多模式深度学习

吉泉 Ngiam[1] jngiam@cs.stanford.edu
Aditya Khosla[1] aditya86@cs.stanford.edu
明宇 Kim[1] minkyu89@cs.stanford.edu
Juhan Nam[1] juhan@ccrma.stanford.edu
宏乐 Lee[2] honglak@eecs.umich.edu
安德鲁●Y. Ng[1] ang@cs.stanford.edu
[1]斯坦福大学计算机科学系，斯坦福，加利福尼亚州94305，美国
[2]密歇根大学计算机科学与工程系，Ann Arbor，MI 48109，USA

摘要

深度网络已成功应用于单一模态（例如，文本，图像或音频）的无监督特征学习。在这项工作中，我们提出了深度网络的新颖应用，以学习多种形式的特征。我们提出了一系列多模式学习任务，并展示了如何训练深入的网络，学习解决这些任务的功能。特别地，我们展示了交叉模态特征学习，其中如果在特征学习时存在多个模态（例如，音频和视频），则可以学习用于一种模态（例如，视频）的更好的特征。此外，我们将展示如何学习模态之间的共享表示，并在独特的任务上对其进行评估，其中分类器使用纯音频数据进行训练，但仅使用纯视频数据进行测试，反之亦然。我们的模型在有关视听语音分类的CUAVE和AVLet-ters数据集上得到验证，展示了AVLetters上最佳发布的视觉语音分类和有效的共享表示学习。

1. 简介

在语音识别中，已知人类整合视听信息以便理解语音。这首先体现在McGurk效应中（麦格劳和麦克唐纳，1976）其中视觉/ ga /具有浊音/ ba /被大多数主体感知为/ da /。特别是，视觉形态提供了信息 - 关于发音和肌肉运动的地方（萨默，1992这通常可以帮助消除具有相似声学的语音之间的歧义（例如，无声辅音/ p /和/ k /）。

多模式学习涉及来自多个来源的信息。例如，图像和3-d深度扫描在一阶相关，因为深度不连续通常表现为图像中的强边缘。相反，用于语音识别的音频和视觉数据在“中级”具有相关性，如音素和视位（嘴唇姿势和动作）；将原始像素与音频波形或频谱图相关联可能很困难。

在本文中，我们对“中级”关系建模感兴趣，因此我们选择使用视听语音分类来验证我们的方法。特别是，我们专注于学习与嘴唇的视频相结合的语音音频的表示。

我们将考虑图中所示的学习设置
1. 总体任务可分为三个阶段
- 功能学习，监督培训和测试。简单的线性分类器用于监督训练和测试，以检查具有多模态数据的不同特征学习模型。特别是，我们考虑三种学习设置 - 多模式融合，交叉模态学习和共享表示学习。

在多模式融合设置中，所有阶段的数据均可在所有阶段获得；这代表了视听语音识别中大多数先前工作中考虑的典型设置（Potamianos等。，2004）。在交叉模态学习中，来自多种模态的数据仅在特征学习期间可用；在监督的训练和测试阶段，仅提供来自单一模态的数据。对于此设置，目的是在给定来自多种模态的未标记数据的情况下学习更好的单一模态表示。最后，我们同意

出现在28 日 国际机器学习会议论文集，美国华盛顿州贝尔维尤，2011年。版权所有者2011 /作者/所有者。

侧面共享表示学习设置，其独特之处在于提供了用于监督训练和测试的不同模态。此设置允许我们评估特征表示是否可以捕获不同模态的相关性。具体来说，研究此设置可以让我们评估所学习的表示是否是模态不变的。

在以下部分中，我们首先描述构建 -

我们模型的块。然后，我们提出了不同的多模式学习模型，从而形成了能够执行各种多模式学习的深层网络 -

任务。最后，我们报告实验结果并得出结论。

2. 背景

最近深度学习的工作（Hinton & Salakhut-d-迪诺夫, 2006; Salakhutdinov和Hinton, 2009）已经研究了可以训练多长的S形网络以产生手写挖掘的有用表示 -

它和文字。关键的想法是逐层使用贪婪

用受限制的玻尔兹曼机器（RBMs）进行训练，然后进行微调。我们使用RBM的扩展（稀疏性）（李等人。, 2007），已被证明可以学习数字和自然图像的有意义的功能。在下一节中，我们将回顾稀疏RBM，它用作模型的分层构建块。

