Layer Normalization

Jimmy Lei Ba, Jamie Ryan Kiros & Geoffrey E. Hinton
University of Toronto

Preprint 2016

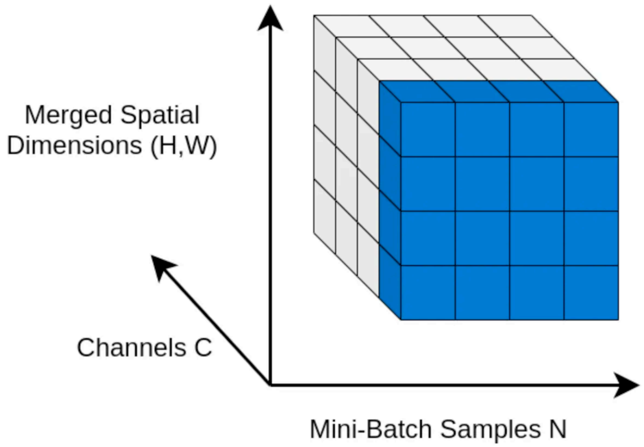
DMAIS@CAU
Yeongon Kim

INDEX

- **Introduction**
  - Batch Normalization
  - Limitations of Batch Normalization
- **Layer Normalization**
- **Experiments**
- **Conclusion**

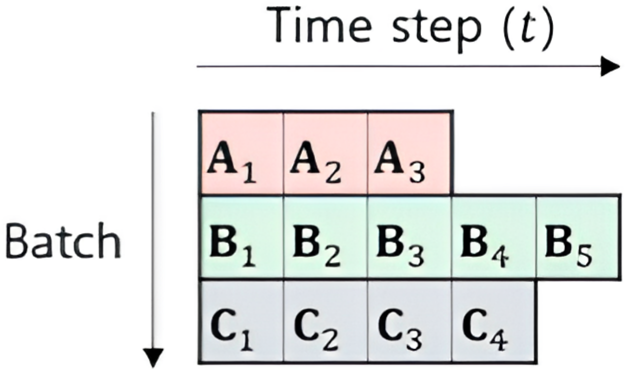
Batch Normalization

- Normalizes each layer’s inputs, feature-wise, using the mini-batch mean and variance
- Reduces internal covariate shift and accelerates training



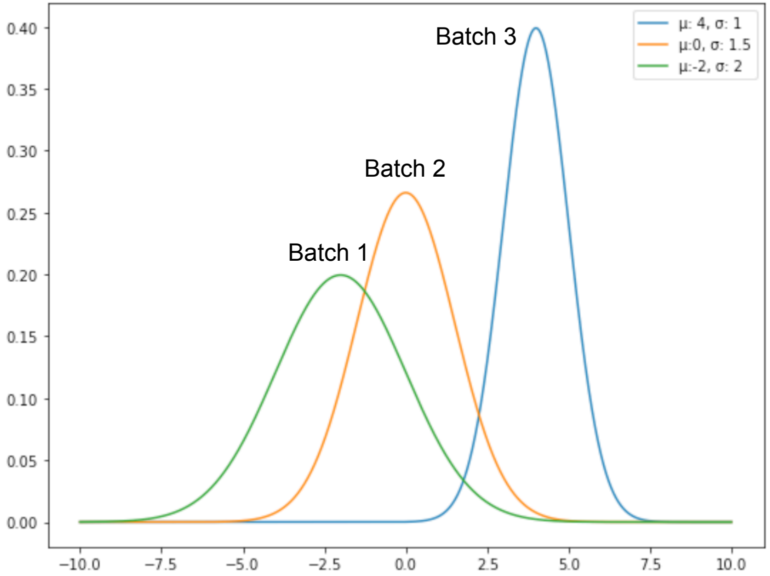
Limitations of Batch Normalization

- Difficult to apply to recurrent neural networks
  - Because sequence lengths vary, problem when a test sequence is longer than the training sequence



