```python
import numpy as np
from .layers import *
from .layer_utils import *


class TwoLayerNet(object):
  """
  A two-layer fully-connected neural network with ReLU nonlinearity and
  softmax loss that uses a modular layer design. We assume an input dimension
  of D, a hidden dimension of H, and perform classification over C classes.

  The architecure should be affine - relu - affine - softmax.

  Note that this class does not implement gradient descent; instead, it
  will interact with a separate Solver object that is responsible for running
  optimization.

  The learnable parameters of the model are stored in the dictionary
  self.params that maps parameter names to numpy arrays.
  """

  def __init__(self, input_dim=3*32*32, hidden_dims=100, num_classes=10,
               dropout=0, weight_scale=1e-3, reg=0.0):
    """
    Initialize a new network.

    Inputs:
    - input_dim: An integer giving the size of the input
    - hidden_dims: An integer giving the size of the hidden layer
    - num_classes: An integer giving the number of classes to classify
    - dropout: Scalar between 0 and 1 giving dropout strength.
    - weight_scale: Scalar giving the standard deviation for random
      initialization of the weights.
    - reg: Scalar giving L2 regularization strength.
    """
    self.params = {}
    self.reg = reg

    # ================================================================ #
    # YOUR CODE HERE:
    #   Initialize W1, W2, b1, and b2.  Store these as self.params['W1'],
    #   self.params['W2'], self.params['b1'] and self.params['b2']. The
    #   biases are initialized to zero and the weights are initialized
    #   so that each parameter has mean 0 and standard deviation weight_scale.
    #   The dimensions of W1 should be (input_dim, hidden_dim) and the
    #   dimensions of W2 should be (hidden_dims, num_classes)
    # ================================================================ #

    self.params['W1'] = weight_scale * np.random.randn(input_dim, hidden_dims) + 0
    self.params['W2'] = weight_scale * np.random.randn(hidden_dims, num_classes) + 0

    self.params['b1'] = np.zeros((hidden_dims, 1))
    self.params['b2'] = np.zeros((num_classes, 1))

    # ================================================================ #
    # END YOUR CODE HERE
```

```python
    # ================================================================ #

  def loss(self, X, y=None):
    """
    Compute loss and gradient for a minibatch of data.

    Inputs:
    - X: Array of input data of shape (N, d_1, ..., d_k)
    - y: Array of labels, of shape (N,). y[i] gives the label for X[i].

    Returns:
    If y is None, then run a test-time forward pass of the model and return:
    - scores: Array of shape (N, C) giving classification scores, where
      scores[i, c] is the classification score for X[i] and class c.

    If y is not None, then run a training-time forward and backward pass and
    return a tuple of:
    - loss: Scalar value giving the loss
    - grads: Dictionary with the same keys as self.params, mapping parameter
      names to gradients of the loss with respect to those parameters.
    """
    scores = None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the forward pass of the two-layer neural network. Store
    #   the class scores as the variable 'scores'.  Be sure to use the layers
    #   you prior implemented.
    # ================================================================ #

    out_l1, cache_l1 = affine_forward(X, self.params['W1'], self.params['b1'])
    out_relu, cache_relu = relu_forward(out_l1)
    scores, cache_l2 = affine_forward(out_relu, self.params['W2'], self.params['b2'])

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    # If y is None then we are in test mode so just return scores
    if y is None:
      return scores

    loss, grads = 0, {}
    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the backward pass of the two-layer neural net.  Store
    #   the loss as the variable 'loss' and store the gradients in the
    #   'grads' dictionary.  For the grads dictionary, grads['W1'] holds
    #   the gradient for W1, grads['b1'] holds the gradient for b1, etc.
    #   i.e., grads[k] holds the gradient for self.params[k].
    #
    #   Add L2 regularization, where there is an added cost 0.5*self.reg*W^2
    #   for each W.  Be sure to include the 0.5 multiplying factor to
    #   match our implementation.
    #
    #   And be sure to use the layers you prior implemented.
    # ================================================================ #

```

```python
    loss, dx2 = softmax_loss(scores, y)
    reg_loss = 0.5 * self.reg * (np.linalg.norm(self.params['W1'], 'fro')**2 + np.linalg.norm(self.params['W2'], 'fro')**2)
    loss += reg_loss

    dh1, dW2, db2 = affine_backward(dx2, cache_l2)
    da = relu_backward(dh1, cache_relu)
    dx1, dW1, db1 = affine_backward(da, cache_l1)

    grads['W1'] = dW1 + self.reg * self.params['W1']
    grads['b1'] = db1.T

    grads['W2'] = dW2 + self.reg * self.params['W2']
    grads['b2'] = db2.T

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    return loss, grads


class FullyConnectedNet(object):
  """
  A fully-connected neural network with an arbitrary number of hidden layers,
  ReLU nonlinearities, and a softmax loss function. This will also implement
  dropout and batch normalization as options. For a network with L layers,
  the architecture will be

  {affine - [batch norm] - relu - [dropout]} x (L - 1) - affine - softmax

  where batch normalization and dropout are optional, and the {...} block is
  repeated L - 1 times.

