# 多模态方法：视觉语音识别（唇读）

计算机视觉与自然语言处理的最新交叉点（上）

这是我们最新出版系列文章的第一部分，介绍了计算机视觉（CV）和自然语言处理（NLP）之间的一些交叉点。我们鼓励读者通过我们的最佳体验来观看这篇文章：

第一部分：视觉语音识别（唇读）

第二部分：图像字幕（从翻译到注意）

第三部分：图像字幕（强化学习与超越）

欢迎通过媒体或直接向提供反馈和意见。谢谢你的阅读！

### 介绍

在这一系列的文章中，我们决定研究计算机视觉（CV）和自然语言处理（NLP）之间的相互作用，这是一个与以前以CV为中心的文章“计算机视觉一年”（可用）相吻合的部分。虽然在一个单一领域内的进步常常令人印象深刻，但就其本质而言，知识是可加的和组合的。这些特点意味着一个领域的改进和突破可能会推动其他领域的进一步发展。通常，两个看似不同的知识体结合在一起，将我们的理解、技术和解决方案推向令人兴奋和无法预见的领域。

在我们之前的工作中，我们简要地概述了计算机视觉在智能方面的主张；建立了能够仅从视觉数据中学习、推断和推理世界的系统。在这里，我们希望在这个讨论中再加上一点：语言在创造和创造智力方面起着什么作用？

这个话题一经提出，就一直是语言学家、神经科学家和人工智能研究人员争论的焦点。我们至少可以说，从进化的角度来看，视觉和语言与人类的经验是密不可分的。一种重在学习的经历。例如，当“猫”的概念在你的脑海中被唤起时，几乎在瞬间，围绕着猫的关系有许多不同的联想。例如：

* 普通猫或特定猫的图像
* 抚摸猫柔软皮毛的感觉
* 字母“c”、“a”和“t”
* 有时自私和基本独立的生物

“猫”这个概念的这些基本经验都告诉我们，猫是什么以及它与我们和世界的关系。“猫”的这种知识是可以接受的；它可以通过直接的经验、思考与猫有关的事情或通过任何媒介获得信息来改变。虽然并非所有的经验在回忆时都需要语言，但经验或思想对自己的表达往往是通过语言。

如果计算机视觉能识别出模式，那么NLP的加入也许能增强这一过程。它可以使机器中的过程类似于人们如何将许多模式和经验与前面提到的“猫”概念联系起来。语言的加入也许最终会为机器提供一种手段，让它们在未来对复杂的概念进行分组、推理和表达。

与我们通过大脑输入的任何形式来迭代、链接和更新概念的方式基本相同，深度学习中的多模式方法也开始崭露头角。下面是CV和NLP之间的一些交叉点：

* 唇读 - 输入为可视；输出为文本
* 图像字幕 - 输入为可视；输出为文本
* 可视问答 - 输入为可视文本；输出为文本
* 从标题生成图像- 输入是文本；输出是可视的

在这些领域中，我们希望能提供深入了解唇读和图像字幕在本系列的进展和技术。而基于字幕的视觉问答和图像生成可能是一些未来工作的主题。

在《深度学习：实践与趋势》（NIPS 2017）中，著名研究人员提供了一个简单的抽象概念，即几乎所有的深度学习方法都可以描述为增强架构或损失函数，或者将先前的方法应用于新的输入/输出组合。虽然过于简单化，但当前深度学习方法的普遍性令人印象深刻。正如我们将看到的那样，这些一般性的做法也限制了新的职权范围。

由于这些方法的普遍性以及它们在各个领域中相对最近的成功，还有一些有趣的二阶效应。尽管处理完全不同的输入和输出格式似乎是个麻烦的问题。研究人员现在可以在许多不同的领域工作，并将他们的技术应用到各个领域，从社会科学到医疗保健，从体育到金融。不管应用程序是什么，在架构和损失函数上收集的技巧和知识可能会被重新利用，并在其他地方重新使用。

在我们对知识的组合本质的评论之后，一些研究孤立的部分瓦解，或者鼓励更多的跨学科工作使用人工智能工具和技术。二阶效应意味着CV研究者经常理解NLP技术，反之亦然。关于深度学习的入门课程和书籍涵盖了NLP、CV、强化学习和生成模型中的用例。

从某种意义上说，我们正在接近一种可推广的人工智能；深度学习中的知识正在整合为一种更具范例性的方法。这种一致性使得来自各个学科的研究人员能够以新的、令人兴奋的方式利用人工智能。也许，一个真正的普遍智慧就在前方，尽管在达到这一点之前，有多少范式必须失衡并重新恢复还不清楚。我们所知道的是，当我们观察技术在多个任务、领域和模式中执行时，在通用模型中的工作继续吸引着我们。

为了与上一期出版物保持一致，我们的目标是尽可能让我们的读者能够接触到人工智能，并为个人提供工具，让他们能够在任何他们想要的深度上了解人工智能。然而，在这篇文章中，我们牺牲了大量的篇幅来深入研究这些领域。我们将继续试验范围和时间表，以了解如何最好地向读者传达主题。对于那些缺乏技术能力的人来说，可能会有一些枯燥乏味的短文，但它们的遗漏不会对非专业读者造成很大影响。我们希望，一个人应该能够拿走一些有价值的东西，不管他们的技能如何。

