Self-Attentional Acoustic Models

Matthias Sperber[1], *Jan Niehues*[1], *Graham Neubig*[2], *Sebastian Stuker*[1], *Alex Waibel*[12]

[1]Karlsruhe Institute of Technology

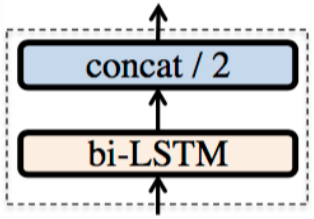
[2]Carnegie Mellon University

Motivation

- RNN suffers from slow computation speed and may not be able to optimally exploit long-range context
- Apply self-attention to acoustic modeling
- Issues:
  - self-attention memory grows quadratically in the sequence length (max 2026; avg 800) -> downsampling
  - previous approaches to incorporate position information are unsuitable -> hybrid self-attention/RNN architecture
  - locality of context plays a special role in acoustic modeling -> apply diagonal Gaussian masks with learnable variance to attention heads

Listen, attend, spell

- an attentional encoder-decoder model
- because acoustic sequences are very long, the encoder performs downsampling to make memory and runtime manageable
- pyramidal LSTM: a stack of LSTM layers where pairs of consecutive outputs of a layer are concatenated before being fed to the next layer



(a) pyramidal (§ 2.1)

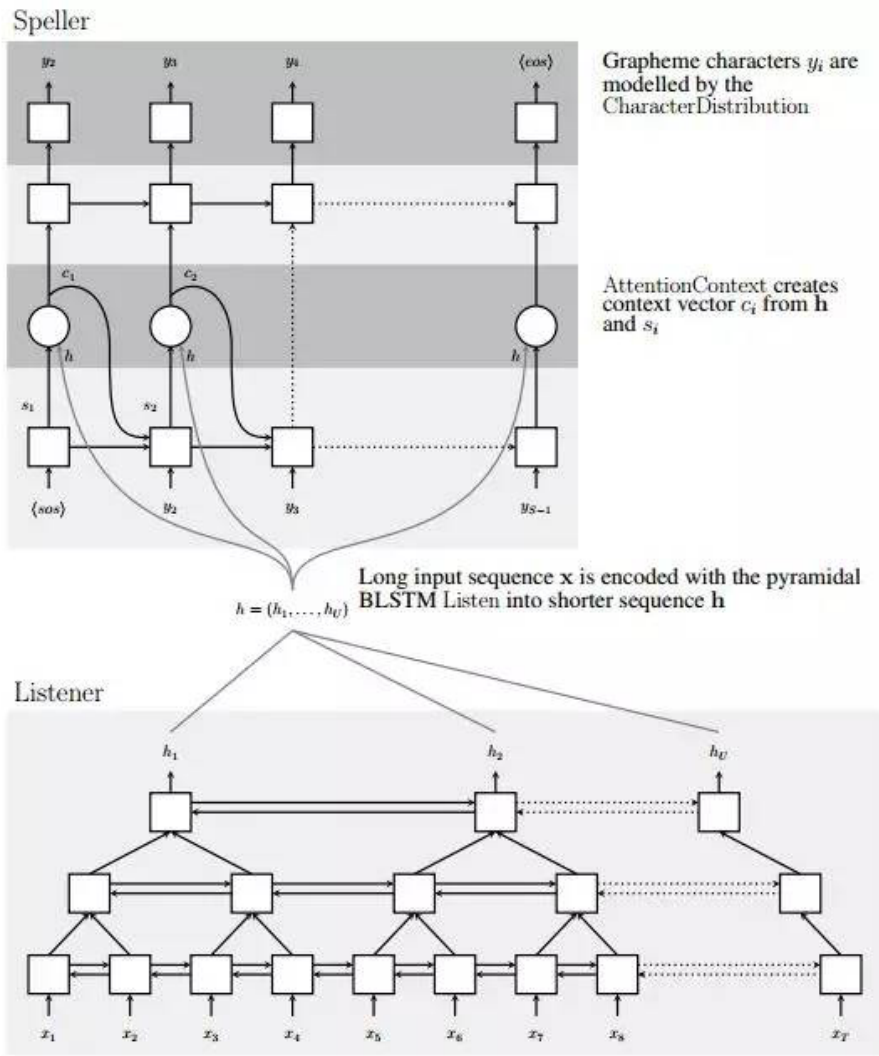


Figure 1: Listen, Attend and Spell (LAS) model: the listener is a pyramidal BLSTM encoding our input sequence $\mathbf{x}$ into high level features $\mathbf{h}$, the speller is an attention-based decoder generating the $\mathbf{y}$ characters from $\mathbf{h}$.