2.1. 稀疏受限的玻尔兹曼机器

RBM是一个无向图形模型，具有隐藏变量（h）和可见变量（v）（图2a）。隐藏变量和可见变量之间存在对称连接（$W_{i,j}$），但隐藏变量或可见变量内没有连接。该模型定义了h，v（方程式）上的概率分布1)。当v或h固定时，这种特殊配置可以很容易地计算条件概率分布（方程式）2).

$$-\log P(\mathbf{v}, \mathbf{h}) \propto E(\mathbf{v}, \mathbf{h}) = \frac{1}{2\sigma^2}\mathbf{v}^T\mathbf{v} - \frac{1}{\sigma^2}\left(\mathbf{c}^T\mathbf{v} + \mathbf{b}^T\mathbf{h} + \mathbf{h}^T W\mathbf{v}\right) \quad (1)$$

$$p(h_j) = sigmoid\left(\frac{1}{\sigma^2}(b_j + \mathbf{w}_j^T\mathbf{v})\right) \quad (2)$$

该公式将可见变量建模为实值单位，将隐藏变量建模为二进制单位。[1] 由于计算对数似然项的梯度是难以处理的，因此我们学习了参数

[1]我们使用高斯可见单位作为连接到输入数据的RBM。在训练更深层时，我们使用二元可见单元。

	特色学习	监督培训	测试
经典深度学习	音频	音频	音频
	视频	视频	视频
多模式融合	A + V	A + V	A + V
	A + V	音频	视频
	A + V	音频	音频

图1：多模式学习设置，其中A + V.
指音频和视频。

模型（$w_{i,j}$，b_j，c_i）使用对比分歧（欣纳 吨, 2002).

规范稀疏模型（李等人。, 2007），我们鼓励每个隐藏单元使用正则化进行预定的预期激活

$$\left\{ \qquad \right\} \qquad j \qquad m \qquad k=1$$

形式为 λ（$\rho - {}^{1}({}^{m}E[h_j| v^k])$）^{2}，其中$v^1, \ldots, v^m$ 是训练集，ρ 确定隐藏单位激活的稀疏性。

3. 学习架构

在本节中，我们描述了用于视听双模特征学习任务的模型，其中模型的音频和视觉输入是连续的音频（频谱图）和视频帧。激励我们的深度自动编码器（Hinton和Salakhutdinov, 2006）模型，我们首先描述几个简单模型及其缺点。

特征学习最直接的方法之一是分别为音频和视频训练RBM模型（图2a，b）。在学习RBM之后，给出可见变量（等式2）的隐藏变量的后验可以用作数据的新表示。我们使用此模型作为基线来比较我们的多模式模型的结果，以及预深度网络的预训练。

为了训练多模态模型，直接的方法是
在串联的音频和视频上训练RBM

数据（图2c）。虽然这种方法联合改造
这是音频和视频数据的分布
仅限于浅层模型。特别是，由于音频和视频数据之间的相关性是高度非线性的，因此RBM很难学习这些相关性并形成多模式表示。在实践中，我们发现学习浅层双峰RBM会导致隐藏单元与各个模态的变量有很强的联系，但很少有单元连接这些模态。

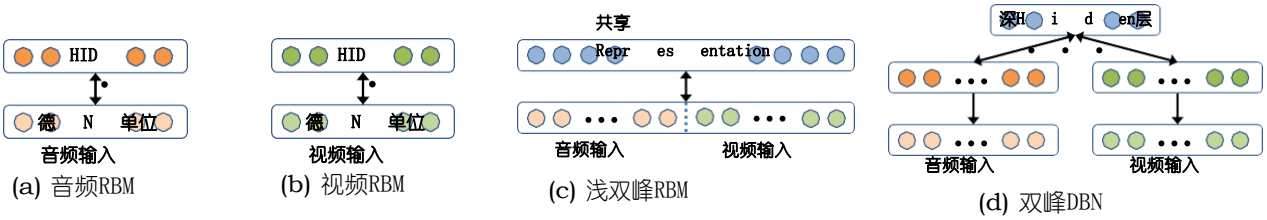

图2：RBM预训练模型。我们将（a）音频和（b）视频的RBM分别作为基线进行训练。浅层模型（c）是有限的，我们发现该模型无法捕获模态之间的相关性。双峰深信念网络（DBN）模型（d）通过首先训练模型（a）和（b）以贪婪的分层方式进行训练。我们稍后“展开”深度模型（d）来训练图3中所示的深度自动编码器模型。

