Batch Normalized Recurrent Highway Networks

Chi Zhang
cxz2081@rit.edu

Thang Nguyen
thn2079@mail.rit.edu

Alexander Loui†
alexander.loui@kodakalaris.com

Shagan Sah
sx4337@rit.edu

Carl Salvaggio
salvaggio@cis.rit.edu

Raymond Ptucha
rwpeec@rit.edu

Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology

† Imaging R&D
Kodak Alaris Inc.
Outline

Introduction

Related Work

Proposed Framework

Experiments

Conclusion
Recurrent Neural Networks (RNN)

Image Stage

Language Stage

CNN

Embed

two

ten

playing

hockey

Embed

START <000000>

two <010000>

men <100000>

playing <000100>

hockey <001000>

Figure:

- CNN feature extractor:
  - a man
  - walks

- Embed:
  - <START>
  - a
  - man

- FC7:
  - <END>
Gradient Flow in Recurrent Networks

$y^{[t]} = f(Wx^{[t]} + Ry^{[t-1]} + b)$

The derivative of the loss $L$ with respect to parameters $\theta$:

$$\frac{dL}{d\theta} = \sum_{1 \leq t_2 \leq T} \frac{dL^{[t_2]}}{d\theta} = \sum_{1 \leq t_2 \leq T} \sum_{1 \leq t_1 \leq t_2} \frac{\partial L^{[t_2]}}{\partial y^{[t_2]}} \frac{\partial y^{[t_2]}}{\partial y^{[t_1]}} \frac{\partial y^{[t_1]}}{\partial \theta}$$

where

$$\frac{\partial y^{[t_2]}}{\partial y^{[t_1]}} = \prod_{t_1 \leq t \leq t_2} \frac{\partial y^{[t]}}{\partial y^{[t-1]}} = \prod_{t_1 \leq t \leq t_2} RT \text{diag}[f'(Ry^{[t-1]})]$$

($Wx^{[t]}$ and $b$ are omitted.)
Let $A \overset{\text{def}}{=} \frac{\partial y[t]}{\partial y[t-1]}$ be the temporal Jacobian, $\gamma$ be a maximal bound on $f'(Ry[t-1])$ and $\sigma_{\text{max}}$ be the largest singular value of $R^T$, we have

$$\|A\| \leq \|\text{diag}[f'(Ry[t-1])]\| \|R^T\| \leq \gamma \sigma_{\text{max}}$$

- Vanishing gradients:
  $$\gamma \sigma_{\text{max}} < 1$$

- Exploding gradients:
  $$\rho > 1$$

where $\rho$ is the spectral radius (supremum in $|\lambda's|$) of $A$, since $\|A\| \geq \rho$. 
Geršgorin Circle Theorem (GCT)

For any square matrix $A \in \mathbb{R}^{n \times n}$

$$\text{spec}(A) \subset \bigcup_{i \in \{1, \ldots, n\}} \{\lambda \in \mathbb{C} \mid \|\lambda - a_{ii}\|_\mathbb{C} \leq \sum_{j=1, i\neq j}^{n} |a_{ij}|\}$$

Possible Solution?

Initialize $R$ with an identity matrix and small random values on the off-diagonals.

c = 1_n, t = 0_n  \implies \lambda_i = 1, \forall i \in \{1, \ldots, n\}

This can be done by coupling C and T: C = 1_n - T
Batch Normalized RHN

\[ h = H(x, W_H) \]
\[ t = T(x, W_T) \]
\[ c = C(x, W_C) \]
\[ s_d^t = h_d^t \odot t_d^t + s_{d-1}^t \odot c_d^t \]
Recall: Batch Normalization

\[
\begin{align*}
\text{Input: } & \text{ Values of } x \text{ over a mini-batch: } B = \{x_1...m\}; \\
\text{Parameters to be learned: } & \gamma, \beta \\
\text{Output: } & \{y_i = \text{BN}_{\gamma,\beta}(x_i)\}
\end{align*}
\]

\[
\begin{align*}
\mu_B & \leftarrow \frac{1}{m} \sum_{i=1}^{m} x_i & \text{// mini-batch mean} \\
\sigma^2_B & \leftarrow \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_B)^2 & \text{// mini-batch variance} \\
\hat{x}_i & \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma^2_B + \epsilon}} & \text{// normalize} \\
y_i & \leftarrow \gamma \hat{x}_i + \beta \equiv \text{BN}_{\gamma,\beta}(x_i) & \text{// scale and shift}
\end{align*}
\]