Self-Attentional Acoustic Models

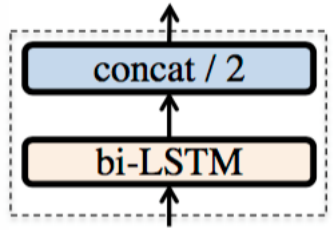
$$Q_i = XW_i^{\mathrm{Q}}, K_i = XW_i^{\mathrm{K}}, V_i = XW_i^{\mathrm{V}} \quad (1)$$

$$\mathrm{head}_i = \mathrm{softmax}\left(\frac{Q_i K_i^{\mathrm{T}}}{\sqrt{d}}\right) V_i \quad \forall i \quad (2)$$

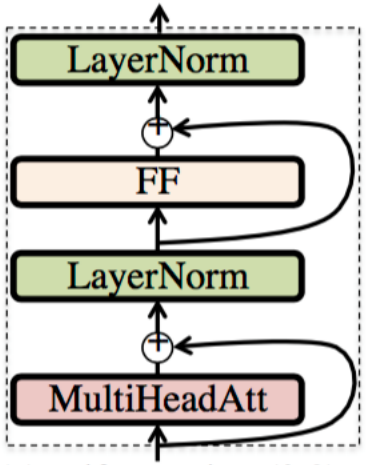
$$\mathrm{MultiHeadAtt} = \mathrm{concat}(\mathrm{head}_1, \mathrm{head}_2, \ldots) \quad (3)$$

$$\mathrm{MidLayer} = \mathrm{LayerNorm}\left[\mathrm{MultiHeadAtt} + X\right] \quad (4)$$

$$\mathrm{SAL} = \mathrm{LayerNorm}\left[\mathrm{FF}\left(\mathrm{MidLayer}\right) + \mathrm{MidLayer}\right] \quad (5)$$



(a) pyramidal (§ 2.1)



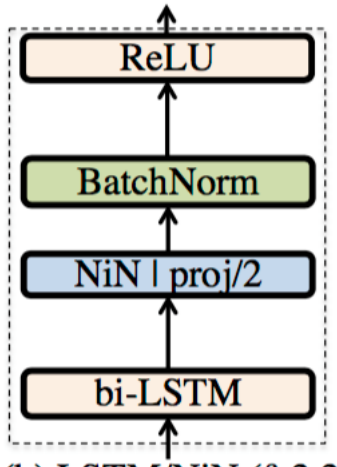
(c) self-attention (§ 3)

Tailoring Self-Attention to Speech

- Downsampling

$$X \in \mathbb{R}^{l \times d} \rightarrow_{\text{reshape}} \hat{X} \in \mathbb{R}^{\frac{l}{a} \times ad}$$

- Position Modeling
  - inputs are fixed feature vectors rather than trainable word embeddings, making it difficult to separate position and content for each state
  - Concatenated Position Representation:
    - 1) concatenate trigonometric encodings to the input feature vectors
    - 2) concatenate learned position embeddings to the input feature vectors
    - 3) concatenate learned position embeddings to the queries Q and keys K
  - Hybrid Models:
    - 1) Stacked hybrid model
    - 2) Interleaved hybrid model: replace ff-layer with an LSTM



(b) LSTM/NiN (§ 2.2)

Tailoring Self-Attention to Speech

- Attention biasing
  - introduce an explicit way of controlling the context range by using a bias matrix

$$\text{head}_i = \text{softmax}(\frac{Q_i K_i^T}{\sqrt{d}} + M)V_i$$

    - Local Masking

$$M_{jk} = \begin{cases} 0 & |j-k| < \frac{b}{2} \\ -\infty & \text{else} \end{cases}.$$

    - Gaussian Bias

$$M_{jk} = \frac{-(j-k)^2}{2\sigma^2}.$$

Experiments

Table 1: *Comparison to baselines. Training speed (char/sec) was measured on a GTX 1080 Ti GPU.*

model	dev WER	test WER	char/sec
pyramidal	15.83	16.16	1.1k
LSTM/NiN	14.57	14.70	1.1k
stacked hybrid	16.38	17.48	2.4k
interleaved hybrid	15.29	16.71	1.5k