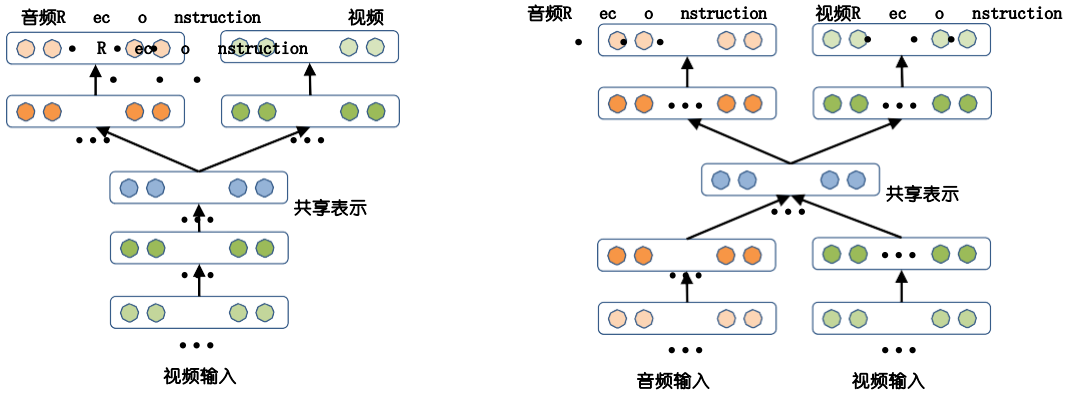

图3：深度自动编码器模型。在（a）中示出了“仅视频”模型，其中模型学习重建两个模态，仅给出视频作为输入。可以为“仅音频”设置绘制类似的模型。我们以去噪方式训练（b）双峰深度自动编码器，使用增强数据集，其中的示例要求网络仅重建两个模态。两种模型都使用稀疏RBM进行预训练（图2d）。由于我们在深层网络中使用sigmoid传递函数，我们可以使用学习的RBM的条件概率分布p（h | v）和p（v | h）来初始化网络。

因此，我们考虑在深度学习方法的推动下，在每个模态的预训练层上贪婪地训练RBM（图2d）。[2] 特别是，后代（方程式2）第一层隐藏变量用作新层的训练数据。通过学习的第一层表示来表示数据，模型可以更容易地学习跨模态的高阶相关性。非正式地，第一层表示对应于音素和视位，第二层表示它们之间的关系。图4显示了来自我们的模型的学习特征的可视化，包括对应于视位的视觉基础的示例。

但是，上述多模式模型仍存在两个问题。首先，模型没有明确的目标来发现整个模型的相关性

[2]可以改为将RBM作为连接两种模态的第一层。但是，由于单层RBM倾向于学习单峰单元，因此为每种模态学习单独的模型效率要高得多。

联系；模型可以找到表示，使得某些隐藏单元仅针对音频进行调谐，而其他隐藏单元仅针对视频进行调谐。其次，模型在交叉模态学习环境中使用是笨拙的，其中在监督训练和测试期间仅存在一种模态。只有一种模态存在，人们需要整合出未观察到的可见变量来进行推理。

因此，我们提出了一个深度自动编码器来解决这两个问题。我们首先考虑交叉模态学习设置，其中在特征学习期间存在两种模态，但是仅使用单一模态用于监督训练和测试。深度自动编码器（图3a）经过训练，在仅给出视频数据时重建两种模态，从而发现模态之间的相关性。类似于韩丁 & Salakhutdinov (2006)，我们基于方程初始化具有双峰DBN权重的深度自动编码器（图2d）2，丢弃任何不再存在的重量。中间层可以用作

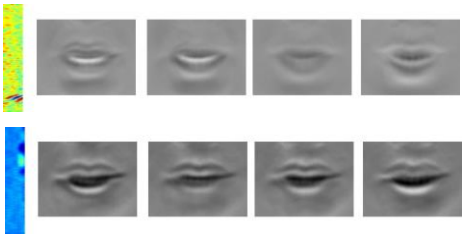

图4：学习表示的可视化。这些图对应于两个深度隐藏单元，其中我们可视化最强连接的第一层特征。这些单元以视听对的形式呈现（我们发现通常很难解释这对之间的连接）。视觉基础捕获唇部运动和关节，包括不同的嘴部关节，嘴的打开和闭合，露出牙齿。

新特征表示。该模型可视为多任务学习的一个实例（卡鲁阿纳, 1997).

