》，载于*NeurIPS* 会议论文集，2013年。[98] S. Gupta, J. Hoffman, 和 J. Malik, 《用于监督迁移的跨模态蒸馏》，载于*CVPR* 会议论文集，2016年。[99] M. Zhao, T. Li, M. Abu Alsheikh, Y. Tian, H. Zhao, A. Torralba, 和 D. Katabi, 《利用无线电信号进行隔墙人体姿态估计》，载于*CVPR* 会议论文集，2018年。[100] Y. Aytar, C. Vondrick, 和 A. Torralba, 《SoundNet：从未标注视频中学习声音表征》，载于 *NeurIPS* 会议论文集，2016年。[101] S. Albanie, A. Nagrani, A. Vedaldi, 和 A. Zisserman, 《利用野外跨模态迁移进行语音情感识别》，载于*ACM Multimedia* 会议论文集，2018年。[102] F. M. Thoker 和 J. Gall, 《用于动作识别的跨模态知识蒸馏》，载于*ICIP* 会议论文集，2019年。[103] N. C. Garcia, P. Morerio, 和 V. Murino, 《利用多流网络进行动作识别的模态蒸馏》，载于*ECCV* 会议论文集，2018年。[104] Y. Tian, D. Krishnan, 和 P. Isola, 《对比表征蒸馏》，载于*ICLR* 会议论文集，2019年。[105] A. H. Owens, 《从配对的音视示例中学习视觉模型》，博士学位论文，麻省理工学院，2016年。[106] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, 和 A. Torralba, 《从声音中学习视觉：环境声音为视觉学习提供监督》，$IJCV$，2018年。[107] H. Alwassel, D. Mahajan, B. Korbar, L. Torresani, B. Ghanem, 和 D. Tran, 《通过跨模态音视频聚类进行自监督学习》，载于*NeurIPS* 会议论文集，2020年。[108] J. Drexler 和 J. Glass, 《无监督语音识别中的视听特征分析》，载于*GLU* 会议论文集，2017年。

[109] W. Havard, J. Chevrot, and L. Besacier, “视觉接地语音模型中的词语识别、竞争与激活”，载于 *CoNLL* 论文集，2019年。[110] S. Scholten, D. Merkx, and O. Scharenborg, “利用视觉接地语音学习识别词语”，载于 *ISCAS* 论文集，2021年。[111] W. Havard, J. Chevrot, and L. Besacier, “视觉接地语音信号模型会关注名词：一项关于英语和日语的双语实验”，载于 *ICASSP* 论文集，2019年。[112] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet：大型分层图像数据库”，载于*CVPR* 论文集，2009年。[113] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, “SUN 数据库：从修道院到动物园的大规模场景识别”，载于 *CVPR* 论文集，2010年。[114] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, “Microsoft COCO：上下文中的常见物体”，载于 *ECCV* 论文集，2014年。[115] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov 等，“Open Images 数据集 V4”，*IJCV*，2020年。[116] K. He, X. Zhang, S. Ren, and J. Sun, “用于图像识别的深度残差学习”，载于 *CVPR* 论文集，2016年。[117] M. Tan and Q. Le, “EfficientNet：重新思考卷积神经网络的模型缩放”，载于 *ICML* 论文集，2019年。[118] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly等，“一张图像胜过 16x16 个单词：用于大规模图像识别的 Transformer”，*arXiv* 预印本 *arXiv:2010.11929*，2021年。[119] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, D. Keysers, J. Uszkoreit, M. Lucic 等，“MLP-Mixer：一种用于视觉的全 MLP 架构”，载于 *NeurIPS* 论文集，2021年。[120] A. Pasad, B. Shi, H. Kamper, and K. Livescu, “视觉和文本监督对低资源语义语音检索的贡献”，载于 *Interspeech* 论文集，2019年。[121] H. McGurk and J. MacDonald, “听见嘴唇，看见声音”，《自然》，1976年。