Table 2: *WER results on position modeling*

model	dev	test
add (trig.)	diverged	
concat (trig.)	30.27	38.60
concat (emb.)	29.81	31.74
stacked hybrid	16.38	17.48
interleaved hybrid	15.29	16.71

Experiments

Table 3: *WER results on attention biasing.*

model	dev	test
stacked hybrid	16.38	17.48
+ local masking	15.42	16.17
+ Gauss mask (init. small)	16.05	16.96
+ Gauss mask (init. large)	14.90	15.89
interleaved hybrid	15.29	16.71
+ local masking	15.44	16.19
+ Gauss mask (init. small)	16.43	16.89
+ Gauss mask (init. large)	15.00	15.82

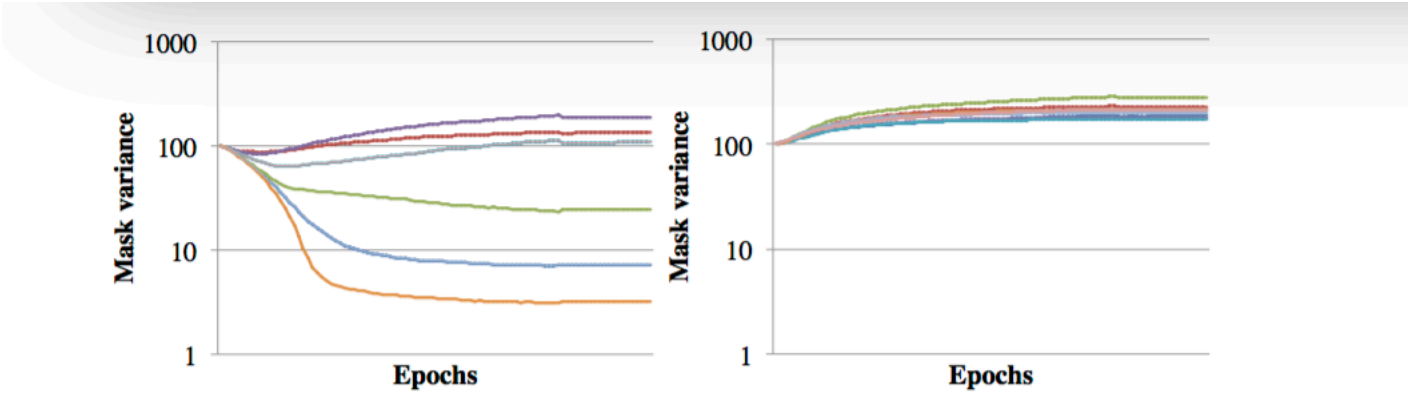


Figure 2: *Evolution of the variance parameters for each of the 8 attention heads over course of training (left: first layer, right: second layer).*

In the first layer, diversity seems to be desirable. In contrast, the second layer does not appear to benefit from limiting its context.

This partly confirms the idea of hierarchical modeling, where the modeling granularity increases across layers, but also shows that even at the bottom layer a controlled amount of long-range context is desirable.

Interpretability of Attention Heads

- Head 2 seems to always focus on the utterance end where we usually expect silence.
- Head 8 is mostly unfocused, which we may interpret as these heads establishing channel and speaker context.

Table 4: *Analysis of function of attention heads. Note that we conducted a small amount of cherry picking by removing 4 outliers that did not seem to fit categories (OY from head 1, ZH from head 3, EH and ER from head 7). Entropy is computed over the correlation scores, truncated below 0.*

i	top phonemes	entropy	comments
1	S, TH, Z	3.7	sibilants
2	</s>	1.9	silence
3	UW, Y, IY, IX	3.6	”you” diphthong
	B, G, D		voiced plosives
	M, NG, N		nasals
4	XM, AW, AA, AY, L, AO, AH	3.2	A, schwa
5	ZH, AXR, R	3.5	R, ZH
6	ZH, Z, S	3.2	sibilants
	IY, IH, Y, UW		”you” diphthong
7	S, </s>, TH CH, SH, F	3.4	fricative, noise
8	mixed	3.7	unfocused