Adaptive Computation Time

Dani Fojo

November 18, 2017

1 Adaptive Computation Time

$$s_t^n = \begin{cases} \mathcal{S}(s_{t-1}, x_t^1) \text{ if } n = 1 \\ \mathcal{S}(s_t^{n-1}, x_t^n) \text{ if } n > 1 \end{cases} \tag{1}$$

$$y_t^n = W_y s_t^n + b_y \tag{2}$$

$$p_t^n = \sigma(W_p s_t^n + b_p) \tag{3}$$

$$N(t) = \min\{k : \sum_{j=1}^{k} p_t^j > 1 - \varepsilon\} \tag{4}$$

$$p_t^{N(t)} = R(t) = 1 - \sum_{j=1}^{N(t)-1} p_t^j \tag{5}$$

$$s_t = \sum_{j=1}^{N(t)} p_t^j s_t^j \tag{6}$$

$$y_t = \sum_{j=1}^{N(t)} p_t^j y_t^j \tag{7}$$

$$\rho_t = N(t) + R(t) \tag{8}$$

$$\rho_t = N(t) + R(t) = N(t) + 1 - \sum_{j=1}^{N(t)-1} p_t^j \tag{9}$$

2 Straight-Through Estimator

$$
\begin{aligned}
f_b \colon \mathbb{R} &\to \{0,1\} && (10)\\
x \mapsto f_b(x) &= \begin{cases} 0 \text{ if } x < 1-\varepsilon \\ 1 \text{ if } x \geq 1-\varepsilon \end{cases} \quad . && (11)
\end{aligned}
$$

$$
s_t^n = \begin{cases} \mathcal{S}(s_{t-1}, x_t^1) \text{ if } n = 1 \\ \mathcal{S}(s_t^{n-1}, x_t^n) \text{ if } n > 1 \end{cases} \tag{12}
$$

$$
y_t^n = W_y s_t^n + b_y \tag{13}
$$

$$
p_t^n = \sigma(W_p s_t^n + b_p) \tag{14}
$$

$$
h_t^n = \sum_{j=1}^{n} p_t^j \tag{15}
$$

$$
N(t) = \min\{k : \sum_{j=1}^{k} f_b(h_t^j) = 1\} = \sum_{j=1}^{\infty} (1 - f_b(h_t^j)) \tag{16}
$$

$$
s_t = \sum_{j=1}^{N(t)} f_b(h_t^j) s_t^j = s_t^{N(t)} \tag{17}
$$

$$
y_t = \sum_{j=1}^{N(t)} f_b(h_t^j) y_t^j = y_t^{N(t)} \tag{18}
$$

$$
\rho_t = N(t) \tag{19}
$$

3 Skip-ACT

$$f_b \colon \mathbb{R} \to \{0,1\} \tag{20}$$

$$x \mapsto f_b(x) = \begin{cases} 0 \text{ if } x < 1-\varepsilon \\ 1 \text{ if } x \geq 1-\varepsilon \end{cases} . \tag{21}$$

$$\frac{\partial f_b}{\partial x} = 1 \tag{22}$$

$$y_t^n = W_y s_t^n + b_y \tag{23}$$

$$p_t^0 = h_{t-1}^{N(t)} - 1 \tag{24}$$

$$p_t^n = \mu\sigma(W_p s_t^n + b_p); \quad \mu \in \mathbb{N} \tag{25}$$

$$h_t^n = \sum_{j=0}^{n} p_t^j \tag{26}$$

$$N(t) = \min\{k : \sum_{j=0}^{k} f_b(h_t^j) = 1\} = \sum_{j=0}^{\infty}(1 - f_b(h_t^j)) \tag{27}$$

$$s_t = f_b(h_t^0)s_{t-1} + (1 - f_b(h_t^0)) \sum_{j=0}^{N(t)} f_b(h_t^j)s_t^j \tag{28}$$

$$y_t = f_b(h_t^0)y_{t-1} + (1 - f_b(y_t^0)) \sum_{j=0}^{N(t)} f_b(h_t^j)y_t^j \tag{29}$$

$$\rho_t = N(t) \tag{30}$$