Batch Normalized Recurrent Highway Networks

Chi Zhang
cxz2081@rit.edu

Shagan Sah
sxs4337@rit.edu

Thang Nguyen
thn2079@mail.rit.edu

Carl Salvaggio
salvaggio@cis.rit.edu

Alexander Loui†
alexander.loui@kodakalaris.com

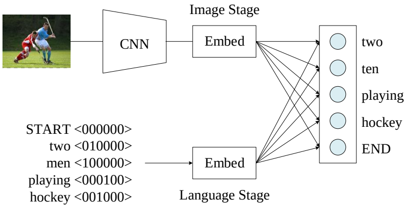
Raymond Ptucha
rwpeec@rit.edu

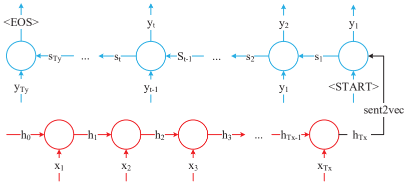
Chester F. Carlson Center for Imaging Science
Rochester Institute of Technology

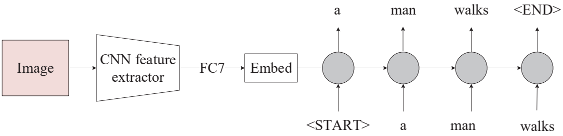
† Imaging R&D
Kodak Alaris Inc.

Outline

Recurrent Neural Networks (RNN)

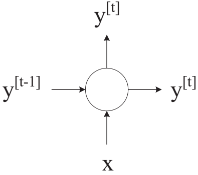






Gradient Flow in Recurrent Networks

$$\mathbf{y}^{[t]} = f(\mathbf{W}\mathbf{x}^{[t]} + \mathbf{R}\mathbf{y}^{[t-1]} + \mathbf{b})$$



The derivative of the loss L with respect to parameters θ:

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

where

$$\frac{\partial y^{[t_2]}}{\partial y^{[t_1]}} = \prod_{t_1 \le t \le t_2} \frac{\partial y^{[t]}}{\partial y^{[t-1]}} = \prod_{t_1 \le t \le t_2} R^T diag[f'(Ry^{[t-1]})]$$

($\mathbf{W}\mathbf{x}^{[t]}$ and $\mathbf{b}$ are omitted.)

Gradient Flow in Recurrent Networks

Let $A \stackrel{\text{def}}{=} \frac{\partial y^{[t]}}{\partial y^{[t-1]}}$ be the temporal Jacobian, γ be a maximal bound on $f'(Ry^{[t-1]})$ and σ_{max} be the largest singular value of R^T, we have

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

- Vanishing gradients:

$$\gamma\sigma_{max} < 1$$

- Exploding gradients:

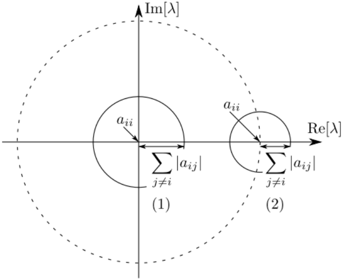
$$\rho > 1$$

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

Geršgorin Circle Theorem (GCT)

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

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



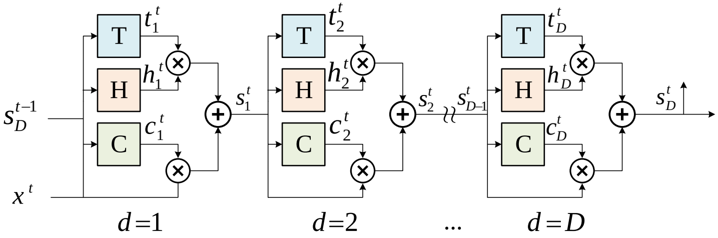
Possible Solution?

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

Zilly *et al.* "Recurrent highway networks." arXiv preprint arXiv:1607.03474 (2016).

