Recurrent Unit and Gated Recurrent Unit

Zhiliang Tian

2016.05.29

1 Introduction

1.1 Recurrent Unit

todo: introduction of Recurrent Unit

1.2 Gated Recurrent Unit

Gated Recurrent Unit is a unit(layer) applying in neural network, called GRU. It is part of recurrent unit family. Comparied with origin recurrent Unit, GRU has gates, which decide how much history information should we use. It is more complex than origin recurrent, but less complex than LSTM(Long short-term memory unit).

2 Feed Forward

2.1 fully connect layer

As we know, a normal fully connect layer in DNN can be described as

$$\mathbf{h} = \phi(W^T * \mathbf{x} + \mathbf{b}) \tag{1}$$

ϕ is some a non-linear transformation, $\mathbf{x}$ is input, $\mathbf{h}$ is output.

2.2 traditional recurrent unit

A traditional recurrent layer can be described as

$$\mathbf{h}^{(t)} = \phi(W_{hx} * \mathbf{x}^{(t)} + W' * \mathbf{h}^{(t-1)} + \mathbf{b}) \tag{2}$$

Compared with Fully connect layer, recurrent unit utilize historical information, adding $(t-1)$-th time's output in unit directly.

2.3 gated recurrent unit

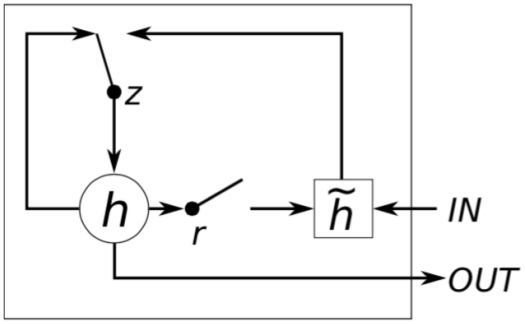

Figure 1: GRU

2.3.1 activation $\mathbf{h}^{(t)}$

$$\mathbf{h}^{(t)} = \mathbf{z}^{(t)} \odot \widetilde{\mathbf{h}^{(t)}} + (\mathbf{1} - \mathbf{z}^{(t)}) \odot \mathbf{h}^{(t-1)} \tag{3}$$

$$h_j^{(t)} = z_j^{(t)} * \widetilde{h_j^{(t)}} + (1 - z_j^{(t)}) * h_j^{(t-1)} \tag{4}$$

GRU's output $\mathbf{h}^{(t)}$ (also called activation of the GRU) combine previous time activation $\mathbf{h}^{(t-1)}$ and current time candidate activation $\widetilde{\mathbf{h}^{(t)}}$. $\odot(\cdot)$ is element-wise multiplication between vectors. $\mathbf{h}$, $\widetilde{\mathbf{h}}$ and $\mathbf{z}$ are column vectors with N dimension.

2.3.2 update gate $\mathbf{z}^{(t)}$

$\mathbf{z}^{(t)}$ is a set of update gates, which decides how much the unit updates its activation. The update gate is computed by

$$\mathbf{z}^{(t)} = \sigma(Wz * \mathbf{x}^{(t)} + Uz * \mathbf{h}^{(t-1)}) \tag{5}$$

$$z_j^{(t)} = \sigma(\sum_i^N Wz_{ij} * x_i^{(t)} + \sum_i^N Uz_{ij} * h_i^{(t-1)}) \tag{6}$$

$\sigma(\cdot)$ is sigmoid function. Wz and Uz is weight matrix with $N * N$ dimension.

2.3.3 candidate activation $\widetilde{\mathbf{h}^{(t)}}$

The candidate activation $\widetilde{\mathbf{h}^{(t)}}$ is computed similar to that of the traditional recurrent unit.

$$\widetilde{\mathbf{h}^{(t)}} = tanh(W\mathbf{x}^{(t)} + U(\mathbf{r}^{(t)} \odot \mathbf{h}^{t-1})) \tag{7}$$

$$\widetilde{h_j^{(t)}} = tanh(\sum_i^N W_{ij}x_i^{(t)} + \sum_i^N U_{ij}(r_i^{(t)} * h_i^{t-1})) \tag{8}$$

