深蓝学院
shenlanxueyuan.com

音频语音与语言处理研究组
SLP
Audio, Speech and Language Processing Group NPU

语音合成：从入门到精通

第四讲：基于序列到序列的声学模型

主讲人 阳 珊

西北工业大学语音与语言处理实验室
syang@nwpu-aslp.org

内容提要

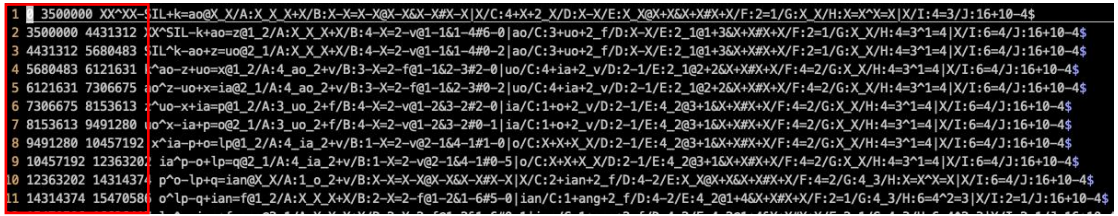
内容提要

概述-文本分析回顾

文本分析过程

- 分词、注音、词性标注、韵律预测等
- 将标准化的抄本表示为音素级别的文本特征
- 文本特征的长度M仅与抄本本身有关，与语音时长无关

概述-文本分析回顾



文本分析过程

示例：欢迎大家参加深蓝学院语音合成课程

声学模型建模目标

建模文本序列X 和语音特征Y之间的映射关系

文本$X = \{x_1, \ldots, x_M\}$, M帧音素

语音$Y = \{y_1, \ldots, y_N\}$， N帧声学特征

建模目标：

$$\hat{\Lambda} = \arg\max_{\Lambda} p(Y|X, \Lambda)$$

文本特征

语音特征

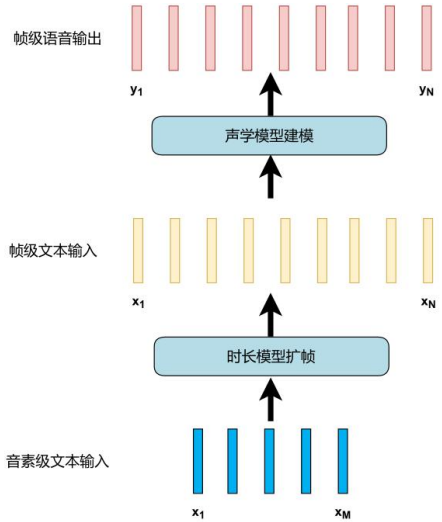
声学模型参数

概述-声学模型回顾

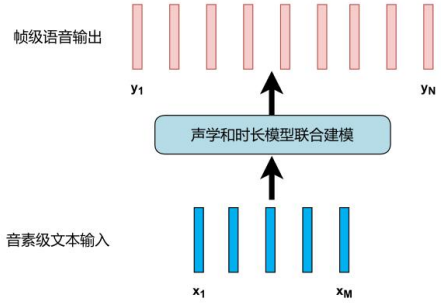
声学模型建模过程

- 帧级声学模型
  - 时长模型 (M帧->N帧)
  - N帧->N帧
- 序列到序列声学模型
  - M帧->N帧

$$\hat{\Lambda} = \arg\max_{\Lambda} p(Y|X,\Lambda)$$



(a) 帧级声学模型



(b) 序列到序列声学模型

概述-声学模型回顾

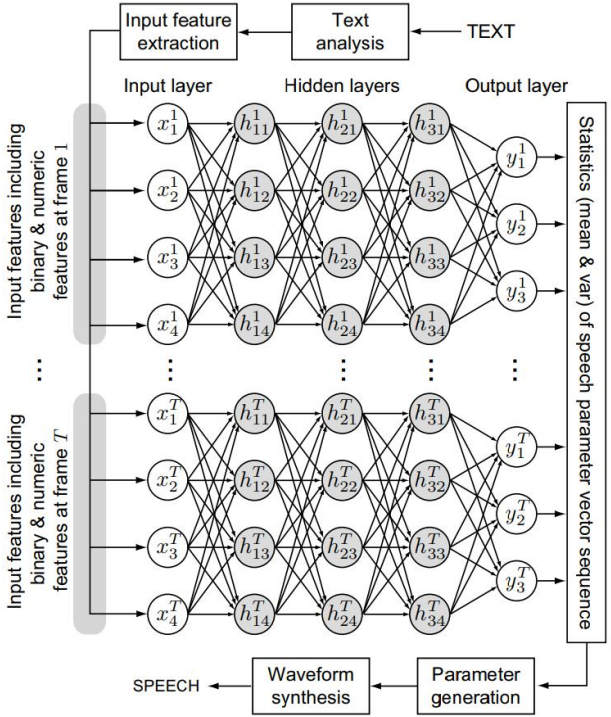
帧级声学模型建模过程

缺点

- 声学模型文本输入长度与语音序列**长度必须一致**。
- 由于文本本身不含有时序信息，因此帧级声学模型通常需要**额外的时长模型**给文本引入时间信息。
- 时长模型的误差影响后续声学模型的效果：**误差累积**问题。

优点

- 由于帧级声学模型是逐帧对应的，其**稳定性**很好。



2013. Heiga et.al. - STATISTICAL PARAMETRIC SPEECH SYNTHESIS USING DEEP NEURAL NETWORKS

内容提要

序列到序列模型

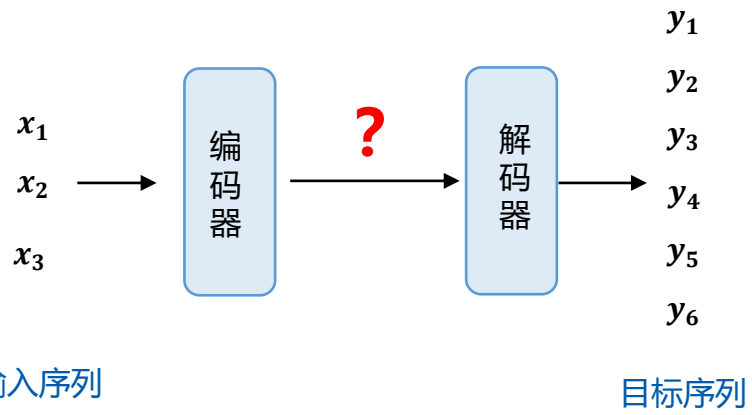
概述

序列到序列模型的核心思想：在学习不等长序列映射关系的同时，建模序列间的时序关系，从而达到序列到序列建模的目的。

典型的序列到序列模型包含一个编码器网络和一个自回归解码器网络，

- 编码器：建模输入序列的内在联系
- 解码器：生成目标序列

核心问题：如何对齐两个不等长序列



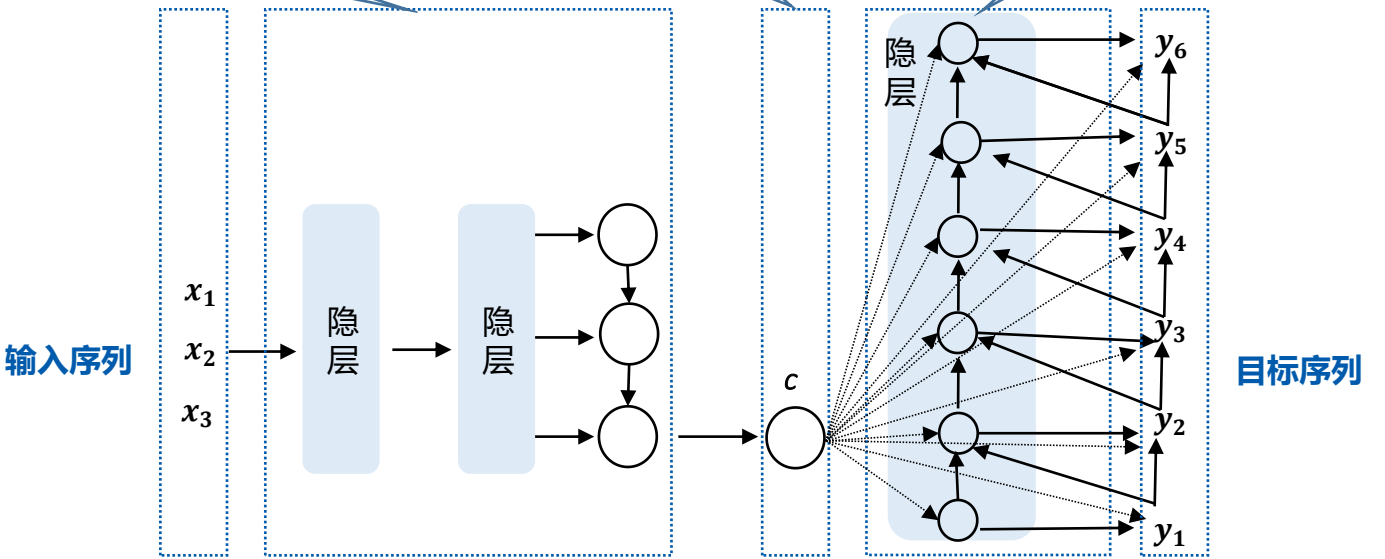
2013. Sutskever et.al. - Sequence to sequence learning with neural networks

序列到序列模型结构

编码器Encoder
作用：学习输入
将其编码成内容向量

内容向量 C
作用：连接编码器和解码器的中间状态向量

解码器Decoder
作用：通过对状态向量的学习来进行输出对应的序列



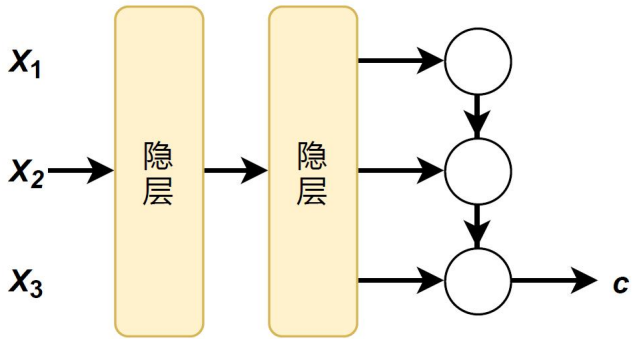
2013. Sutskever et.al. - Sequence to sequence learning with neural networks