[122] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, “利用神经网络整合声学和视觉语音信号”，*IEEE* 通信杂志，1989年。[123] W. H. Sumby and I. Pollack, “噪声环境下视觉对语音可懂度的贡献”，美国声学学会期刊，1954年。[124] E. D. Petajan, “通过自动唇读增强语音识别”，载于 全球电信会议论文集，1984年。[125] E. Petajan, B. Bischoff, D. Bodoff, and N. M. Brooke, “一种改进的自动唇读系统，用于增强语音识别”，载于 人因计算机系统会议论文集，1988年。[126] P. Ma, S. Petridis, and M. Pantic, “基于 Conformer 的端到端视听语音识别”，载于 *ICASSP* 论文集，2021年。[127] K. Prajwal, L. Momeni, T. Afouras, and A. Zisserman, “利用注意力进行视觉关键词检测”，载于 英国机器视觉会议论文集，2021年。[128] C. Yu and L. B. Smith, “通过跨情境统计在不确定性下快速习得词汇”，心理科学，2007年。[129] E. D. Thiessen, “视觉信息对成年人和婴儿听觉统计学习的影响”，认知科学，2010年。[130] T. Cunillera, E. Camara, M. Laine, and A. Rodriguez-Fornells, “视觉信息对成年人和婴儿听觉统计学习的影响”，实验心理学季刊，2010年。[131] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “从图像描述到视觉指称：用于事件描述语义推理的新型相似度指标”，计算语言学协会汇刊，2014年。[132] P. Sharma, N. Ding, S. Goodman, and R. Soricut, “Conceptual Captions：用于自动图像描述生成的清洗后、经过上位词处理的图像替代文本数据集”，载于*ACL* 论文集，2018年。[133] C. Liu, J. Mao, F. Sha, and A. Yuille, “神经图像描述生成中的注意力正确性”，载于 *AAAI* 论文集，2017年。[134] J. Lu, J. Yang, D. Batra, and D. Parikh, “Neural Baby Talk”，载于 *CVPR* 论文集，2018年。[135] A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, “图像描述生成中的对象幻觉”，载于 *EMNLP* 论文集，2018年。

[136] R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, D. Batra, 和 D. Parikh, 《提供提示：利用解释使视觉与语言模型更加接地》，载于*Proc.* 国际计算机视觉会议，2019年。[137] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, 和 S. Lazebnik, 《Flickr30k实体：收集区域到短语对应关系，以构建更丰富的图像到句子模型》，*IJCV*，2015年。[138] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, 和 V. Ferrari, 《通过定位叙事连接视觉与语言》，载于*Proc. ECCV*，2020年。[139] G. Varol, L. Momeni, S. Albanie, T. Afouras, 和 A. Zisserman, 《借助手语词典扩大手语识别的规模》，*IJCV*，2022年。[140] J. Wilpon, L. Rabiner, C.-H. Lee, 和 E. Goldman, 《使用隐马尔可夫模型自动识别非受限语音中的关键词》，*IEEE/ACM Trans.Acoust. Speech Signal Process.*，1990年。[141] I. Szoke, P. Schwarz, P. Matejka, L. Burget, M. Karafi´at, M. Fapso, 和 J. Cernocky, 《非正式连续语音关键词检出方法比较》，载于*Proc. Interspeech*，2005年。[142] A. Garcia 和 H. Gish, 《利用最少语音资源检出任意单词》，载于*Proc. ICASSP*，2006年。[143] C. Chelba, T. J. Hazen, 和 M. Saraclar, 《口语内容的检索与浏览》，*IEEE Signal Proc. Mag.*，2008年。[144] H.-Y. Lee, T.-H. Wen, 和 L.-S. Lee, 《通过声学相似性图增强语言模型，改进口语内容的语义检索》，载于*Proc.* 口语语言技术研讨会，2012年。[145] Y.-C. Li, H.-y. Lee, C.-T. Chung, C.-a. Chan, 和 L.-s. Lee, 《利用基于自动发现的声学模式的查询扩展，实现口语内容的无监督语义检索》，载于*Proc. ASRU*，2013年。[146] L.-s. Lee, J. Glass, H.-y. Lee, 和 C.-a. Chan, 《口语内容检索——超越级联式语音识别与文本检索》，*IEEE/ACM Trans. Acoust., Speech,Signal Process.*，2015年。[147] ´E. L. Ferrand, S. Bird, 和 L. Besacier, 《在原住民社区实现交互式转录》，载于*Proc. COLING*，2020年。[148] E. van der Westhuizen, H. Kamper, R. Menon, J. Quinn, 和 T. Niesler, 《面向低资源语言的高效无自动语音识别关键词检出的特征学习》，*Comput.Speech Lang.*，2022年。