我们在设置中使用深度自动编码器（图3a）模型，其中在监督训练和测试中仅存在单个模态。另一方面，当多个模态可用于任务时（例如，多模式融合），如何使用模型则不太清楚，因为需要为每种模态训练深度自动编码器。一个简单的解决方案是训练网络，使解码权重相关联。然而，这种方法不能很好地扩展 – 如果我们允许在测试时存在或不存在任何模态组合，我们将需要训练指数数量的模型。

受到去噪自动编码器的启发（文森特等人。，2008），我们建议使用增强但有噪声的数据集训练双峰深自动编码器（图3b），其中附加示例仅具有单模态作为输入。实际上，我们添加了一个输入模态（例如视频）和其他输入模态（例如音频）的原始值零值的示例，但仍需要网络重建两种模态（音频和视频）。因此，三分之一的训练数据仅具有用于输入的视频，而另外三分之一的数据仅具有音频，并且最后三分之一的数据具有音频和视频。

由于使用稀疏RBM进行初始化，我们发现隐藏单元即使在深度自动编码器训练之后也具有低预期激活。因此，当其中一个输入模态设置为零时，第一层表示也接近于零。在这种情况下，我们基本上是训练一个特定模态的深度自动编码器网络（图3a）。实际上，该方法学习的模型对于没有模态的输入是鲁棒的。

4. 实验和结果

我们评估我们的孤立字母和数字的视听语音分类方法。使用交叉验证选择稀疏参数 ρ，而所有其他参数（包括隐藏层大小和权重正则化）保持固定。[3]

4.1. 数据预处理

我们使用其频谱图表示音频信号[4] 利用时间导数，得到483维向量，其用PCA白化减少到100维。10个连续的音频帧被用作我们模型的输入。

对于视频，我们预处理帧以便仅提取包含嘴的感兴趣区域（ROI）。[5] 每个口的投资回报率都被重新调整为
60×80像素，进一步减少到32维，[6] 使用PCA美白。还使用了还原载体上的时间衍生物。我们使用4个连续的视频帧进行输入，因为它具有与10个音频帧大致相同的持续时间。

对于这两种方式，我们还进行了特征均值归一化（Potamianos等。，2004），类似于从每个例子中删除DC分量。我们还注意到，在表示中添加时间导数已被广泛用于文献中，因为它有助于模拟动态语音信息（Potamianos等。，2004; 赵和巴纳德, 2009).使用归一化线性斜率计算时间导数，使得导数特征的动态范围与原始信号相当。

4.2. 数据集和任务

由于无监督特征学习仅需要未标记的数据，因此我们将不同的数据集（如下所列）组合在一起以学习特征。AVLetters和CUAVE进一步用于监督分类。我们确保没有测试数据用于无监督的特征学习。所有深度自动编码器模型都使用所有可用的未标记音频和视频数据进行训练。

[3]我们交叉验证 ρ 超过0.01, 0.03, 0.05, 0.07。第一层功能是视频（1536单位）的4倍过度完成和音频（1500单位）的1.5倍过度完成。第二层是组合的第一层（4554单位）的1.5倍。

[4]每个频谱图帧（161个频率区间）具有20ms窗口，重叠10ms。

[5]我们使用了现成的物体探测器（达拉尔& 特里格斯, 2005）随着时间的推移用中值滤波来提取口腔区域。

[6]类似于（Duchnowski等人。，1994）我们发现32个维度足够并且表现良好。

cu（Patterson等人。，2002）。36位发言者说数字0到9。我们使用数据集的正常部分，其中包含正面的发言者说每个数字5次。我们在独立于扬声器的设置中评估了CUAVE数据集上的数字分类。由于尚未有针对此数据集进行评估的固定协议，因此我们选择使用奇数编号的扬声器作为测试集，使用偶数编号的扬声器作为训练集。

Av信（马修斯等人。，2002）。10个发言者说字母A到Z，每个三次。数据集提供了60个像素的预提取唇区域。由于原始音频不适用于此数据集，我们将其用于仅视觉唇读任务的评估（第4.3节）。我们报告使用的第三个测试设置的结果赵和巴纳德(2009）和马修斯 等。(2002）进行比较。