今后几年将进一步深入更多的领域，以越来越快的速度部署更好的技术。理解我们的假设可能是错误的，关于人工智能能做什么和不能做什么，对社会来说是重要的一步。归根结底，这些技术旨在模拟和改进我们在世界各地导航的过程。为了学习他们自己的元结构，我们每天都在潜意识中部署。

如果人类从来没有接受过对我们能力的限制，为什么我们会认为机械化的智力在某种程度上会受到内在的限制？随着新的、不可预见的突破，任何人都能预测技术的长期未来的假设，充其量可能是站不住脚的。最好的策略可能是尽可能多地了解情况，积极参与即将到来的进步。

谢谢你，



### 第一部分：视觉语音识别（唇读）

该小组先前的工作详细介绍了计算机视觉领域的许多进展。在实践中，研究并不是孤立地进行的，基于此，我们对计算机视觉（CV）和自然语言处理（NLP）之间的交叉点进行了简短的探索，即视觉语音识别，也称为唇读。

与计算机视觉领域的进步类似，NLP作为一个领域已经出现了类似的大量涌入和采用深度学习技术的情况，特别是随着单词嵌入和递归神经网络（RNNs）等技术的发展。此外，使用输入组合来解决复杂的跨域问题的驱动力已经产生了许多令人兴奋的事情。令我们兴奋的一个原因是，在准确度排名中，我们看到了唇读技能从人类的主导地位转变为机器的主导地位。另一个仍然来自于实现这一点的方法。

就在不久前，唇读被认为是一个困难的问题，很像围棋游戏所带来的困难；尽管它并不那么出名。除了解决这个问题外，唇读的进步还可能带来一些新的应用。例如，在嘈杂的环境中听写信息，更好地处理多个同时发言的人，以及总体上提高语音识别系统的性能。相反，仅从视频中提取对话可能是未来关注的一个领域。

人们希望，我们对这个利基应用的关注既能说明问题，又能提供信息。在唇读方面，相对较小的深度学习工作足以打乱受过专业训练的唇读者的传统优势。同时，人工智能研究的组合性质和这些进步的核心技术以闪烁的方式融合了场之间的界限。在哪里，如果有，这样的进步高原是每个人的嘴唇上的问题。

### 提出问题

众所周知，预测技术创新和进步的任务相当困难，最好留给同事和朋友之间的小赌注。在进行评估时，人们通常会将机器的性能与人类已经擅长的任务进行比较，例如行走、写作、运动等。在唇读方面，我们学到了两件事，这让我们感到惊讶。首先，这些机器最近成功地超越了专家级的人类，其次，专家级的人类一开始就不那么准确。

不管专家设置了什么样的标准，我们认为最好深入研究是什么使这成为一个难以掌握的挑战。虎钳，类似于构成唇读字母表的唇部动作，对那些试图使用它们的人构成了明显的挑战。也就是说，多个声音共享同一个形状。辅音之间存在一定程度的歧义，这与Fisher在他对视觉符号的广泛研究中所记载的问题没有关系。

图1：虎钳示例

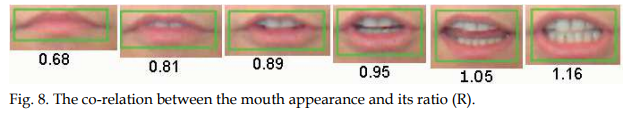


因为只有这么多的形状，一个人的嘴可以在清晰度，映射说的形状精确到潜在的单词是有挑战性的。尤其是当许多交流更多地依赖于声音而不是视觉信息时；声音交流是依赖于声音的。因此，对于人和机器来说，在没有语音上下文的情况下实现高精度是极其困难的。

### 早期结果

有了这些限制，早期的研究集中在问题的简化版本就不足为奇了。最初，特征工程通过在嘴周围放置边界框的面部识别模型进行改进，并提取出与面部方向无关的嘴唇模型。一些常用的特征是用于检测嘴的包围盒的宽高比、舌的出现（图像的红色通道中的像素强度）和图像中的“白度”的齿数的近似值。

图2:\*\*\*将嘴唇提取为特征



这些方法获得了令人印象深刻的结果（超过70%的单词准确率）测试与分类器训练的同一个说话人，他们的测试。但是，当你试图从那些不在训练范围内的人那里口述时，他们的表现受到了严重的损害。胡子男性的唇部检测也比较困难，因此在此类病例中表现较差。因此，特征工程方法虽然是一种改进，但最终未能很好地推广。

在此之后，使用不同的viseme分类方法和定义的语言模型改进了最新的性能。语言模型帮助过滤明显不正确的结果，并通过仅从合理的选项中选择来改进结果，例如“soon”中的第4个字符的“n”，而不是“soow”或“soog”。通过“微调”用于音素分类的viseme分类器，仍然有了更大的改进，这使他们能够处理在相似间隔内包含相同viseme的单词的多个可能的解决方案。这提高了准确性，并且性能比以前的方法更好。