序列到序列模型

编码器

学习输入序列X的固定长度隐层表示向量 c $\quad c = \text{Encoder}(X), c \in \mathbb{R}^D$

- 通常采用RNN结构，其最后一个时刻的输出作为输入序列的固定长度表示
- 固定长度的隐层向量可以看做是整个输入序列X中所有信息的压缩表示
- 该表示（内容向量）直接用来控制解码器的生成过程



2013. Sutskever et.al. - Sequence to sequence learning with neural networks

序列到序列模型

自回归模型

- 经典的生成式模型，建模目标数据X的分布$P(x)$
- 通过相邻点或前一个点来直接建模下一个点的分布，从而获得整个数据X的分布
- 给定初始数据点，可以无限制（长度不确定）、无约束（内容不确定）的生成无穷个符合分布$P(x)$的数据点 → 无条件自回归过程
- 针对语音来讲，符合语音的单调时序生成特点（语音的每一个采样点）

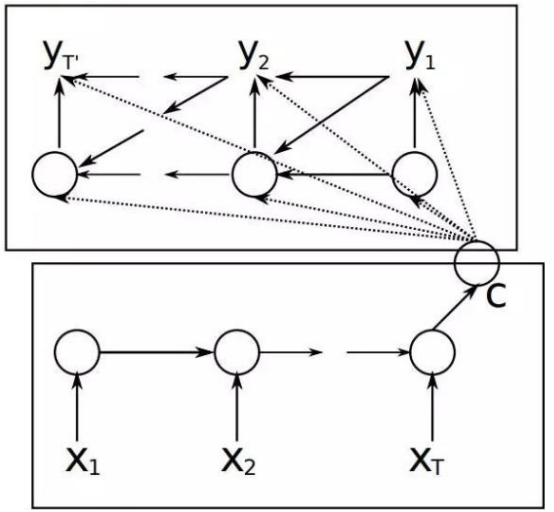
$$p_{data}(x) = \prod_{i=1}^{n} p(x_i | x_1, x_2, \ldots, x_{i-1})$$

序列到序列模型

解码器

- 自回归地生成目标序列每一帧
  - 条件自回归模型 $p(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{t=1}^{N} p\left(\boldsymbol{y}_t \mid \boldsymbol{c}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}\right)$
  - 利用解码器的初始状态s_0和编码器的内容向量c生成目标序列第一帧
  - 通过自回归过程循环迭代生成目标序列Y的每一帧，直到合成到EOS为止
  - 每一个时刻都需要使用编码器学习到的内容向量

$$\hat{\boldsymbol{y}}_t = \text{Decoder}(\boldsymbol{c}, \boldsymbol{y}_{t-1}, \boldsymbol{s}_t)$$



2013. Sutskever et.al. - Sequence to sequence learning with neural networks

序列到序列模型

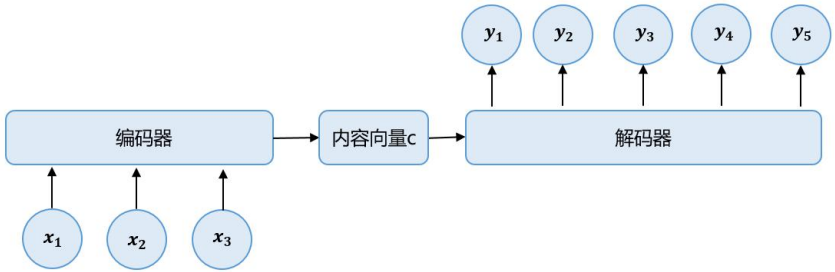
经典的序列到序列模型

- 给定输入序列X和目标序列Y，最大化给定输入X下生成Y的概率
  - 条件自回归 $p(\boldsymbol{Y}|\boldsymbol{X}) = \prod_{t=1}^{N} p\left(\boldsymbol{y}_t \mid \boldsymbol{c}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{t-1}\right)$
- 编码器：学习输入序列X的D维的固定长度隐层表示向量c

$$\boldsymbol{c} = \text{Encoder}(X), \boldsymbol{c} \in \mathbb{R}^D$$

- 解码器：根据内容向量c 和t时刻的隐状态s，自回归地生成目标序列每一帧

$$\hat{\boldsymbol{y}}_t = \text{Decoder}(\boldsymbol{c}, \boldsymbol{y}_{t-1}, \boldsymbol{s}_t)$$



2013. Sutskever et.al. - Sequence to sequence learning with neural networks

序列到序列模型

经典的序列到序列模型

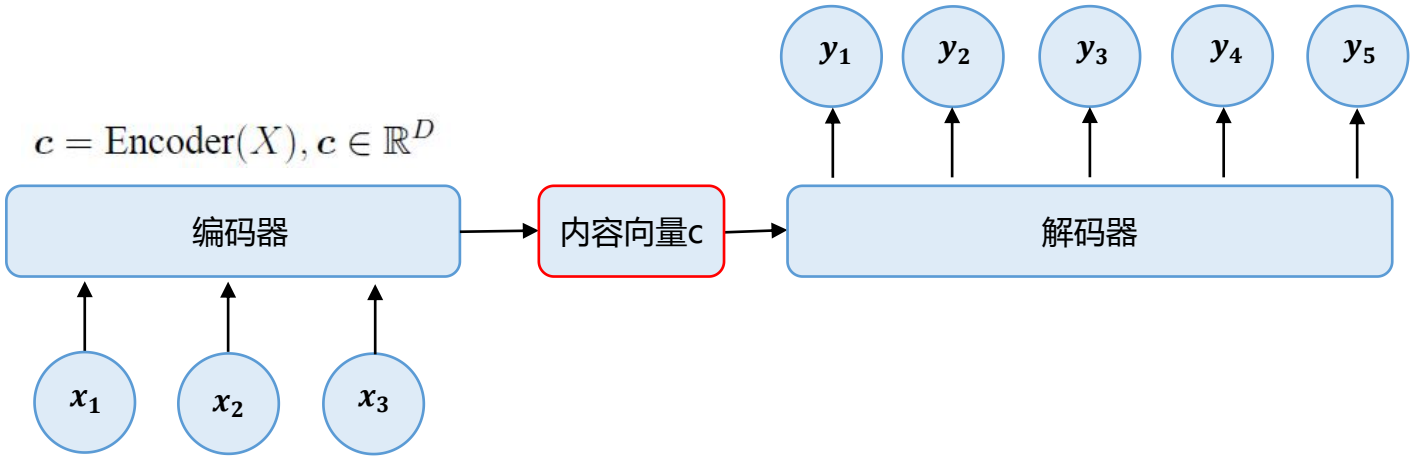
- 优点
  - 不再依赖额外的时长模型，直接通过解码器的自回归生成过程来获取变长的目标序列；
  - 避免了时长模型和声学模型级联时的累积误差；
  - 简化了语音合成的流程。

- 缺点
  - 编码器对M帧输入序列X压缩成一帧向量，过度压缩存在信息损失；
  - 解码器的每一个时刻使用的是同一个内容向量，无法考虑输入序列中不同部分对当前时刻贡献，例如语音合成中一段语音的内容主要是由某一个字所决定的。

2013. Sutskever et.al. - Sequence to sequence learning with neural networks

经典的序列到序列模型

- 解决方法
  - 编码器不再压缩输入序列，并且解码器生成过程中的不同时刻使用特定的内容向量
- 固定的内容向量
  - 在编码器学习输入序列X的隐层表示时，直接获得内容向量



基于注意力机制的序列到序列模型

- 变化的内容向量
- 核心问题是求解对齐矩阵

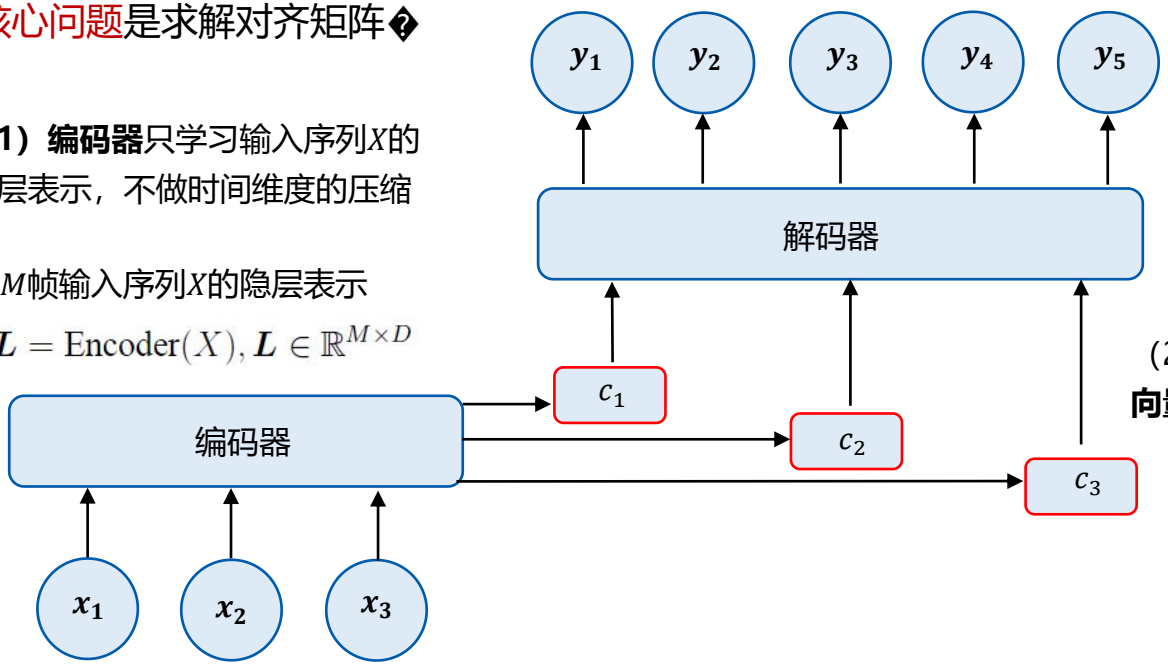
（1）编码器只学习输入序列X的隐层表示，不做时间维度的压缩

- M帧输入序列X的隐层表示

$$\boldsymbol{L} = \text{Encoder}(X), \boldsymbol{L} \in \mathbb{R}^{M \times D}$$