AVLetters2（考克斯等人。，2008）。5个发言者说字母A到Z，每个七次。这是AVLetters数据集的新高定义版本。我们仅将此数据集用于无监督培训。

斯坦福数据集。23名志愿者讲述了数字0到9，字母A到Z以及来自TIMIT数据集的所选句子。我们以与CUAVE数据集类似的方式收集此数据，并仅将其用于无监督培训。

TIMIT（费舍尔等人。，1986）。我们将此数据集用于无监督音频特征预训练。

我们注意到，在所有数据集中，嘴唇在外观，方向和大小方面存在差异。对于每个音频 - 视频剪辑，从重叠的帧序列中提取特征。由于示例具有不同的持续时间，因此我们将每个示例划分为S个相等的切片，并在每个切片上执行平均池化。随后将所有切片的特征连接在一起。具体来说，我们使用S = 1和S = 3组合特征，以形成用线性SVM进行分类的最终特征表示。

4.3. 跨模式学习

在交叉模态学习实验中，我们评估当在特征学习期间给定多个模态（例如，音频和视频）时我们是否可以学习一种模态（例如，视频）的更好表示。

在AVLetters数据集（表1a）中，我们的深度自动编码器模型显示出与先前工作中的手工设计功能相比的显着改进。仅视频深度自动编码器在数据集上表现最佳，分类准确度达到64.4%，优于之前公布的最佳结果。

在CUAVE数据集（表1b）上，通过学习两个视频的视频功能有了改进和仅与视频数据学习功能相比的音频（虽然表现不如现有技术）。在我们的模型中，我们选择使用一个非常简单的前端，它只提取边界框，而不需要对方向或透视变化进行任何修正。相比之下，最近的AAM模型（帕帕- dreou等。，2009训练以准确地跟踪说话者的脸部并进一步用平均脸部模板登记脸部，消除形状变形。将这些复杂的视觉前端与我们的功能相结合，可以做得更好。

表1：（a）AVLetters和（b）CUAVE上的视觉语音分类的分类性能。深度自动编码器表现最佳，并展示有效的交叉模态学习。在指示的情况下，误差条显示由于随机初始化引起的变化（2sd）。结果是连续语音识别性能，但我们注意到CUAVE的正常部分有发言者说孤立的数字。这些模型使用的视觉前端系统比我们的系统复杂得多，并且使用不同的列车/测试分割。

特征表示	准确性
基线预处理视频	46.2%
RBM视频（图2b）	54.2%±3.3%
仅视频深度自动编码器（图3a）	**64.4%**±2.4%
双峰深自动编码器（图3b）	59.2%
多尺度空间分析（马修斯等人。，2002）	44.6%
本地二进制模式（赵和巴纳德，2009）	58.85%

(a) Av快报

特征表示	准确性
基线预处理视频	58.5%
RBM视频（图2b）	65.4%±0.6%
仅视频深度自动编码器（图3a）	**68.7%**±1.8%
双峰深自动编码器（图3b）	66.7%
离散余弦变换（Gurban & Thiran，2009）	64% †§
主动表现模型（帕潘德里欧等人。，2007）	75.7% †
主动表现模型（Pitsikalis等。，2006）	68.7% †
融合整体+补丁（Lucey & Sridharan，2006）	77.08% †
视觉 Aam（帕潘德里欧等人。，2009）	83% †§

(b) CUAVE视频

表2：在清洁和嘈杂条件下CUAVE上的双峰语音分类的数字分类性能。我们在0 dB SNR下将白高斯噪声添加到原始音频信号。由于添加到音频数据的随机噪声，误差条反映了结果的变化（2 sd）。我们比较了双峰深度自动编码器模型的性能与最佳音频功能（音频RBM）和最佳视频功能（仅视频深度自动编码器）。

特征表示	准确性（清洁音频）	准确性（嘈杂的音频）
（a）音频RBM（图2a）	**95.8%**	75.8% ± 2.0%
（b）仅视频深度自动编码器（图3a）	68.7%	68.7%
（c）双峰深自动编码器（图3b）	90.0%	77.3% ± 1.4%
（d）双峰+音频RBM	94.4%	**82.2%** ± 1.2%
（e）仅视频深AE +音频RBM	87.0%	76.6% ± 0.8%