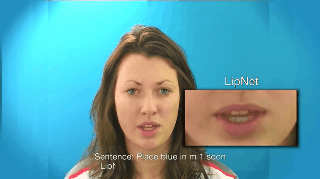
这些早期的技术使得在一个看不见的测试集上的准确率达到了19%，比之前由一个听力受损的唇读样本所产生的17%（+/-12%）的准确率的最佳水平有所提高。平均表现优于普通人群的样本群

麦格克和麦克唐纳在他们1976年的论文中争论说，语音最好理解为双峰，即同时接受视觉和听觉输入，如果这两个领域中的任何一个不存在，个人的理解能力可能会受到损害。直觉上，我们中的许多人都能回忆起在电话中听错的言语，或是在嘈杂的环境中把声音和嘴唇配对时遇到的困难。双模输入的需求，以及上下文的限制，阻碍了人们和机器准确阅读嘴唇的能力。这表明有必要进一步研究这些综合信息来源的使用。一个将我们带进唇读的最新时代的方向。

### 深度学习的到来

正是有了这一点，我们介绍了阿塞尔等人最近的工作。（2016） - “LipNet：端到端的句子级唇读。”[17]“LipNet”介绍了第一种句子级端到端唇读算法的方法。Wand、Koutník和Schmidhuber[18]早期的工作将LSTMs[19]应用于该任务，但仅用于单词分类。然而，他们早期的进步，包括端到端的可训练性，无疑对空间工作的主体是有价值的。对于那些希望了解LSTMs及其变体的人，Christopher Olah在这里提供了一个直观而详细的使用说明

图3:\*\*LipNet句子级示例



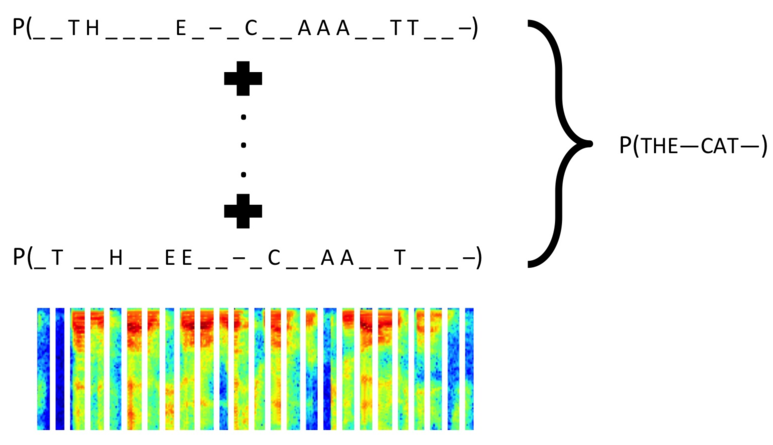
在该体系结构的高层次上，从视频序列中提取的帧在卷积神经网络（CNN）中以小集合进行处理，而LSTM变量在CNN输出上按顺序运行以生成输出字符。更准确地说，10帧序列被分组在一个块中（宽度x高度x 10），序列长度可能变化，但是这些帧的连续性质创建了时空CNN。

然后，这种称为门控递归单元（GRU）的LSTM变体的输出被多层感知器（MLP）处理以输出来自时空CNN的不同字符的值。最后，连接主义时间分类（CTC）提供了对序列输出的最终处理，以使其更易于理解精确的输出，即单词和句子。这种方法允许信息通过包含单词和句子的时间段传递，从而提高网络预测的准确性。

作者注意到，LipNet解决了说话人之间的泛化问题，即早期方法中的方差问题，以及运动特征的提取，最初在Zhou等人中被归类为开放性问题。（2014年）。[25][26]，我们认为，LipNet中的方法在精度测量的狭窄范围之外是有趣和令人兴奋的。CNN和RNN在网络中的结合，本身就是我们对类似乐高的深度学习研究方法的评论的一个回响，也许，更多的证据表明，差分程序设计很快将占据主导地位。深入学习是磨砺磨砺。维维可微规划[27]

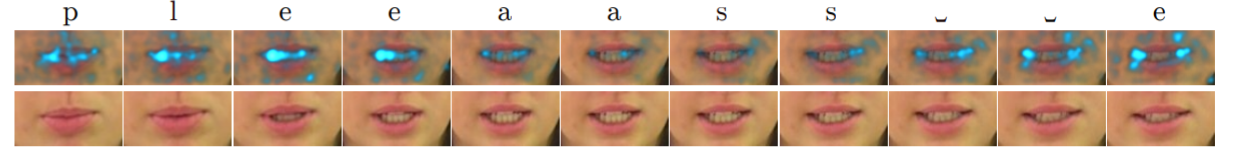
LipNet还利用了语音识别系统中通常使用的附加算法 - 连接主义时间分类（CTC）输出。在对框架字符进行分类（与定义输出序列的更多字符结合）之后，c t c可以将多个序列（例如“cúaaúu t t”和“ccaaaaúu t”）的概率分组为同一候选单词（在本例中为“cat”），以便进行最后的句子预测。因此该算法是无对齐的。CTC解决了时序可变的序列匹配问题。

图4：反恐委员会的行动



通过预测字母表字符和附加的“u”（空格）字符，可以通过删除重复的字母和空格来生成单词预测，如图5所示，用于单词“please”的分类。实际上，这意味着拉长的发音、重音和时间的变化，以及音节和单词之间的停顿，仍然可以使用CTC产生一致的预测结果。

