Demystifying Dropout

Hongchang Gao[1,2], Jian Pei[3,4] and Heng Huang[1,2]

[1] JD Finance America Corporation
[2] Department of Electrical and Computer Engineering, University of Pittsburgh, USA
[3] JD.com
[4] School of Computing Science, Simon Fraser University, Canada

June 13, 2019

Outline

1. **Motivation**

2. **Forward and Backward Dropout**
   - Definition
   - Observations

3. **Augmented Dropout**
   - Augmented Dropout
   - Results

Motivation

1. Dropout is a popular technique to allievate overfitting.
2. What is the underlying reason for its performance gain?
   - Feature augmentation, Regularization
3. **Dropout happens in both the forward and backward pass**
   - Forward pass: feature augmentation
   - Backward pass: noisy gradient
   - Which pass accounts for performance gain of dropout?

1 Motivation

2 Forward and Backward Dropout
- Definition
- Observations

3 Augmented Dropout
- Augmented Dropout
- Results

Definition

Forward Dropout

forward dropout randomly drops features in the forward pass but it does not drop features and backpropagated errors in the backward pass.

- Forward pass

$$\begin{aligned} z^{(l+1)} &= W_l(h^{(l)} \odot \epsilon^{(l)}) + b^{(l)} \ , \\ h^{(l+1)} &= f_l(z^{(l+1)}) \ . \end{aligned} \tag{1}$$

- Backward pass

$$\begin{aligned} \frac{\partial J}{\partial W^{(l)}} &= \delta^{(l+1)} h^{(l)T} \ , \\ \delta^{(l)} \triangleq \frac{\partial J}{\partial z^{(l)}} &= (W^{(l)T} \delta^{(l+1)}) \odot f_l'(z^{(l)}) \ . \end{aligned} \tag{2}$$

Definition

Backward Dropout

Backward dropout keeps all features in the forward pass while drops features andback propagated errors as the standard dropout in the back-ward pass.

- Forward pass

$$\begin{aligned} z^{(l+1)} &= W_l h^{(l)} + b^{(l)} \ , \\ h^{(l+1)} &= f_l(z^{(l+1)}) \ . \end{aligned} \tag{3}$$

- Backward pass

$$\begin{aligned} \frac{\partial J}{\partial W^{(l)}} &= \delta^{(l+1)}(h^{(l)} \odot \epsilon^{(l)})^T \ , \\ \delta^{(l)} \triangleq \frac{\partial J}{\partial z^{(l)}} &= ((W^{(l)T}\delta^{(l+1)}) \odot f_l'(z^{(l)})) \odot \epsilon^{(l)} \ . \end{aligned} \tag{4}$$

1. Motivation

2. Forward and Backward Dropout
   - Definition
   - Observations

3. Augmented Dropout
   - Augmented Dropout
   - Results

Observations

Table: The test accuracy of ConvNet-Quick for CIFAR10

Dropout Ratio	0.3	0.2	0.1	0.05	0.01	0.005
Plain	0.7579					
Standard Dropout	0.7523	0.7617	0.7657	0.7647	0.7626	0.7608
Forward Dropout	0.6908	0.7211	0.7482	0.7627	0.7612	0.7578
Backward Dropout	0.7433	0.7557	0.7585	0.7593	0.7583	0.7599

Table: The test accuracy of ResNet-20 for CIFAR10

Dropout Ratio	0.3	0.2	0.1	0.05	0.01	0.005
Plain	0.9143					
Standard Dropout	0.9163	0.9176	0.9193	0.9174	0.9141	0.9154
Forward Dropout	0.9007	0.9093	0.9169	0.9171	0.9141	0.9142
Backward Dropout	0.9109	0.9130	0.9140	0.9142	0.9147	0.9146

Observations

1. When the dropout ratio is large ($p = 0.2$), the training loss of the forward dropout are much larger than those of other methods. In other words, when p is large, due to interrupting features heavily, the forward dropout can increase the model bias and cause underfitting. As a result, it degrades the performance of the plain network. As for the standard dropout, although it employs the same dropout ratio with the forward dropout, yet its loss and accuracy are better than those of the forward dropout. The possible reason is the implicit regularization of the noisy gradient in the backward pass, which is helpful to escape local minima. When it comes to the backward dropout, although it can arrive at a smaller training loss compared with the standard dropout, it cannot outperform the standard one. The possible reason is that there is no data augmentation in the forward pass as the standard dropout. Hence, its generalization performance is worse than the standard dropout.