这些视频分类结果表明，深度自动编码器通过在给定附加音频数据时发现更好的视频表示来实现交叉模态学习。特别是，即使AVLetters数据集没有任何音频数据，我们也可以通过使用其他额外的未标记音频和视频数据学习更好的视频功能来提高性能。

然而，双模式深度自动编码器的性能不如仅视频深度自动编码器：虽然仅视频自动编码器只学习视频功能（这也有利于音频重建），但双模自动编码器只学习音频，仅限视频和不变的功能。因此，当手头的任务仅具有视觉输入时，由双模自动编码器学习的特征集可能不是最佳的。

我们还注意到，与使用音频RBM功能相比，音频的交叉模态学习并没有改善分类结果;音频功能对语音分类具有高度的辨别力，添加视频信息有时会影响性能。

4.4. 多模式融合结果

虽然单独使用音频信息可以很好地进行语音识别，但融合音频和视频信息可以显着提高性能，尤其是当音频因噪声而降级时（古尔- 班和蒂兰，2009; 帕潘德里欧等人。，2007; 皮西卡利斯等。，2006; 帕潘德里欧等人。，2009）。特别是，通常会发现音频功能本身表现良好，并且连接视频功能有时会损害性能。因此，我们在干净和嘈杂的音频设置中评估我们的模型。

视频模态通过提供诸如发音位置之类的信息来补充音频模态，这可以帮助区分类似的发声语音。然而，当简单地连接音频和视觉特征时（表2e），与仅使用音频特征相比，性能通常更差（表2a）。由于我们的模型能够学习多模式功能不仅仅是简单地连接音频和视觉功能，我们建议将音频功能与我们的多模式功能相结合（表2d）。当最佳音频功能与双峰功能连接时，它优于其他功能组合。这表明所学习的多模态特征能够更好地补充音频特征。

4.5. 麦格劳克效应

表3：McGurk效应

音频/视频设置	模型预测		
	/ga/ m a1	*BA*	*/dae/*
视觉/ ga /，音频/ ga /	82.6%	2.2%	15.2%
视觉/ ba /，音频/ ba /	4.4%	89.1%	6.5%
视觉/ ga /，音频/ ba /	28.3%	13.0%	58.7%

麦格劳克效应（麦格劳和麦克唐纳，1976）是指视听感知现象，其中视觉/ ga /具有音频/ ba /被大多数主体感知为/ da /。由于我们的模型学习了多模态表示，因此观察模型是否能够复制类似的效果会很有趣。

我们从23名志愿者那里获得了5次重复/ ga /，/ ba /和/ da /的数据。双模式深度自动编码器功能[7] 用于在这个3路分类任务上训练线性SVM。该模型在模拟McGurk效应的三个条件下进行了测试。当视觉和音频数据在测试时匹配时，模型能够预测正确的等级/ ba /和/ ga /，分别具有82.6％和89.1％的准确度。另一方面，当在测试时混合有视觉/ ga /浊音/ ba /时，模型最有可能预测/ da /，即使/ da /既没有出现在视觉输入中也没有出现，与麦格劳克对人的影响。使用双峰DBN（图2d）或连接音频和视频RBM功能未观察到相同的效果。

[7] / ga /，/ ba /和/ da / data未用于训练双峰深度自动编码器。

4.6. 共享表示学习

表4：CUAVE上的共享表示学习。机会表现为10%。

火车/测试	方法	准确性
音频视频	Raw-CCA	41.9%
	RBM-CCA功能	**57.3%**
	双峰深AE	30.7%
视频/音频	Raw-CCA	42.9%
	RBM-CCA功能	**91.7%**
	双峰深AE	24.3%

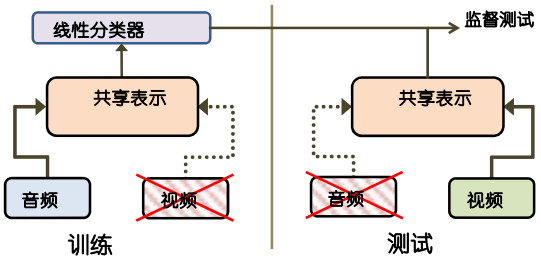

图5：用于评估共享表示的“听力观察”设置（音频训练，视频测试）。

在这个实验中，我们提出了一种新颖的设置，它检查是否可以通过音频和视频语音数据学习共享表示。在监督训练期间，算法仅从一种模态（例如，音频）提供数据，并且稍后仅在另一种形式（例如视频），如图5所示。实质上，我们告诉受监督的学习者数字“1”，“2”等是如何发声的，同时要求它根据如何区分它们他们是视觉上说的 - “听到要看”。如果我们能够在共享表示中捕获模态中的相关性，那么模型将很好地执行此任务。