  Similar to the TwoLayerNet above, learnable parameters are stored in the
  self.params dictionary and will be learned using the Solver class.
  """

  def __init__(self, hidden_dims, input_dim=3*32*32, num_classes=10,
               dropout=0, use_batchnorm=False, reg=0.0,
               weight_scale=1e-2, dtype=np.float32, seed=None):
    """
    Initialize a new FullyConnectedNet.

    Inputs:
    - hidden_dims: A list of integers giving the size of each hidden layer.
    - input_dim: An integer giving the size of the input.
    - num_classes: An integer giving the number of classes to classify.
    - dropout: Scalar between 0 and 1 giving dropout strength. If dropout=0 then
      the network should not use dropout at all.
    - use_batchnorm: Whether or not the network should use batch normalization.
    - reg: Scalar giving L2 regularization strength.
    - weight_scale: Scalar giving the standard deviation for random
      initialization of the weights.
    - dtype: A numpy datatype object; all computations will be performed using
      this datatype. float32 is faster but less accurate, so you should use
      float64 for numeric gradient checking.
```

```python
    - seed: If not None, then pass this random seed to the dropout layers. This
      will make the dropout layers deteriminstic so we can gradient check the
      model.
    """
    self.use_batchnorm = use_batchnorm
    self.use_dropout = dropout > 0
    self.reg = reg
    self.num_layers = 1 + len(hidden_dims)
    self.dtype = dtype
    self.params = {}

    # ================================================================ #
    # YOUR CODE HERE:
    #   Initialize all parameters of the network in the self.params dictionary.
    #   The weights and biases of layer 1 are W1 and b1; and in general the
    #   weights and biases of layer i are Wi and bi. The
    #   biases are initialized to zero and the weights are initialized
    #   so that each parameter has mean 0 and standard deviation weight_scale.
    #
    #   BATCHNORM: Initialize the gammas of each layer to 1 and the beta
    #   parameters to zero.  The gamma and beta parameters for layer 1 should
    #   be self.params['gamma1'] and self.params['beta1'].  For layer 2, they
    #   should be gamma2 and beta2, etc. Only use batchnorm if self.use_batchnorm
    #   is true and DO NOT do batch normalize the output scores.
    # ================================================================ #

    dims = []
    dims.append(input_dim)
    dims.extend(hidden_dims)
    dims.append(num_classes)

    for i in np.arange(self.num_layers):
      num = str(i+1)
      self.params['W'+num] = weight_scale * np.random.randn(dims[i], dims[i+1]) + 0
      self.params['b'+num] = np.zeros((dims[i+1]))

      if i == (self.num_layers - 1):
        break

      if self.use_batchnorm:
        self.params['gamma'+num] = np.ones((dims[i+1]))
        self.params['beta'+num] = np.zeros((dims[i+1]))

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    # When using dropout we need to pass a dropout_param dictionary to each
    # dropout layer so that the layer knows the dropout probability and the mode
    # (train / test). You can pass the same dropout_param to each dropout layer.
    self.dropout_param = {}
    if self.use_dropout:
      self.dropout_param = {'mode': 'train', 'p': dropout}
```

```
      if seed is not None:
        self.dropout_param['seed'] = seed

    # With batch normalization we need to keep track of running means and
    # variances, so we need to pass a special bn_param object to each batch
    # normalization layer. You should pass self.bn_params[0] to the forward pass
    # of the first batch normalization layer, self.bn_params[1] to the forward
    # pass of the second batch normalization layer, etc.
    self.bn_params = []
    if self.use_batchnorm:
      self.bn_params = [{'mode': 'train'} for i in np.arange(self.num_layers - 1)]

    # Cast all parameters to the correct datatype
    for k, v in self.params.items():
      self.params[k] = v.astype(dtype)


  def loss(self, X, y=None):
    """
    Compute loss and gradient for the fully-connected net.

