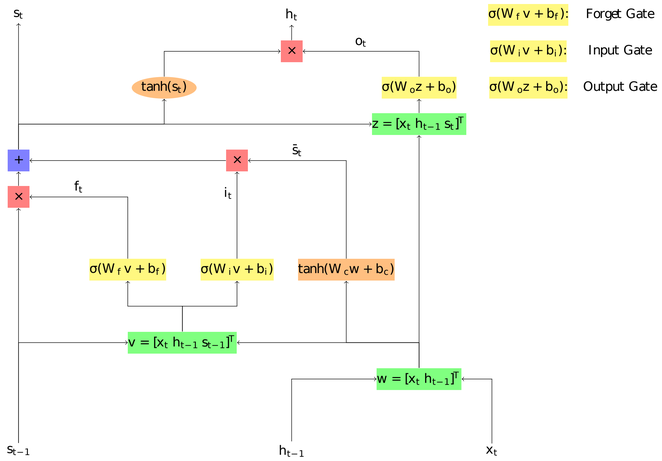
BUILDING AN END-TO-END SPEECH RECOGNIZER

Moritz Wolter
Department ESAT, KU Leuven, Belgium
moritzalexander.wolter@student.kuleuven.be

1. Goals

- Transcribe speech to text.
- Use an end to end system.
- Find suitable parameters.
- Explore regularization options.

2. Long Short Term Memory

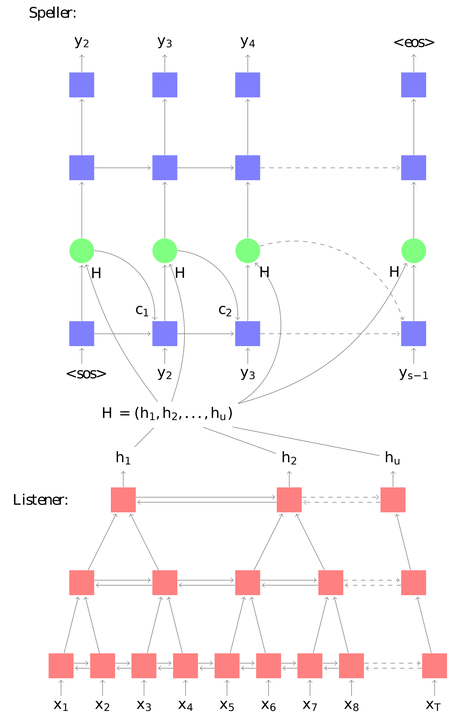


- Differentiable versions of computer memory chips.
- Gates $\mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t$, state $\mathbf{s}_t$ and output $\mathbf{h}_t$.

$$\mathbf{i}_t = \sigma(\mathbf{W}_i[\mathbf{x}_t\ \mathbf{h}_{t-1}\ \mathbf{s}_{t-1}]^T + \mathbf{b}_i), \quad (1)$$
$$\mathbf{f}_t = \sigma(\mathbf{W}_f[\mathbf{x}_t\ \mathbf{h}_{t-1}\ \mathbf{s}_{t-1}]^T + \mathbf{b}_f), \quad (2)$$
$$\mathbf{s}_t = \mathbf{f}_t\mathbf{s}_{t-1} + \mathbf{i}_t \tanh(\mathbf{W}_s[\mathbf{x}_t\ \mathbf{h}_{t-1}]^T + \mathbf{b}_s), \quad (3)$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_o[\mathbf{x}_t\ \mathbf{h}_{t-1}\ \mathbf{s}_t]^T + \mathbf{b}_o), \quad (4)$$
$$\mathbf{h}_t = \mathbf{o}_t \tanh(\mathbf{s}_t). \quad (5)$$

- Most important equation is the one for the state $\mathbf{s}_t$, the state functions as cell memory.

3. Listen Attend and Spell



- Mel feature inputs.
- Listener consits of one bidirectional and one ore more pyramidal bidirectional LSTM layer.
  - pyramidal layers concatenate features over the feature dimension in time:

$$\mathbf{h}_t^n = \text{BLSTM}(\mathbf{h}_{t-1}^n, [\mathbf{h}_{2t}^{n-1}, \mathbf{h}_{2t+1}^{n-1}]). \quad (6)$$

- Speller consits of an attend and spell cell.
  - The attent and spell function evaluates:

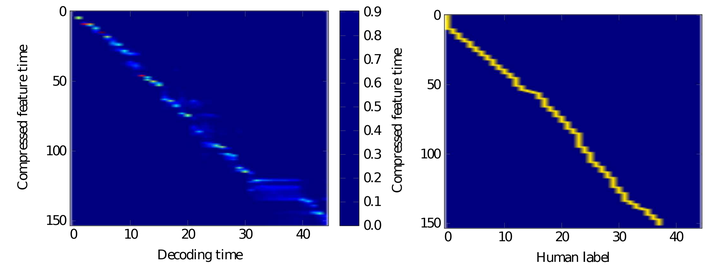
$$\mathbf{s}_i = \text{RNN}(\mathbf{s}_{i-1}, \mathbf{y}_{i-1}, \mathbf{c}_{i-1}), \quad (7)$$
$$\mathbf{c}_i = \text{AttentionContext}(\mathbf{s}_i, \mathbf{H}), \quad (8)$$
$$P(\mathbf{y}_i|\mathbf{x}, \mathbf{y}_{<i}) = \text{CharacterDistribution}(\mathbf{s}_i, \mathbf{c}_i). \quad (9)$$

  - The attention context function computes:

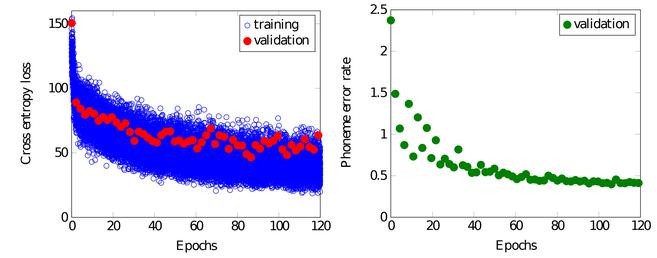
$$e_{i,u} = \phi(\mathbf{s}_i)^T \psi(\mathbf{h}_u), \quad (10)$$
$$\alpha_{i,u} = \frac{\exp(e_{i,u})}{\sum_u \exp(e_{i,u})}, \quad (11)$$
$$\mathbf{c}_i = \sum_u \alpha_{i,u}\mathbf{h}_u. \quad (12)$$

  - ϕ and ψ are MLPs.

4. Results



- Network alignment and human alignment resemble each other.
- Result obtained using greedy decoding test set error rate 0.55.



- Improved results where found using a larger network, beam search and dropout, test set error rate 0.45.
- Utterance `fmld0_sx295`:
  - Human labeling:
    <sos> sil ih f sil k eh r l sil k ah m z sil t ah m aa r ah hh
    ae v er r ey n jh f er m iy dx iy ng ih sil t uw sil <eos>
  - Network labeling:
    <sos> sil hh ih f sil k ih r ow sil k ah m sil sil t ah m aa aa hh
    hh v v er ey n n sil f f er m iy iy iy iy sil sil t uw sil sil <eos>
  - Error ratio 0.36 for this utterance.

5. Literature

[1] W. Chan, N. Jaitly, Q.V. Le, and O. Vinyals, "Listen attend and Spell." *arXiv preprint*, pages, 2015.

[2] A. Graves. "Supervised Sequence Labelling with Recurrent Neural Networks." *Springer, Berlin Heidelberg*, 2012.