RECURRENT NEURAL NETWORK IN NATURAL LANGUAGE PROCESSING

Sütő Evelyne [1]

[1]Babeş-Bolyai University, Computer Science, Cluj-Napoca

- **Purpose:** Present a model that is appropriate for sequence generation and processing
- **Selected architecture:** Recurrent neural network and more precisely Long Short-Term memory network
- Shortcoming of RNN architecture
- Why LSTM?
- How does LSTM work?
- Improvements of LSTM over the years

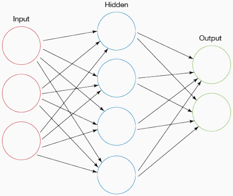


Figure 1: The representation of a simple Artificial Neural Network with one hidden layer. Image source: `https://en.wikipedia.org/wiki/Artificial_neural_network`

$$output = \sigma(\sum_{n=1}^{N}(w_i * x_i + b_i)) \tag{1}$$

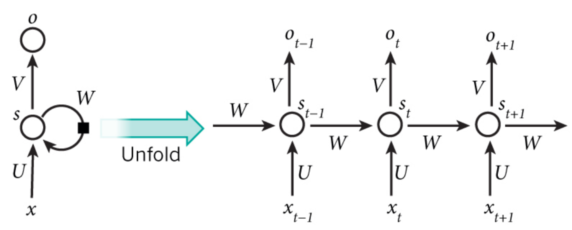


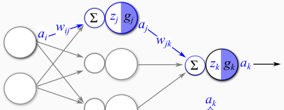
Figure 2: Recurrent Neural Network .
Image source: `https://magenta.tensorflow.org`

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h) \quad (2)$$

$$y_t = W_{hy}h_t + b_y \quad (3)$$

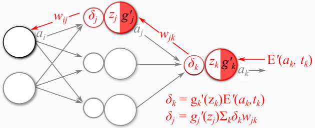
BACKPROPAGATION

I. Forward-propagate Input Signal



$$a_k = \overbrace{g_k(\overbrace{b_k + \Sigma_j \underbrace{g_j(\underbrace{b_j + \Sigma_i a_i w_{ij}}_{z_j})}_{a_j} w_{jk}}^{z_k})}^{a_k}$$

