Towards Binary-Valued Gates for Robust LSTM Training

Zhuohan Li, Di He, Fei Tian, Wei Chen, Tao Qin, Liwei Wang, Tie-Yan Liu
Peking University & Microsoft Research Asia

北京大學 PEKING UNIVERSITY

Microsoft

1. Long Short-Term Memory (LSTM) RNN

- $h_t, c_t = \mathrm{LSTM}(h_{t-1}, c_{t-1}, x_t)$
- $f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$
- $i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$
- $g_t = \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g)$
- $o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$
- $c_t = f_t \odot c_{t-1} + i_t \odot g_t$
- $h_t = o_t \odot \tanh(c_t)$

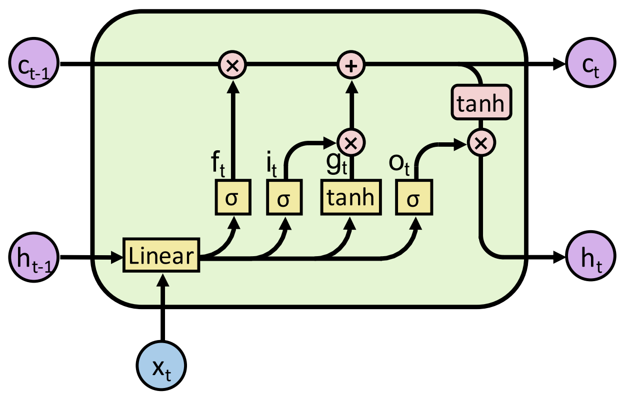
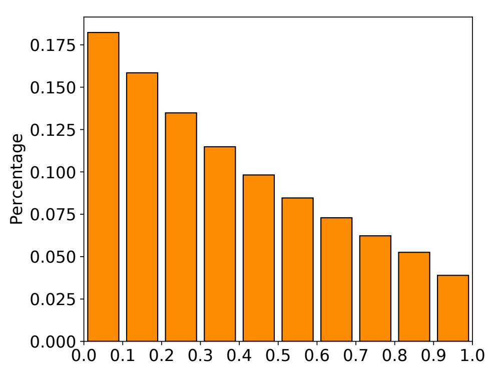
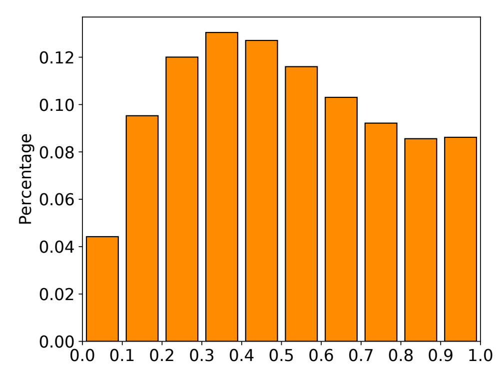


Figure credit to: Christopher Olah, "Understanding LSTM Networks"

2. Histograms of Gate Distributions in LSTM



LSTM Input gates

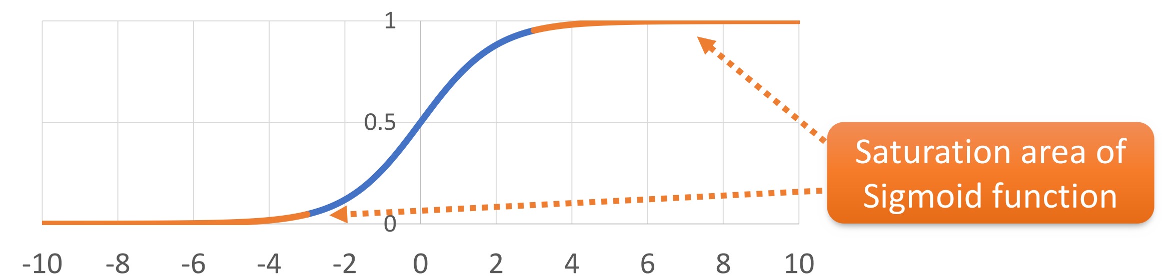


LSTM Forget gates

3. Training LSTM Gates Towards Binary Values

Push the **gate values** to the **boundary** of range (0, 1)