（3）让**解码器**每一个时刻关注特定的内容向量

$$\boldsymbol{y_t} = f(\boldsymbol{c_t}, \boldsymbol{h_t})$$



（2）在解码器的t时刻提供**不同的内容向量**c_t，其中c_t为L中每一帧的加权和

$$\boldsymbol{c_t} = \sum_{s=1}^{M} \alpha_{ts} \boldsymbol{l_s}$$

对齐矩阵

注意力机制

- 目标
  - 求解对齐矩阵α，以获得每一个解码器时刻所对应的内容向量

$$\alpha_{ts} = \frac{\exp\left(\text{score}\left(\boldsymbol{h}_t, \boldsymbol{l}_s\right)\right)}{\sum_{s'=1}^{M} \exp\left(\text{score}\left(\boldsymbol{h}_t, \boldsymbol{l}_{s'}\right)\right)}$$

$$\boldsymbol{c}_t = \sum_{s=1}^{M} \alpha_{ts} \boldsymbol{l}_s$$

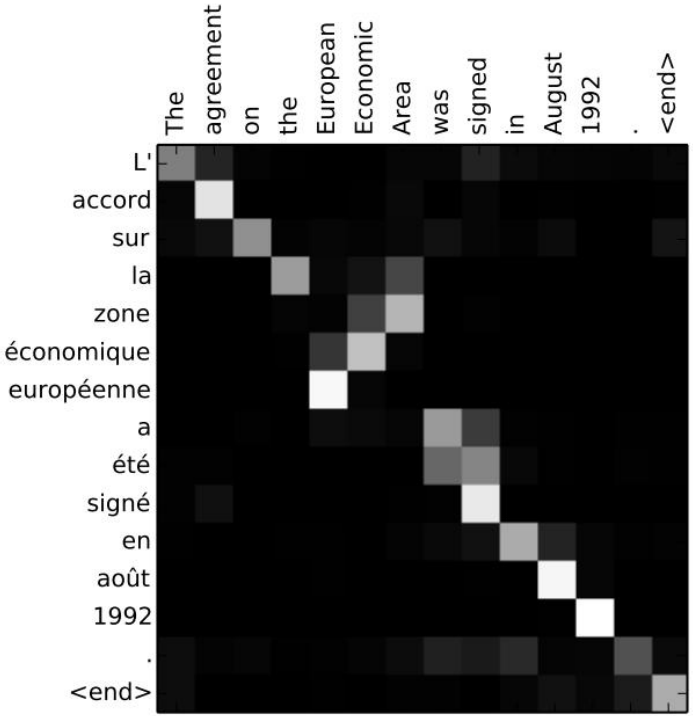
  - 给定目标序列Y及解码器的隐层输出$\boldsymbol{h}$，输入序列X的隐层表示 L
    - 计算每一个时刻 $\boldsymbol{h}_t$在每一帧 $\boldsymbol{l}_s$上的得分 （打分函数）