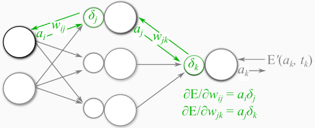
II. Back-propagate Error Signals



$\delta_k = g_k'(z_k)E'(a_k,t_k)$

$\delta_j = g_j'(z_j)\Sigma_k\delta_k w_{jk}$

III. Calculate Parameter Gradients



$\partial E/\partial w_{ij} = a_i\delta_j$

$\partial E/\partial w_{jk} = a_j\delta_k$

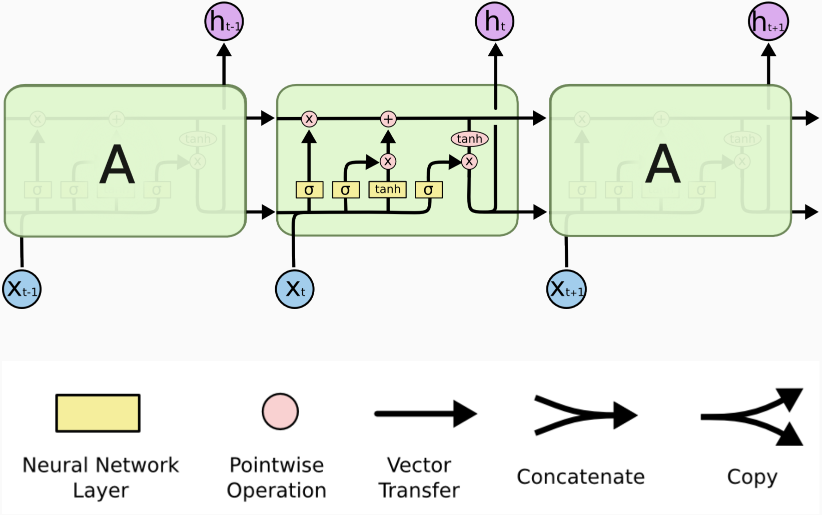
IV. Update Parameters

$w_{ij} = w_{ij} - \eta(\partial E/\partial\, w_{ij})$

$w_{jk} = w_{jk} - \eta(\partial E/\partial w_{jk})$

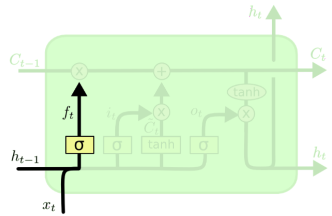
for learning rate η

LSTM



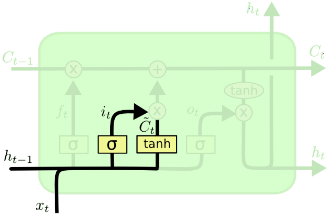
h_{t-1}
h_t
h_{t+1}
A
σ
tanh
A
X_{t-1}
X_t
X_{t+1}
Neural Network Layer
Pointwise Operation
Vector Transfer
Concatenate
Copy

LSTM



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

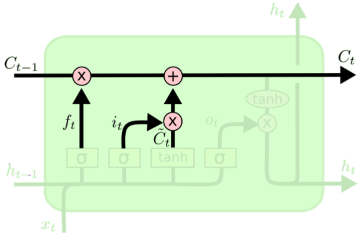
Figure 4: Image source:
`http://colah.github.io/posts/2015-08-Understanding-LSTMs/`



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

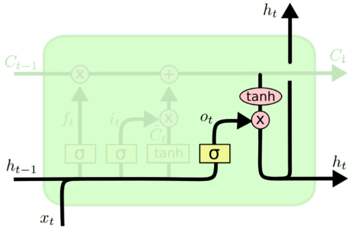
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

Figure 5: Image source:
`http://colah.github.io/posts/2015-08-Understanding-LSTMs/`



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Figure 6: Image source:
`http://colah.github.io/posts/2015-08-Understanding-LSTMs/`



$$o_t = \sigma\left(W_o \,[h_{t-1}, x_t] \;+\; b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Figure 7: Image source:
`http://colah.github.io/posts/2015-08-Understanding-LSTMs/`

VARIANTS OF LSTM

- The first variation of LSTM introduced in 1997 included only input and output gate withan internal state.
- LSTM was not able to reset its own internal state. To solve this the forget gate has been introduced.
- Later this model was expanded with peephole connections, connections from the cells to the gates, which can control the gates in order to make precise timings easier to learn.
- LSTM networks were improved further by replacing the mixture of Real Time Recurrent Learning and Backpropagation Through Time with full Backpropagation Trough Time training.
- Generating new LSTM cells with evolutionary algorithms.
- Hybrid networks e.g. LSTM with Convolutional networks.

NATURAL LANGUAGE PROCESSING

- Neural Machine translation
- Text summarization
- Speech to text
- Paraphrase generation

LTSM

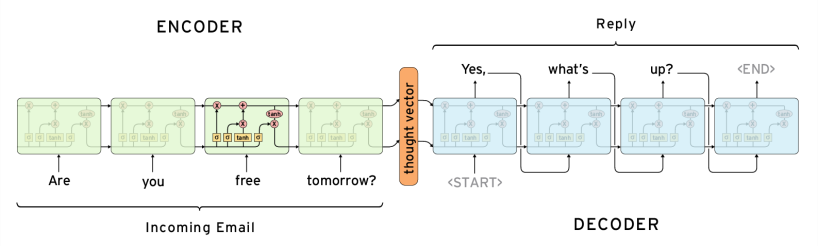


Figure 8: Image source:
`https://medium.com/botsupply/generative-model-chatbots/`

