

# Dropout

February 13, 2025

## 1 Dropout

In this notebook, you will implement dropout. Then we will ask you to train a network with batchnorm and dropout, and achieve over 55% accuracy on CIFAR-10.

CS231n has built a solid API for building these modular frameworks and training them, and we will use their very well implemented framework as opposed to “reinventing the wheel.” This includes using their Solver, various utility functions, and their layer structure. This also includes `nndl.fc_net`, `nndl.layers`, and `nndl.layer_utils`. As in prior assignments, we thank Serena Yeung & Justin Johnson for permission to use code written for the CS 231n class ([cs231n.stanford.edu](http://cs231n.stanford.edu)).

```
[4]: ## Import and setups

import time
import numpy as np
import matplotlib.pyplot as plt
from nndl.fc_net import *
from nndl.layers import *
from utils.data_utils import get_CIFAR10_data
from utils.gradient_check import eval_numerical_gradient,   
    ↪eval_numerical_gradient_array
from utils.solver import Solver

%matplotlib inline
plt.rcParams['figure.figsize'] = (10.0, 8.0) # set default size of plots
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['image.cmap'] = 'gray'

# for auto-reloading external modules
# see http://stackoverflow.com/questions/1907993/
    ↪autoreload-of-modules-in-ipython
%load_ext autoreload
%autoreload 2

def rel_error(x, y):
    """ returns relative error """
    return np.max(np.abs(x - y) / (np.maximum(1e-8, np.abs(x) + np.abs(y))))
```

```
[6]: # Load the (preprocessed) CIFAR10 data.

data = get_CIFAR10_data()
for k in data.keys():
    print('{}: {}'.format(k, data[k].shape))
```

```
X_train: (49000, 3, 32, 32)
y_train: (49000,)
X_val: (1000, 3, 32, 32)
y_val: (1000,)
X_test: (1000, 3, 32, 32)
y_test: (1000,)
```

## 1.1 Dropout forward pass

Implement the training and test time dropout forward pass, `dropout_forward`, in `nndl/layers.py`. After that, test your implementation by running the following cell.

```
[8]: x = np.random.randn(500, 500) + 10

for p in [0.3, 0.6, 0.75]:
    out, _ = dropout_forward(x, {'mode': 'train', 'p': p})
    out_test, _ = dropout_forward(x, {'mode': 'test', 'p': p})

    print('Running tests with p = ', p)
    print('Mean of input: ', x.mean())
    print('Mean of train-time output: ', out.mean())
    print('Mean of test-time output: ', out_test.mean())
    print('Fraction of train-time output set to zero: ', (out == 0).mean())
    print('Fraction of test-time output set to zero: ', (out_test == 0).mean())
```

```
Running tests with p = 0.3
Mean of input: 9.999539269119458
Mean of train-time output: 9.964057520340884
Mean of test-time output: 9.999539269119458
Fraction of train-time output set to zero: 0.700972
Fraction of test-time output set to zero: 0.0
Running tests with p = 0.6
Mean of input: 9.999539269119458
Mean of train-time output: 9.975847670174648
Mean of test-time output: 9.999539269119458
Fraction of train-time output set to zero: 0.401384
Fraction of test-time output set to zero: 0.0
Running tests with p = 0.75
Mean of input: 9.999539269119458
Mean of train-time output: 10.009570251722403
Mean of test-time output: 9.999539269119458
Fraction of train-time output set to zero: 0.249264
Fraction of test-time output set to zero: 0.0
```

## 1.2 Dropout backward pass

Implement the backward pass, `dropout_backward`, in `nndl/layers.py`. After that, test your gradients by running the following cell:

```
[12]: x = np.random.randn(10, 10) + 10
dout = np.random.randn(*x.shape)

dropout_param = {'mode': 'train', 'p': 0.8, 'seed': 123}
out, cache = dropout_forward(x, dropout_param)
dx = dropout_backward(dout, cache)
dx_num = eval_numerical_gradient_array(lambda xx: dropout_forward(xx,
    dropout_param)[0], x, dout)

print('dx relative error: ', rel_error(dx, dx_num))
```

dx relative error: 5.4456136849927844e-11

## 1.3 Implement a fully connected neural network with dropout layers

Modify the `FullyConnectedNet()` class in `nndl/fc_net.py` to incorporate dropout. A dropout layer should be incorporated after every ReLU layer. Concretely, there shouldn't be a dropout at the output layer since there is no ReLU at the output layer. You will need to modify the class in the following areas:

- (1) In the forward pass, you will need to incorporate a dropout layer after every relu layer.
- (2) In the backward pass, you will need to incorporate a dropout backward pass layer.

