Generating Sequences With Recurrent Neural Networks

December 8, 2023

1 Summary

0. This paper shows how LSTM networks can be used to generate complex sequences with long-range structure, simply by predicting one data point at a time. The approach is demonstrated for text (where the data are discrete) and online handwriting (where the data are real-valued). It is then extended to handwriting synthesis by allowing the network to condition its predictions on a text sequence to obtain realistic results.

1. **The model.** A vector $\mathbf{x} = (x_1, \cdots, x_T)$ is passed through N recurrently connected hidden layers $\mathbf{h}^n = (h_1^n, \cdots, h_T^n)$ to output $\mathbf{y} = (y_1, \cdots, y_T)$.

   The paper claims that the skip connections from inputs to all hidden layers and hidden layers to output helps in training by mitigating vanishing gradients problem.

   Hidden layer activations and output are computed as,

   $$\begin{aligned} h_t^1 &= \mathcal{H}(W_{ih^1} x_t + W_{h^1 h^1} h_{t-1}^1 + b_h^1) \\ h_t^n &= \mathcal{H}(W_{ih^n} x_t + W_{h^{n-1} h^n} h_{t-1}^{n-1} + W_{h^n h^n} h_{t-1}^{n-1} + b_h^n) \\ \hat{y}_t &= b_y + \sum_{n=1}^{N} W_{h^n y} h_t^n \\ y_t &= \mathcal{Y}(\hat{y}_t) \end{aligned}$$

   Output y_t is used to predict distribution $Pr(x_{t+1}|y_t)$ for the next input. The probability given by the network for the input $\mathbf{x}$ is given by $\Pi_{t=1}^{T} Pr((x_{t+1}|y_t)$ and log of this is used to define the loss function.

   An LSTM memory cell was used for $\mathcal{H}$ function. A full gradient was used for backpropagation (as opposed to the original paper) and gradient clipping employed to deal with exploding gradients.

2. **Text prediction.** To increase generative flexibility the paper also uses character level prediction although it is known that it performs slightly worse than word level prediction.

Both approaches overfit Penn Treebank data and regularization was used. The results are presented in a table to show that word level prediction slightly outperformed to that of character level.

A larger network was used to predict Wikipedia data. This approach outperformed existing winner of a competition. Dynamic weight adjusted proved to be more effective.

3. **Handwriting prediction.** With minimal preprocessing (as opposed to papers before it) this paper builds a handwriting prediction model from online data. Thus demonstrating that LSTM networks work successfully with real valued data too. Mixture Density Networks are used to parametrize probability distribution to predict next input. This approach offers more flexibility to model the data.

4. **Handwriting synthesis.** Above RNN is used with the modificatio that an added input from the character sequence, mediated by the window layer. Convolution over the text is used to generate the character input.

2 Analysis

The paper successfully uses LSTM to generate seuquences with long dependencies. It is well rounded paper that introduces simple but significant techniques to achieve novel results.