GRID LSTM

A UNIFIED EXTENSION OF LSTM TO DEEP NETWORK

Presented by Kaiwen Wu
ROADMAP

Arch | LSTM | Highway | Grid

Expr | Trans
RECAP LSTM

\[ g^u = \sigma(W^u H) \]
\[ g^f = \sigma(W^f H) \]
\[ g^o = \sigma(W^o H) \]
\[ c = \tanh(W^c H) \]
\[ m := g^f \odot m + g^u \odot c \]
\[ h := \tanh(g^o \odot m) \]

- input gate
- forget gate
- output gate
- modulated input
- update memory
- output new hidden

No sigmoid-like function when updating memory. Therefore no vanishing gradient in time axis.
A natural extension of LSTM is *Stacked LSTM*

However, it loses the dynamic routing ability along depth axis

The depth cannot be large to prevent vanishing gradient
Applying LSTM to depth axis rather than time axis yields highway network.
GRID LSTM

- LSTM connection along any or all dimensions
- Inputs to any or all dimensions
- Each LSTM dimension maintains its own memory
TRICKS

- Evaluation order of each dimension (*Priority Dimensions*)
- Feedforward connection along some dimensions (*Non-LSTM Dimensions*)
- Weight sharing
EXPERIMENT: TRANSLATION

- Existing work: seq2seq (bottleneck problem) => attention
- 3D Grid LSTM in translation:

Figure 7: Illustration of the 3-LSTM neural translation model.
What other applications of Grid LSTM you can think of?