Check your implementation by running the following code. Our  $W_1$  gradient relative error is on the order of  $1e-6$  (the largest of all the relative errors).

```
[15]: N, D, H1, H2, C = 2, 15, 20, 30, 10
X = np.random.randn(N, D)
y = np.random.randint(C, size=(N,))

for dropout in [0.5, 0.75, 1.0]:
    print('Running check with dropout = ', dropout)
    model = FullyConnectedNet([H1, H2], input_dim=D, num_classes=C,
                              weight_scale=5e-2, dtype=np.float64,
                              dropout=dropout, seed=123)

    loss, grads = model.loss(X, y)
    print('Initial loss: ', loss)

    for name in sorted(grads):
        f = lambda _: model.loss(X, y)[0]
        grad_num = eval_numerical_gradient(f, model.params[name], verbose=False,
            h=1e-5)
```

```

    print('{} relative error: {}'.format(name, rel_error(grad_num,
↪grads[name])))
    print('\n')

```

```

Running check with dropout = 0.5
Initial loss: 2.309771209610118
W1 relative error: 1.7941888539955072e-07
W2 relative error: 7.439246140739186e-08
W3 relative error: 1.9103712157213065e-08
b1 relative error: 4.112890360182144e-09
b2 relative error: 5.756217724722137e-10
b3 relative error: 1.3204470857080166e-10

```

```

Running check with dropout = 0.75
Initial loss: 2.306133548427975
W1 relative error: 8.729861444757663e-08
W2 relative error: 2.9777308030455514e-07
W3 relative error: 1.8832779990857742e-08
b1 relative error: 5.379484079091421e-08
b2 relative error: 3.6529948054257565e-09
b3 relative error: 9.987242764516995e-11

```

```

Running check with dropout = 1.0
Initial loss: 2.3053332250963194
W1 relative error: 1.274409704172877e-06
W2 relative error: 4.678743286048728e-07
W3 relative error: 6.915242087138281e-08
b1 relative error: 4.085353848305475e-08
b2 relative error: 1.9513425433873213e-09
b3 relative error: 9.387142701440351e-11

```

## 1.4 Dropout as a regularizer

In class, we claimed that dropout acts as a regularizer by effectively bagging. To check this, we will train two small networks, one with dropout and one without dropout.

```
[17]: # Train two identical nets, one with dropout and one without
```

```

num_train = 500
small_data = {
    'X_train': data['X_train'][:num_train],
    'y_train': data['y_train'][:num_train],
    'X_val': data['X_val'],
    'y_val': data['y_val'],

```

```

}

solvers = {}
dropout_choices = [0.6, 1.0]
for dropout in dropout_choices:
    model = FullyConnectedNet([100, 100, 100], dropout=dropout)

    solver = Solver(model, small_data,
                    num_epochs=25, batch_size=100,
                    update_rule='adam',
                    optim_config={
                        'learning_rate': 5e-4,
                    },
                    verbose=True, print_every=100)

    solver.train()
    solvers[dropout] = solver