Greedy Decoder

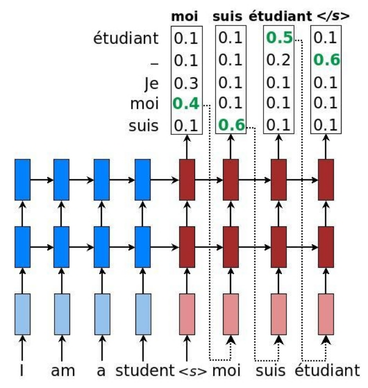


Figure 9: Image source: `https://www.packtpub.com/mapt/book/big_data_and_business_intelligence/`

ENCODER-DECODER (SEQUENCE TO SEQUENCE) MODEL

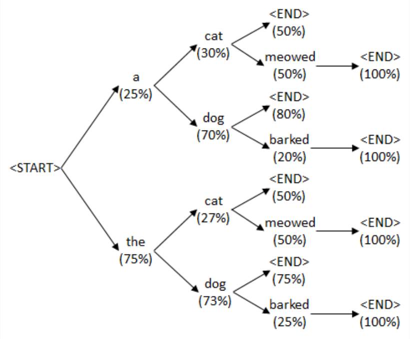


Figure 10: Image source:
`https://geekyisawesome.blogspot.com/2016/10/`

NEURAL MACHINE TRANSLATION

- Researchers have applied this model to English-French translation [Sutskever et al., 2014] in 2014
- Conclusions
  - LSTM sequence to sequence models are able to map even very long sentences to the translation language
  - Deep LSTMs can significantly outperform shallow LSTMs
  - Unoptimized model was able to produce state of the art results with relatively short training
  - State of the art result: 37 BLEU score
  - This model's result: 36.5
- Researchers improved this unoptimized version [Wu et al., 2016] in 2016 to use for Google's translation system

- Another application is shown in [Graves and Jaitly, 2014] done in 2014 for speech recognition system.

- Conclusions
    - Bi-Directional LSTMs are needed to exploit future context as well.
    - Connectionist Temporal Classification
    - Model that does not need data preprocessing.
    - Their best score on 81 hour dataset (Wall street journal) is 8.2 while the baseline was 7.8 (word error rate/character error rate)

TEXT SUMMARIZATION

- Sequence to sequence model was used for text summarization in [Nallapati et al., 2016] in 2016

- Conclusions
  - Encoder-Decoder model with a bi-directional GRU-RNN as the encoder and a uni-directional GRU-RNN.
  - Address the problem of rarewords
  - They have reached state of the art results.

PARAPHRASE GENERATION

- Paraphrasing, the act to express the same meaning in different possible ways.
- Paraphrase generation has been researched in [Prakash et al., 2016] in 2016.

- Conclusions
  - Encoder-Decoder model with deep LSTM networks.
  - They have reached state of the art results.

CONCLUSION

- Recurrent networks are ideal for sequence modeling.
- LSTM is an appropriate approach when long term dependencies have to be modeled.
- LSTMs are still popular research topics.
- LSTMs sometimes can be too complicated to train.
- The literature is still looking for simpler architectures to replace them e.g. 2D Convolutional networks.

Graves, A. and Jaitly, N. (2014).
Towards end-to-end speech recognition with recurrent neural networks.
In *International Conference on Machine Learning*, pages 1764–1772.

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. (2016).
Abstractive text summarization using sequence-to-sequence rnns and beyond.
arXiv preprint arXiv:1602.06023.

Prakash, A., Hasan, S. A., Lee, K., Datla, V., Qadir, A., Liu, J., and Farri, O. (2016).
Neural paraphrase generation with stacked residual lstm networks.
arXiv preprint arXiv:1610.03098.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014).
Sequence to sequence learning with neural networks.

In *Advances in neural information processing systems*, pages 3104–3112.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016).
Google's neural machine translation system: Bridging the gap between human and machine translation.
arXiv preprint arXiv:1609.08144.