图5：“请”的显著性图



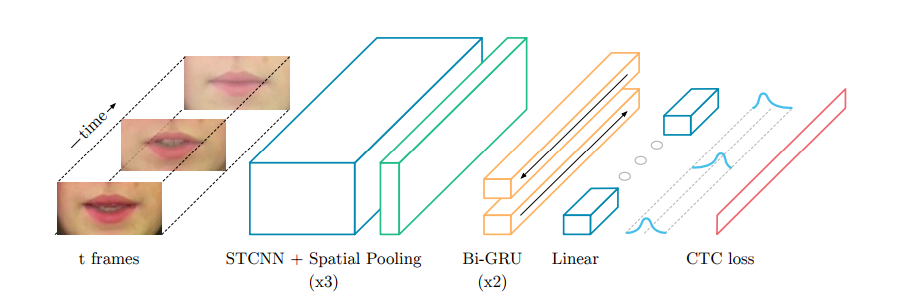
CTC是一个输出校准功能和基于该校准的损耗校正功能，独立于CNN和LSTM变量。由于将网络的原始输出（例如，原始类分数或本例中的字符）转换为预期输出（例如，概率分布或本例中的单词和句子），人们也可以认为CTC类似于softmax。CTC使得将单个字符输出与字级匹配成为可能。Awni Hannun提供了一个优秀的动态出版物，解释CTC的操作；可用。

有一个很好的视频，它涵盖了LipNet的一些功能，以及一个在自主车辆中运行的特定用例。看到LipNet投入运行，我们已经讨论过的关于这个系统的很多东西联系在一起了

### 架构和结果

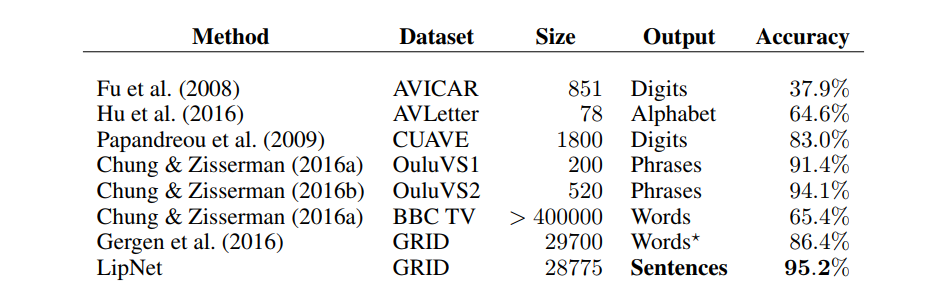
这种方法的一个特点是输出标签之间没有相互制约的关系。例如，“c a t”中的字母“a”不以“c”或“t”为条件。取而代之的是用三个时空卷积来提取这种关系，然后用两个GRU来处理一组输入图像。然后GRUs的输出通过一个MLP来计算CTC损失（见图6）。

图6:LipNet架构



LipNet的架构被认为是一个经验上的成功，对网格数据集（一个用于研究的视听句子语料库）中的句子实现了95.2%的预测准确率。[31]然而，关于深度语音识别的文献（Amodei等人，2015）[32]表明，随着更多的数据和更大的模型。评论员们回忆起之前在概括性和胡子处理上遇到的困难，对形成LipNet示例视频的来自GRID的不寻常句子表示关注。网格的有限性产生了对过度拟合的恐惧；但是LipNet在现实世界中的表现如何呢？

图7:LipNet和其他方法



### 更丰富数据的到来

在LipNet发布后不久，DeepMind发布了“野生唇读句子”，并解决了一些关于LipNet通用性的问题。从视觉特征提取的CNNs和语音转录的LSTMs两个方面，提出了一种新颖的唇读方法。通过为每个输入类型添加单独的注意机制，然后将它们组合起来生成字符输出，实现了对原始LipNet体系结构的准确性和通用性的改进。

注意机制，在这篇文章的第二部分中详细讨论，是指一种将注意力集中在神经网络中输入或前一层的特定部分的技术。最近的一种技术，从早期的工作中获得灵感，但在2013/2014年被亚历克斯·格雷夫斯的研究所普及，它的使用部分来自于他与记忆相关的工作：现在著名的序列生成论文[36]以及他在神经图灵机上的工作。[37]

注意力机制是最近在深度学习中取得成功的一个推动者；这是由于数据处理更加高效和聪明。它还允许这些模型具有更高的可解释性，即如果询问为什么网络认为某个图像是一只狗，通常很难查看和理解网络的内部结构来找出原因。注意允许网络突出显示在其预测中使用的图像的突出部分，例如鼻子和尖耳朵。注意力已经成为一种非常普遍的技术，它催生了诸如“注意力就是你所需要的一切”这样的论文，这完全避免了卷积和递归技术用于机器翻译的问题。

