There are several interesting extensions to consider: first, the attention mechanism that provides the alignment is unconstrained, yet in fact always must move monotonically from left to right. This monotonicity could be incorporated as a soft or hard constraint; second, the sequence to sequence model is used in batch mode – decoding a sentence given the entire corresponding lip sequence. Instead, a more on-line architecture could be used, where the decoder does not have access to the part of the lip sequence in the future; finally, it is possible that research of this type could discern important discriminative cues that are beneficial for teaching lip reading to the deaf or hard of hearing.

内容整理

问题：不像先前的工作只是注重识别有限数量的单词和短语，我们将唇读作为一个开放世界的问题——在无限制的自然语言句子和随便选出的视频中进行识别。

贡献：

1. 构建了一个“Watch，Listen，Attend and Spell”（WLAS）模型，可以从视频中将嘴部动作转化为字符
2. 一个课程学习的策略加速训练
3. 建立了一个“Lip Reading Sentences”数据库，包括100000个来此英国电视的自然语句

介绍

问题：同唇形多意思

应用：誊写和重配音档案无声电影，多说话者的同声传译

该模型用了序列到序列（包含注意力机制的编码器-解码器）翻译体系结构，早被用于语音识别和机器翻译。

相关工作

**唇读：**一些论文用CNN预测音素和视素，为了识别整句，用LSTM分类器在discrete cosine transform (DCT) and deep bottleneck features (DBF)上。LSTM和HOG识别短语。视频包含单词少，最多60词。最近用LSTM和CTC来计算标签，但是预测句子不长。

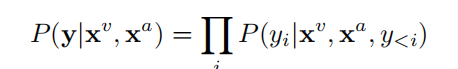
**语音视觉语音识别：**问题与唇读类似，前馈DNN进行音素分类，使用非公共视听集。同时用HMM在预处理视觉特征上。CNN用于预处理。

**语音识别：**先前工作主要分为两种类别

1. CTC，预测逐帧标签，做视频与预测的对齐，输出标签不相互制约。
2. 序列到序列的模型，读取所有输入序列来预测句子。最近的成功用了预定的采样和注意机制。

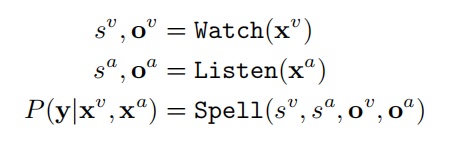
体系结构（Watch，Listen，Attend，Spell）

用图片和音频向量作输入



Encoder-decoder

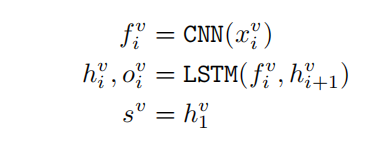
该模型包括三个关键部分，，图像编码器Watch，声音编码器Listen，字符解码器Spell。每个编码器将输入变为固定维度的状态向量s和编码器的输入序列o，解码器提取状态和注意力向量从编码器和编码器在输出上产生的概率分布。



