Regularizing and Optimizing LSTM Language models

Stephen Merity, Nitish Shirish Keskar, Richard Socher. ICLR 2018

Soyoung Yoon, Jaeyoung Hwang, Dongmin Seo

Motivation

Regularization Techniques

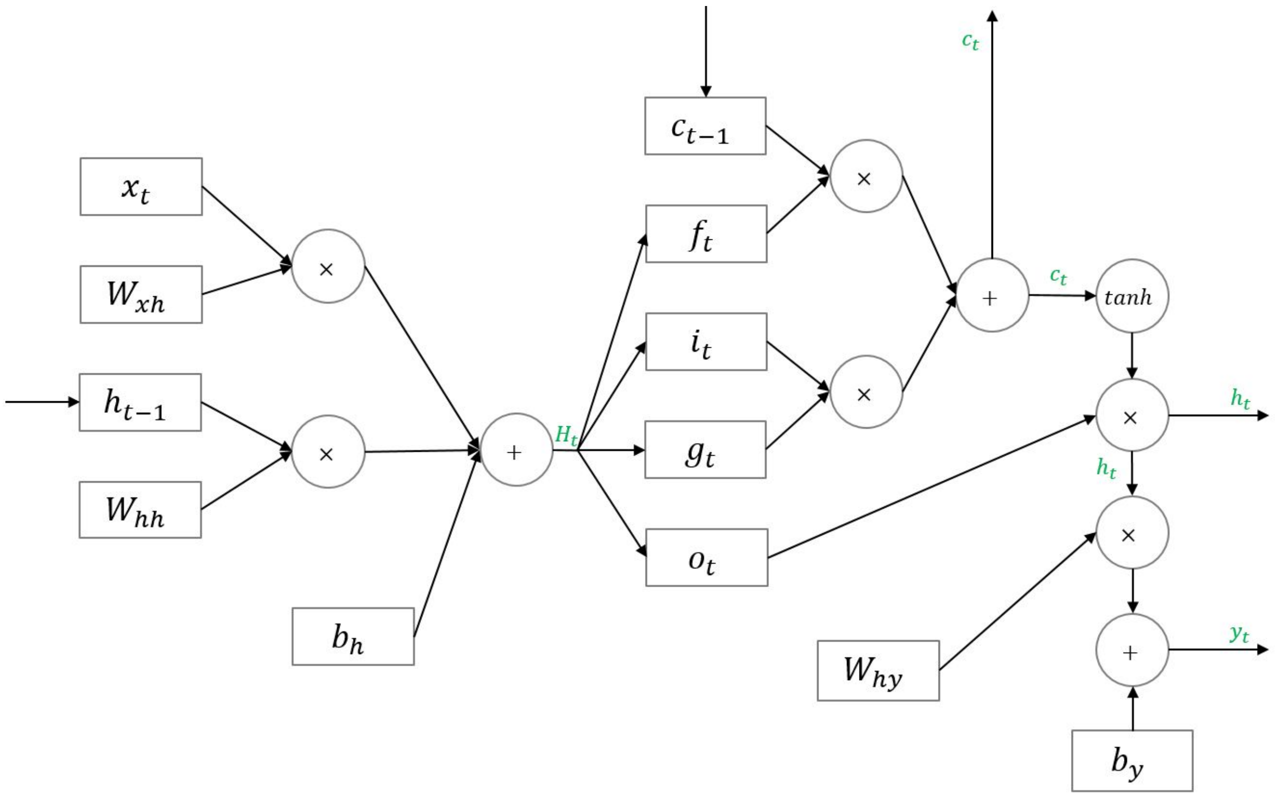
- Dropout
- Batch Normalization

-> *Only effective for feed-forward and conv net*

Previous studies increase training parameter

We want regularization strategies can be used with
no modification to existing LSTM implementations

