Recurrent Attention Unit with Step-down Optimized by NADAM

Enrique Boswell Nueve IV

March 26, 2019

In this document, the author will be providing full deviations for a Recurrent Attention Unit[2]. This unit has been modified with a Step-down Neural network to allow for differences in dimensionality between the inputs and targets. The parameters for the RAU are being updated with the optimization technique NADAM (Nesterov-accelerated adaptive moment estimation)[1].

References

[1] Dozat, Timothy. "Incorporating nesterov momentum into adam." (2016).

[2] Zhong, Guoqiang, Guohua Yue, and Xiao Ling. "Recurrent Attention Unit." arXiv preprint arXiv:1810.12754 (2018).

Contents

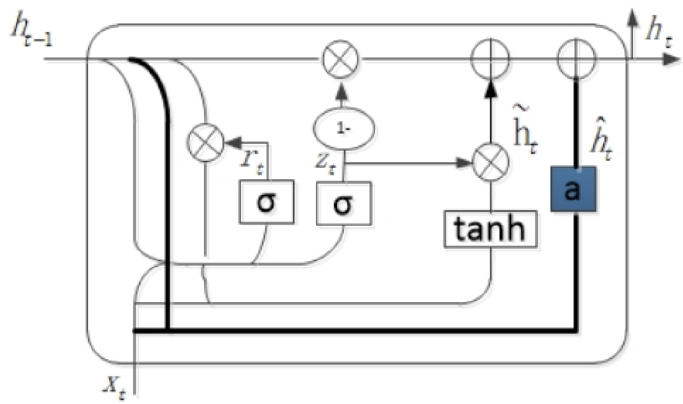

Figure 1: Recurrent Attention Unit Diagram[2].

1 Forward Pass at Time t

1.1 Update Gate

$$z_t = \sigma\big(W_z x_t + U_z h_{t-1}\big)$$

1.2 Reset Gate

$$r_t = \sigma\big(W_r x_t + U_r h_{t-1}\big)$$

1.3 Candidate Gate

$$\tilde{h_t} = \tanh\big(W_c x_t + (U_c h_{t-1}) \odot r_t\big)$$

1.4 Attention Gate

$$\hat{h_t} = \tanh\big(W_a \gamma_t\big)$$

1.5 Probabilistic Vector

$$\gamma_t = \text{softmax}\big(\alpha_t\big)$$

1.6 Learning Function: General

$$\alpha_t = W_{\alpha_t} h_{t-1}^T x_t$$

1.7 State Vector

$$h_t = \big(1 - z_t\big)h_{t-1} + \big(\frac{z_t}{2}\big)\big(\tilde{h_t} + \hat{h_t}\big)$$

1.8 Step Down Network

$$y = \sigma\big(W^T h_t\big)$$

1.9 Error Calculation

$$E = \frac{1}{2}\big(\hat{y_t} - y_t\big)^2$$

2 Backpropagation of Weights at Time t

2.1 Update Gate Present Weights

$$\frac{\delta E_t}{\delta W_z} = \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_t} \cdot \frac{\delta h_t}{\delta z_t^o} \cdot \frac{\delta z_t^o}{\delta z_t^i} \cdot \frac{\delta z_t^i}{\delta W_z}$$

$$\frac{\delta E_t}{\delta W_z} = x_t[\big(W[(y - \hat{y}) \odot \sigma(W^T h_t) \odot (1 - \sigma(W^T h_t))]\big) \odot (\frac{\tilde{h_t}}{2} + \frac{\hat{h_t}}{2} - h_{t-1}) \odot (\sigma(W_z x_t + U_z x_t)$$
$$\odot (1 - \sigma(W_z x_t + U_z h_{t-1})))]^T$$

2.2 Update Gate Memory Weights

$$\frac{\delta E_t}{\delta U_z} = \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_t} \cdot \frac{\delta h_t}{\delta z_t^o} \cdot \frac{\delta z_t^o}{\delta z_t^i} \cdot \frac{\delta z_t^i}{\delta U_z}$$

$$\frac{\delta E_t}{\delta U_z} = h_{t-1}[\big(W[(y - \hat{y}) \odot (W^T h_t) \odot (1 - (W^T h_t))]\big) \odot (\frac{\tilde{h_t}}{2} + \frac{\hat{h_t}}{2} - h_{t-1}) \odot (\sigma(W_z x_t + U_z x_t)$$
$$\odot (1 - \sigma(W_z x_t + U_z h_{t-1})))]^T$$