```

```

(Iteration 1 / 125) loss: 2.300199
(Epoch 0 / 25) train acc: 0.158000; val_acc: 0.127000
(Epoch 1 / 25) train acc: 0.132000; val_acc: 0.121000
(Epoch 2 / 25) train acc: 0.204000; val_acc: 0.170000
(Epoch 3 / 25) train acc: 0.240000; val_acc: 0.192000
(Epoch 4 / 25) train acc: 0.312000; val_acc: 0.274000
(Epoch 5 / 25) train acc: 0.314000; val_acc: 0.269000
(Epoch 6 / 25) train acc: 0.364000; val_acc: 0.252000
(Epoch 7 / 25) train acc: 0.390000; val_acc: 0.281000
(Epoch 8 / 25) train acc: 0.386000; val_acc: 0.290000
(Epoch 9 / 25) train acc: 0.372000; val_acc: 0.267000
(Epoch 10 / 25) train acc: 0.424000; val_acc: 0.286000
(Epoch 11 / 25) train acc: 0.396000; val_acc: 0.275000
(Epoch 12 / 25) train acc: 0.458000; val_acc: 0.299000
(Epoch 13 / 25) train acc: 0.496000; val_acc: 0.305000
(Epoch 14 / 25) train acc: 0.492000; val_acc: 0.299000
(Epoch 15 / 25) train acc: 0.550000; val_acc: 0.296000
(Epoch 16 / 25) train acc: 0.584000; val_acc: 0.297000
(Epoch 17 / 25) train acc: 0.582000; val_acc: 0.309000
(Epoch 18 / 25) train acc: 0.612000; val_acc: 0.306000
(Epoch 19 / 25) train acc: 0.628000; val_acc: 0.323000
(Epoch 20 / 25) train acc: 0.608000; val_acc: 0.324000
(Iteration 101 / 125) loss: 1.369535
(Epoch 21 / 25) train acc: 0.644000; val_acc: 0.330000
(Epoch 22 / 25) train acc: 0.708000; val_acc: 0.341000
(Epoch 23 / 25) train acc: 0.690000; val_acc: 0.297000
(Epoch 24 / 25) train acc: 0.740000; val_acc: 0.300000
(Epoch 25 / 25) train acc: 0.750000; val_acc: 0.329000
(Iteration 1 / 125) loss: 2.300607
(Epoch 0 / 25) train acc: 0.172000; val_acc: 0.167000
(Epoch 1 / 25) train acc: 0.210000; val_acc: 0.197000

```

```

(Epoch 2 / 25) train acc: 0.284000; val_acc: 0.240000
(Epoch 3 / 25) train acc: 0.302000; val_acc: 0.246000
(Epoch 4 / 25) train acc: 0.392000; val_acc: 0.289000
(Epoch 5 / 25) train acc: 0.420000; val_acc: 0.274000
(Epoch 6 / 25) train acc: 0.420000; val_acc: 0.304000
(Epoch 7 / 25) train acc: 0.474000; val_acc: 0.293000
(Epoch 8 / 25) train acc: 0.516000; val_acc: 0.330000
(Epoch 9 / 25) train acc: 0.566000; val_acc: 0.322000
(Epoch 10 / 25) train acc: 0.620000; val_acc: 0.321000
(Epoch 11 / 25) train acc: 0.656000; val_acc: 0.317000
(Epoch 12 / 25) train acc: 0.676000; val_acc: 0.319000
(Epoch 13 / 25) train acc: 0.680000; val_acc: 0.304000
(Epoch 14 / 25) train acc: 0.752000; val_acc: 0.323000
(Epoch 15 / 25) train acc: 0.802000; val_acc: 0.321000
(Epoch 16 / 25) train acc: 0.804000; val_acc: 0.300000
(Epoch 17 / 25) train acc: 0.868000; val_acc: 0.303000
(Epoch 18 / 25) train acc: 0.894000; val_acc: 0.298000
(Epoch 19 / 25) train acc: 0.910000; val_acc: 0.282000
(Epoch 20 / 25) train acc: 0.926000; val_acc: 0.316000
(Iteration 101 / 125) loss: 0.245816
(Epoch 21 / 25) train acc: 0.950000; val_acc: 0.282000
(Epoch 22 / 25) train acc: 0.958000; val_acc: 0.292000
(Epoch 23 / 25) train acc: 0.966000; val_acc: 0.307000
(Epoch 24 / 25) train acc: 0.966000; val_acc: 0.285000
(Epoch 25 / 25) train acc: 0.970000; val_acc: 0.284000