2014. Bahdanau et.al. - Neural machine translation by jointly learning to align and translate

序列到序列模型

注意力机制

- 对齐矩阵示例

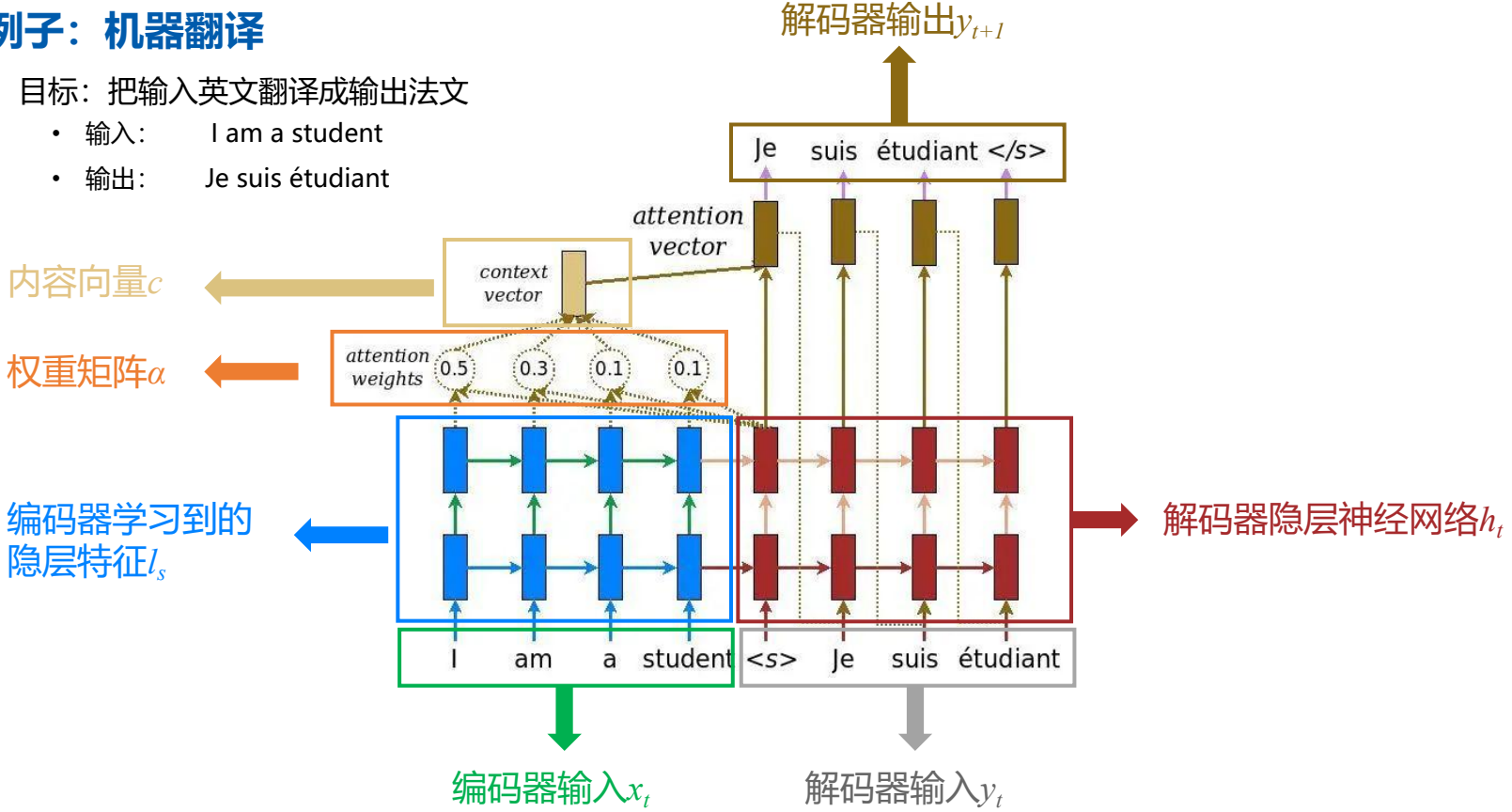


2014. Bahdanau et.al. - Neural machine translation by jointly learning to align and translate

基于注意力机制的序列到序列模型

例子：机器翻译

- 目标：把输入英文翻译成输出法文
  - 输入：　I am a student
  - 输出：　Je suis étudiant

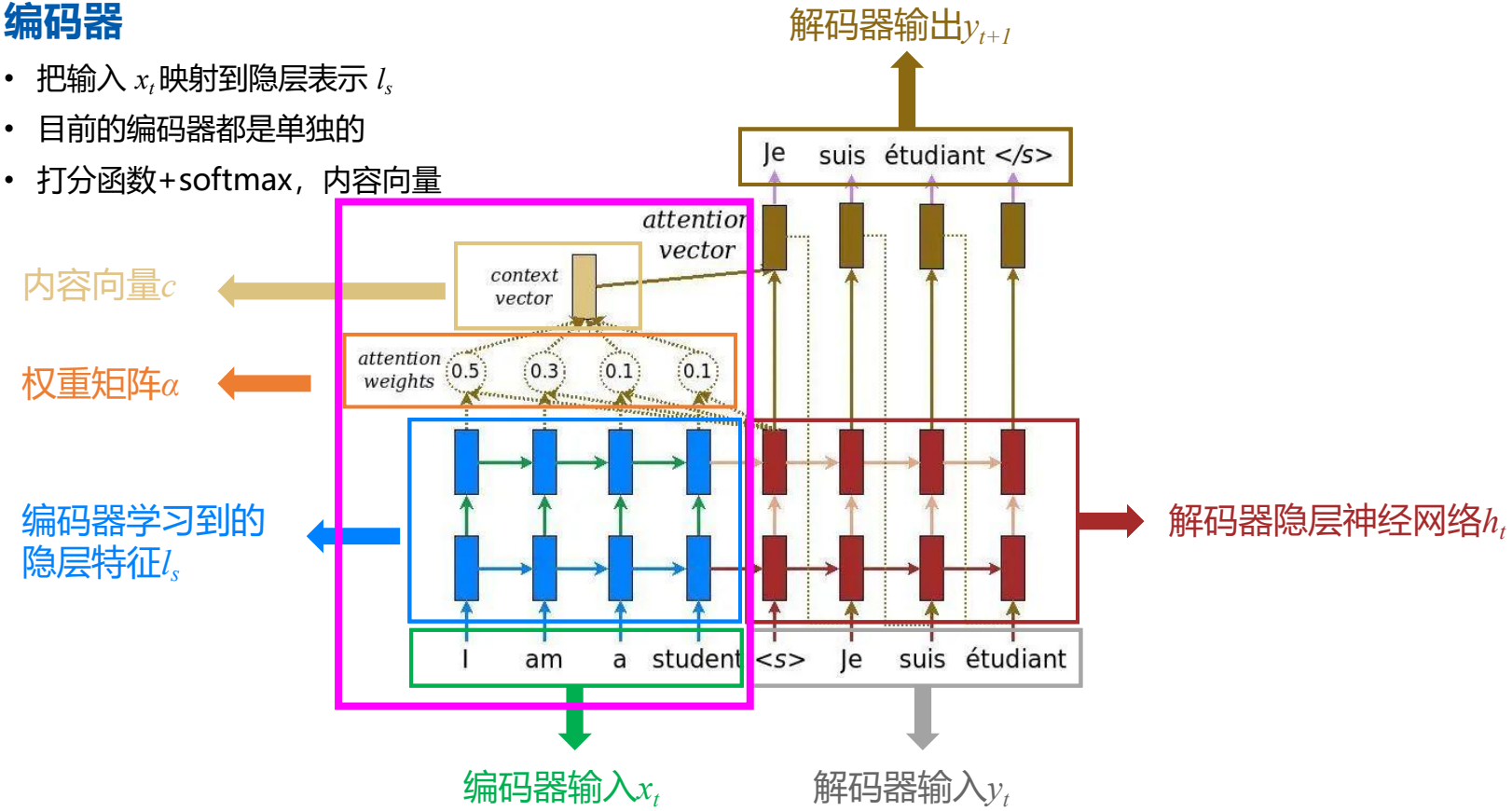


2014. Bahdanau et.al. - Neural machine translation by jointly learning to align and translate

基于注意力机制的序列到序列模型

编码器

- 把输入 x_t 映射到隐层表示 l_s
- 目前的编码器都是单独的
- 打分函数+softmax，内容向量

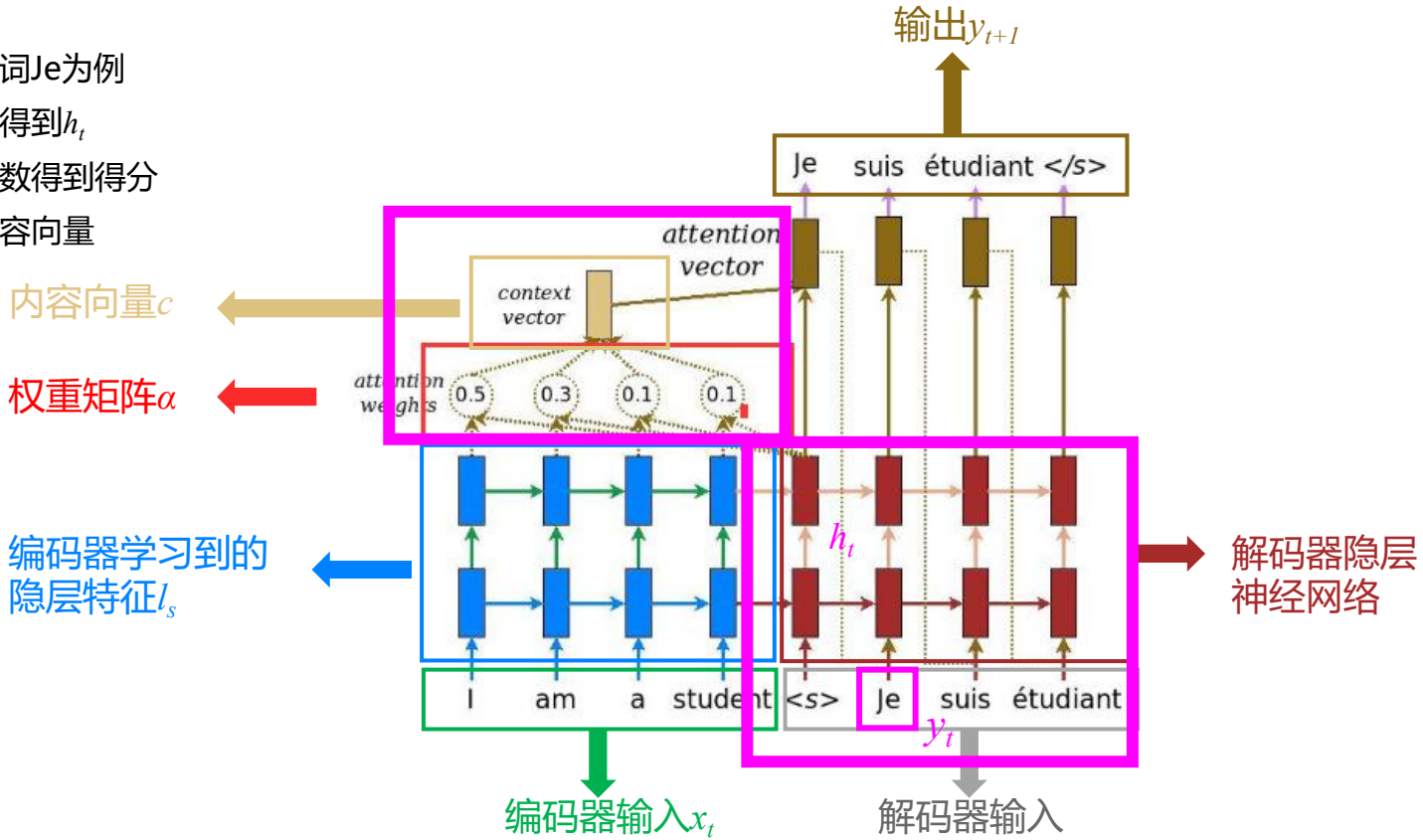


2014. Bahdanau et.al. - Neural machine translation by jointly learning to align and translate

基于注意力机制的序列到序列模型

解码器

- 以第一个单词Je为例
- 隐层变换，得到h_t
- 通过打分函数得到得分
- 相乘得到内容向量

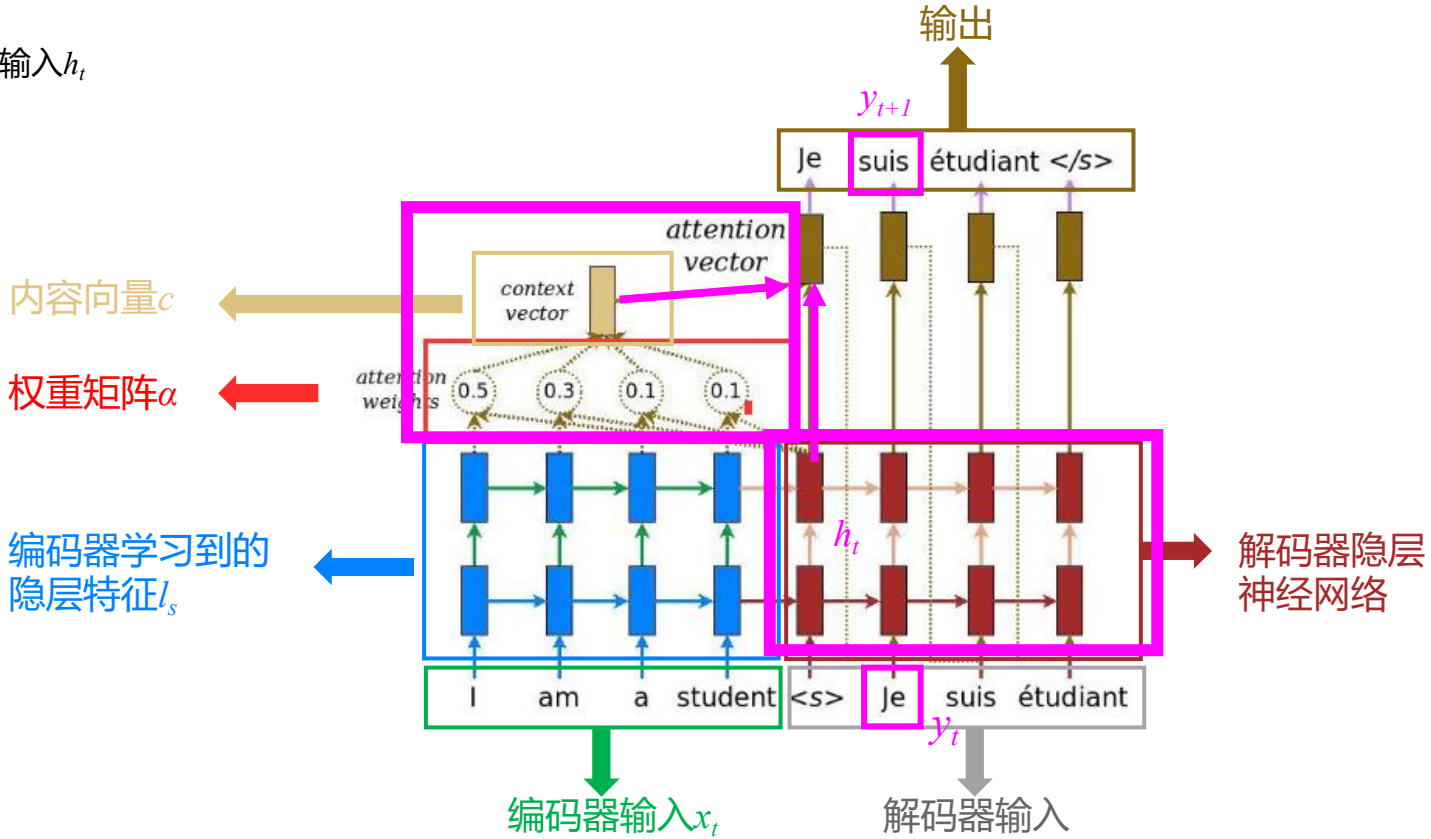


2014. Bahdanau et.al. - Neural machine translation by jointly learning to align and translate

基于注意力机制的序列到序列模型

解码器

- 解码器隐层输入h_t
- 内容向量$\boldsymbol{c}$
- → 输出 y_{t+1}

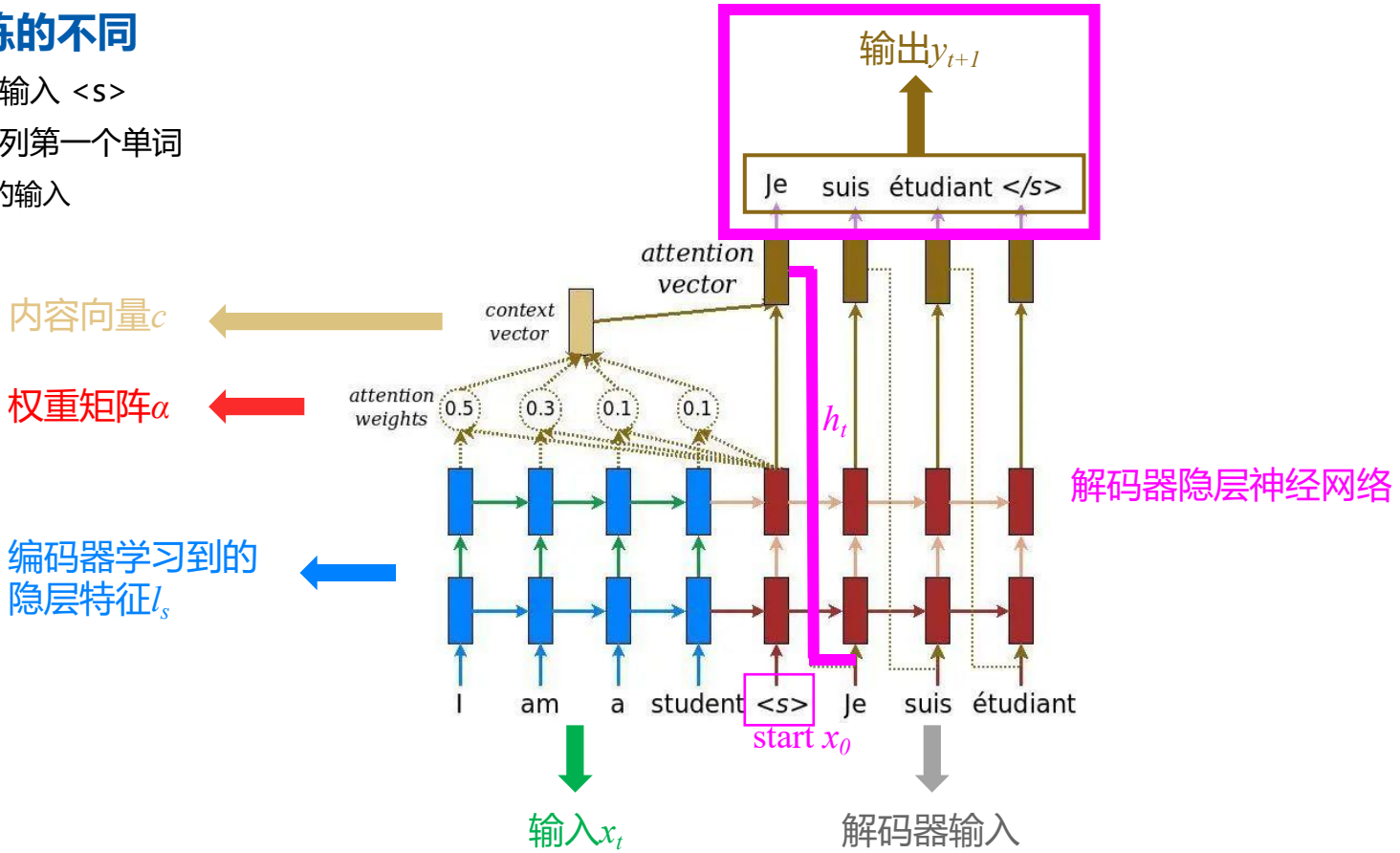


2014. Bahdanau et.al. - Neural machine translation by jointly learning to align and translate

基于注意力机制的序列到序列模型

合成与训练的不同

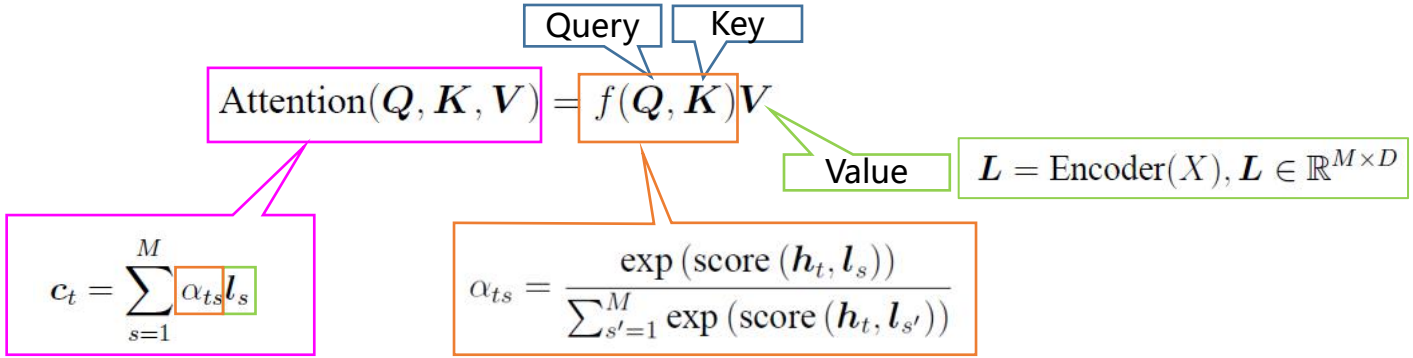
- 解码器初始输入 <s>
- 得到目标序列第一个单词
- 作为下一个的输入



2014. Bahdanau et.al. - Neural machine translation by jointly learning to align and translate

注意力机制概述

- 给定解码器t时刻的向量，求解权重矩阵α_t，从而获得其对应的内容向量c_t
  - Value：值向量，由编码器编码获得，矩阵大小$M * D_1$，M为输入序列长度， D_1为维度
  - Query：查询向量，一般为解码器隐层输出，其矩阵大小为$N \times D_2$，N为目标序列长度，D_2为维度
  - Key：键向量，作为计算权重矩阵的索引，一般由Value通过前馈网络获得，矩阵大小$D_2 \times M$，
  - 最后计算得到的内容向量矩阵C的大小为$N \times D_1$，其长度和目标序列一致



注意力机制代码讲解

https://github.com/tensorflow/tensorflow/blob/r1.12/tensorflow/contrib/seq2seq/python/ops/attention_wrapper.py

注意力机制分类

- Hard (stochastic) attention