等转为“野生唇读句子”，钟等。（2017）展示他们的WLAS网络。由三个主要的子模块组成（看，听拼写） - ，注意力分散到拼写模块中。系统如下：

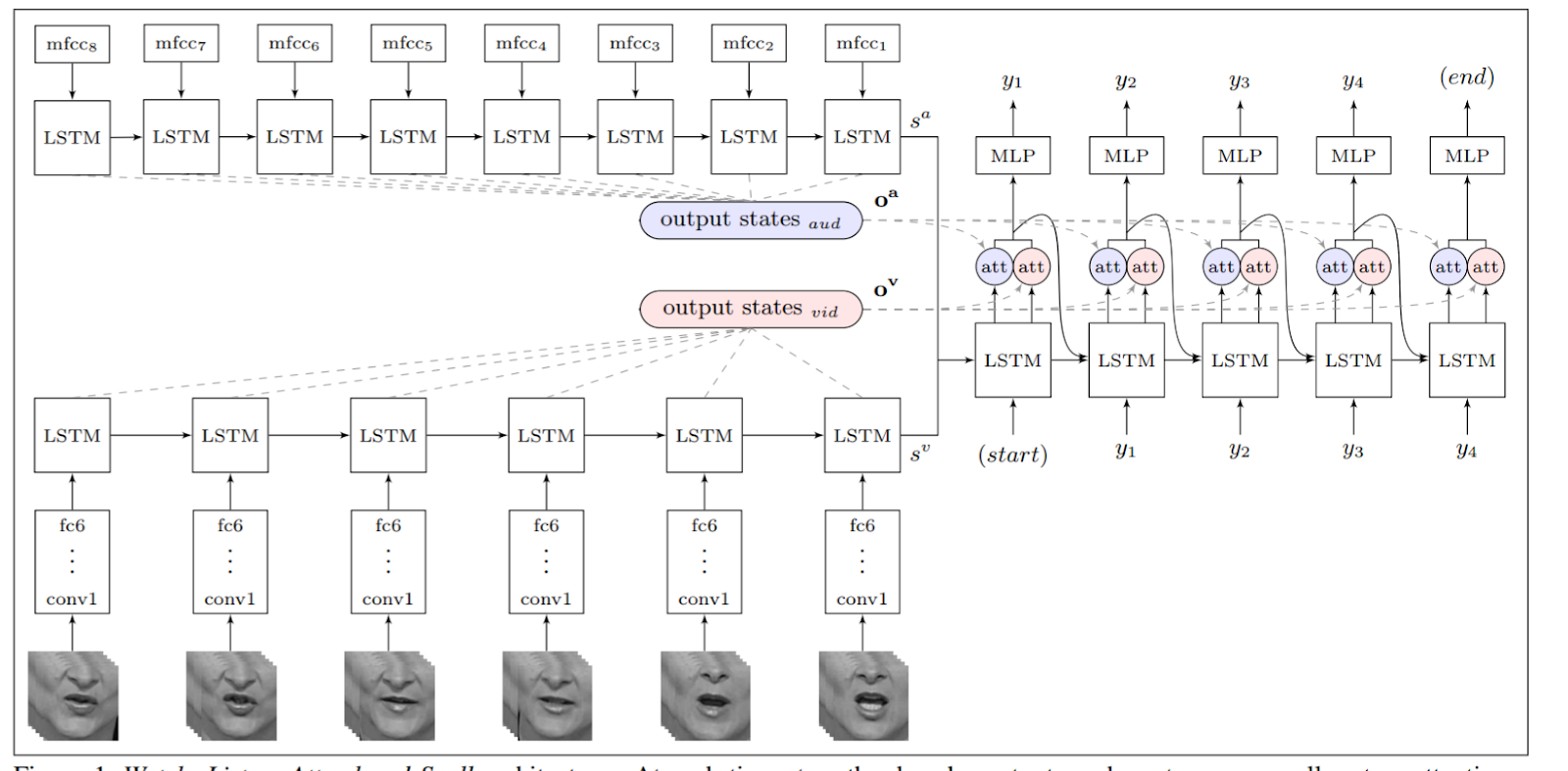
* Watch（图像编码器）：获取图像并将其编码为深层表示，以供进一步的模块处理。
* 监听（音频编码器）：允许系统接收音频格式作为唇读的可选帮助。这直接处理13维MFCC特征（见下一节）。
* 拼写（字符解码器）：这个模块包含了以前所有模块的信息。上面的每个编码器将其各自的输入序列转换为固定维状态向量和编码器输出序列。字符解码器是一个LSTM转换器，然后从两个编码器中读取固定状态和注意向量，并在输出字符序列上产生概率分布。最后，注意向量与输出状态融合，产生包含产生下一步输出所需信息的上下文向量。
* 注意（音频和视频注意机制的独立调节）：注意每个特定输入信号/流中的重要内容，即音频或视频。如果不加注意，该模型的字错误率会超过100%，而且似乎会忘记输入信号。这表明双重注意机制确实允许这种技术端到端地工作。它还允许网络处理不同步的音频/视频（不同的采样率），包括缺少的流。

### WLAS功能；更多数据中的更多细节

Watch是一个VGG mt，它提取LSTM要使用的逐帧特征表示，LSTM生成状态向量和输出信号。观看模块查看视频中的每个帧，并提取模块已学会查找的相关特征，即某些嘴唇移动/位置。这是由一个普通的VGG-M CNN完成的，它为每个帧输出一个特征表示。