```

```

[19]: # Plot train and validation accuracies of the two models

train_accs = []
val_accs = []
for dropout in dropout_choices:
    solver = solvers[dropout]
    train_accs.append(solver.train_acc_history[-1])
    val_accs.append(solver.val_acc_history[-1])

plt.subplot(3, 1, 1)
for dropout in dropout_choices:
    plt.plot(solvers[dropout].train_acc_history, 'o', label='%.2f dropout' % dropout)
plt.title('Train accuracy')
plt.xlabel('Epoch')
plt.ylabel('Accuracy')
plt.legend(ncol=2, loc='lower right')

plt.subplot(3, 1, 2)
for dropout in dropout_choices:

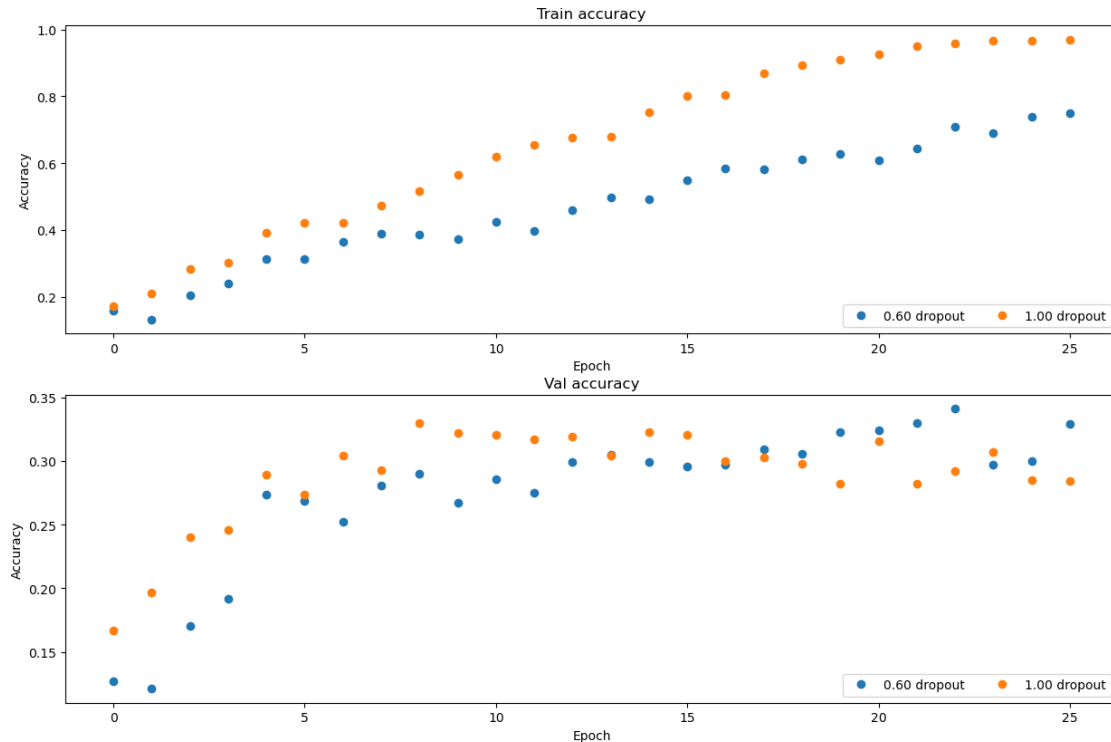
```

```

plt.plot(solvers[dropout].val_acc_history, 'o', label='%.2f dropout' % dropout)
plt.title('Val accuracy')
plt.xlabel('Epoch')
plt.ylabel('Accuracy')
plt.legend(ncol=2, loc='lower right')

plt.gcf().set_size_inches(15, 15)
plt.show()

```



## 1.5 Question

Based off the results of this experiment, is dropout performing regularization? Explain your answer.

## 1.6 Answer:

Yes, dropout is performing regularization. Based on the results of this experiment, the dropout is not performing effective regularization when it has a dropout of 1.00 but in comparison 0.60 does better. As we can see in the graphs above, the 0.60 dropout higher training accuracy and 1.00 dropout has a much higher training accuracy. A lower training accuracy is consistent with the expected behavior of dropout in that randomly deactivating neurons will prevent the model from overfitting to the training data. With the 1.00, we see that the training accuracy increases and gets really close to 1, likely indicating overfitting. However, when we look at the validation accuracies, the model with the 0.60 dropout has a very similar performance to the 1.00 dropout

model. So, this might mean that a high dropout rate might be good on the training data but may be suppressing the model's ability to learn meaningful patterns due to overfitting. Basically, the dropout is acting as a regularizer, but the dropout rate of 1.00 might be too aggressive so we do not see an improved validation accuracy or generalization.