2.3 Candidate Gate Present Weights

$$\frac{\delta E_t}{W_{\tilde{h_t}}} = \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_t} \cdot \frac{\delta h_t}{\tilde{h_t^o}} \cdot \frac{\delta \tilde{h_t^o}}{\delta \tilde{h_t^i}} \cdot \frac{\delta \tilde{h_t^i}}{W_{\tilde{h_t}}}$$

$$\frac{\delta E_t}{W_{\tilde{h_t}}} = x_t[\big(W[(y-\hat{y}) \odot \sigma(W^T h_t) \odot (1-\sigma(W^T h_t))]\big) \odot (\frac{z_t}{2}) \odot (1-\tanh^2(W_{\tilde{h_t}} x_t + (U_{\tilde{h_t}} h_{t-1}) \odot r_t)]^T$$

2.4 Candidate Gate Memory Weights

$$\frac{\delta E_t}{U_{\tilde{h}}} = \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_t} \cdot \frac{\delta h_t}{\tilde{h_t^o}} \cdot \frac{\delta \tilde{h_t^o}}{\delta \tilde{h_t^i}} \cdot \frac{\delta \tilde{h_t^i}}{U_{\tilde{h_t}}}$$

$$\frac{\delta E_t}{U_{\tilde{h_t}}} = (h_{t-1} \odot r_t)[\big(W[(y-\hat{y}) \odot \sigma(W^T h_t) \odot (1-\sigma(W^T h_t))]\big) \odot (\frac{z_t}{2})$$
$$\odot (1-\tanh^2(W_{\tilde{h_t}} x_t + (U_{\tilde{h_t}} h_{t-1}) \odot r_t))]^T$$

2.5 Reset Gate Present Weights

$$\frac{\delta E_t}{\delta W_r} = \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_t} \cdot \frac{\delta h_t}{\delta \tilde{h_t^o}} \cdot \frac{\delta \tilde{h_t^o}}{\delta \tilde{h_t^i}} \cdot \frac{\delta \tilde{h_t^i}}{\delta r_t^o} \cdot \frac{\delta r_t^o}{\delta r_t^i} \cdot \frac{\delta r_t^i}{\delta W_r}$$

$$\frac{\delta E_t}{\delta W_r} = x_t[\big(W[(y-\hat{y}) \odot \sigma(W^T h_t) \odot (1-\sigma(W^T h_t))]\big) \odot (\frac{z_t}{2}) \odot (1-\tanh^2(W_{\tilde{h_t}} x_t + (U_{\tilde{h}} h_{t-1}) \odot r_t))$$
$$\odot (U_{\tilde{h_t}} h_{t-1}) \odot (\sigma(W_r x_t + U_r h_{t-1}) \odot (1-\sigma(W_r x_t + U_r h_{t-1})))]^T$$

2.6 Reset Gate Memory Weights

$$\frac{\delta E_t}{\delta U_r} = \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_t} \cdot \frac{\delta h_t}{\delta \tilde{h_t^o}} \cdot \frac{\delta \tilde{h_t^o}}{\delta \tilde{h_t^i}} \cdot \frac{\delta \tilde{h_t^i}}{\delta r_t^o} \cdot \frac{\delta r_t^o}{\delta r_t^i} \cdot \frac{\delta r_t^i}{\delta U_r}$$

$$\frac{\delta E_t}{\delta U_r} = h_{t-1}[\big(W[(y-\hat{y}) \odot \sigma(W^T h_t) \odot (1-\sigma(W^T h_t))]\big) \odot (\frac{z_t}{2}) \odot (1-\tanh^2(W_{\tilde{h_t}} x_t + (U_{\tilde{h_t}} h_{t-1}) \odot r_t))$$
$$\odot (U_{\tilde{h_t}} h_{t-1}) \odot (\sigma(W_r x_t + U_r h_{t-1}) \odot (1-\sigma(W_r x_t + U_r h_{t-1})))]^T$$

2.7 Attention Gate Weight

$$\frac{\delta E_t}{\delta W_a} = \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_t} \cdot \frac{\delta h_t}{\delta \hat{h_t}} \cdot \frac{\delta \hat{h_t}}{\delta W_a}$$

$$\frac{\delta E_t}{\delta W_a} = \gamma_t[\big(W[(y-\hat{y}) \odot \sigma(W^T h_t) \odot (1-\sigma(W^T h_t))]\big) \odot (\frac{z_t}{2}) \odot (1-\tanh^2(W_a \gamma_t))]^T$$