然后将这些特征表示序列输入一个规则的LSTM，LSTM生成一个状态向量（或单元状态）和一个输出信号。对于LSTM s和GRUs，下一个LSTM单元有一个输出和一个“状态”输入。输出是字符预测（或预测字符的概率分布），而状态是编码“过去”的内容，即LSTM计算/存储过去的内容，用于预测下一个输出。

图8：观看、收听、观看和拼写架构



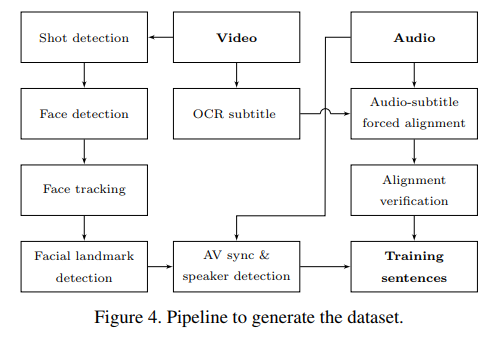
侦听模块使用Mel频率倒谱系数（MFCCs）作为输入。这些参数定义了基于信号变换的声音短期功率谱的表示。MFCC确保转换被缩放到模拟人类听力范围的频率。随后，针对每个音频和视频输入的参与模块中的独立注意机制被组合。然后依次通过咒语模块。使用多层感知器（MLP）在每个时间步上，LSTM的输出以softmax结束，以定义输出字符的概率。

有了这些，我们将回到之前工作中提到的类似的进展主题：数据可用性和网络堆栈能力。基于神经网络的方法通常具有数据量大的特点。伴随着唇读的进步，我们创建了一个独特的数据集，用于训练和测试唇读网络。以前，对唇读的研究受到现有数据集和它们的小词汇的阻碍。我们只需看看由Stafylakis和Tzimiropoulos（2017，第2页）表示的Chung等人（2016/2017）数据集、LRW和LRS的理想特征，就可以理解这些数据在改进研究工作中的价值：

“我们选择使用LRW数据库进行实验，因为它结合了许多吸引人的特性，例如大尺寸（∼500K剪辑）、说话人的高可变性、姿势和照明、野外环境中的非实验室以及目标词作为整个话语的一部分而不是孤立的。”[40]

Chung等人。（2017年）创建了一个管道，从BBC录音以及包含的闭路字幕中自动生成数据集，从而在数据密集型研究领域取得进展。他们的发明是一个用于视觉语音识别的“唇读句子”（LRS）数据集，由来自英国电视台的10万多个自然句子组成

图9：生成LRW/LRS数据集的管道



作者还用风暴噪声（即天气风暴）破坏了上述数据，证明了网络能够使用失真的低音量数据，或者完全丢弃音频用于预测目的。判断听力预测是否有价值。对于那些希望看到更多的人，钟俊森介绍了作者在CVPR工作的精彩概述。

尽管朝着降低数据需求的方向发展是迫在眉睫的，但这种模式尚未转变；而且很可能在未来一段时间内仍将保持这种方式。至于堆叠性，野生结构中的唇线和唇读的本质，说明了神经网络的乐高性质，例如，CNN通过注意力技术插入RNN中。[44]虽然这确实是一个严重的过度简化，作为一个启发，我们发现它在解释和理解许多现有和新的人工智能研究的快速进展中越来越有用。

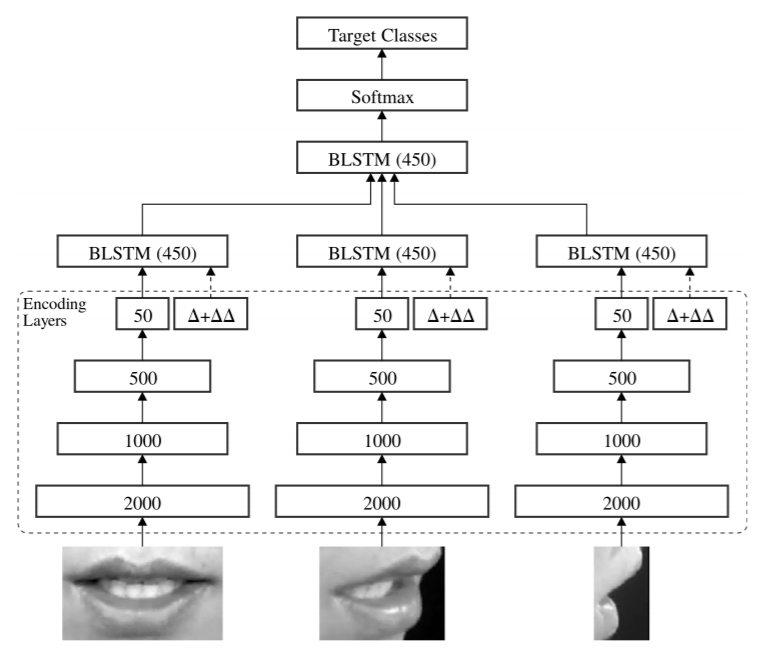
在这里，这最后一点扩展到架构本身之外，也将潜在的输入堆栈记录到我们的启发式中。这些作品的一个重大贡献是创建了一个端到端的体系结构，能够使用音频、视频或两者的组合作为输入来生成文本预测作为输出：创建一个真正的多模态模型。产生单一输出序列的多个输入序列。解决这个多模态问题，以及其他类似的问题，可能会为连接视频、音频和语言系统开辟新的途径。