这三种模型共同训练。

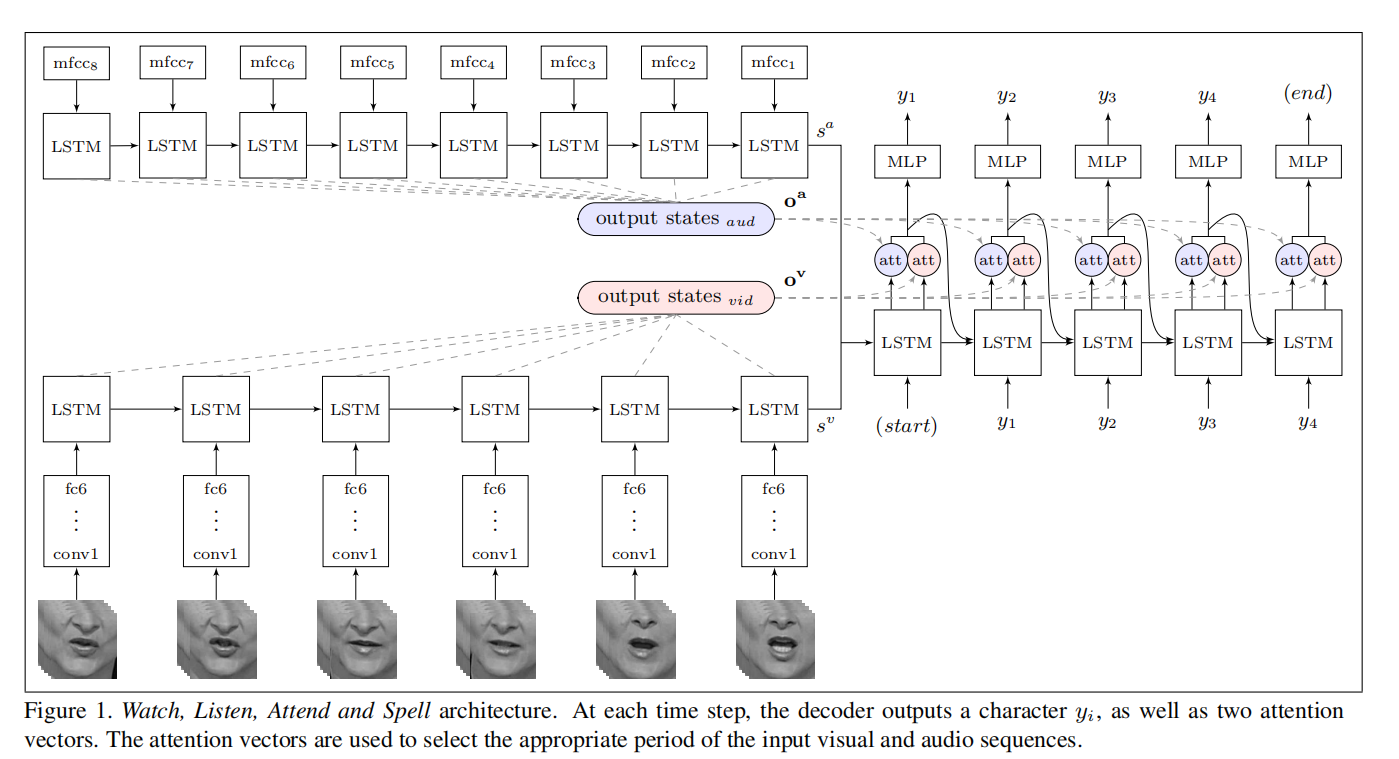
Watch：图像编码器

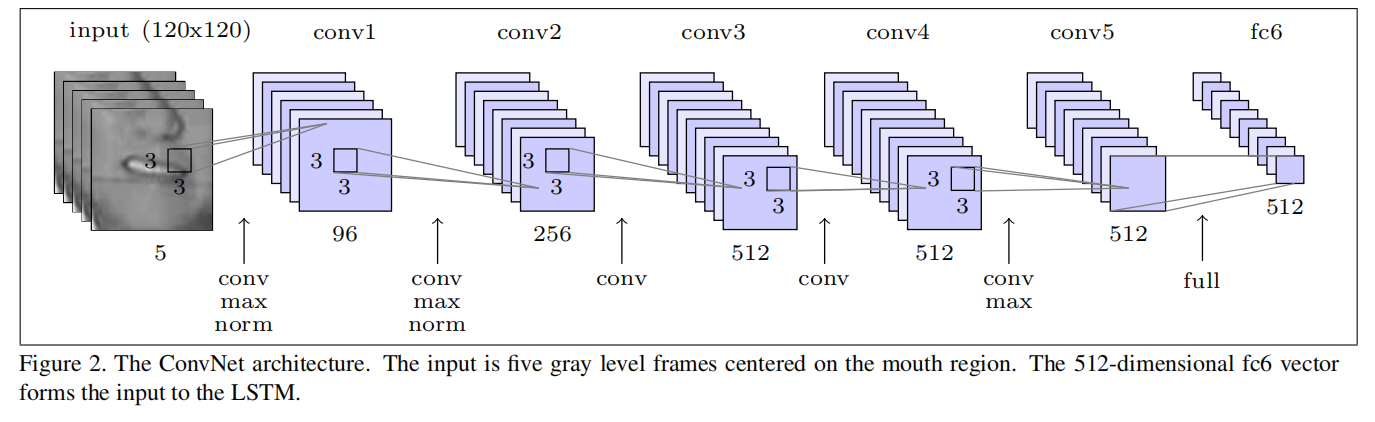
包括卷积模块，循环模块



卷积模块基于VGG=M模型，内存高效，快速训练，的题分类性能。

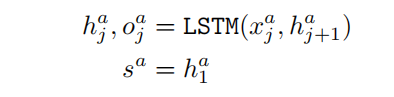
LSTM消耗输入特征，产生固定维度向量sv，每个时间步会有一个输出o，通过相反的时间序列从输入中提取信息。





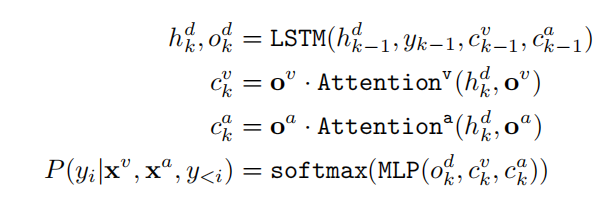
Listen:音频编码器

LSTM直接使用13维度的MFCC特征以相反的时间序列



Spell:字符解码器

该模型基于LSTM传感器，加入了双注意力机制。在每个输出步k，解码器LSTM产生LSTM状态hkd，和输出向量okd，从先前步的上下文向量ck-1v，ck-1a，输出yk-1和编码状态hk-1d。注意力向量由注意力机制产生。注意力基于视听并可以处理由不同取样率造成的异步输出。注意力向量与输出向量结合产生上下文向量，这个向量浓缩了产生下步输出的信息。输出字符的概率分布由MLP和softmax处理后得出