2.8 Learning Function Weight

$$
\begin{aligned}
\frac{\delta E_t}{W_\alpha} &= \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_t} \cdot \frac{\delta h_t}{\delta \hat{h_t}} \cdot \frac{\delta \hat{h_t}}{\delta \gamma_t} \cdot \frac{\delta \gamma_t}{\delta \alpha_t} \cdot \frac{\delta \alpha_t}{\delta W_\alpha} \\
\frac{\delta E_t}{\delta W_\alpha} &= h_{t-1}^T x_t [(\big(W[(y-\hat{y}) \odot \sigma(W^T h_t) \odot (1-\sigma(W^T h_t))]\big) \odot (\frac{z_t}{2}) \odot (1-\tanh^2(W_a \gamma_t)) \\
&\quad \odot softmax'(\alpha_t)]^T
\end{aligned}
$$

2.9 Step-down Network Weight

$$
\begin{aligned}
\frac{\delta E_t}{\delta W} &= \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta W} \\
\frac{\delta E_t}{\delta W} &= h_t [(y-\hat{y}) \odot \sigma(W^T h_t) \odot (1-\sigma(W^T h_t))]^T
\end{aligned}
$$

3 Backpropagation of State at time t

3.1 Past State of Update Gate

$$
\begin{aligned}
\frac{\delta E_t}{\delta h_{t-1}} &= \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_t} \cdot \frac{\delta h_t}{\delta z_t^o} \cdot \frac{\delta z_t^o}{\delta z_t^i} \cdot \frac{\delta z_t^i}{\delta h_{t-1}} \\
\frac{\delta E_t}{\delta h_{t-1}} &= U_z[\big(W[(y-\hat{y}) \odot \sigma(W^T h_t) \odot (1-\sigma(W^T h_t))]\big) \odot (\frac{\tilde{h_t}}{2} + \frac{\hat{h_t}}{2} - h_{t-1}) \odot (\sigma(W_z x_t + U_z x_t) \\
&\quad \odot (1-\sigma(W_z x_t + U_z h_{t-1}))))]
\end{aligned}
$$

3.2 Past State of Candidate Gate

$$
\begin{aligned}
\frac{\delta E_t}{h_{t-1}} &= \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_t} \cdot \frac{\delta h_t}{\tilde{h_t^o}} \cdot \frac{\delta \tilde{h_t^o}}{\delta \tilde{h_t^i}} \cdot \frac{\delta \tilde{h_t^i}}{h_{t-1}} \\
\frac{\delta E_t}{h_{t-1}} &= (U_{\tilde{h_t}} \odot r_t)[\big(W[(y-\hat{y}) \odot \sigma(W^T h_t) \odot (1-\sigma(W^T h_t))]\big) \odot (\frac{z_t}{2}) \\
&\quad \odot (1-\tanh^2(W_{\tilde{h_t}} x_t + (U_{\tilde{h_t}} h_{t-1}) \odot r_t)]
\end{aligned}
$$

3.3 Past State of Reset Gate

$$
\begin{aligned}
\frac{\delta E_t}{\delta h_{t-1}} &= \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_t} \cdot \frac{\delta h_t}{\delta \tilde{h_t^o}} \cdot \frac{\delta \tilde{h_t^o}}{\delta \tilde{h_t^i}} \cdot \frac{\delta \tilde{h_t^i}}{\delta r_t^o} \cdot \frac{\delta r_t^o}{\delta r_t^i} \cdot \frac{\delta r_t^i}{\delta h_{t-1}} \\
\frac{\delta E_t}{\delta h_{t-1}} &= U_r[\big(W[(y-\hat{y}) \odot \sigma(W^T h_t) \odot (1-\sigma(W^T h_t))]\big) \odot (\frac{z_t}{2}) \odot (1-\tanh^2(W_{\tilde{h_t}} x_t + (U_{\tilde{h_t}} h_{t-1}) \odot r_t)) \\
&\quad \odot (U_{\tilde{h_t}} h_{t-1}) \odot (\sigma(W_r x_t + U_r h_{t-1}) \odot (1-\sigma(W_r x_t + U_r h_{t-1}))))]
\end{aligned}
$$