2.3.4 candidate activation $\widetilde{\mathbf{h}^{(t)}}$

$\mathbf{r}^{(t)}$ is a set of update gates. When r_j close to 0, it means gate turn off, then GRU forget all previous state(called reset). Reset gate decide how much the candidate utilize history information. The reset gate is computed similar to the update gate:

$$\mathbf{r}^{(t)} = \sigma(Wr * \mathbf{x}^{(t)} + Ur * \mathbf{h}^{(t-1)}) \tag{9}$$

$$r_j^{(t)} = \sigma(\sum_i^N Wr_{ij} * x_i^{(t)} + \sum_i^N Ur_{ij} * h_i^{(t-1)}) \tag{10}$$

2.3.5 summary of GRU's feedforward

$$\mathbf{h}^{(t)} = \mathbf{z}^{(t)} \odot \widetilde{\mathbf{h}^{(t)}} + (\mathbf{1} - \mathbf{z}^{(t)}) \odot \mathbf{h}^{(t-1)} \tag{11}$$

$$\mathbf{z}^{(t)} = \sigma(Wz * \mathbf{x}^{(t)} + Uz * \mathbf{h}^{(t-1)}) \tag{12}$$

$$\widetilde{\mathbf{h}^{(t)}} = tanh(W\mathbf{x}^{(t)} + U(\mathbf{r}^{(t)} \odot \mathbf{h}^{t-1})) \tag{13}$$

$$\mathbf{r}^{(t)} = \sigma(Wr * \mathbf{x}^{(t)} + Ur * \mathbf{h}^{(t-1)}) \tag{14}$$

2.4 some discuss

Todo

3 Back Propagation

3.1 gradient of non-linear transformation

3.1.1 $tanh(\cdot)$

$$t = tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \tag{15}$$

$$dtanh(x) = \frac{\alpha(tanh(x))}{\alpha(x)} = \frac{4}{(\exp(x) + \exp(-x))^2} \tag{16}$$

We call it $dtanh(x)$.

3.1.2 $\sigma(\cdot)$

$$t = \sigma(x) = sigmoid(x) = \frac{1}{1 + \exp(-x)} \tag{17}$$

$$d\sigma(x) = \frac{\alpha(\sigma(x))}{\alpha(x)} = \frac{\exp(-x)}{(1 + \exp(-x))^2} \tag{18}$$

We call it $d\sigma(x)$.

3.2 traditional recurrent unit

Todo

3.3 gated recurrent unit

As many complex topology structure, we compute the gradient from easy to difficult.

3.3.1 $\widetilde{\mathbf{h}^{(t)}}$, W and U

$\widetilde{\mathbf{h}^{(t)}}$:

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\widetilde{h^{(t)}}_j)} = \frac{\alpha(h_j^{(t)})}{\alpha(\widetilde{h^{(t)}}_j)} = z_j^{(t)} \tag{19}$$

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\widetilde{\mathbf{h}^{(t)}})} = \mathbf{z}^{(t)} \tag{20}$$

W:

$$\begin{aligned} &\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(W)} = \frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\widetilde{\mathbf{h}^{(t)}})} \frac{\alpha(tanh(\mathbf{t}))}{\alpha(\mathbf{t})} \frac{\alpha(W\mathbf{x}^{(t)} + U(\mathbf{r}^{(t)} \odot \mathbf{h}^{t-1}))}{\alpha(W)} \\ &\mathbf{t} = W\mathbf{x}^{(t)} + U(\mathbf{r}^{(t)} \odot \mathbf{h}^{t-1}) \end{aligned} \tag{21}$$

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(W_{ij})} = \frac{\alpha(\mathbf{h}_j^{(t)})}{\alpha(W_{ij})} = \frac{\alpha(\mathbf{h}_j^{(t)})}{\alpha(\widetilde{\mathbf{h}_j^{(t)}})} \frac{\alpha(tanh(\mathbf{t}_j))}{\alpha(\mathbf{t}_j)} \frac{\alpha(\sum_i^N W_j \mathbf{x}_i^{(t)})}{\alpha(W_{ij})} \tag{22}$$

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(W_{ij})} = z_j^{(t)} * dtanh(t_j) * x_i^{(t)} \tag{23}$$