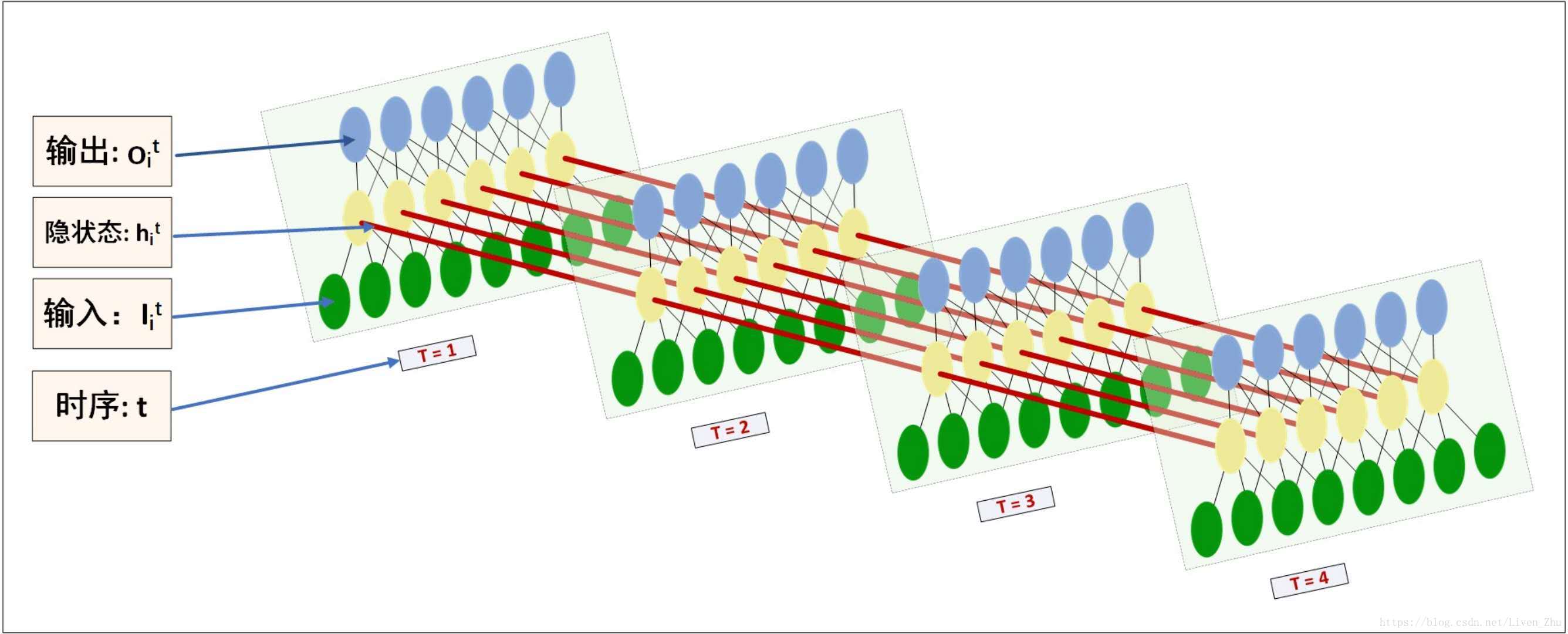


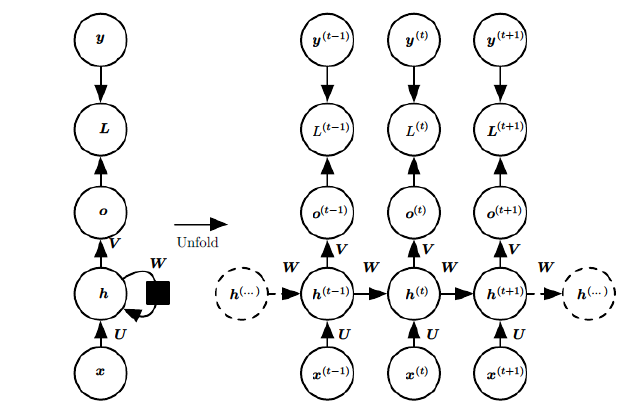
K=1时用最终编码器状态sl和sa作为先前解码器状态的输入

插入机器学习方面的知识

RNN

RNN网络总体



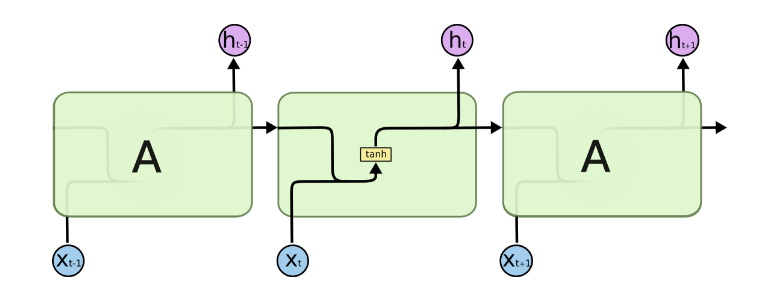


RNN网络每一时序的运行

现在将RNN的运算部分包装起来，并把相应变量修改为惯例字母。其实普通RNN内部并非为多层网络，仅仅是一个tanh层。



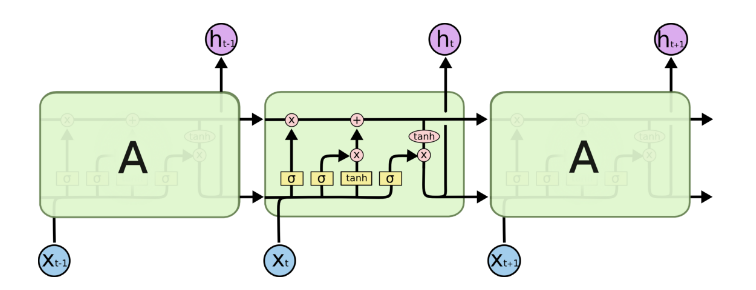
这个图按时间轴展开，即为常见的RNN结构图



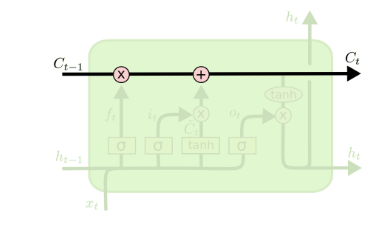
图中可以很清晰看出在隐藏状态h(t)由x(t)和h(t−1)得到。得到h(t)后一方面用于当前层的模型损失计算，另一方面用于计算下一层的h(t+1)。

LSTM

由于RNN梯度消失的问题，大牛们对于序列索引位置t的隐藏结构做了改进，这样的特殊RNN就是我们的LSTM。由于LSTM有很多的变种，这里我们以最常见的LSTM为例讲述。LSTM的结构如下图：



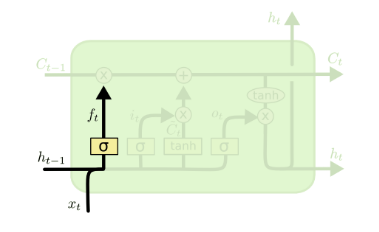
从上图中可以看出，在每个序列索引位置t时刻向前传播的除了和RNN一样的隐藏状态h(t)h(t)，还多了另一个隐藏状态，如图中上面的长横线。这个隐藏状态我们一般称为细胞状态(Cell State)，记为C(t)。如下图所示：