  - 依靠概率采样编码器的一部分来计算内容向量
  - 蒙特卡洛采样方法来计算梯度
- Soft (deterministic) attention （主流）
  - 计算编码器表示中每一帧的确定性得分来获取内容向量
  - 直接利用反向传播来计算梯度

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = f(\boldsymbol{Q}, \boldsymbol{K})\boldsymbol{V}$$

Key

Query

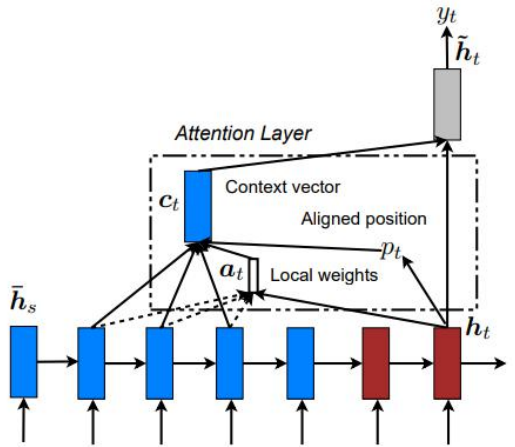
Value

2015. Kelvin et.al. - Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

注意力机制分类

- Global attention
  - 解码器每一个时刻的内容向量计算过程均考虑编码器的所有输入信息
  - 能够考虑输入序列X的所有信息
  - 序列X过长时计算量开销较大

$$\begin{aligned} \boldsymbol{a}_t(s) &= \text{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) \\ &= \frac{\exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'} \exp\left(\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)} \end{aligned}$$



2015. Luong et.al. - Effective Approaches to Attention-based Neural Machine Translation

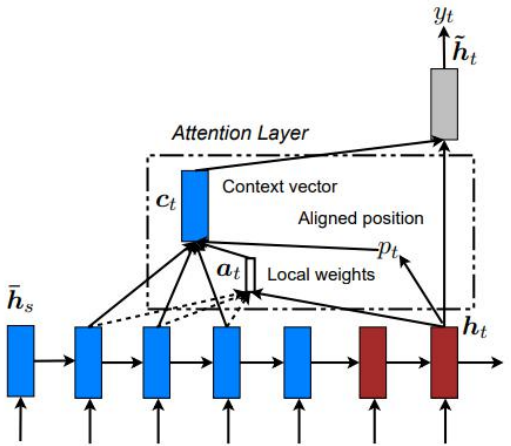
注意力机制分类

- Local attention
  - 解码器t时刻只考虑编码器输出的部分信息 $[p_t - D, p_t + D]$
  - 可以看做介于soft 和hard attention之间，更易训练

$$p_t = S \cdot \text{sigmoid}(\boldsymbol{v}_p^\top \tanh(\boldsymbol{W_p} \boldsymbol{h}_t))$$

$$\boldsymbol{a}_t(s) = \text{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) \exp\left(-\frac{(s - p_t)^2}{2\sigma^2}\right)$$

  - Etc. (GMM based)



2015. Luong et.al. - Effective Approaches to Attention-based Neural Machine Translation

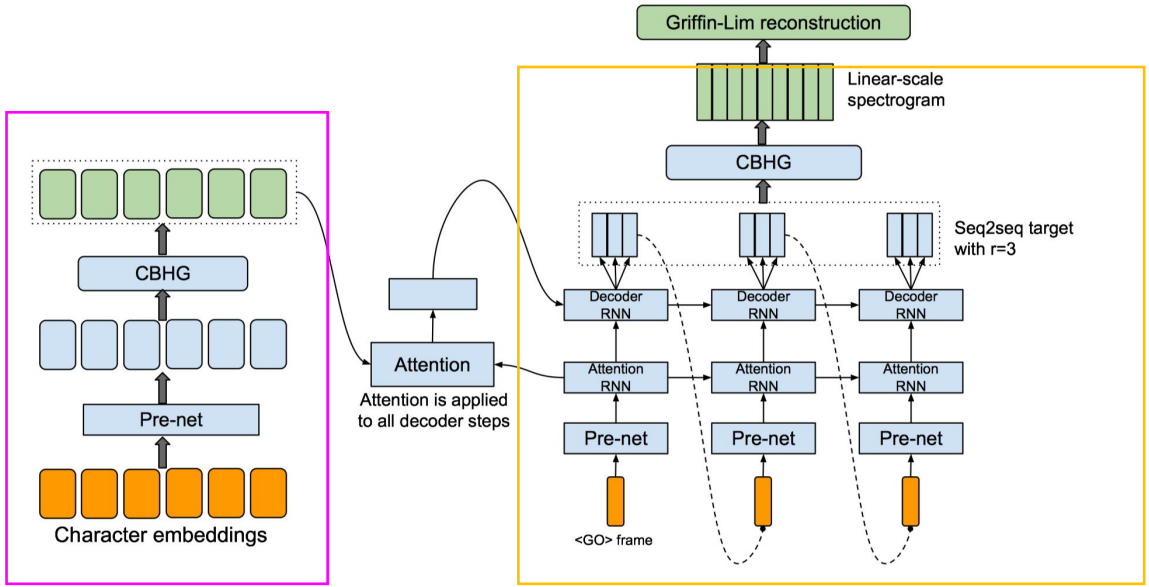
内容提要

序列到序列声学模型Tacotron

Tacotron模型框架结构

后处理：
Griffin-Lim声码器，把梅尔谱还原成语音信号

编码器：
字符编码作为输入，学习文本表示



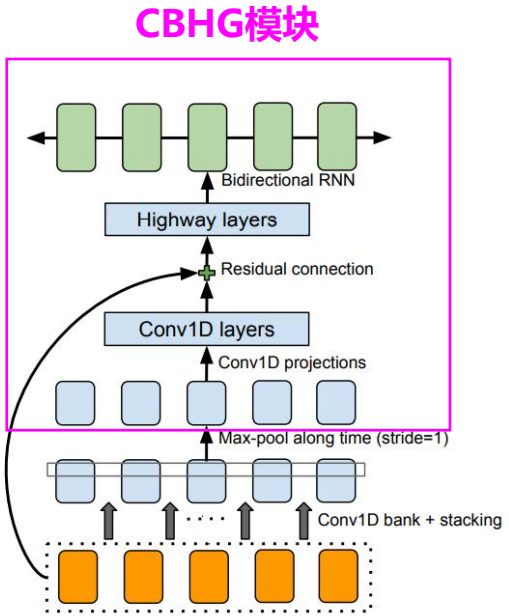
解码器：
自回归地生成语音的梅尔谱

2017. Wang et.al. - Tacotron: Towards end-to-end speech synthesis.

序列到序列声学模型Tacotron

编码器

- Conv bank
  - 目的：学习不同尺度context的文本表示
  - 组成：多个卷积核大小不一的卷积神经网络
- 输出拼接stacking
- Max-pool along time
- CBHG模块
  - 一维卷积
  - Highway network
  - 双向GRU

