Implicitly-Defined Neural Networks for Sequence Labeling

WeiYang

weiyang@godweiyang.com

www.godweiyang.com

East China Normal University
Department of Computer Science

2017.11.03

Outline

Formula

- input sequence

$$[\xi_1, \xi_2, \ldots, \xi_n] \tag{1}$$

- states production

$$\begin{aligned} h_1 &= f(\xi_1, h_s) \\ h_2 &= f(\xi_2, h_1) \\ &\ldots \\ h_n &= f(\xi_n, h_{n-1}) \end{aligned} \tag{2}$$

- LSTM & GRU

Structure

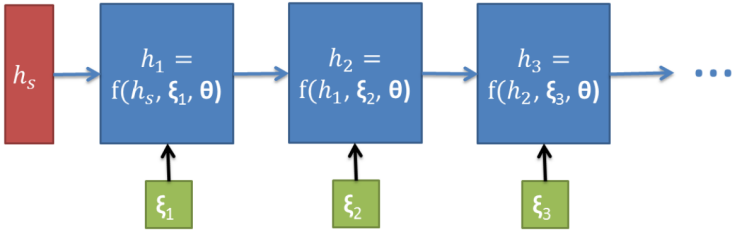

Figure 1: Traditional RNN structure.

Formula

- implicit hidden layer

$$H = [h_1, h_2, \ldots, h_n] \tag{3}$$

where

$$\begin{aligned} h_1 &= f(h_s, h_2, \xi_1) \\ &\ldots \\ h_i &= f(h_{i-1}, h_{i+1}, \xi_i) \\ &\ldots \\ h_n &= f(h_{n-1}, h_e, \xi_n) \end{aligned} \tag{4}$$

- INN

$$\begin{aligned} \xi &= g(\theta, X) \\ H &= F(\theta, \xi, H) \\ L &= \ell(\theta, H, Y) \end{aligned} \tag{5}$$

Structure

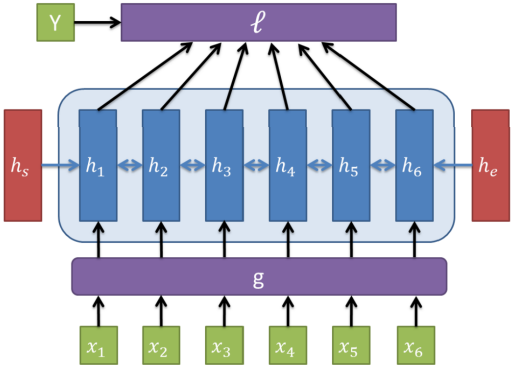

Figure 2: Proposed INN Architecture

Forward propagation

- solve the equation

$$H = F(H) \tag{6}$$

- approximate Newton solve

$$H_{n+1} = H_n - (I - \nabla_H F)^{-1}(H_n - F(H_n)) \tag{7}$$

- Krylov subspace methods(BiCG-STAB method)

Backward propagation

- gradient of the loss function

$$\nabla_\theta L = \nabla_\theta \ell + \nabla_H \ell \nabla_\theta H \tag{8}$$

where

$$\nabla_\theta H = (I - \nabla_H F)^{-1}(\nabla_\theta F + \nabla_\xi F \nabla_\theta \xi) \tag{9}$$

so

$$\nabla_\theta L = \nabla_\theta \ell + \nabla_H \ell (I - \nabla_H F)^{-1}(\nabla_\theta F + \nabla_\xi F \nabla_\theta \xi) \tag{10}$$

- Krylov subspace methods(BiCG-STAB method)

Transition Functions

$$
\begin{aligned}
h_t &= (1 - z_t)\hat{h}_t + z_t\tilde{h}_t \\
\tilde{h}_t &= \tanh(\, Wx_t + U(r_t\hat{h}_t) + \tilde{b}) \\
z_t &= \sigma(\, W_z x_t + U_z\hat{h}_t + b_z) \\
r_t &= \sigma(\, W_r x_t + U_r\hat{h}_t + b_r)
\end{aligned}
\tag{11}
$$

where

$$
\begin{aligned}
\hat{h}_t &= sh_{t-1} + (1 - s)h_{t+1} \\
s &= \tfrac{s_p}{s_p + s_n} \\
s_p &= \sigma(\, W_p x_t + U_p h_{t-1} + b_p) \\
s_n &= \sigma(\, W_n x_t + U_n h_{t+1} + b_n)
\end{aligned}
\tag{12}
$$

POS

Architecture	WSJ Accuracy
GRU	96.43
LSTM	96.47
Bidirectional GRU	97.28
b-LSTM	97.25
INN	**97.37**
Stanford POS Tagger	97.33

Table 2: Tagging performance relative to recurrent architectures and Stanford POS Tagger.