学习共享表示的一种方法是找到最大化的模态的变换
相关性。特别是，我们建议使用canoni-校准相关分析（CCA）（Hardoon等。，2004），它找到音频和视频数据的线性变换，以形成共享表示。[8] 在原始数据上学习CCA共享表示会产生令人惊讶的良好性能（表4：Raw-CCA）。然而，学习第一层特征上的CCA表示（即，音频RBM和视频RBM特征）导致显着更好的性能，与使用用于监督分类的原始模态（表4：RBM-CCA特征）相比。由于对音频执行测试，这尤其令人惊讶

[8]给定音频数据a和视频数据v，CCA找到矩阵P和Q，使得Pa和Qv具有最大相关性。

比视频测试更好，即使模型是在视频数据上训练的。这些结果表明，捕获模态之间的关系需要至少一个非线性阶段才能成功。当从两种模态中学习到良好的特征时，线性模型可以很好地捕捉关系。然而，重要的是要注意CCA，一种线性变换，对诸如跨模态学习等其他任务没有帮助。

我们进一步使用此任务来检查双峰深度自动编码器的特征是否捕获了模态之间的相关性。[9] 虽然双峰深度自动编码器模型的表现不如CCA，但结果表明我们的学习表示对输入模态是部分不变的。

4.7. 其他控制实验

仅视频深度自动编码器具有音频作为训练提示和多个隐藏层（图3a）。我们首先考虑通过训练一个不重建音频数据的类似深度自动编码器来删除音频作为提示；CUAVE的表现下降了7.7%，AVLetters的表现下降了14.3%。接下来，我们训练了一个只有视频的浅自动编码器和一个隐藏层来重建音频和视频[10]；CUAVE的性能下降了2.1%，AVLetters的性能下降了5.0%。因此，作为提示和深度的音频都是仅视频深度自动编码器表现良好的重要因素。

我们还比较了使用双模DBN而不将其作为自动编码器进行训练的性能。在仅存在一种模态的情况下，我们使用与双峰深自动编码器相同的方法，将缺席模态设置为零。[11] 双模DBN在交叉模态和共享表示任务中表现更差，并且没有显示McGurk效应。它在多模式融合任务上的表现相当。[12]

[9]对于双峰深度自动编码器，我们在计算共享表示时将缺席模态的值设置为零，这与特征学习阶段一致。

[10]单个隐藏层将视频作为输入并重建音频和视频。

[11]我们也尝试交替使用Gibbs取样来获得后验，但结果更差。

[12]对于仅视频设置，双峰DBN在CUAVE数据集上执行率下降4.9%，在AVLetters数据集上执行率下降5.0%。它是“听见”任务的偶然机会，并且在“看到听到”时获得了28.1%。

5. 相关工作

虽然我们提出了用于多模态学习的神经网络的特殊情况，但我们注意到先前的视听语音识别工作（Duchnowski等人。，1994; Yuhas等。，1989; 梅尔等人。，1996; Bregler和Konig，1994）还探讨了神经网络的使用。Yuhas等。（1989）训练神经网络预测给定视觉输入的听觉信号。当他们将预测的听觉信号（来自使用视觉输入的网络）与嘈杂的听觉信号相结合时，他们在嘈杂的环境中表现出改善的性能。Duchnowski等人。（1994）和梅尔等人。（1996训练单独的网络来模拟音素和视位，并将预测结合在语音层以预测口语音素。

与这些方法相比，我们使用隐藏单元来构建数据的新表示。此外，我们不对音素或视位进行建模，这需要昂贵的标签工作。最后，我们通过对学习的浅表示中的相关性建模来构建深度双峰表示。

6. 讨论

手工工程任务特定的功能通常很困难且耗时。例如，目前尚不清楚唇读的适当特征是什么（仅视觉数据）。对于多模态数据，这种困难更加明显，因为这些特征必须涉及多个数据源。在这项工作中，我们展示了如何将深度学习应用于发现多模态特征的这一具有挑战性的任务。