除了细胞状态，LSTM图中还有很多门控结构(Gate)。LSTM在在每个序列索引位置t的门一般包括遗忘门，输入门和输出门三种。下面我们就来研究上图中LSTM的遗忘门，输入门和输出门以及细胞状态。

1. 遗忘门

遗忘门（forget gate）顾名思义，是控制是否遗忘的，在LSTM中即以一定的概率控制是否遗忘上一层的隐藏细胞状态。遗忘门子结构如下图所示：

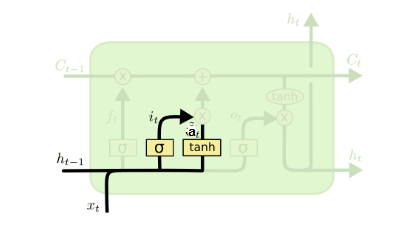


图中输入的有上一序列的隐藏状态h(t−1)和本序列数据x(t)，通过一个激活函数，一般是sigmoid，得到遗忘门的输出f(t)。由于sigmoid的输出f(t)在[0,1]之间，因此这里的输出f^{(t)}代表了遗忘上一层隐藏细胞状态的概率。用数学表达式即为：

= σ(++)

其中Wf,Uf,bf为线性关系的系数和偏倚，和RNN中的类似。σ为sigmoid激活函数。

1. 输入门



从图中可以看到输入门由两部分组成，第一部分使用了sigmoid激活函数，输出为i(t),第二部分使用了tanh激活函数，输出为a(t), 两者的结果后面会相乘再去更新细胞状态。用数学表达式即为：

i(t)=σ(Wih(t−1)+Uix(t)+bi)

a(t)=tanh(Wah(t−1)+Uax(t)+ba)

ps：tanh是双曲正切，上减下加

1. 细胞状态更新

细胞状态C(t)由两部分组成，第一部分是C(t−1)和遗忘门输出f(t)的乘积，第二部分是输入门的i(t)和a(t)的乘积，即：

C(t)=C(t−1)⊙f(t)+i(t)⊙a(t)

