Practical Deep Neural Networks

GPU computing perspective
Recurrent Neural Networks

Yuhuang Hu    Chu Kiong Loo

Advanced Robotic Lab
Department of Artificial Intelligence
Faculty of Computer Science & IT
University of Malaya
Outline

1. Introduction
2. SRN
3. LSTM
4. Sequence Modeling
5. Q&A
Outline

1. Introduction
2. SRN
3. LSTM
4. Sequence Modeling
5. Q&A
Assumed prerequisites

- Neural Computation (DL book chapter 4)
- Machine Learning Basics (DL book chapter 5)
- MLP Networks (DL book chapter 6)
Deep Learning book Chapter 10: Sequence Modeling: Recurrent Recursive Nets
CS224d: GRUs and LSTMs – for machine translation
The Unreasonable Effectiveness of Recurrent Neural Networks
LSTM: A Search Space Odyssey
Supervised Sequence Labelling with Recurrent Neural Networks
SRN architecture
SRN architecture

\[
y_{h}^{t} = f_{h}(W_{i}x^{t} + W_{h}y^{t-1})
\]
\[
y_{o}^{t} = f_{o}(W_{o}y_{h}^{t})
\]

where \( W_{h}, W_{i}, o \) are the hidden, input and output weight matrices, \( x^{t} \) is the input vector, and \( y_{h}^{t} \) is a vector representing the activation of hidden units at time step \( t \). Functions \( f_{h}(\cdot) \) and \( f_{o}(\cdot) \) are non-linear functions.
Outline

1. Introduction
2. SRN
3. LSTM
4. Sequence Modeling
5. Q&A
LSTM architecture
LSTM architecture

\[ z^t = g(W_z x^t + R_z y^{t-1} + b_z) \quad \text{block input} \]
\[ i^t = \sigma(W_i x^t + R_i y^{t-1} + p_i \odot c^{t-1} + b_i) \quad \text{input gate} \]
\[ f^t = \sigma(W_f x^t + R_f y^{t-1} + p_f \odot c^{t-1} + b_f) \quad \text{forget gate} \]
\[ c^t = i^t \odot z^t + f^t \odot c^{t-1} \quad \text{cell state} \]
\[ o^t = \sigma(W_o x^t + R_o y^{t-1} + p_o \odot c^t + b_o) \quad \text{output gate} \]
\[ y^t = o^t \cdot h(c^t) \quad \text{block output} \]

Here \( x^t \) is the input vector at time \( t \), the \( W \) are rectangular matrices, the \( R \) are square recurrent weight matrices, the \( p \) are peehole weights vectors and \( b \) are bias vectors. Functions \( \sigma \), \( g \) and \( h \) are point-wise non-linear activation functions: logistic sigmoid is used for as activation function of the gates and hyperbolic tangent is usually used as the block input and output activation function. The point-wise multiplication of two vectors is denoted with \( \odot \)
Outline

1. Introduction
2. SRN
3. LSTM
4. Sequence Modeling
5. Q&A
Modes of Processing

Left to right: (a) fixed-size input to fixed-size output (e.g. image classification); (b) sequence output (e.g. image captioning); (c) sequence input (e.g. sentiment analysis); (d) sequence input and sequence output (e.g. machine translation); (e) synced sequence input and output (e.g. video classification)
Example: character prediction

**Figure:** Predict “hello”
Example: image captioning

Sequence Modeling

Yuhuang Hu, Chu Kiong Loo (UM)
Outline

1. Introduction
2. SRN
3. LSTM
4. Sequence Modeling
5. Q&A
Q&A

**Hey Bert, ask if it has a favourite colour.**

N. Harding