[149] H. Kamper, A. Jansen, 和 S. Goldwater, 《使用分段贝叶斯模型进行完全无监督的小词汇量语音识别》, 载于*Proc. Interspeech*, 2015年。[150] H. Kamper, M. Elsner, A. Jansen, 和 S. Goldwater, 《利用弱自顶向下约束进行无监督神经网络特征提取》, 载于*Proc. ICASSP*, 2015年。[151] D. Renshaw, H. Kamper, A. Jansen, 和 S. Goldwater, 《零资源语音挑战赛中无监督表征学习神经网络方法的比较》, 载于*Proc. Interspeech*, 2015年。[152] K. Olaleye, D. Oneata, 和 H. Kamper, 《使用视觉接地语音模型进行未转录语音中的关键词定位》, *IEEE JSTSP*, 2022年。[153] J. Amores, 《多示例分类：综述、分类体系与比较研究》, 人工智能, 2013年。[154] M.-A. Carbonneau, V. Cheplygina, E. Granger, 和 G. Gagnon, 《多实例学习：问题特征与应用综述》, 模式识别, 2018年。[155]J. Choe, S. J. Oh, S. Chun, Z. Akata, 和 H. Shim, 《弱监督目标定位的评估：协议、度量指标与数据集》, *arXiv*预印本 *arXiv:2007.04178*, 2020年。[156] D. Zhang, J. Han, G. Cheng, 和 M.-H. Yang, 《弱监督目标定位与检测：综述》, *PAMI*, 2021年。[157] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, 和 A. Torralba, 《学习用于判别式定位的深层特征》, 载于*Proc. CVPR*, 2016年。[158] A. J. Bency, H. Kwon, H. Lee, S. Karthikeyan, 和 B. Manjunath, 《使用深层特征图进行弱监督定位》, 载于*Proc. ECCV*, 2016年。[159] M. D. Zeiler 和 R. Fergus, 《卷积网络的可视化与理解》, 载于*Proc. ECCV*, 2014年。[160] L. Bazzani, A. Bergamo, D. Anguelov, 和 L. Torresani, 《利用深度网络进行自监督目标定-位》, 载于*Proc. WACV*, 2016年。[161] D. W. Oard 和 A. R. Diekema, 《跨语言信息检索》。*ARIST*, 1998年。[162] S. Bansal, H. Kamper, A. Lopez, 和 S. Goldwater, 《迈向无需语音识别的语音到文本翻译》, 载于*Proc. EACL*, 2017年。