致谢

我们感谢克莱姆森大学提供CUAVE数据集和萨里大学提供的AVLet-ters2数据集。我们还要感谢Quoc Le，Andrew Saxe，Andrew Maas和Adam Coates进行富有洞察力的讨论，感谢匿名审稿人提供了有用的评论。这项工作得到DARPA深度学习计划的支持，合同号为FA8650-10-C-7020。

参考

Bregler，C。和Konig，Y。“Eigenlips”用于强大的语音识别。在ICASSP，1994年。

Caruana，R。多任务学习。机器学习，28（1）：41-75,1997。

Cox，S.，Harvey，R.，Lan，Y。和Newman，J。多音调唇读的挑战。在听觉 - 视觉语音处理国际会议上，2008年。

Dalal，N。和Triggs，B。用于人体检测的定向梯度的直方图。在CVPR，2005年。

Duchnowski，P.，Meier，U。和Waibel，A。见到我，听我说：集成自动语音识别和唇读。在ICSLP，pp.547-550,1994。

Fisher，W.，Doddington，G。和Marshall，Goudie。该DARPA语音识别研究数据库：规范和状态。在DARPA语音识别研讨会，第249-249页，1986年。

Gurban，M。和Thiran，JP用于视听语音识别的信息理论特征提取。IEEE Trans。在Sig上。Proc。，57（12）：4765-4776,2009。

Hardoon，David R.，Szedmak，Sandor R.和Shawe-taylor，John R.典型相关分析：应用于学习方法的概述。Neural Computation，16：2639-2664,2004。

Hinton，G。通过最小化对比差异来培训专家产品。神经计算，2002。

Hinton，G。和Salakhutdinov，R。用神经网络减少数据的维数。Science，313（5786）：504-507,2006。

Lee，H.，Ekanadham，C。和Ng，A。Sparse视觉区域V2的深信念网模型。在NIPS，2007年。

Lucey，P。和Sridharan，S.Patch为基础的视觉语音表示。在HCSNet关于在人机交互中使用视觉的研讨会，2006年。

Matthews，I.，Cootes，TF，Bangham，JA和Cox，S。提取唇部的视觉特征。PAMI，24：198-213,2002。

McGurk，H。和MacDonald，J。听到嘴唇和看到声音。Nature，264（5588）：746-748,1976。

Meier，U。，Hu ¨rst，W。和Duchnowski，P。Adaptive Bimodal Sensor Fusion For Automatic Speechreading。在ICASSP，pp.833-836,1996。

Papandreou，G.，Katsamanis，A.，Pitsikalis，V。和Maragos，P。多模式融合和学习，具有不确定的特征，应用于视听语音识别。在MMSP，第264-267页，2007年。

Papandreou，G.，Katsamanis，A.，Pitsikalis，V。和Maragos，P。自适应多模式融合，通过不确定性补偿应用于视听语音识别。IEEE TASLP，17（3）：423-435,2009。

Patterson，E.，Gurbuz，S.，Tufekci，Z。和Gowdy，J。CUAVE：一种用于多模式人机界面研究的新型视听数据库。2：2017-2020,2002。

Pitsikalis，V.，Katsamanis，A.，Papandreou，G。和Maragos，P。自适应多模态融合的不确定性补偿。在ICSLP，第2458-2461页，2006年。

Potamianos，G.，Neti，C.，Luettin，J。和Matthews，I.视听自动语音识别：概述。在视觉和视听语音处理中的问题。麻省理工学院出版社，2004。

Salakhutdinov，R。和Hinton，G。Semantic hashing。IJAR，50（7）：969-978,2009。

Summerfield，Q。唇读和视听语音感知。跨。R. Soc。Lond。，pp.71-78,1992。

Vincent，P.，Larochelle，H.，Bengio，Y。和Manzagol，PA使用去噪自动编码器提取和组合强大的功能。在ICML中，第1096-1103页。ACM，2008。

Yuhas，BP，Goldstein，MH和Sejnowski，TJ使用神经网络整合声学和视觉语音信号。IEEE通讯杂志，第65-71页，1989年。

Zhao，G。和Barnard，M。使用当地时空描述符进行Lipreading。IEEE Transactions on Multimedia，11（7）：1254-1265,2009。