其中，⊙为Hadamard积

1. 输出门

从图中可以看出，隐藏状态h(t)的更新由两部分组成，第一部分是o(t), 它由上一序列的隐藏状态h(t−1)和本序列数据x(t)，以及激活函数sigmoid得到，第二部分由隐藏状态C(t)和tanh激活函数组成, 即：

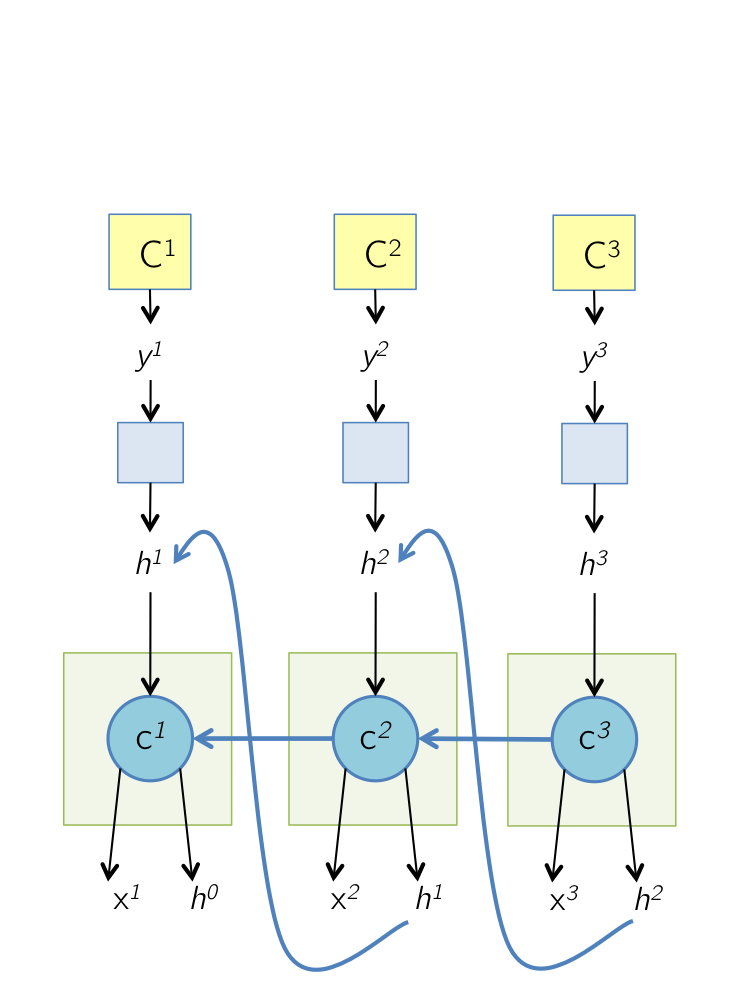
o(t)=σ(Woh(t−1)+Uox(t)+bo)

h(t)=o(t)⊙tanh(C(t))

1. 更新当前序列索引预测输出

ŷ (t)=σ(Vh(t)+c)

反向传播LSTM



通过梯度下降法进行反向传播

训练策略

课程学习

LSTM在大的世间步长的情况下收敛速度很慢，因为解码器在开始很难从所有输入步中提取出相关的信息。

引入了一个新的策略，开始学习时只用单个单词样例，之后慢慢增加序列长度。更快，更少过拟合，因为这是种自然增加数据的方式。

预定抽样

当训练循环神经网络时，常常用先前时间步的ground truth作为下一个时间步的输入，可以使模型学习目标tokens的语言模型。然而由于干涉，先前步的ground-truth不能用，会低效，因为模型不能处理某些时间步的坏预测。

于是使用预定抽样法，训练时，随机从先前输出取样，并不全用ground-truth，ground-truth在短子句中用，整句需要随着步数增加从0-0.25增加输出取样的概率。超过0.25便不能稳定学习了。