Limitations of Batch Normalization

- ❑ Unsuitable for small batch sizes
  - Since the statistics rely on only a few samples, the distribution varies from batch to batch



Limitations of Batch Normalization

- Unsuitable for small batch sizes
  - Unstable statistics from small batches inaccurately update the running statistics used for inference
  - Due to a mismatch between the dataset statistics and the running statistics, performance degrades

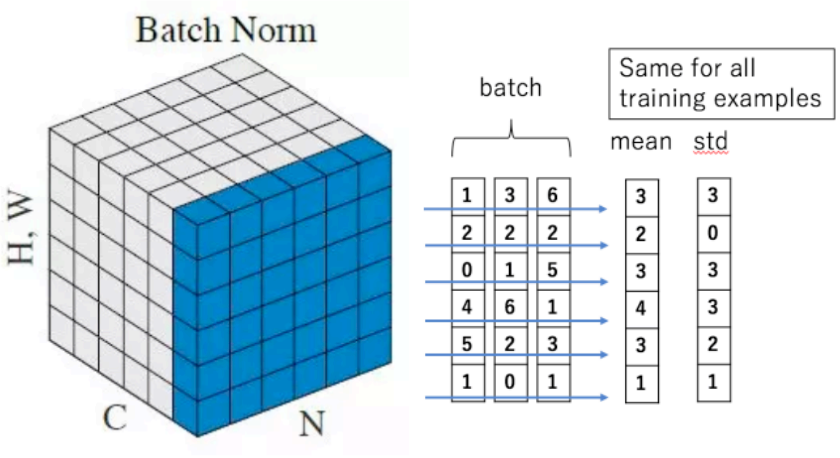
$$\overline{\mu}_t = m\,\overline{\mu}_{t-1} + (1-m)\,\mu_{B,t},$$
$$\overline{\sigma}_t^2 = m\,\overline{\sigma}_{t-1}^2 + (1-m)\,\sigma_{B,t}^2,$$

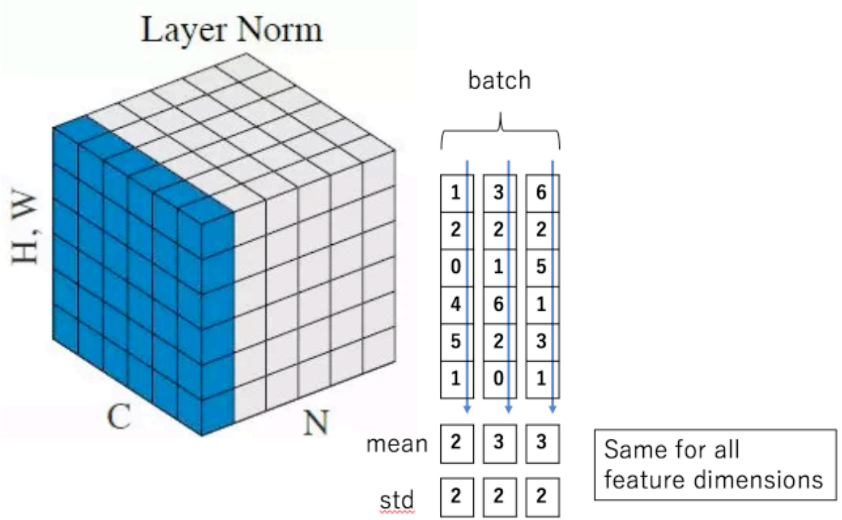
INDEX

- Introduction
  - Batch Normalization
  - Limitations of Batch Normalization
- **Layer Normalization**
- **Experiments**
- **Conclusion**