    Input / output: Same as TwoLayerNet above.
    """
    X = X.astype(self.dtype)
    mode = 'test' if y is None else 'train'

    # Set train/test mode for batchnorm params and dropout param since they
    # behave differently during training and testing.
    if self.dropout_param is not None:
      self.dropout_param['mode'] = mode
    if self.use_batchnorm:
      for bn_param in self.bn_params:
        bn_param['mode'] = mode

    scores = None

    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the forward pass of the FC net and store the output
    #   scores as the variable "scores".
    #
    #   BATCHNORM: If self.use_batchnorm is true, insert a bathnorm layer
    #   between the affine_forward and relu_forward layers.  You may
    #   also write an affine_batchnorm_relu() function in layer_utils.py.
    #
    #   DROPOUT: If dropout is non-zero, insert a dropout layer after
    #   every ReLU layer.
    # ================================================================ #

    fc_outs = {}
    fc_caches = {}

    batchnorm_outs = {}
    batchnorm_caches = {}

    h = {}
    h[0] = X
```

```
    relu_caches = {}

    dropout_caches = {}

    for i in np.arange(self.num_layers):
      num = str(i+1)
      # fc
      fc_outs[i+1], fc_caches[i+1] = affine_forward(h[i], self.params['W'+num], self.params['b'+num])
      if i == (self.num_layers - 1):
        break
      relu_input = fc_outs[i+1]

      # batch-norm
      if self.use_batchnorm:
        batchnorm_outs[i+1], batchnorm_caches[i+1] = batchnorm_forward(fc_outs[i+1], self.params['gamma'+num], self.params['beta'+num], self.bn_params[i])
        relu_input = batchnorm_outs[i+1]

      # relu
      h[i+1], relu_caches[i+1] = relu_forward(relu_input)

      # dropout
      if self.use_dropout:
        h[i+1], dropout_caches[i+1] = dropout_forward(h[i+1], self.dropout_param)

    scores = fc_outs[self.num_layers]

    # ================================================================ #
    # END YOUR CODE HERE
    # ================================================================ #

    # If test mode return early
    if mode == 'test':
      return scores

    loss, grads = 0.0, {}
    # ================================================================ #
    # YOUR CODE HERE:
    #   Implement the backwards pass of the FC net and store the gradients
    #   in the grads dict, so that grads[k] is the gradient of self.params[k]
    #   Be sure your L2 regularization includes a 0.5 factor.
    #
    #   BATCHNORM: Incorporate the backward pass of the batchnorm.
    #
    #   DROPOUT: Incorporate the backward pass of dropout.
    # ================================================================ #

    loss, dx = softmax_loss(scores, y)
    reg_loss_sum = 0
    for i in np.arange(self.num_layers):
      num = str(i+1)
      reg_loss_sum += np.linalg.norm(self.params['W'+num], 'fro')**2
    loss += 0.5 * self.reg * reg_loss_sum

    dict_dW = {}
    dict_db = {}

```

```
    dict_dgamma = {}
    dict_dbeta = {}

    # fc - last layer
    dh, dW, db = affine_backward(dx, fc_caches[self.num_layers])
    dict_dW[self.num_layers] = dW
    dict_db[self.num_layers] = db

    for i in np.arange(self.num_layers - 1, 0, -1):
      #dropout
      if self.use_dropout:
        dh = dropout_backward(dh, dropout_caches[i])

      # relu
      dx_relu = relu_backward(dh, relu_caches[i])
      fc_input = dx_relu

      # batch-norm
      if self.use_batchnorm:
        dx_batchnorm, dgamma, dbeta = batchnorm_backward(dx_relu, batchnorm_caches[i])
        dict_dgamma[i] = dgamma
        dict_dbeta[i] = dbeta
        fc_input = dx_batchnorm

      # fc
      dh, dW, db = affine_backward(fc_input, fc_caches[i])
      dict_dW[i] = dW
      dict_db[i] = db

    for i in np.arange(self.num_layers):
      num = str(i+1)
      grads['W'+num] = dict_dW[i+1] + self.reg * self.params['W'+num]
      grads['b'+num] = dict_db[i+1]
      if i == (self.num_layers - 1):
        break
      if self.use_batchnorm:
        grads['gamma'+num] = dict_dgamma[i+1]
        grads['beta'+num] = dict_dbeta[i+1]

    # ============================================================ #
    # END YOUR CODE HERE
    # ============================================================ #

    return loss, grads
```