### 新的探索道路

至于如何遵循前面详述的方法，我们将注意力转向这一领域的一些最新工作。虽然不是详尽无遗，但以下是我们在这一领域中遇到的一些最佳改进：

* 将残差网络与LSTMs相结合用于唇读：在原有LRW纸张的基础上，采用时空卷积网络、残差网络和堆叠式双向短时记忆网络（BiLSTMs）进行改进。后者像前面提到的BiGRU一样，按正向和反向顺序处理输入特性。该方法在LRW文件上的字准确率从76.2%提高到83%，同时也提高了网格数据集上的字准确率。目前，它们还没有扩展到LRS数据集上的句子。
* 通过领域对抗训练改进与说话人无关的唇读：有助于在只有少量数据的情况下提高目标说话人的性能。在模型是基于大量数据训练的情况下，它没有那么有效。因此，我们将有兴趣看到其性能的LRS有1000多个扬声器。
* 端到端多视图唇读：实现非正面唇读的分类，也利用双向长-短记忆（BiLSTM）。对于每个视点，作者创建一个相同的编码MLP架构（视频处理流），使得网络能够同时训练多个视点（参见图10）。

图10：架构



* 使用噪声不变训练的视觉语音增强：通过提供在噪声环境中增强可视扬声器的语音的方法来解决一些相关的问题。该方法使用之前看到的视听输入，通过匹配嘴唇运动，将声音与背景噪声分离。尽管它不同于其他方法，但这个想法本身是新颖的，坦率地说，相当酷。尤其是因为它使用了一个唇读数据集来完成这个任务。

“视觉语音增强用于在嘈杂环境中拍摄的视频，以增强可视扬声器的声音并降低背景噪声。虽然大多数现有的方法只使用音频输入，但我们为此目的提出了一个视听神经网络模型。可见的嘴部运动用来将说话人的声音与背景声音分开。”