多模型训练

多模型输入常常被一个模型所主导，本例中音频信号主导。为了防止这个，对于每个示例，在训练时间统一选择以下输入类型之一：（1）仅音频; （2）只有嘴唇; （3）音频和嘴唇。

很少比例的句子能够与面部表情对应。用其他句子作为补充提升音频识别的效能并且构建一个丰富的语言模型来提升泛化性。

用有噪声的音频训练

开始用纯净的音频作训练，之后加入信噪比10dB和0dB的高斯白噪声分别训练。

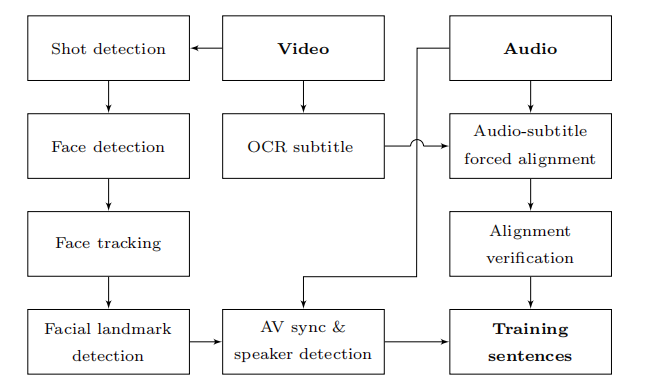
实现

数据集

在本节中，我们描述了用于自动生成用于视听语音识别的大规模数据集的多级流水线，本文用了再2010到2016间的BBC节目的记录。

选择节目的特点：

1. 在新闻和访谈中有多个讲话者出现
2. 很少有屏幕切换，可以得到完整的语句和其相关的面部表情的对应



视频准备

首先，通过比较连续帧的颜色直方图来检测镜头边界，之后将基于HOG的人脸检测作用于视频的每一帧上。用KLT追踪器来检测同一个人的人脸，使用回归树集合从稀疏的像素强度子集中提取面部标志。

音频和文字准备

BBC视频的字幕和音频不是同步播放的。The Penn Phonetics Lab Forced Aligner被用于对齐字幕和音频信号。对齐时会有错误，因为转录不是逐字进行的，对齐后的标签需要用the commercial IBM Watson Speech to Text service做检查。

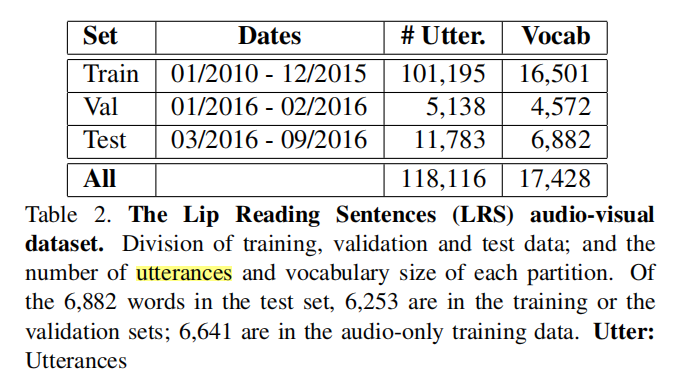
音频视频同步和讲话者检测

BBC音频和视频流会错同步大概1s，对面部和句子间的对应造成影响。用two-stream network 进行同步。同一网络也用于确定谁在视频中发言，如果是画外音，则拒绝剪辑。

句子抽取

在转录中，使用标点将视频划分为句子和短语。句子会被句号，逗号，问号划分。由于GPU内存限制，我们最多剪辑100个字符或者10s，对单词数量没有限制。

训练，验证，测试集根据播报时间来划分。



**只有音频的数据**

除了音频和视觉的数据集，本文还准备了只有音频的辅助训练集。这些也是从BBC中收录的，但没有对应的面部动作。这个集只用于训练，不用于测试。