Observations

2. When the dropout ratio becomes moderately small ($p = 0.05$), the training loss of the forward and backward dropout approach to that of the standard dropout, and their accuracy is a little better than that of the plain network. Thus, the data augmentation in the forward and the noisy gradient in the backward pass caused by the mild noise are helpful to improve the generalization performance. Additionally, the forward dropout outperforms the backward one. In other words, mild noise is more helpful in the forward pass.
3. When the dropout ratio is very small ($p = 0.005$), it is intuitive that the forward dropout has little effect on the performance of the plain deep neural networks. Surprisingly, the backward dropout has better performance than the plain network, which means that the noisy gradient caused by the small noise in the backward pass contributes to improving the generalization performance.

1. Motivation

2. Forward and Backward Dropout
   - Definition
   - Observations

3. Augmented Dropout
   - Augmented Dropout
   - Results

Augmented Dropout

Based on aforementioned observations, we propose the augmented dropout, which employs different dropout strategy for two passes.

- Forward pass

$$\begin{aligned} \hat{h}^{(l)}_{forward} &= h^{(l)} \odot \epsilon^{(l)}_{foward} \ , \\ z^{(l+1)} &= W_l \hat{h}^{(l)}_{forward} + b^{(l)} \ , \\ h^{(l+1)} &= f_l(z^{(l+1)}) \ , \end{aligned} \tag{5}$$

- Backward pass

$$\begin{aligned} \hat{h}^{(l)}_{backward} &= h^{(l)} \odot \epsilon^{(l)}_{backward} \ , \\ \frac{\partial J}{\partial W^{(l)}} &= \delta^{(l+1)} \hat{h}^{(l)T}_{backward} \ , \\ \delta^{(l)} \triangleq \frac{\partial J}{\partial z^{(l)}} &= (W^{(l)T} \delta^{(l+1)}) \odot (f'_l(z^{(l)}) \odot \epsilon^{(l)}_{backward}) \ , \end{aligned} \tag{6}$$

1 Motivation

2 Forward and Backward Dropout

- Definition
- Observations

3 Augmented Dropout

- Augmented Dropout
- Results

Results

Table: The test accuracy of ConvNet-Quick

Datasets	Standard		Augmented	
	Dropout Ratio	Acc	Acc	Dropout Ratio
SVHN	0.1	0.9240	0.9248	0.1/0.0002
	0.2	0.9231	0.9258	0.2/0.0002
	0.3	0.9249	0.9250	0.3/0.0002
CIFAR10	0.1	0.7657	0.7674	0.1/0.0002
	0.2	0.7617	0.7655	0.2/0.0002
	0.3	0.7523	0.7606	0.3/0.0002

Results

Table: The test accuracy of ResNet-20

Datasets	Standard		Augmented	
	Dropout Ratio	Acc	Acc	Dropout Ratio
SVHN	0.1	0.9618	0.9627	0.1/0.0002
	0.2	0.9626	0.9648	0.2/0.0002
	0.3	0.9655	0.9667	0.3/0.0002
CIFAR10	0.1	0.9193	0.9196	0.1/0.0001
	0.2	0.9176	0.9195	0.2/0.0001
	0.3	0.9163	0.9177	0.3/0.0001
CIFAR100	0.1	0.6762	0.6786	0.1/0.0001
	0.2	0.6748	0.6770	0.2/0.0001
	0.3	0.6686	0.6688	0.3/0.0001

Thank You