Windowed Attention Mechanisms for Speech Recognition

Shucong Zhang, Erfan Loweimi, Peter Bell, Steve Renals
Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK

Introduction

- For speech recognition, one output label is often related to a small span of input frames.
- The input frames and the output labels usually follow a monotonic left-to-right order.
- The usual attention mechanisms consider all the input frames and do not guarantee monotonic alignment.
- To address this, rule-based sliding window methods restrict attention mechanisms attend inputs within a large window. But they often have inferior results.
- We propose a fully-trainable windowed attention mechanism. It has advantages in both efficiency and accuracy.

The window shift and window size

- The window shift is estimated by an MLP:

$$\boldsymbol{s_i} = \boldsymbol{N} \cdot \boldsymbol{\sigma}(\mathbf{MLP}(\boldsymbol{q_i}))$$

- q_i: the decoder hidden state at time step i.
  N : the maximum allowed step size; set as a hyperparameter.
- The window size D_i is learned in the same manner by a separate MLP.

The Gaussian location score

$$\boldsymbol{l_{ij}} = \begin{cases} \exp\left(-\frac{(j-m_i)^2}{2(D_{jl})^2}\right), & \boldsymbol{m_i - D_{jl} \le j < m_i} \\ \exp\left(-\frac{(j-m_i)^2}{2(D_{jr})^2}\right), & \boldsymbol{m_i \le j < m_i + D_{jr}} \end{cases}$$

- m_i is a function of the window shifts:

$$\boldsymbol{m_i = m_{i-1} + s_i}$$

- j is the index of the encoder hidden state. D_{jl} and D_{jr} are the left window size and the right window size. The window can be asymmetric if they are learned by two separate MLPs.
- It encourage high scores around the window centre.

The sigmoid location score

$$\boldsymbol{l_{ij}} = \begin{cases} \boldsymbol{\sigma}(\boldsymbol{k}(\boldsymbol{j} - \boldsymbol{m_i}) + \boldsymbol{b}), & \boldsymbol{m_i - D_{jl} \le j < m_i} \\ \boldsymbol{\sigma}(\boldsymbol{k}(\boldsymbol{m_i} - \boldsymbol{j}) + \boldsymbol{b}), & \boldsymbol{m_i \le j < m_i + D_{jr}} \end{cases}$$

- The hyperparameter k and b are chosen to make the location score almost uniformly distributed within the window.
- The propose of this score function is to make the window shift trainable.
- It does not provide any information except the window location.

Fully-trainable windowed attention

- The location score makes the window fully-trainable and it provides location information to the attention mechanism.
- We also use content-based attention to compute a content score $\boldsymbol{e_{ij}}$.
- The final attention score is the product of the location score and the content score:

$$\boldsymbol{\alpha_{ij}} = \frac{\exp(\boldsymbol{e_{ij}}) \cdot \boldsymbol{l_{ij}}}{\sum_{k=m_i-D_{jl}}^{m_i+D_{jr}} \exp(\boldsymbol{e_{ik}}) \cdot \boldsymbol{l_{ik}}}$$

Experimental Results

Function Type	N, D_l, D_r	PER (TIMIT Test)
Baseline: content-based attention		20.1%
Gaussian- Fixed length window	5, 3, 3	17.0%
	5, 4, 2	17.3%
	5, 2, 4	17.3%
Gaussian-one window MLP	5, 12, 12	16.8%
Gaussian-two window MLP	4, 6, 6	**16.7%**
sigmoid(±1.5x + 3)	5, 4, 4	17.8%
sigmoid(±1.5x + 7)	5, 4, 4	23.4%
sigmoid(±1.5x + 7)	5, 7, 7	19.1%

Table 1: Phone error rate on TIMIT test set. N, D_l , D_r denote the maximum allowed step size, left window size and right window size. One unit of the step/window size is 0.04s.

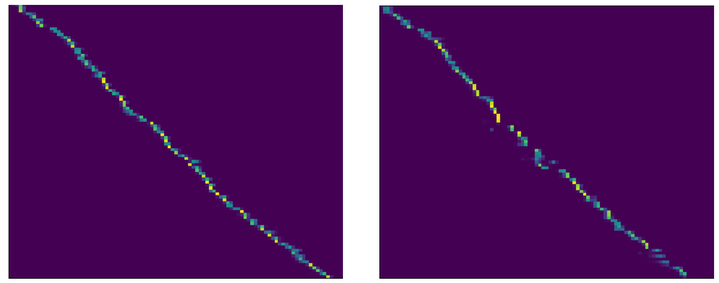

Figure 1: Attention vectors generated by the Gaussian location function (left) and the sigmoid location function (right).

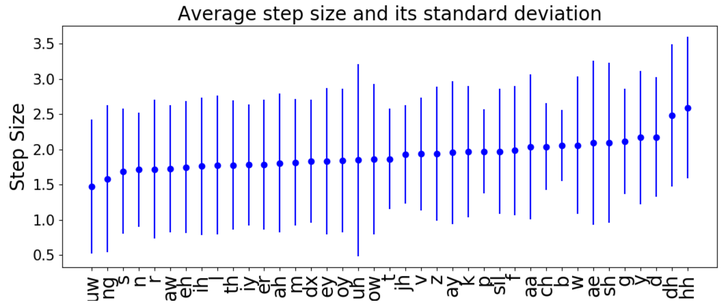

Figure 2: The average learned step size for each phoneme and its standard deviation. The data is collected on TIMIT development set and test set.

Model (Train)	CER(Dev)	CER (Test)
train_si284 eval92	dev93	eval92
Baseline: content-based attention	11.1%	8.9%
location-based attention	9.6%	6.9%
Gaussian-two window MLP	9.0%	6.5%
CTC-attention	**7.7%**	5.9%
CTC-Gaussian	7.8%	**5.8%**
train_si284 subset (30K)	dev93	eval92
location-based attention	9.9%	7.9%
Gaussian-two window MLP	9.5%	7.2%
CTC-attention	**9.1%**	**6.9%**
train_si284 subset (15K)	dev93	eval92
location-based attention	15.7%	13.7%
Gaussian- two window MLP	13.2%	9.6%
CTC-attention	**10.8%**	**8.3%**

Table 2: Character error rates on WSJ. The max step/window size is 1.32s. However, the learned step/window sizes are small except the silence parts.