- Well aligns with **the original purpose** of gates: to get the information in or skip by "opening" or "closing"
- **Ready for further compression** by pushing the activation function to be binarized
- Enables **better generalization**



Output falls in the saturation area → Parameters in the gates perturb → Change to **the output of the gates** will be small → Change to **the final loss** will also be little

Robust to **model compression**

Better test performance
Flat region generalize better

4. Gumbel-Softmax Estimator

- Straight forward idea: **sharpen the Sigmoid function** by using a smaller temperature $\tau < 1$

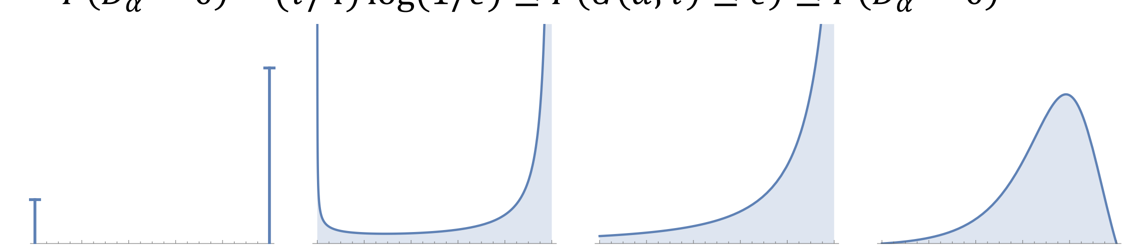
$$f_{W,b}(x) = \sigma((Wx+b)/\tau) = \sigma((W/\tau)x + (b/\tau))$$

- We leverage the **Gumbel-Softmax estimator** to estimate the Bernoulli distribution $D_\alpha \sim B(\sigma(\alpha))$ with prob. $\sigma(\alpha)$. Define

$$G(\alpha, \tau) = \sigma\left(\frac{\alpha + \log U - \log(1-U)}{\tau}\right),$$

where $U \sim \mathrm{Uniform}(0,1)$, then the following holds for $\epsilon \in (0, 1/2)$:

$$P(D_\alpha = 1) - (\tau/4)\log(1/\epsilon) \le P(G(\alpha,\tau) \ge 1-\epsilon) \le P(D_\alpha = 1)$$
$$P(D_\alpha = 0) - (\tau/4)\log(1/\epsilon) \le P(G(\alpha,\tau) \le \epsilon) \le P(D_\alpha = 0)$$



$\tau = 0$ $\quad\tau = 1/2$ $\quad\tau = 1$ $\quad\tau = 2$

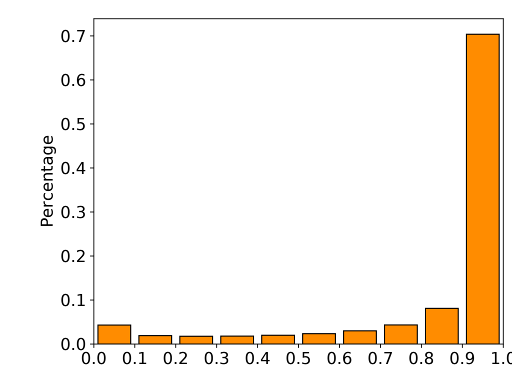
5. Gumbel-Gate LSTM (G²-LSTM)

- $h_t, c_t = \mathrm{LSTM}(h_{t-1}, c_{t-1}, x_t)$
- $f_t = G(W_{xf}x_t + W_{hf}h_{t-1} + b_f, \tau)$
- $i_t = G(W_{xi}x_t + W_{hi}h_{t-1} + b_i, \tau)$
- $g_t = \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g)$
- $o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$
- $c_t = f_t \odot c_{t-1} + i_t \odot g_t$
- $h_t = o_t \odot \tanh(c_t)$