Weight-dropped LSTM



Optimization : NT-ASGD

SGD Loss

$$\min_{w} \quad \frac{1}{N}\sum_{i=1}^{N} f_i(w)$$

SGD update

$$w_{k+1} = w_k - \gamma_k \hat{\nabla} f(w_k)$$

Last update in ASGD

$$\frac{\sum_{i=T}^{k} w_i}{(k-T+1)}$$

Regularization techniques

1. Variable length backpropagation sequences
   - randomly select sequence length

2. Variational dropout
   - First, binary dropout mask. Locked dropout for others.

3. Embedding dropout
   - drop out on embedding matrix

Regularization techniques

4. Weight tying

- Share weights between embedding & softmax

5. Independent embedding size and hidden size

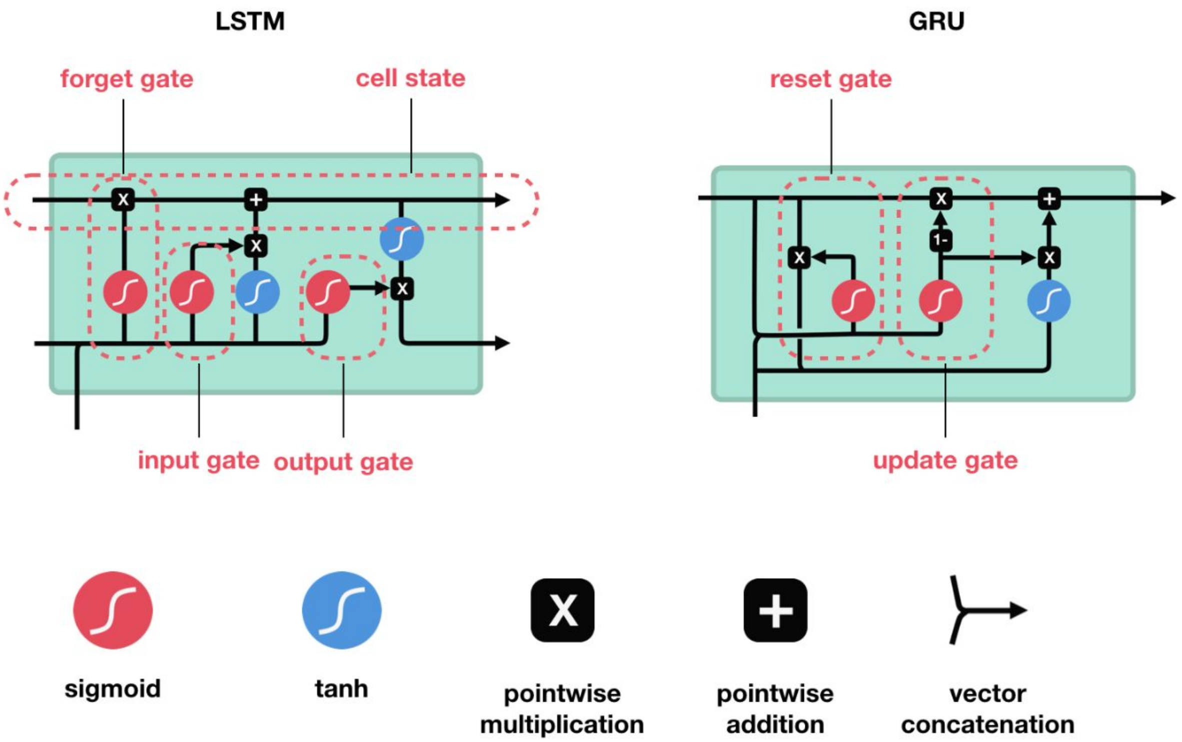
- LSTM input & output dimension = reduced embedding size