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(W)} = \frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\widetilde{\mathbf{h}^{(t)}})} \odot \frac{\alpha(tanh(\mathbf{t}))}{\alpha(\mathbf{t})} \frac{\alpha(W\mathbf{x}^{(t)})}{\alpha(W)} \tag{24}$$

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(W)} = (\mathbf{z}^{(t)} \odot dtanh(\mathbf{t}))\{\mathbf{x}^{(t)}, \mathbf{x}^{(t)}, ..., \mathbf{x}^{(t)}\} \tag{25}$$

Notice that $\mathbf{z}^{(t)}$, $dtanh(\mathbf{t})$ and $\mathbf{x}^{(t)}$ are column vectors with N dimension, $\{\mathbf{x}^{(t)}, \mathbf{x}^{(t)}, ..., \mathbf{x}^{(t)}\}$ is a $N * N$ matrix consisting of N column vector $\mathbf{x}^{(t)}$.

U:

Similar to W:

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(U)} = (\mathbf{z}^{(t)} \odot dtanh(\mathbf{t}))\{\mathbf{r}^{(t)} \odot \mathbf{h}^{(t-1)}, \mathbf{r}^{(t)} \odot \mathbf{h}^{(t-1)}, ..., \mathbf{r}^{(t)} \odot \mathbf{h}^{(t-1)}\} \tag{26}$$

$\mathbf{r}^{(t)} \odot \mathbf{h}^{(t-1)}$ is a column vector.

3.3.2 $\mathbf{z}^{(t)}$, Wz and Uz

Similar to $\widetilde{\mathbf{h}^{(t)}}$, W and U:

$\mathbf{z}^{(t)}$:

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(z_j^{(t)})} = \frac{\alpha(h_j^{(t)})}{\alpha(z_j^{(t)})} = \widetilde{h^{(t)}}_j - h_j^{(t-1)} \tag{27}$$

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{z}^{(t)})} = \widetilde{\mathbf{h}^{(t)}} - \mathbf{h}^{(t-1)} \tag{28}$$

Wz:

$$\begin{aligned} &\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(Wz)} = \frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{z}^{(t)})} \frac{\alpha(\sigma(\mathbf{tz}))}{\alpha(\mathbf{tz})} \frac{\alpha(Wz\mathbf{x}^{(t)} + Uz(\mathbf{h}^{t-1}))}{\alpha(Wz)} \\ &\mathbf{tz} = Wz\mathbf{x}^{(t)} + Uz(\mathbf{h}^{t-1}) \end{aligned} \tag{29}$$

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(Wz)} = ((\widetilde{\mathbf{h}^{(t)}} - \mathbf{h}^{(t-1)}) \odot d\sigma(\mathbf{tz}))\{\mathbf{x}^{(t)}, \mathbf{x}^{(t)}, ..., \mathbf{x}^{(t)}\} \tag{30}$$

Uz:

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(Uz)} = ((\widetilde{\mathbf{h}^{(t)}} - \mathbf{h}^{(t-1)}) \odot d\sigma(\mathbf{tz}))\{\mathbf{h}^{(t-1)}, \mathbf{h}^{(t-1)}, ..., \mathbf{h}^{(t-1)}\} \tag{31}$$

$\mathbf{h}^{(t-1)}$ is a column vector.

3.3.3 $\mathbf{r}^{(t)}$, Wr and Ur

We compute the gradient based on the gradient of $\widetilde{\mathbf{h}^{(t)}}$

$\mathbf{r}^{(t)}$:

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{r}^{(t)})} = \frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\widetilde{\mathbf{h}^{(t)}})} \odot \frac{\alpha(tanh(\mathbf{t}))}{\alpha(\mathbf{t})} \frac{\alpha(W\mathbf{x}^{(t)} + U(\mathbf{r}^{(t)} \odot \mathbf{h}^{t-1}))}{\alpha(\mathbf{r}^{(t)})} \tag{32}$$

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{r}^{(t)})} = (\mathbf{z}^{(t)} \odot dtanh(\mathbf{t})) \frac{\alpha(U(\mathbf{r}^{(t)} \odot \mathbf{h}^{t-1}))}{\alpha(\mathbf{r}^{(t)} \odot \mathbf{h}^{t-1}))} \frac{\alpha(\mathbf{r}^{(t)} \odot \mathbf{h}^{t-1}))}{\alpha(\mathbf{r}^{(t)})} \tag{33}$$