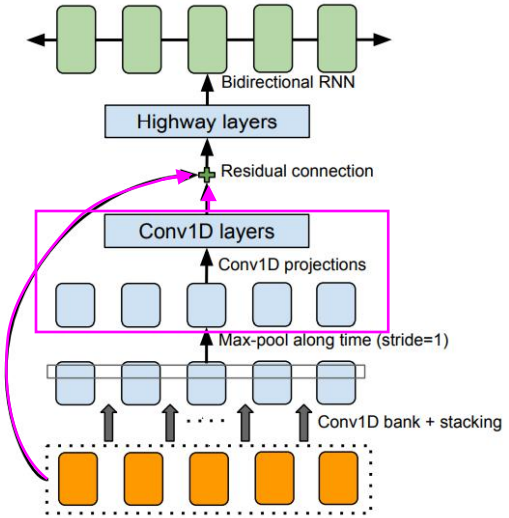


2017. Wang et.al. - Tacotron: Towards end-to-end speech synthesis.

序列到序列声学模型Tacotron

CBHG模块

- 一维卷积
  - 原因：多个卷积神经网络的输出信息维度过高
  - 目的：压缩融合多个CNN的高维输出信息
- 残差连接
  - 输入信息直接送到隐层网络
  - 目的：减少随机梯度下降过程中的优化难度（深层网络梯度消失）

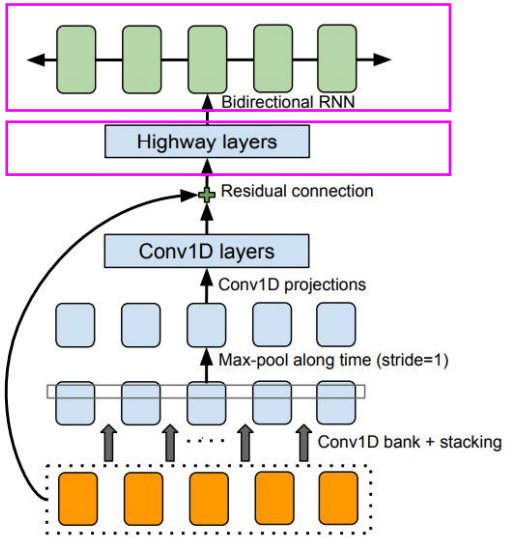


2017. Wang et.al. - Tacotron: Towards end-to-end speech synthesis.

序列到序列声学模型Tacotron

CBHG模块

- Highway network $y_i = H_i(\mathbf{x}) * T_i(\mathbf{x}) + x_i * (1 - T_i(\mathbf{x}))$
  - 隐层输入x的一部分经过非线性变换，一部分直接传递给 y
  - $T(x)$作为门控单元，$T(x) = 0$时输入信息直接传给 y
  - 反向传播时，可以让更多的信息直接回流到输入，不需要经过非线性变换
  - 深度网络的收敛性能好
- Bi-GRU：
  - 双向RNN网络，学习序列的上下文依赖

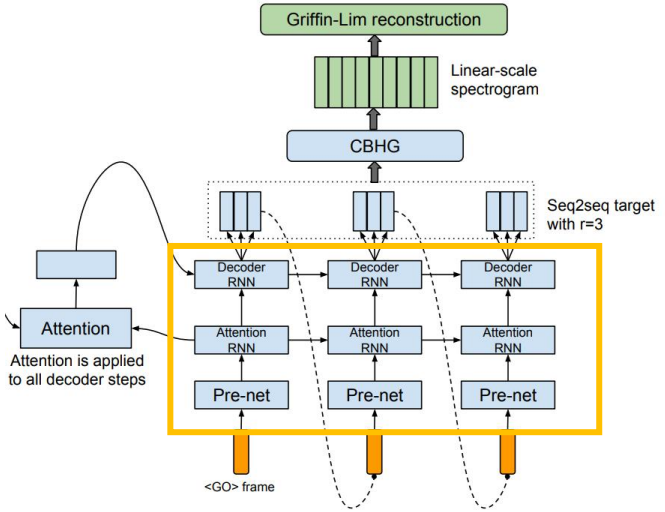


2017. Wang et.al. - Tacotron: Towards end-to-end speech synthesis.

序列到序列声学模型Tacotron

自回归解码器网络

- Pre-net
  - 自回归训练阶段，t时刻使用$t-1$时刻的真实梅尔谱
  - 生成阶段，t时刻使用$t-1$时刻模型输出的梅尔谱
  - 利用dropout提升模型泛化能力
- Attention RNN：
  - 隐层中使用RNN建模序列的时序依赖
- Decoder RNN:
  - 建模内容向量和隐层特征的时序依赖

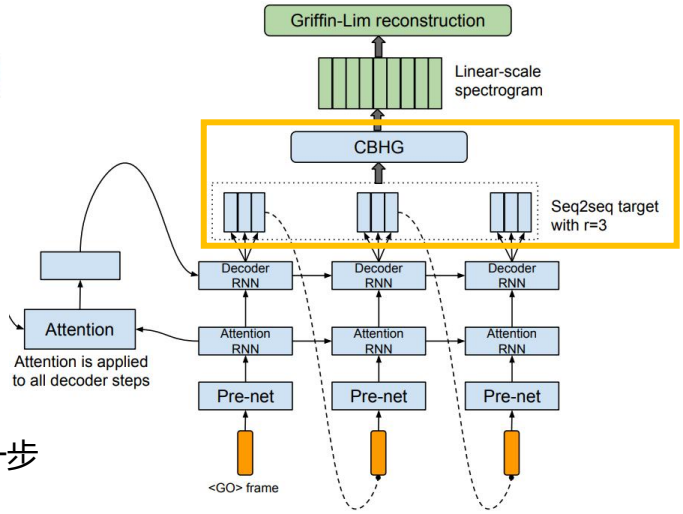


2017. Wang et.al. - Tacotron: Towards end-to-end speech synthesis.

序列到序列声学模型Tacotron

自回归解码器网络

- 注意力机制
  - Global soft attention $\text{score}(\boldsymbol{h}_t, \boldsymbol{l}_s) = \boldsymbol{v}^\top \tanh(\boldsymbol{W}_1 \boldsymbol{h}_t + \boldsymbol{W}_2 \boldsymbol{l}_s)$
- 多帧预测
  - 文本序列长度和语音特征长度差异过大
  - 减少输入和输出序列长度差异，易收敛
- CBHG后处理
  - 自回归只能考虑之前时刻的上下文信息，后处理网络能够进一步考虑整个序列信息，最后生成线性谱



$$[\hat{\boldsymbol{y}}_{t+1}, \hat{\boldsymbol{y}}_{t+2}, \ldots, \hat{\boldsymbol{y}}_{t+K}]^\top = \text{DecRNN}(\boldsymbol{c}_t, \boldsymbol{h}_t)$$

2017. Wang et.al. - Tacotron: Towards end-to-end speech synthesis.

序列到序列声学模型Tacotron

实验

- 文本：英文字符输入，中文音素、韵律边界等
- 语音表示：梅尔谱，50ms帧长，12.5ms帧移
- 多帧预测：一次预测两帧
- 语音还原：Griffin-Lim

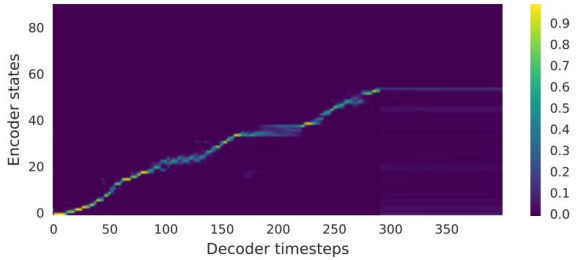
	mean opinion score
Tacotron	3.82 ± 0.085
Parametric	3.69 ± 0.109
Concatenative	4.09 ± 0.119

Spectral analysis	*pre-emphasis*: 0.97; *frame length*: 50 ms; *frame shift*: 12.5 ms; *window type*: Hann
Character embedding	256-D
Encoder CBHG	*Conv1D bank*: K=16, conv-k-128-ReLU *Max pooling*: stride=1, width=2 *Conv1D projections*: conv-3-128-ReLU $\rightarrow$ conv-3-128-Linear *Highway net*: 4 layers of FC-128-ReLU *Bidirectional GRU*: 128 cells
Encoder pre-net	FC-256-ReLU $\rightarrow$ Dropout(0.5) $\rightarrow$ FC-128-ReLU $\rightarrow$ Dropout(0.5)
Decoder pre-net	FC-256-ReLU $\rightarrow$ Dropout(0.5)$\rightarrow$ FC-128-ReLU $\rightarrow$ Dropout(0.5)
Decoder RNN	2-layer residual GRU (256 cells)
Attention RNN	1-layer GRU (256 cells)
Post-processing net CBHG	*Conv1D bank*: K=8, conv-k-128-ReLU *Max pooling*: stride=1, width=2 *Conv1D projections*: conv-3-256-ReLU $\rightarrow$ conv-3-80-Linear *Highway net*: 4 layers of FC-128-ReLU *Bidirectional GRU*: 128 cells
Reduction factor (r)	2

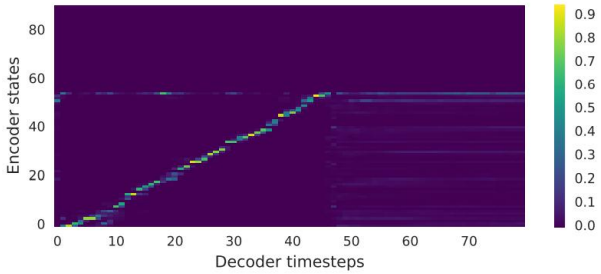
2017. Wang et.al. - Tacotron: Towards end-to-end speech synthesis.

序列到序列声学模型Tacotron

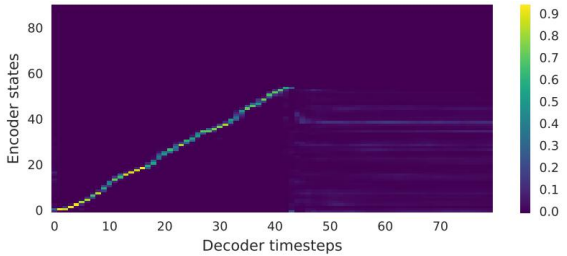
注意力机制分析（一次预测5帧）



(a) Vanilla seq2seq + scheduled sampling



(b) GRU encoder



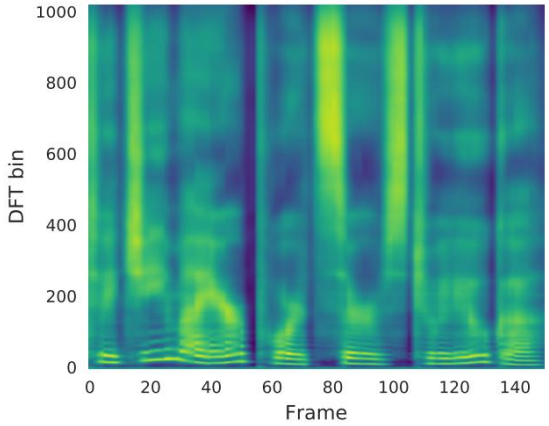
(c) Tacotron (proposed)

Samples: https://google.github.io/tacotron/publications/tacotron/index.html

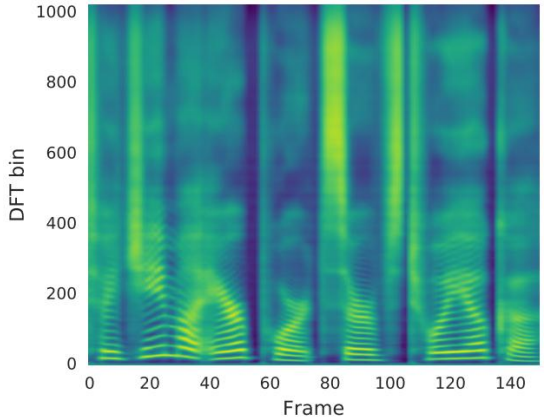
2017. Wang et.al. - Tacotron: Towards end-to-end speech synthesis.

序列到序列声学模型Tacotron

Post-net效果分析



(a) Without post-processing net



(b) With post-processing net

2017. Wang et.al. - Tacotron: Towards end-to-end speech synthesis.

序列到序列声学模型Tacotron

Tacotron优势

- 无需复杂的文本前端分析模块
- 无需额外的时长模型
- 极大地简化了声学模型构建流程，减少了语音合成任务对领域知识的依赖

2017. Wang et.al. - Tacotron: Towards end-to-end speech synthesis.

序列到序列声学模型Tacotron

Tacotron存在的问题

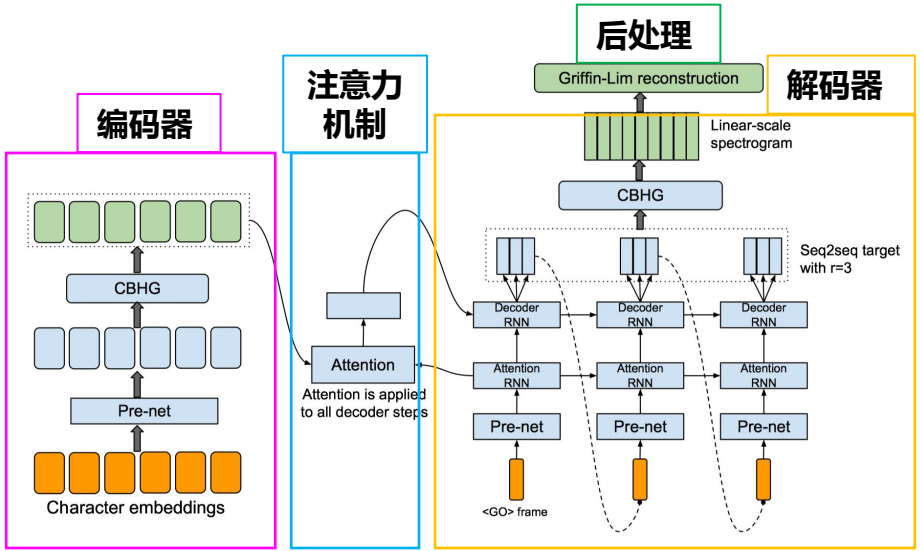
- （1）CBHG网络比较复杂，**参数量**相对较大
- （2）**Griffin-Lim**重建过程中相位问题导致的**语音失真**
- （3）自回归解码器生成过程中**无法停止**
- （4）Global soft **attention** 的问题
  - 针对语音合成任务稳定性较差
  - 模型训练过程中，每个时刻预测的语音帧数越少越难训练

2017. Wang et.al. - Tacotron: Towards end-to-end speech synthesis.

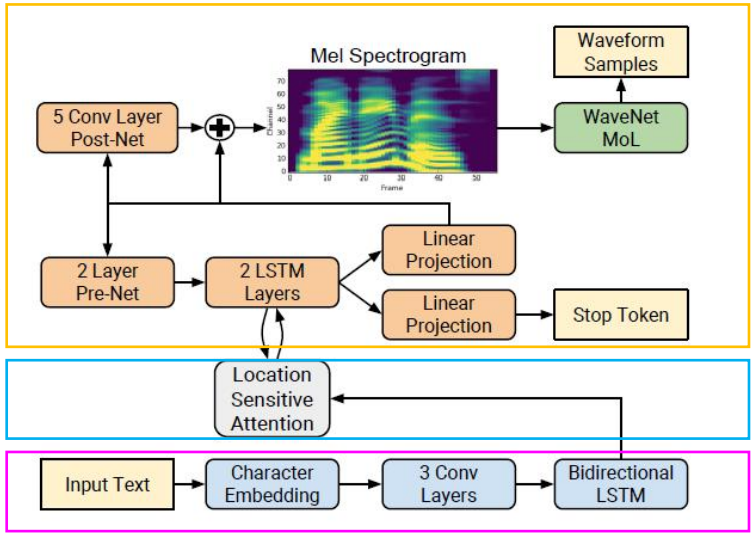
内容提要

序列到序列声学模型Tacotron

Tacotron模型框架结构



Tacotron2 模型框架结构



(1) CBHG导致的参数量问题

(2) Griffin-Lim语音失真问题

(3) 解码器生成过程中无法停止的问题

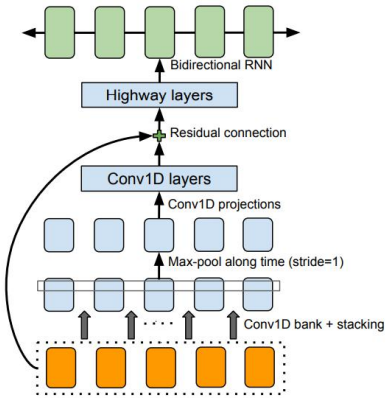
(4) Global soft attention稳定性问题

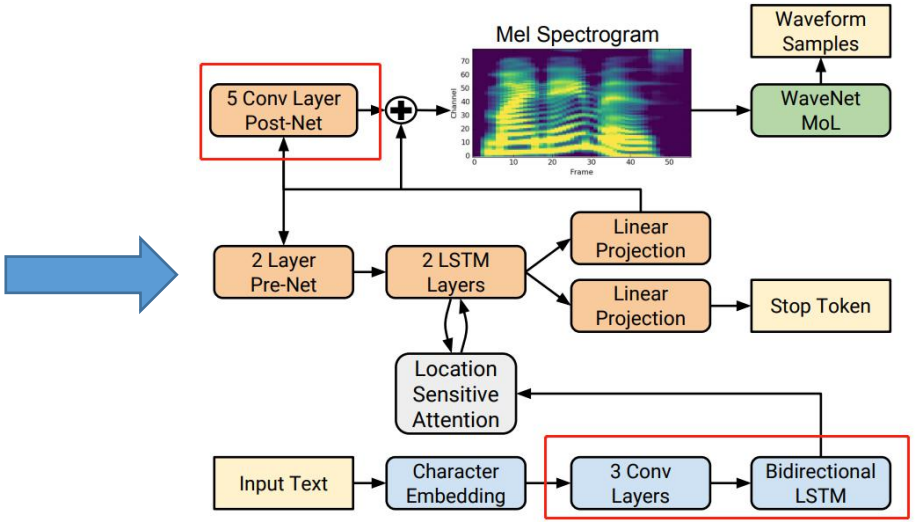
2017. Wang et.al. - Tacotron: Towards end-to-end speech synthesis.