3.4 Past State of Attention Gate

$$\frac{\delta E_t}{\delta h_{t-1}} = \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_t} \cdot \frac{\delta h_t}{\delta \hat{h_t}} \cdot \frac{\delta \hat{h_t}}{\delta h_{t-1}}$$
$$\frac{\delta E_t}{\delta h_{t-1}} = W_{\alpha_t} x_t [\big(W[(y-\hat{y}) \odot \sigma(W^T h_t) \odot (1-\sigma(W^T h_t))]\big) \odot (\frac{z_t}{2}) \odot (1-\tanh^2(W_a \gamma_t)) \odot softmax'(\alpha_t))]$$

3.5 Past State of State Vector

$$\frac{\delta E_t}{\delta h_{t-1}} = \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_{t-1}}$$
$$\frac{\delta E_t}{\delta h_{t-1}} = \big(W[(y-\hat{y}) \odot \sigma(W^T h_t) \odot (1-\sigma(W^T h_t)) \odot (1-z_t)$$

3.6 Backpropagation State Value

$$\frac{\delta E_t}{\delta h_{t-1}} = \frac{\delta E_t}{\delta z_t^i} \cdot \frac{\delta z_t^i}{\delta h_{t-1}} + \frac{\delta E_t}{\delta \tilde{h_t^i}} \cdot \frac{\delta \tilde{h_t^i}}{h_{t-1}} + \frac{\delta E_t}{\delta r_t^i} \cdot \frac{\delta r_t^i}{\delta h_{t-1}} + \frac{\delta E_t}{\delta \hat{h_t}} \cdot \frac{\delta \hat{h_t}}{\delta h_{t-1}} + \frac{\delta E_t}{\delta y_t} \cdot \frac{\delta y_t}{\delta h_{t-1}}$$

4 Parameter Update

4.1 Update Update Gate

$$\nabla_{Wz} E = \sum_{t=0}^{T} \frac{\delta E_t}{\delta W_z} \qquad \nabla_{Uz} E = \sum_{t=0}^{T} \frac{\delta E_t}{\delta U_z}$$

4.2 Update Reset Gate

$$\nabla_{Wr} E = \sum_{t=0}^{T} \frac{\delta E_t}{\delta W_r} \qquad \nabla_{Ur} E = \sum_{t=0}^{T} \frac{\delta E_t}{\delta U_r}$$

4.3 Update Candidate Gate

$$\nabla_{Wc} E = \sum_{t=0}^{T} \frac{\delta E_t}{\delta W_c} \qquad \nabla_{Uc} E = \sum_{t=0}^{T} \frac{\delta E_t}{\delta U_c}$$

4.4 Update Attention Weight

$$\nabla_{W\,a} E = \sum_{t=0}^{T} \frac{\delta E_t}{\delta W_a}$$

4.5 Update Learning Function Weight

$$\nabla_{W\,\alpha} E = \sum_{t=0}^{T} \frac{\delta E_t}{\delta W_\alpha}$$

4.6 Optimization Technique: NADAM

Algorithm 1 Nesterov-accelerated adaptive moment estimation

$g_t \leftarrow \nabla_{\theta_{t-1}} f(\theta_{t-1})$
$\hat{g}_t \leftarrow \frac{g_t}{1-\prod_{i=1}^{t} \mu_i}$
$m_t \leftarrow \mu m_{t-1} + (1-\mu) g_t$ ▷ μ=.99
$\hat{m}_t \leftarrow \frac{m_t}{1-\prod_{i=1}^{t+1} \mu_i}$
$n_t \leftarrow v n_{t-1} + (1-v) g_t^2$ ▷ v=.9
$\hat{n}_t \leftarrow \frac{n_t}{1-v^t}$
$\bar{m}_t \leftarrow (1-\mu_t)\hat{g}_t + \mu_{t+1}\hat{m}_t$
$\theta_t \leftarrow \theta_{t-1} - \eta \frac{\bar{m}_t}{\sqrt{\hat{n}_t}+\varepsilon}$ ▷ $\varepsilon = 10e^-8$

5 Activation Functions

5.1 Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

5.2 Hyperbolic Tangent

$$\tanh(x) = \frac{e^x + e^{-x}}{e^x - e^{-x}}$$

5.3 Softmax

$$softmax(x)_j = \frac{e^{x_j}}{\sum_{k=1}^{K} e^{x_k}} \quad \text{for j=1,...,k}$$