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{r}^{(t)})} = (\mathbf{z}^{(t)} \odot dtanh(\mathbf{t})) \frac{\alpha(U(\mathbf{r}^{(t)} \odot \mathbf{h}^{t-1}))}{\alpha(\mathbf{r}^{(t)} \odot \mathbf{h}^{t-1}))} \odot \mathbf{h}^{t-1} \tag{34}$$

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{r}^{(t)})} = (U^T(\mathbf{z}^{(t)} \odot dtanh(\mathbf{t}))) \odot \mathbf{h}^{t-1} \tag{35}$$

From equation 34 to equation 35, we revise the law of matrix's gradient. For a column vector $\mathbf{y}$ with M dimension, a column vector $\mathbf{x}$ with N dimension and a $N * M$ matrix W:

$$\mathbf{y} = W^T\mathbf{x} \tag{36}$$

The gradient:

$$\frac{\alpha(\mathbf{y}_k)}{\alpha(\mathbf{x}_i)} = W_{ki} \tag{37}$$

$$\frac{\alpha(\mathbf{y})}{\alpha(\mathbf{x})} = W^T \tag{38}$$

$$\Delta(\mathbf{y}) = \frac{\alpha(\mathbf{y})}{\alpha(\mathbf{x})}\Delta(\mathbf{x}) = W^T\Delta(\mathbf{x}) \tag{39}$$

But $\frac{\alpha(\mathbf{y})}{\alpha(\mathbf{x})}$ is a matrix, how can we use it. The matrix will multiply with y's gradient. For example, in a model, J is loss from model, the gradient of J respect to $\mathbf{x}$ can be computed.

Using equation 10

$$W * \Delta(\mathbf{y}) = \Delta(\mathbf{x}) \tag{40}$$

$$\frac{\alpha(J)}{\alpha(\mathbf{x})} = \Delta(\mathbf{x}) = W * \Delta(\mathbf{y}) = W\frac{\alpha(J)}{\alpha(\mathbf{y})} \tag{41}$$

Wr: Similar to Wz:

$$\begin{aligned}&\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(Wz)} = \frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{r}^{(t)})}\frac{\alpha(\sigma(\mathbf{tr}))}{\alpha(\mathbf{tr})}\frac{\alpha(Wr\mathbf{x}^{(t)}+Ur(\mathbf{h}^{t-1}))}{\alpha(Wr)}\\&\mathbf{tr} = Wr\mathbf{x}^{(t)} + Ur(\mathbf{h}^{t-1})\end{aligned} \tag{42}$$

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(Wr)} = (((U^T(\mathbf{z}^{(t)} \odot dtanh(\mathbf{t}))) \odot \mathbf{h}^{t-1}) \odot d\sigma(\mathbf{tr}))\{\mathbf{x}^{(t)}, \mathbf{x}^{(t)}, ..., \mathbf{x}^{(t)}\} \tag{43}$$

Ur:

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(Ur)} = (((U^T(\mathbf{z}^{(t)} \odot dtanh(\mathbf{t}))) \odot \mathbf{h}^{t-1}) \odot d\sigma(\mathbf{tr}))\{\mathbf{h}^{(t-1)}, \mathbf{h}^{(t-1)}, ..., \mathbf{h}^{(t-1)}\} \tag{44}$$

$\mathbf{h}^{(t-1)}$ is a column vector.

3.3.4 $\mathbf{x}^{(t)}$

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{x}^{(t)})} = \frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{z}^{(t)})} * \frac{\alpha(\mathbf{z}^{(t)})}{\alpha(\mathbf{x}^{(t)})} + \frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{r}^{(t)})} * \frac{\alpha(\mathbf{r}^{(t)})}{\alpha(\mathbf{x}^{(t)})} + \frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\widetilde{\mathbf{h}^{(t)}})} * \frac{\alpha(\widetilde{\mathbf{h}^{(t)}})}{\alpha(\mathbf{x}^{(t)})} \tag{45}$$