## 1.7 Final part of the assignment

Get over 55% validation accuracy on CIFAR-10 by using the layers you have implemented. You will be graded according to the following equation:

$\min(\text{floor}((X - 32\%) / 23\%, 1)$  where if you get 55% or higher validation accuracy, you get full points.

```
[28]: # ===== #
# YOUR CODE HERE:
# Implement a FC-net that achieves at least 55% validation accuracy
# on CIFAR-10.
# ===== #
model = FullyConnectedNet([600, 400, 200, 50], dropout=0.75, weight_scale=4e-2,
    ↪use_batchnorm=True)
solver = Solver(model, data,
                num_epochs=20, batch_size=500,
                update_rule='adam',
                optim_config={'learning_rate': 3e-3},
                lr_decay=0.95,
                verbose=True, print_every=100)
solver.train()
# ===== #
# END YOUR CODE HERE
# ===== #
```

```
(Iteration 1 / 1960) loss: 2.316019
(Epoch 0 / 20) train acc: 0.202000; val_acc: 0.204000
(Epoch 1 / 20) train acc: 0.445000; val_acc: 0.451000
(Iteration 101 / 1960) loss: 1.553881
(Epoch 2 / 20) train acc: 0.525000; val_acc: 0.492000
(Iteration 201 / 1960) loss: 1.462293
(Epoch 3 / 20) train acc: 0.546000; val_acc: 0.530000
(Iteration 301 / 1960) loss: 1.395289
(Epoch 4 / 20) train acc: 0.601000; val_acc: 0.536000
(Iteration 401 / 1960) loss: 1.324262
(Epoch 5 / 20) train acc: 0.593000; val_acc: 0.550000
(Iteration 501 / 1960) loss: 1.377314
(Epoch 6 / 20) train acc: 0.606000; val_acc: 0.565000
(Iteration 601 / 1960) loss: 1.239717
(Epoch 7 / 20) train acc: 0.630000; val_acc: 0.558000
(Iteration 701 / 1960) loss: 1.167200
(Epoch 8 / 20) train acc: 0.658000; val_acc: 0.585000
(Iteration 801 / 1960) loss: 1.114987
```



(Epoch 9 / 20) train acc: 0.669000; val\_acc: 0.583000  
(Iteration 901 / 1960) loss: 1.073143  
(Epoch 10 / 20) train acc: 0.670000; val\_acc: 0.591000  
(Iteration 1001 / 1960) loss: 1.015183  
(Epoch 11 / 20) train acc: 0.700000; val\_acc: 0.583000  
(Iteration 1101 / 1960) loss: 1.046773  
(Epoch 12 / 20) train acc: 0.691000; val\_acc: 0.584000  
(Iteration 1201 / 1960) loss: 1.055535  
(Epoch 13 / 20) train acc: 0.728000; val\_acc: 0.588000  
(Iteration 1301 / 1960) loss: 0.958463  
(Epoch 14 / 20) train acc: 0.732000; val\_acc: 0.588000  
(Iteration 1401 / 1960) loss: 0.835853  
(Epoch 15 / 20) train acc: 0.732000; val\_acc: 0.585000  
(Iteration 1501 / 1960) loss: 0.902360  
(Epoch 16 / 20) train acc: 0.761000; val\_acc: 0.575000  
(Iteration 1601 / 1960) loss: 0.886267  
(Epoch 17 / 20) train acc: 0.761000; val\_acc: 0.578000  
(Iteration 1701 / 1960) loss: 0.976805  
(Epoch 18 / 20) train acc: 0.763000; val\_acc: 0.597000  
(Iteration 1801 / 1960) loss: 0.772563  
(Epoch 19 / 20) train acc: 0.781000; val\_acc: 0.591000  
(Iteration 1901 / 1960) loss: 0.815448  
(Epoch 20 / 20) train acc: 0.820000; val\_acc: 0.603000

[ ]:

[ ]: