STATE-OF-THE-ART SPEECH RECOGNITION WITH SEQUENCE-TO-SEQUENCE MODELS

7107018016 翁婉庭

2018 IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP)

Introduction

- **Traditional:**

  ASR(automatic speech recognition):
  AM(acoustic model) , PM(pronunciation model) , LM(language model)

- **Now:**

  Attention-based encorder-decorder architectures ：
  Sequence-to-sequence models
  ex:LAS(Listen,Attend and Spell) ,RNN-T,RNA.....

- **Goal:**

  explore various structure and optimization improvements to allow sequence-to-sequence models to significantly outperform a conventional ASR system on a voice search task.

  ➡ minimum WER(word error rate)

Model

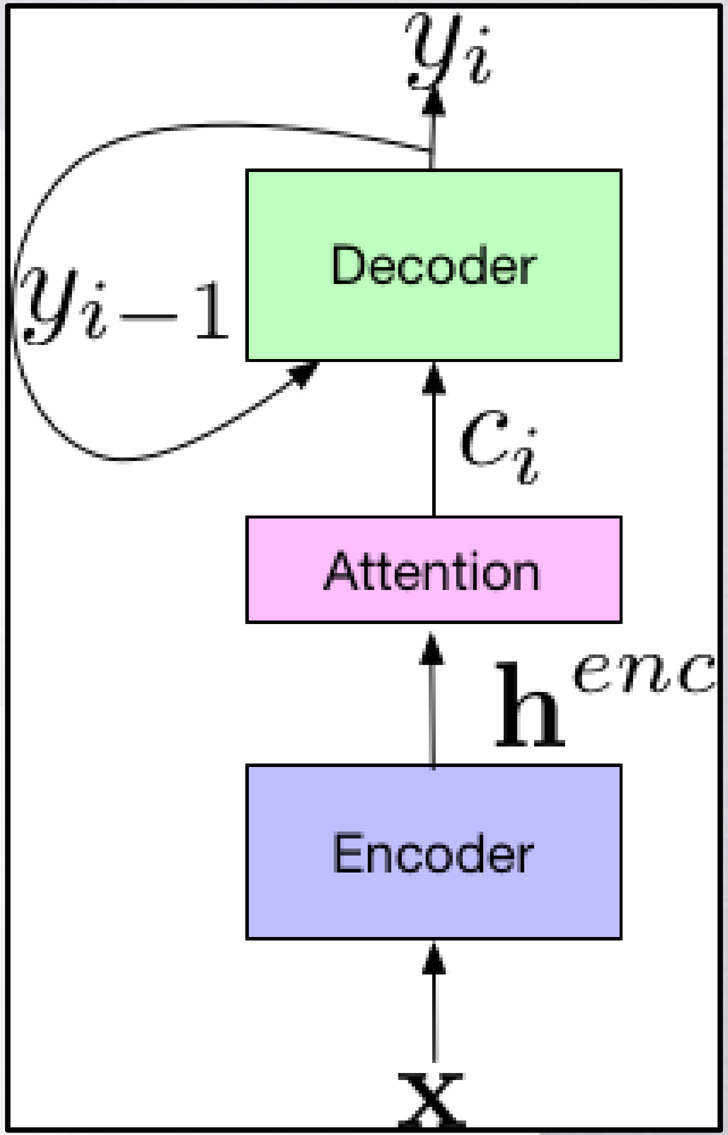
- Basic LAS Model

3 modules:

1.**listener (encoder編碼器)module**
input x ,
map them to a higher-level feature representation, h^{enc}

2. **Attention module**
input h^{enc}
output attention context, c_i

3.**Speller (decorder解碼器)module**
take c_i and embede previous prediction y_{i-1},
in order to produce a probability distribution, $\mathrm{P}(y_i|y_{i-1}, \dots, y_0)$,
over the current sub-word unit, y_i



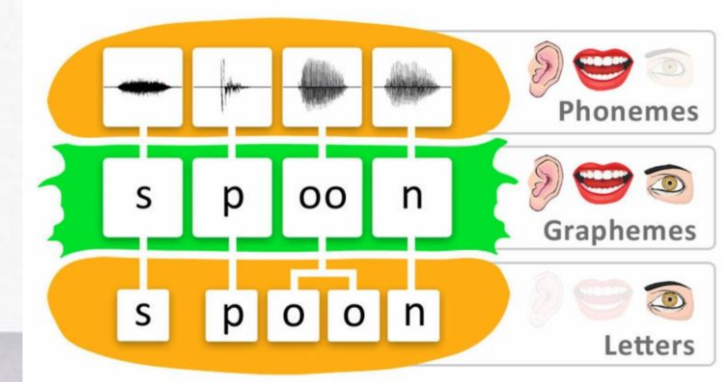
Structure Improvements

1. Wordpiece models(WPM)

traditional: graphemes(字形)

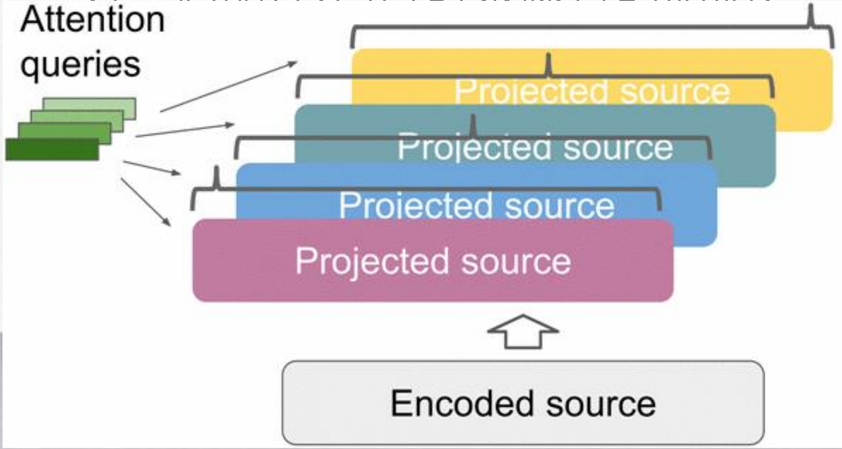
now: wordpiece

✓ 有較高準確度
✓ 能記住常出現的單詞發音
✓ 減少解碼步驟



2. Multi-headed attention(MHA)

(single-headed)
(ex只關注在話語信號上)
↓
(MHA)
(ex一個關注:話語信號上，
另一個關注:包含背景噪音的話語)



Optimization Improvements

1. Minimum Word Error Rate(MWER) Training

Loss function

$$\mathcal{L}_{\text{MWER}} = \mathbb{E}_{P(\mathbf{y}|\mathbf{x})}[\mathcal{W}(\mathbf{y}, \mathbf{y}^*)] + \lambda \mathcal{L}_{\text{CE}}$$

N-best list $\text{NBest}(\mathbf{x}, N) = \{\mathbf{y}_1, \cdots, \mathbf{y}_N\}$

$$\mathcal{L}_{\text{MWER}}^{\text{N-best}} = \frac{1}{N} \sum_{\mathbf{y}_i \in \text{NBest}(\mathbf{x}, N)} [\mathcal{W}(\mathbf{y}_i, \mathbf{y}^*) - \widehat{\mathcal{W}}] \hat{P}(\mathbf{y}_i|\mathbf{x}) + \lambda \mathcal{L}_{\text{CE}}$$

Re-normalized (N-best hypotheses)

$$\widehat{P}(\mathbf{y}_i|\mathbf{x}) = \frac{P(\mathbf{y}_i|\mathbf{x})}{\sum_{\mathbf{y}_i \in \text{NBest}(\mathbf{x}, N)} P(\mathbf{y}_i|\mathbf{x})},$$

- $\mathcal{W}(y, y^*)$: the number of word errors
- y*: the ground-label sequence
- $\mathcal{W}(y_i, y^*) - \widehat{\mathcal{W}}$: variance reduction
- **CE: cross-entropy**

求 $\mathbb{E}_{p(y|x)}$:sampling or restricting the summation to an N-best list of decoded hypothese

Optimization Improvements

2. Scheduled Sampling

在訓練的時候 ，提供先前的標記(token)作為下次預測的標記

→ 縮小訓練及推理(inference)之間的差距

3. Asynchronous and Synchronous

異步訓練:不同副本(replica)各自運行反向傳播的過程，並獨立地更新參數
同步訓練:所有的副本同時讀取參數的取值，並且當反向傳播算法完成之後，
計算出不同副本上參數梯度的平均值，最後再根據平均值對參數進行更新

→ 同步訓練:提供更快的收斂速度及更好的模型

4. Label smoothing

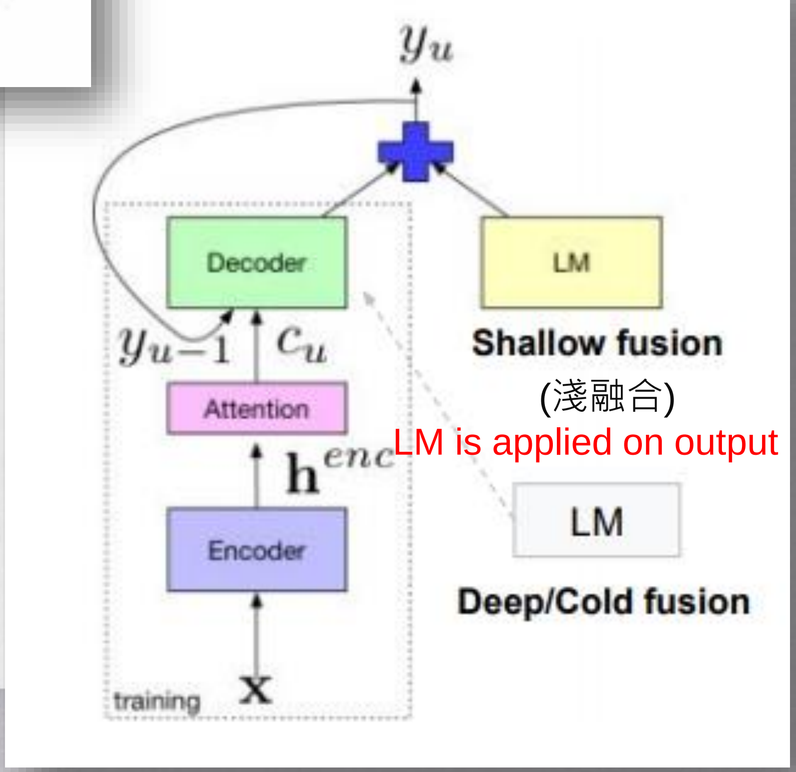
→ 正規化機制，防止模型過度自信的預測
(在訓練時即假設標籤可能存在錯誤，避免 “過分” 相信訓練樣本的標籤)

Second-Pass Rescoring

透過對數線性插值(Log-Linear interpolation)將LM合併到第二次校正模型

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \log P(\mathbf{y}|\mathbf{x}) + \lambda \log P_{LM}(\mathbf{y}) + \gamma \mathrm{len}(\mathbf{y})$$

合併外部的LM:
Shallow fusion(淺融合):
通常只在推理時執行



Motivation:
- ✓ LAS model requires audio-text pairs(音頻文本):only have text
- ✓ 一些語音搜索的錯誤可以從更多的純文本數據上訓練的良好的LM來修復

Reference	LAS model output
What language is built into electrical circuitry of a computer?	what language is built into electrical circuit tree of a computer
Leona Lewis believe	vienna lewis believe
Suns-Timberwolves score	sun's timberwolves score

Experimental

- Training set:

  15million English utterances from Google Voice Search Traffic.

  +Varying degrees of noise and reverberation which from YouTube and daily life noisy environmental recordings.

- Evaluation:

  the resulting model which trained with only voice search data

Results

Exp-ID	Model	VS/D	1st pass Model Size
E8	Proposed	**5.6/4.1**	**0.4 GB**
E9	Conventional LFR system	6.7/5.0	0.1 GB (AM) + 2.2 GB (PM) + 4.9 GB (LM) = 7.2GB

Table 5: Resulting WER on voice search (VS)/dictation (D). The improved LAS outperforms the conventional LFR system while being more compact. Both models use second-pass rescoring.