序列到序列声学模型变体（一）

Tacotron 2

- 参数量的减少
  - CBHG -> 3 Conv layers + BLSTM 或 5 Conv layers
  - 去除Attention RNN



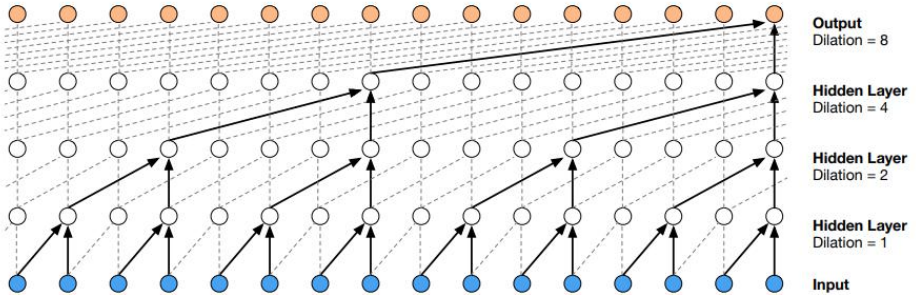


2017. Shen et.al. - ***Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions***

序列到序列声学模型变体（一）

Tacotron 2

- Griffin-Lim导致的语音失真问题
- WaveNet
  - 条件自回归模型 (梅尔谱作为条件)
  - 直接建模语音采样点
  - 因果卷积：语音信号的单调特性
  - 膨胀卷积：考虑更多的上文信息

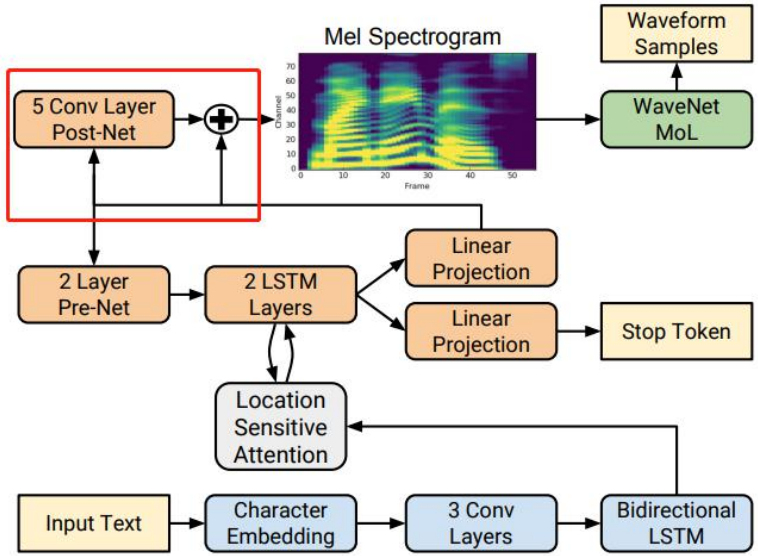


2017. Shen et.al. - Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

序列到序列声学模型变体（一）

Tacotron 2

- PostNet的改进
  - CBHG -> 5层卷积
  - 后处理网络的输入和输出均和真实梅尔谱计算L2误差
  - 残差连接，帮助模型收敛

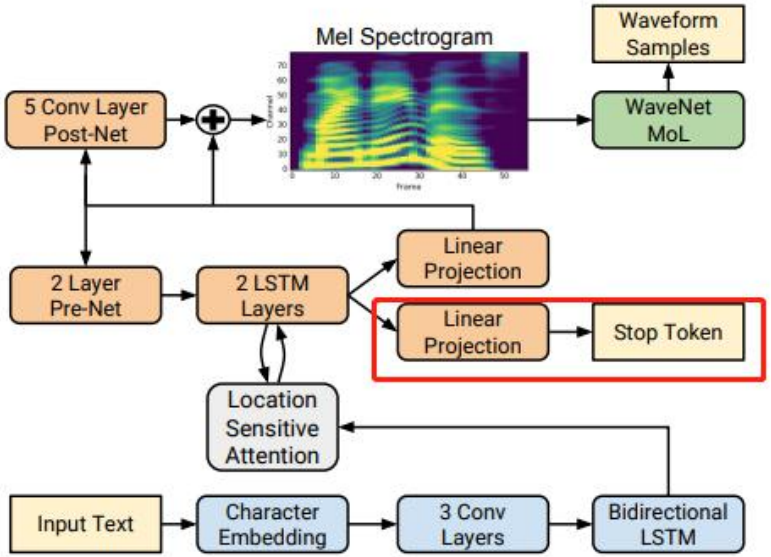


2017. Shen et.al. - Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

序列到序列声学模型变体（一）

Tacotron 2

- 自回归生成过程中无法停止的问题
  - 解码的每个时刻预测是否应该停止（stop token）
  - 通过decoder RNN的输出和注意力机制的内容向量，使用一层Sigmoid激活的前馈神经网络预测[0,1]之间的停止概率
  - 解码时，设定阈值判断是否停止生成



2017. Shen et.al. - Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

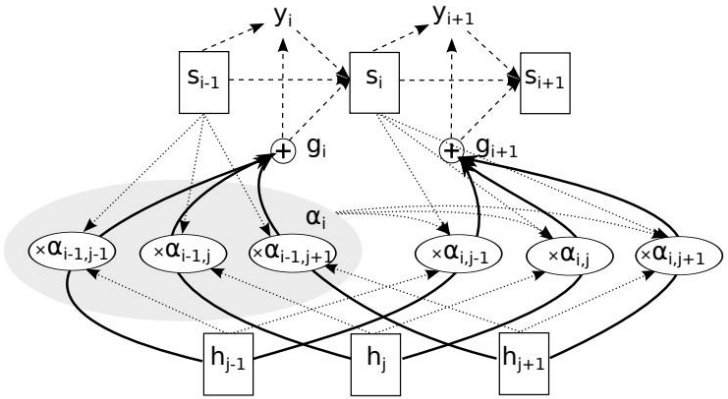
序列到序列声学模型变体（一）

Tacotron 2

- 注意力机制稳定性问题
  - 传统注意力机制中，每个时刻使用编码器哪些部分是无约束的
  - 语音的产生过程是单调的
- Location sensitive attention （LSA）
  - 在解码器t时刻考虑前面时刻的对齐矩阵

$$e_{i,j} = Score(s_{i-1}, h_j)$$

$$\alpha_{i,j} = \exp(e_{i,j}) \Big/ \sum_{j=1}^{L} \exp(e_{i,j})$$



$$f_i = F * \alpha_{i-1}$$

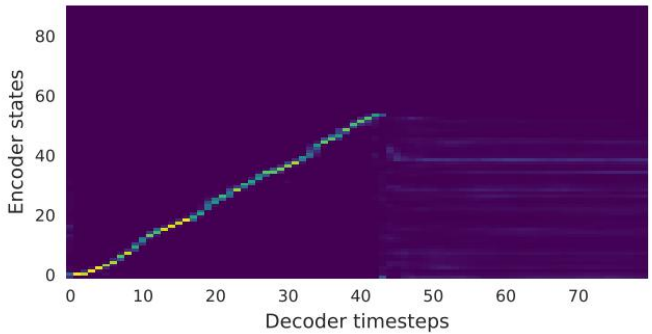
$$e_{i,j} = w^\top \tanh(Ws_{i-1} + Vh_j + b) \Rightarrow e_{i,j} = w^\top \tanh(Ws_{i-1} + Vh_j + Uf_{i,j} + b)$$

2015. Chorowski *et.al. - Attention-based models for speech recognition*

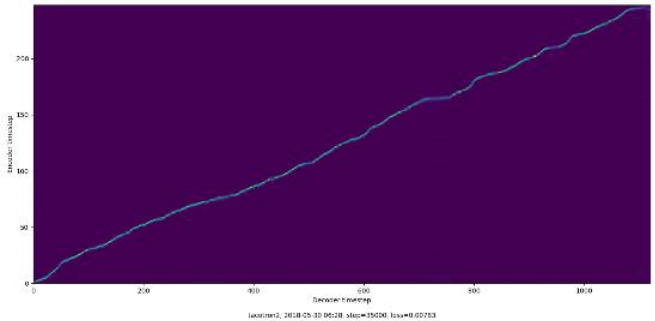
序列到序列声学模型变体（一）

Tacotron 2

- 不同注意力机制对比：



(a) 传统注意力机制



(b) LSA

2015. Chorowski *et.al. - Attention-based models for speech recognition*

序列到序列声学模型变体（一）

Tacotron 2

- 实验

System	MOS
Parametric	3.492 ± 0.096
Tacotron (Griffin-Lim)	4.001 ± 0.087
Concatenative	4.166 ± 0.091
WaveNet (Linguistic)	4.341 ± 0.051
Ground truth	4.582 ± 0.053
Tacotron 2 (this paper)	$\mathbf{4.526 \pm 0.066}$

Table 1. Mean Opinion Score (MOS) evaluations with 95% confidence intervals computed from the t-distribution for various systems.

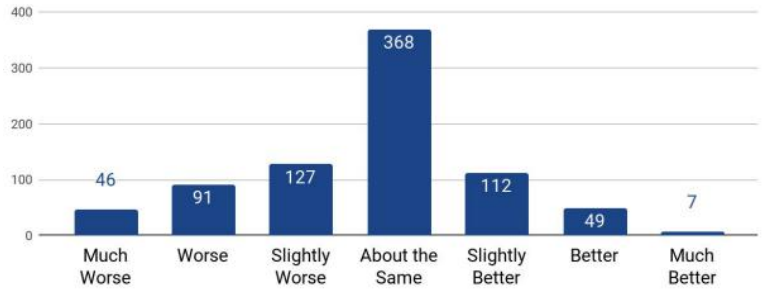


Fig. 2. Synthesized vs. ground truth: 800 ratings on 100 items.

2017. Shen et.al. - Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

序列到序列声学模型变体（一）

Tacotron 2

- 声学模型与声码器对接时的不匹配问题
  - 声学模型训练过程中，每个时刻解码器输入使用真实的梅尔谱 $\hat{y}_t = \text{Decoder}(c, y_{t-1}, s_t)$
  - 声学模型解码过程中，每个时刻使用解码器预测的梅尔谱
  - 声码器WaveNet训练时，使用的是真实的梅尔谱
  - 声码器合成时，使用的是声学模型预测出的梅尔谱

2017. Shen et.al. - Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

序列到序列声学模型变体（一）

Tacotron 2

- GTA生成（Ground-truth alignment）
  - 声学模型训练完成后，使用该模型对训练数据做一次解码
  - 解码时每个时刻使用上一时刻真实的Mel谱，保证生成谱和真实谱的长度一致
  - 缓解声学模型与声码器对接中的特征不匹配

	Synthesis	
Training	Predicted	Ground truth
Predicted	4.526 ± 0.066	4.449 ± 0.060
Ground truth	4.362 ± 0.066	4.522 ± 0.055

2017. Shen et.al. - Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions

总结

- 序列到序列声学模型在发音自然度方面优于传统帧级声学模型
- 序列到序列声学模型能够大大简化语音合成系统的构建流程，但是其训练效率相对帧级模型较低，参数量也较大。
- 基于序列到序列声学模型能够便捷地应用到风格合成、说话人自适应等方面
- 就目前来看，序列到序列声学模型的一个关键问题是生成稳定性问题，其根源为注意力机制本身的稳定性问题

内容提要

实战：基于中文Tacotron系统

尝试按照README.md中的步骤，实现关键代码，从而构建中文Tacotron合成系统

针对开源标贝中文数据集，已经给出准备的文本特征。首先，针对文本对应的语音数据提取语音特征，实现Tacotron系统中的CBHG模块，完成模型训练和测试。

Repo：https://github.com/nwpuaslp/TTS_Course

深蓝学院
shenlanxueyuan.com

音频语音与语言处理研究组
SLP
Audio, Speech and Language Processing Group NPU

感谢聆听！

Thanks for Listening