Layer Normalization: Idea

- Transpose the axes and apply normalization sample-wise





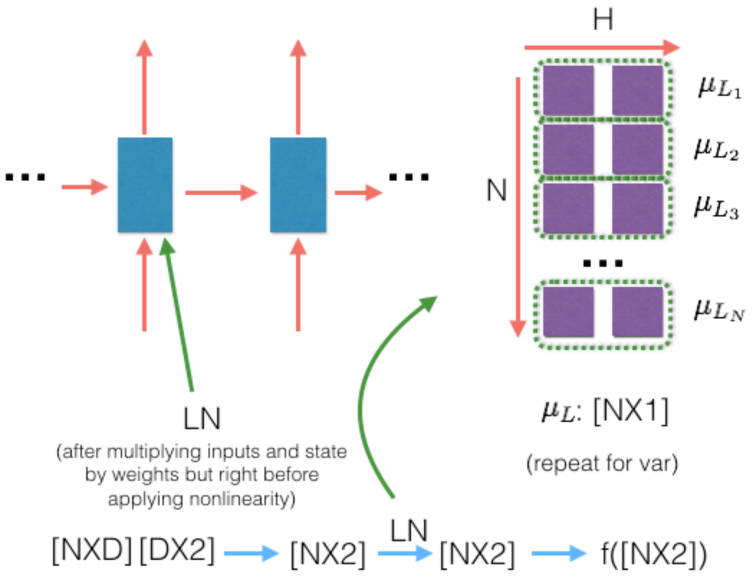
Layer Normalization: Methodology

- Normalization statistics is computed from inputs
- Each neuron has its own adaptive bias and gain

$$\mathbf{h}^t = f\left[\frac{\mathbf{g}}{\sigma^t} \odot \left(\mathbf{a}^t - \mu^t\right) + \mathbf{b}\right] \qquad \mu^t = \frac{1}{H}\sum_{i=1}^{H} a_i^t \qquad \sigma^t = \sqrt{\frac{1}{H}\sum_{i=1}^{H}\left(a_i^t - \mu^t\right)^2}$$

Benefits of Layer Normalization

- Performs exactly the same computation at training and test times
- Not dependent on the mini-batch size
- Applicable to RNNs



INDEX

- **Introduction**
  - Batch Normalization
  - Limitations of Batch Normalization
- **Layer Normalization**
- **Experiments**
- **Conclusion**

Experiments

- Teaching machines to read and comprehend

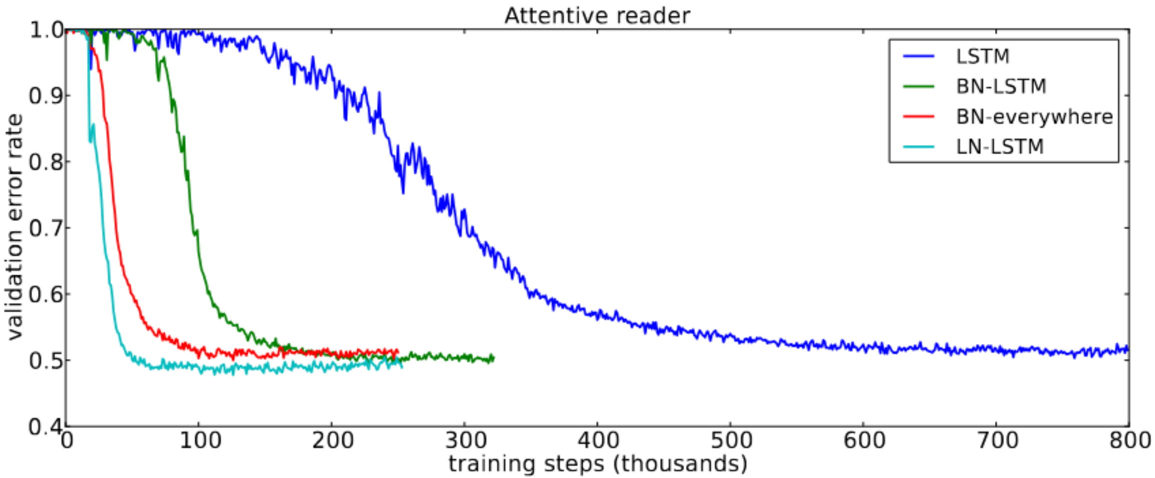


Figure 2: Validation curves for the attentive reader model. BN results are taken from [Cooijmans et al., 2016].