Image Captioning Results

Table: Evaluation metrics on MSCOCO dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>LSTM</th>
<th>RHN</th>
<th>BN_RHN</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU-1</td>
<td>0.706</td>
<td>0.618</td>
<td>0.710</td>
</tr>
<tr>
<td>BLEU-2</td>
<td>0.533</td>
<td>0.430</td>
<td>0.539</td>
</tr>
<tr>
<td>BLEU-3</td>
<td>0.397</td>
<td>0.291</td>
<td>0.404</td>
</tr>
<tr>
<td>BLEU-4</td>
<td>0.298</td>
<td>0.196</td>
<td>0.305</td>
</tr>
<tr>
<td>ROUGE-L</td>
<td>0.524</td>
<td>0.451</td>
<td>0.531</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.248</td>
<td>0.181</td>
<td>0.252</td>
</tr>
<tr>
<td>CIDEr</td>
<td>0.917</td>
<td>0.520</td>
<td>0.964</td>
</tr>
</tbody>
</table>

Figure: The total loss change vs. training steps.
Image Captioning Results

(LSTM) a group of people standing around a parking meter
(RHN) a group of people standing next to each other
(BNRHN) a young man riding a skateboard down a street
(G.T.) a person is doing a trick on a skateboard

(LSTM) a red stop sign sitting on top of a metal pole
(RHN) a red stop sign sitting on the side of a road
(BNRHN) a stop sign with a street sign attached to it
(G.T.) Street corner signs above a red stop sign
Image Captioning Results

(LSTM) a box with a donut and a cup of coffee
(RHN) a birthday cake with a picture of a dog on it
(BNRHN) a plate with a doughnut and a cup of coffee
(G.T.) A bag with a hot dog inside of it

(LSTM) a large brown dog sitting on top of a wooden bench
(RHN) a statue of a cow with a bird on top of it
(BNRHN) a statue of a cow standing on top of a wooden bench
(G.T.) A giant chair with a horse statue on it
Image Captioning Results

(LSTM) a bus driving down a street next to a tall building
(RHN) a group of people riding bikes down a street
(BNRHN) a city street filled with lots of traffic
(G.T.) A group of people walking down a sidewalk near a bus

(LSTM) a cat sitting on a chair in a kitchen
(RHN) a cat sitting on a chair in a room
(BNRHN) a black and white dog standing in a kitchen
(G.T.) A puppy is looking at a paper bag in the kitchen
Image Captioning Results – Negative

(LSTM) a rear view mirror of a car in the side view mirror
(RHN) a rear view mirror on the side of a car
(BNRHN) a rear view mirror with a dog in the side mirror
(G.T.) A guy takes a picture of his car’s rear view mirror

(LSTM) a person sitting on a bench in a park
(RHN) a wooden bench sitting on top of a lush green field
(BNRHN) a person sitting on a bench in a park
(G.T.) A woman standing next to a group of horses on a field
Conclusion

- We introduce a novel recurrent neural network model that is based on batch normalization and recurrent highway networks.
- The analyses provide insight into the ability of the batch normalized recurrent highway model to dynamically control the gradient flow across time steps.
- This model takes advantages of faster convergence compared to the original RHN.
- Experimental results on image captioning task reveals that our proposed model achieves high METEOR and BLEU scores compared to previous models on a modern dataset.
Please feel free to contact us if you have any question.

Chi Zhang
cxz2081@rit.edu

Thang Nguyen
thn2079@mail.rit.edu

Alexander Loui†
alexander.loui@kodakalaris.com

Shagan Sah
sxs4337@rit.edu

Carl Salvaggio
salvaggio@cis.rit.edu

Raymond Ptucha
rwpeec@rit.edu