### 工具书类

1. [^]（#4a99）M油箱。（2017年）。计算机视觉一年。[在线]TheMTank.com。网址：http://www.themtank.org/a-year-in-computer-vision
2. [^]（#215e）奥利奥葡萄酒和斯科特里德（2017）。深入学习：实践与趋势（NIPS 2017教程，第一部分和第二部分）。[在线视频]史蒂文·范·瓦伦伯格（www.youtube.com）。可用：
3. [^]（#5571）Kaiser等人。（2017年）。一个模型来学习它们。[在线]arXiv:1706.05137。可用：
4. [^]（#5571）桥本等人。（2017年）。联合多任务模型：为多个NLP任务生成一个神经网络。[在线]arXiv:1611.01587。可用：
5. [^]（#5571）Zamir等人。（2018年）。任务单。[网站]Taskonomy/斯坦福大学。可用：
6. [^]（#abc1）Ruder，S.（2016年）。关于单词嵌入 - 第1部分。[在线]塞巴斯蒂安·鲁德尔博客（Ruder.io）。可用：
7. [^]（#abc1）Karpathy，A.（2015年）。递归神经网络的不合理有效性。[在线]Andrej Karpathy博客（）。可用：
8. [^]（#9685）费舍尔，C.G.（1968）。视觉感知辅音之间的混淆。《言语、语言和听力研究杂志》，12月，第11卷，第796-804页。
9. [^]（#b677）
10. [^]（#40b8）运动、姿势（全正面视图（0°）、角度视图（45°）和侧视图）、多人、视频条件/分辨率/照明、语音方法（口音、风格和语音速率）。摘自：Bear，H.L.（2017年，第25页）。解码虎钳：提高机器唇读。[在线]arXiv:1710.01288。可用：。（最初于2016年从IEEE会议记录中获得）。
11. [^]（#40b8）言语语境：主体、声音、时间、地点。
12. [^]（#1796）哈萨纳，A.B.A.（2011）。视觉语音识别。[在线]arXiv:1409.1411。可用：。
13. [^]（#1796）同上
14. [^]（#e3aa）贝尔，H.L.（2017）。解码虎钳：提高机器唇读。[在线]arXiv:1710.01288。可用：。（最初于2016年从IEEE会议记录中获得）。
15. [^]（#0a54）伊斯顿，R.D.，巴萨拉，M.（1982）。唇读时的知觉支配。知觉与心理物理学，32（6），第562-570页。
16. [^]（#64bd）麦克唐纳，J.，麦高克，H.（1976）。《倾听嘴唇和看见声音》，《自然》，第264卷，12月，第746-748页。可用：
17. [^]（#86d8）Assael等人。（2016年）。Lipnet：端到端的句子级唇读。[在线]arXiv:1611.01599。可用：
18. [^]（#86d8）Wand等人。（2016年）。有长期短期记忆的唇读。[在线]arXiv:1601.08188。可用：
19. [^]（#86d8）长期短期记忆（LSTM）是一种在递归神经网络（RNN）中的单位，负责在一段时间内记忆和传递值。见：Hochreiter，S.，Schmidhuber，J.（1997）。长期短期记忆。神经计算，第9卷（8），第1735-1780页。可用：
20. [^]（#86d8）端到端，我们定义为，神经网络或系统所有部分内所有参数的联合训练。而不是单独训练。
21. [^]（#86d8）奥拉，C.（2015）。了解LSTM网络。[在线]Colah的博客（）。可用：
22. [^]（#554e）阿塞尔，Y.（2016）。你认为唇读有多容易？[在线视频]Yannis Assael（）。可用：
23. [^]（#cad0）卡恩，U.（2016）。卷积神经网络的直观解释。[在线]Ujjwal Karn博客（）。可用：
24. [^]（#1edf）Cho等人。（2014年）。使用RNN编码解码器学习短语表示进行统计机器翻译。[在线]arXiv:1406.1078。可用：
25. [^]（#9f6c）见Zhou等人。（2014年）。视觉语音解码的最新进展。图像和视觉计算，第32卷（9），第590-605页。
26. [^]（#9f6c）引述/引用自：Assael等人。（2016年）。Lipnet：端到端的句子级唇读。[在线]arXiv:1611.01599。可用：
27. [^]（#9f6c）法语是谷歌的神经机器翻译服务提供的。见Yann LeCun的帖子：
28. [^]（#9f01）Brueckner，R.（2016年）。使用开源Warp-CTC加速机器学习。[在线]在HPC内部（Inside HPC.com）。可用：
29. [^]（#b279）汉农，A.（2017）。用CTC进行序列建模。[在线]提取（）。可用：
30. [^]（#bbb5）阿塞尔，Y.（2016）。LipNet在自主汽车中的应用| CES 2017。[在线视频]Yannis Assael（）。可用：
31. [^]（#75aa）巴克等人。（2018年）。网格视听句子语料库。[在线]谢菲尔德大学。提供：（上次更新，2013年3月18日）。
32. [^]（#75aa）Amodei等人。（2015年）。深度语音2：英语和汉语中的端到端语音识别。[在线]arXiv:1512.02595。可用：
33. [^]（#9239）Chung等人。（2017年）。在野外读唇语的句子。[在线]arXiv:1611.05358。可用：
34. [^]（#9239）Simonyan，K.，Zisserman，A.（2015年）。用于大规模图像识别的非常深卷积网络。[在线]arXiv:1409.1556。可用：
35. [^]（#9239）格雷夫斯，A.，贾伊里，N.（2014）。用递归神经网络进行端到端语音识别，第31届机器学习国际会议论文集，北京，中国，JMLR:W&CP第32卷。可用：
36. [^]（#3dc7）格雷夫斯，A.（2014）。用递归神经网络生成序列。[在线]arXiv:1308.0850。可用：
37. [^]（#3dc7）Graves等人。（2014年）。神经图灵机器。[在线]arXiv:1410.5401。可用：
38. [^]（#aa0c）查特菲尔德等人。（2014年）。魔鬼归来的细节：深究卷积网。[在线]arXiv:1405.3531。可用：
39. [^]（#c750）实用密码。（2018年）。Mel频率倒谱系数（MFCC）教程。[在线]实用密码术（）。提供：（访问日期：2018年2月20日）。
40. [^]（#ea01）Stafylakis，T.，Tzimiropoulos，G.（2017年）。结合残差网络和LSTMs进行唇读。[联机]arXiv:1703.04105v4。可用：
41. [^]（#6289）唇读单词（LRW）和唇读句子（LRS）
42. [^]（#b6f9）多模式培训。音频流是最容易学习的。在训练过程中，他们随机选择视频、音频或两者的训练组合。他们还添加噪音的音频，以阻止它支配学习过程，因为唇读是一个更难的问题。
43. [^]（#9b46）Chung，J.S.（2017年）。野生唇读句子（唇读句子数据集），CVPR 2017。[在线视频]保留知识（）。可用：
44. [^]（#22eb）本文使用的另一个技巧是课程学习，这涉及到在培训过程中逐步显示网络更难学习的数据。这样可以大大加快训练过程，减少过度拟合。在这种情况下，首先用单个单词训练模型，然后是2个单词、3个单词，最后是完整的句子。
45. [^]（#cb19）Stafylakis，T.，Tzimiropoulos，G.（2017年）。结合残差网络和LSTMs进行唇读。[联机]arXiv:1703.04105v4。可用：
46. [^]（#1ddb）字精度=1字错误率
47. [^]（#1ddb）Wand，M.，Schmidhuber，J.（2017年）。利用领域对抗训练提高说话人独立唇读能力。[在线]arXiv:1708.01565。可用：
48. [^](#0eeb) Petridis et al. (2017). End-to-End Multi-View Lipreading. [Online] arXiv: 1709.00443. Available: **arXiv:1709.00443v1**
49. [^](#7eac) The idea behind it is that the derivatives of image features are associated with feature extractors (e.g. HoG, Sobel filter, etc.) https://www.learnopencv.com/histogram-of-oriented-gradients/
50. [^](#7eac) Gabbay et al. (2017). Visual Speech Enhancement using Noise-Invariant Training. [Online] arXiv: 1711.08789. Available: [arXiv:1711.08789v2](https://arxiv.org/abs/1711.08789v2)