Experiments

❑ Handwriting sequence generation

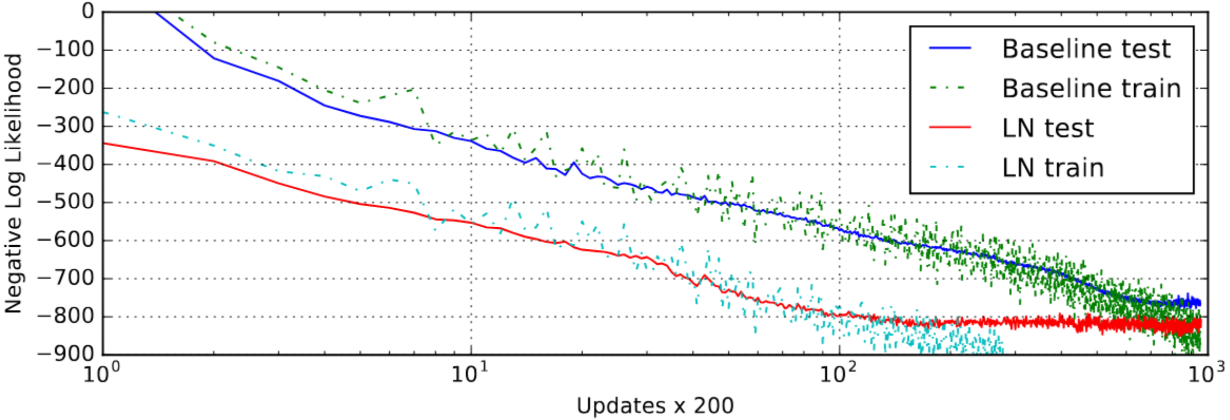


Figure 5: Handwriting sequence generation model negative log likelihood with and without layer normalization. The models are trained with mini-batch size of 8 and sequence length of 500,

Experiments

❑ MNIST Classification

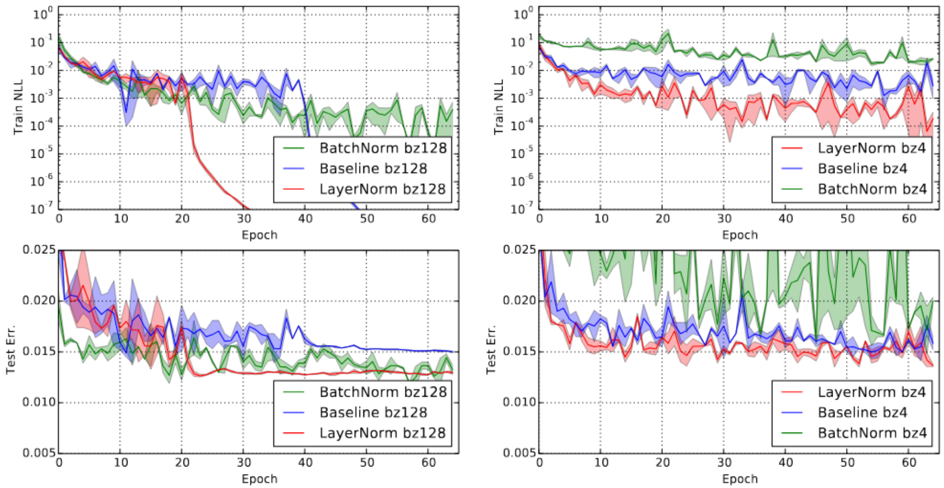


Figure 6: Permutation invariant MNIST 784-1000-1000-10 model negative log likelihood and test error with layer normalization and batch normalization. (Left) The models are trained with batch-size of 128. (Right) The models are trained with batch-size of 4.

Conclusion

- Layer norm reduces training time by normalizing each training example individually

- Not dependent on the mini-batch size and applicable to RNNs

Thank you!