6. Activation Regularization (AR) and Temporal Activation Regularization (TAR)

Result

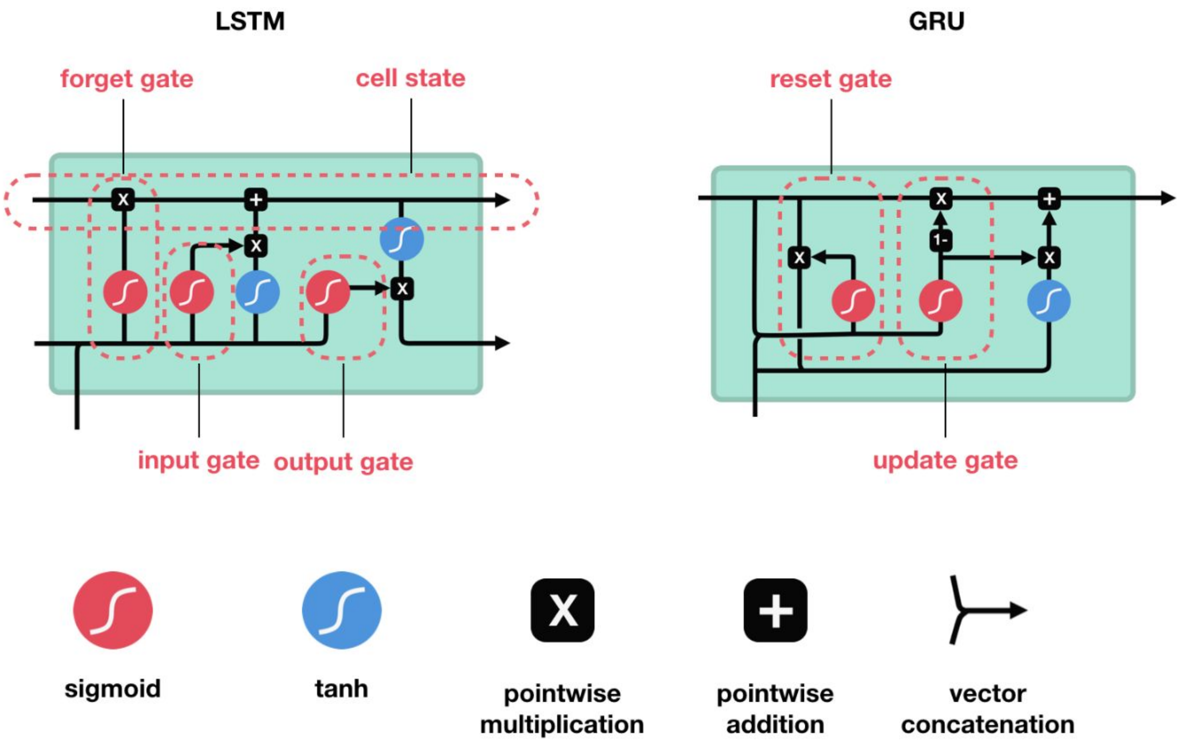
Model	PTB Validation	PTB Test	WT2 Validation	WT2 Test
AWD-LSTM (tied)	60.0	57.3	68.6	65.8
– fine-tuning	60.7	58.8	69.1	66.0
– NT-ASGD	66.3	63.7	73.3	69.7
– variable sequence lengths	61.3	58.9	69.3	66.2
– embedding dropout	65.1	62.7	71.1	68.1
– weight decay	63.7	61.0	71.9	68.7
– AR/TAR	62.7	60.3	73.2	70.1
– full sized embedding	68.0	65.6	73.7	70.7
– weight-dropping	71.1	68.9	78.4	74.9

What to Change : **LSTM -> GRU**



https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

What to Change : **LSTM -> GRU**



Difference of GRU

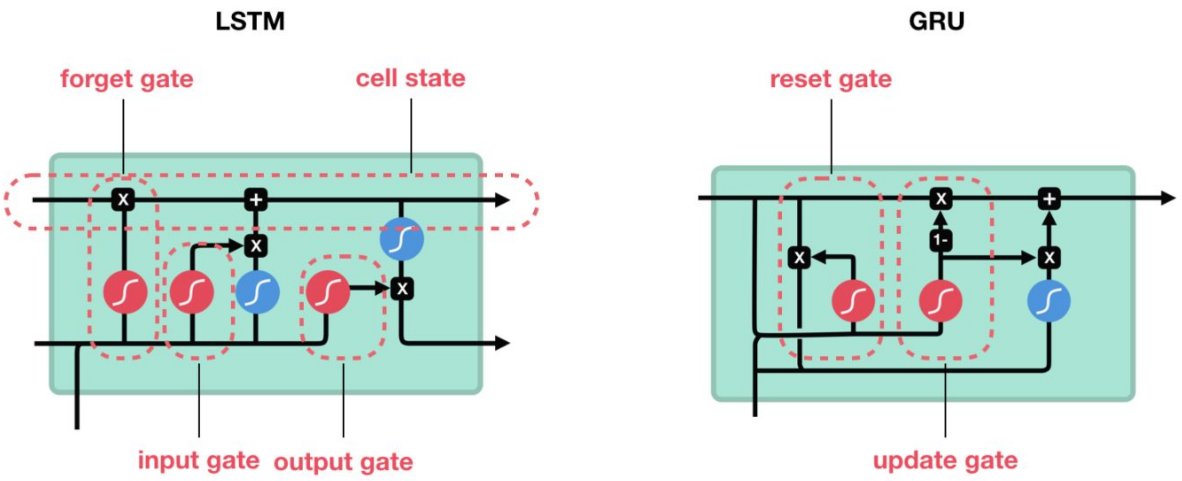
1. **3->2 gates**
   input, forget -> update

2. **don't use memory unit**
   use hidden state

-> simpler!

https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

What to Change : **LSTM -> GRU**



GRU have less parameters:

expect to achieve
faster training time and
compact model size

https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21