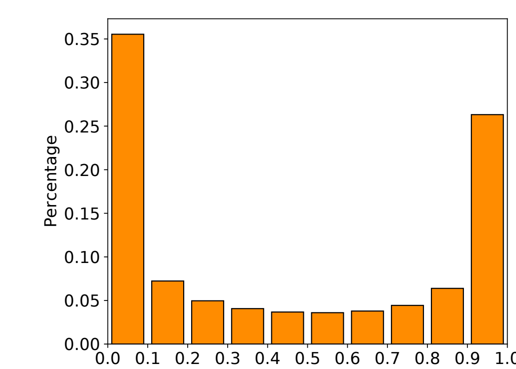
6. Experimental Results

Model	Result	Round	Round & Clip	SVD	SVD+
Penn Treebank (Perpelexity)					
Baseline	52.8	53.2 (+0.4)	53.6 (+0.8)	56.6 (+3.8)	65.5 (+12.7)
Sharpened Sigmoid	53.2	53.5 (+0.3)	53.6 (**+0.4**)	54.6 (+1.4)	60.0 (+6.8)
G²-LSTM	**52.1**	**52.2 (+0.1)**	**52.8** (+0.7)	**53.3 (+1.2)**	**56.0 (+3.9)**
IWSLT'14 German→English (BLEU)					
Baseline	31.00	28.65 (-2.35)	21.97 (-9.03)	30.52 (-0.48)	29.56 (-1.44)
Sharpened Sigmoid	29.73	27.08 (-2.65)	25.14 (-4.59)	29.17 (-0.53)	28.82 (-0.91)
G²-LSTM	**31.95**	**31.44 (-0.51)**	**31.44 (-0.51)**	**31.62 (-0.33)**	**31.28 (-0.67)**
WMT'14 English→German (BLEU)					
Baseline	21.89	16.22 (-5.67)	16.03 (-5.86)	21.15 (-0.74)	19.99 (-1.90)
Sharpened Sigmoid	21.64	16.85 (-4.79)	16.72 (-4.92)	20.98 (-0.66)	19.87 (-1.77)
G²-LSTM	**22.43**	**20.15 (-2.28)**	**20.29 (-2.14)**	**22.16 (-0.27)**	**21.84 (-0.51)**

7. Histograms of Gate Distributions in G²-LSTM

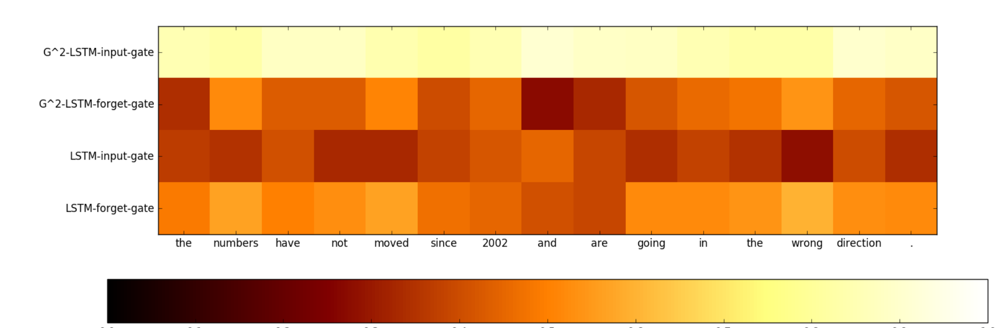


G²-LSTM Input gates



G²-LSTM Forget gates

8. Visualization of Average Gate Values



Zhuohan Li is **applying for a Ph.D.** in Fall 2018
Please contact if you are interested!
Email: lizhuohan@pku.edu.cn
https://zhuohan.li

Microsoft Research Asia
Contact: Tao Qin
Email: taoqin@Microsoft.com
http://research.microsoft.com/~taoqin

ICML | 2018