[163] A. B´erard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, “有声读物的端到端自动语音翻译”，载于*Proc. ICASSP*，2018年。[164] D. Hershcovich, S. Frank, H. C. Lent, M. de Lhoneux, M. Abdou, S. Brandl, E. Bugliarello, L. C. Piqueras, I. Chalkidis, R. Cui, C. Fierro, K. Margatina, P. Rust, and A. Sogaard, “跨文化自然语言处理中的挑战与策略”，载于*Proc. ACL*，2022年。[165] F. Gao，“日语：一种承载浓厚文化内涵的语言”，*J. Intercultural Commun.*，2005年。[166] T. Larina，“文化特定的交际风格：解读语言和文化特质的框架”，*Int. Rev. Pragmatics*，2015年。[167] J. House and G. Kasper，英语和德语中的礼貌标记。会话惯例，2011年。[168] A. G. Ryder, J. Yang, X. Zhu, S. Yao, J. Yi, S. J. Heine, and R. M. Bagby，“抑郁症的文化塑造：中国的躯体症状与北美的心理症状？”，*J. Abnormal Psychol.*，2008年。[169] L. J. Kirmayer，“抑郁症和焦虑症临床表现的文化变异：对诊断和治疗的启示。”*J. Clin. Psychiatry*，2001年。[170] S. Hareli, K. Kafetsios, and U. Hess，“关于情绪表达与社会规范习得的跨文化研究”，*Frontiers Psychol.*，2015年。[171] K. Loveys, J. Torrez, A. Fine, G. Moriarty, and G. Coppersmith，“网络抑郁症语言标记的跨文化差异”，载于*Proc. Workshop Comput.Linguistics and Clin. Psychol.*，2018年。[172] D. Elliott, S. Frank, K. Sima’an, and L. Specia，“Multi30k：多语言英语-德语图像描述”，*ACL Vis. and Lang. Workshop*，2016年。[173] X. Li, W. Lan, J. Dong, and H. Liu，“为图像添加中文图注”，载于*Proc.ICMR*，2016年。[174] K. Simonyan and A. Zisserman，“用于大规模图像识别的超深卷积网络”，*arXiv preprint ar Xiv:1409.1556*，2014年。[175] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le，“Specaugment：一种简单的自动语音识别数据增强方法”，载于*Proc. Interspeech*，2019年。

[176] D. Kingma and J. Ba，“Adam：一种随机优化方法”，载于 Proc. ICLR，2015年。[177] R. Sanabria, H. Tang, and S. Goldwater，“利用注意力进行词语切分的难点”，载于 Proc. EMNLP，2021年。[178] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand，“显著性模型接下来应该关注哪里？”，载于 Proc. ECCV，2016年。[179] J. Zhang, Z. L. Lin, J. Brandt, X. Shen, and S. Sclaroff，“通过激励反向传播实现自上而下的神经注意力”，IJCV，2018年。[180] S.-A. Rebuffi, R. Fong, X. Ji, and A. Vedaldi，“往返再探：重新审视反向传播显著性方法”，载于 Proc. CVPR，2020年。[181] A. Graves, G. Wayne, and I. Danihelka，“神经图灵机”，arXiv preprint arXiv:1410.5401，2014年。[182] M.-T. Luong, H. Pham, and C. D. Manning，“基于注意力的神经机器翻译的有效方法”，载于 Proc. EMNLP，2015年。[183] M. Ilse, J. Tomczak, and M. Welling，“基于注意力的深度多示例学习”，载于 Proc. ICML，2018年。[184] P. Viola, M. Jones et al.，“稳健的实时目标检测”，IJCV，2001年。[185] E. van der Westhuizen, H. Kamper, R. Menon, J. Quinn, and T. Niesler，“用于低资源语言高效无自动语音识别关键词检出的特征学习”，Comput. Speech Lang.，2022年。[186] P. Boersma and V. Van Heuven，“用 PRAAT 说与不说”，Glot Int.，2001年。[187] J. B. Wood，《约鲁巴语构建札记》。埃克塞特大学，1879年。[188] I. Orife, D. I. Adelani, T. E. Fasubaa, V. Williamson, W. F. Oyewusi, O. Wa- hab, and K. Tubosun，“改进约鲁巴语变音符号恢复”，arXiv preprint arXiv:2003.10564，2020年。[189] 维基百科，“科恩κ系数”，2022年，在线；访问日期：2022年10月21日。 [在线]。网址：https://en.wikipedia.org/wiki/Cohen%27s kappa[190] K. Olaleye, B. van Niekerk, and H. Kamper，“利用弱监督实现语音关键词定位”，SAS NeurIPS研讨会，2020年。

[191] L. Nortje 和 H. Kamper，《面向零资源口语的视觉提示关键词定位》，载于 *IEEE SLT* 论文集，2023年。[192] Y.-J. Shih、H.-F. Wang、H.-J. Chang、L. Berry、H.-y. Lee 和 D. Harwath，《Speech-clip：将语音与预训练视觉与语言模型相结合》，载于 *IEEE SLT* 论文集，2023年。[193] P. Peng 和 D. Harwath，《视觉接地的自监督语音模型中的词语发现》，载于 *Interspeech* 论文集，2022年。