$$\begin{aligned}&\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{x}^{(t)})} =\\&((\widetilde{\mathbf{h}^{(t)}} - \mathbf{h}^{(t-1)}) \odot d\sigma(\mathbf{tz})) * \frac{\alpha(Wz\mathbf{x}^{(t)})}{\alpha(\mathbf{x})}\\&+(U^T(\mathbf{z}^{(t)} \odot dtanh(\mathbf{t}))) \odot \mathbf{h}^{t-1} \odot d\sigma(\mathbf{tr})) * \frac{\alpha(Wr\mathbf{x}^{(t)})}{\alpha(\mathbf{x})}\\&+(\mathbf{z}^{(t)} \odot dtanh(\mathbf{t}))\frac{\alpha(W\mathbf{x}^{(t)})}{\alpha(\mathbf{x})}\end{aligned} \tag{46}$$

$$\begin{aligned}&\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{x}^{(t)})} = Wz * ((\widetilde{\mathbf{h}^{(t)}} - \mathbf{h}^{(t-1)}) \odot d\sigma(\mathbf{tz}))\\&+Wr * (U^T(\mathbf{z}^{(t)} \odot dtanh(\mathbf{t}))) \odot \mathbf{h}^{t-1} \odot d\sigma(\mathbf{tr})) + W * (\mathbf{z}^{(t)} \odot dtanh(\mathbf{t}))\end{aligned} \tag{47}$$

3.3.5 $\mathbf{h}^{(t-1)}$

Similar to $\mathbf{x}^{(t)}$:

$$\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{h}^{(t-1)})} = \frac{\alpha((\mathbf{1} - \mathbf{z}^{(t)}) \odot \mathbf{h}^{(t-1)})}{\alpha(\mathbf{h}^{(t-1)})} + \frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{z}^{(t)})} * \frac{\alpha(\mathbf{z}^{(t)})}{\alpha(\mathbf{h}^{(t-1)})} + \frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{r}^{(t)})} * \frac{\alpha(\mathbf{r}^{(t)})}{\alpha(\mathbf{h}^{(t-1)})} + \frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\widetilde{\mathbf{h}^{(t)}})} * \frac{\alpha(\widetilde{\mathbf{h}^{(t)}})}{\alpha(\mathbf{h}^{(t-1)})} \tag{48}$$

$$
\begin{aligned}
&\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{h}^{(t-1)})} = \\
&(\mathbf{1}-\mathbf{z}^{(t)}) + ((\widetilde{\mathbf{h}^{(t)}} - \mathbf{h}^{(t-1)}) \odot d\sigma(\mathbf{tz})) * \frac{\alpha(Uz\mathbf{h}^{(t-1)})}{\alpha(\mathbf{h}^{(t-1)})} \\
&+(U^T(\mathbf{z}^{(t)} \odot dtanh(\mathbf{t}))) \odot \mathbf{h}^{t-1} \odot d\sigma(\mathbf{tr}) * \frac{\alpha(Ur\mathbf{h}^{(t-1)})}{\alpha(\mathbf{h}^{(t-1)})} \\
&+(\mathbf{z}^{(t)} \odot dtanh(\mathbf{t}))\frac{\alpha(U(\mathbf{r}^{(t)} \odot \mathbf{h}^{t-1}))}{\alpha(\mathbf{r}^{(t)} \odot \mathbf{h}^{t-1})} * \frac{\alpha(\mathbf{r}^{(t)} \odot \mathbf{h}^{t-1})}{\mathbf{h}^{t-1}}
\end{aligned}
\tag{49}
$$

$$
\begin{aligned}
&\frac{\alpha(\mathbf{h}^{(t)})}{\alpha(\mathbf{h}^{(t-1)})} = \\
&(\mathbf{1}-\mathbf{z}^{(t)}) + Uz * ((\widetilde{\mathbf{h}^{(t)}} - \mathbf{h}^{(t-1)}) \odot d\sigma(\mathbf{tz})) \\
&+Ur * ((U^T(\mathbf{z}^{(t)} \odot dtanh(\mathbf{t}))) \odot \mathbf{h}^{t-1} \odot d\sigma(\mathbf{tr})) \\
&+(U * (\mathbf{z}^{(t)} \odot dtanh(\mathbf{t}))) * \mathbf{r}^{(t)}
\end{aligned}
\tag{50}
$$

3.4 some discuss

Todo