Kodak alaris

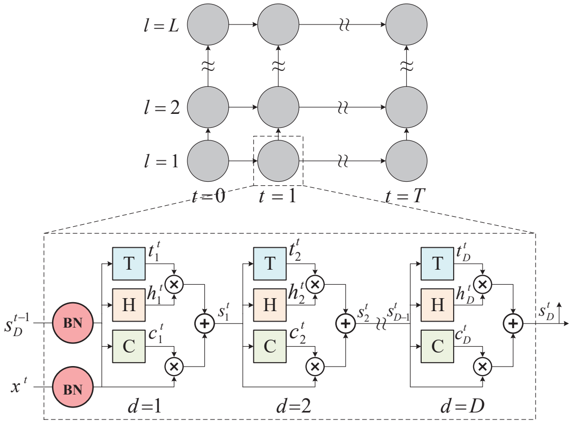
Recurrent Highway Networks



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

This can be done by coupling C and T: $C = 1_n - T$

Batch Normalized RHN



$$\mathbf{h} = H(\mathbf{x}, \mathbf{W}_H)$$
$$\mathbf{t} = T(\mathbf{x}, \mathbf{W}_T)$$
$$\mathbf{c} = C(\mathbf{x}, \mathbf{W}_C)$$
$$\mathbf{s}_d^t = \mathbf{h}_d^t \odot \mathbf{t}_d^t + \mathbf{s}_{d-1}^t \odot \mathbf{c}_d^t$$

Recall: Batch Normalization

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_\mathcal{B} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_\mathcal{B}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_\mathcal{B})^2 \qquad \text{// mini-batch variance}$$

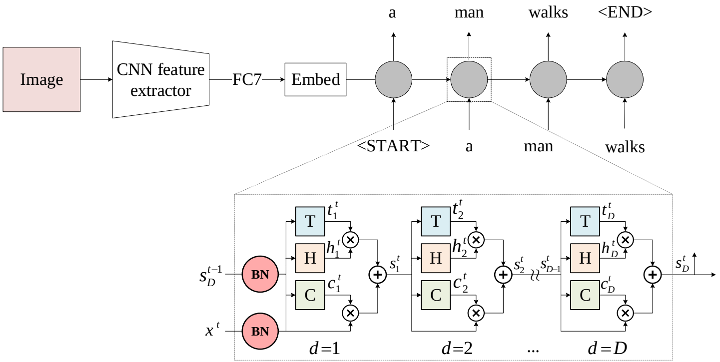
$$\widehat{x}_i \leftarrow \frac{x_i - \mu_\mathcal{B}}{\sqrt{\sigma_\mathcal{B}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Ioffe *et al.* ,"Batch normalization: Accelerating deep network training by reducing internal covariate shift." ICML, 2015.

Kodak alaris

Image Captioning



Kodak alaris

Image Captioning Results

Table: Evaluation metrics on MSCOCO dataset.

Model	LSTM	RHN	BN_RHN
BLEU-1	0.706	0.618	**0.710**
BLEU-2	0.533	0.430	**0.539**
BLEU-3	0.397	0.291	**0.404**
BLEU-4	0.298	0.196	**0.305**
ROUGE-L	0.524	0.451	**0.531**
METEOR	0.248	0.181	**0.252**
CIDEr	0.917	0.520	**0.964**

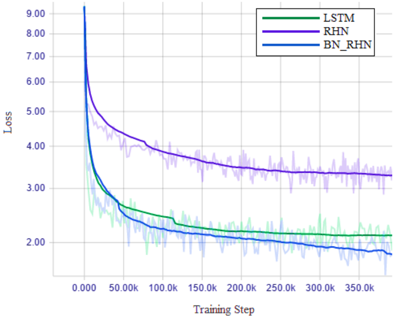


Figure: The total loss change vs. training steps.

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

Kodak alaris

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

Kodak alaris

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