Activation Function Implementations:

Implementation of `activations.Linear`:

```
class Linear(Activation):
    def __init__(self):
        super().__init__()

    def forward(self, Z: np.ndarray) -> np.ndarray:
        """Forward pass for f(z) = z.

        Parameters
        ----------
        Z  input pre-activations (any shape)

        Returns
        -------
        f(z) as described above applied elementwise to `Z`
        """
        return Z

    def backward(self, Z: np.ndarray, dY: np.ndarray) -> np.ndarray:
        """Backward pass for f(z) = z.

        Parameters
        ----------
        Z   input to `forward` method
        dY  derivative of loss w.r.t. the output of this layer
            same shape as `Z`

        Returns
        -------
        derivative of loss w.r.t. input of this layer
        """
        return dY
```

Implementation of `activations.Sigmoid`:

```python
class Sigmoid(Activation):
    def __init__(self):
        super().__init__()

    def forward(self, Z: np.ndarray) -> np.ndarray:
        """Forward pass for sigmoid function:
        f(z) = 1 / (1 + exp(-z))

        Parameters
        ----------
        Z  input pre-activations (any shape)

        Returns
        -------
        f(z) as described above applied elementwise to `Z`
        """
        ### YOUR CODE HERE ###
        return ...

    def backward(self, Z: np.ndarray, dY: np.ndarray) -> np.ndarray:
        """Backward pass for sigmoid.

        Parameters
        ----------
        Z   input to `forward` method
        dY  derivative of loss w.r.t. the output of this layer
            same shape as `Z`

        Returns
        -------
        derivative of loss w.r.t. input of this layer
        """
        ### YOUR CODE HERE ###
        return ...
```

Implementation of `activations.ReLU`:

```
class ReLU(Activation):
    def __init__(self):
        super().__init__()

    def forward(self, Z: np.ndarray) -> np.ndarray:
        """Forward pass for relu activation:
        f(z) = z if z >= 0
               0 otherwise

        Parameters
        ----------
        Z  input pre-activations (any shape)

        Returns
        -------
        f(z) as described above applied elementwise to `Z`
        """
        return np.maximum(0, Z)

    def backward(self, Z: np.ndarray, dY: np.ndarray) -> np.ndarray:
        """Backward pass for relu activation.

        Parameters
        ----------
        Z   input to `forward` method
        dY  derivative of loss w.r.t. the output of this layer
            same shape as `Z`

        Returns
        -------
        derivative of loss w.r.t. input of this layer
        """
        dYdZ = Z
        dYdZ[Z >= 0] = 1
        dYdZ[Z < 0] = 0
        return dY * dYdZ
```

Implementation of `activations.SoftMax`:

```
class SoftMax(Activation):
    def __init__(self):
        super().__init__()

    def forward(self, Z: np.ndarray) -> np.ndarray:
        """Forward pass for softmax activation.
        Hint: The naive implementation might not be numerically stable.

        Parameters
        ----------
        Z  input pre-activations (any shape)

        Returns
        -------
        f(z) as described above applied elementwise to `Z`
        """
        ### YOUR CODE HERE ###
        stableZ = Z - np.max(Z, axis=-1, keepdims=True)
        exp = np.exp(stableZ)
        return np.divide(exp, np.sum(exp, axis=-1, keepdims=True))

    def backward(self, Z: np.ndarray, dY: np.ndarray) -> np.ndarray:
        """Backward pass for softmax activation.

        Parameters
        ----------
        Z   input to `forward` method
        dY  derivative of loss w.r.t. the output of this layer
            same shape as `Z`

        Returns
        -------
        derivative of loss w.r.t. input of this layer
        """
        ### YOUR CODE HERE ###
        probs = self.forward(Z)
        backward = []
        for idx, x in enumerate(probs):
            diag = np.diagflat(x)
            x = x.reshape((-1, 1))
            J = diag - np.dot(x, x.T)
            backward.append(dY[idx] @ J)
        final = np.array(backward)
        return final
```

Layer Implementations:

Implementation of `layers.FullyConnected`:

```
class FullyConnected(Layer):
    """A fully-connected layer multiplies its input by a weight matrix, adds
    a bias, and then applies an activation function.
    """

    def __init__(
        self, n_out: int, activation: str, weight_init="xavier_uniform"
    ) -> None:

        super().__init__()
        self.n_in = None
        self.n_out = n_out
        self.activation = initialize_activation(activation)

        # instantiate the weight initializer
        self.init_weights = initialize_weights(weight_init, activation=activation)
```

```
        self.init_weights = initialize_weights(weight_init, activation=activation)

    def _init_parameters(self, X_shape: Tuple[int, int]) -> None:
        """Initialize all layer parameters (weights, biases)."""
        self.n_in = X_shape[1]

        ### BEGIN YOUR CODE ###

        W = self.init_weights((self.n_in, self.n_out))
        b = np.zeros((1, self.n_out))

        self.parameters = OrderedDict({"W": W, "b": b})
        self.cache: OrderedDict = ({"Z": [], "X": []})  # cache for backprop
        self.gradients: OrderedDict = ({"W": np.zeros_like(W), "b": np.zeros_like(b)})  # parameter gradients initialized to zero
                                           # MUST HAVE THE SAME KEYS AS `self.parameters`

        ### END YOUR CODE ###

    def forward(self, X: np.ndarray) -> np.ndarray:
        """Forward pass: multiply by a weight matrix, add a bias, apply activation.
        Also, store all necessary intermediate results in the `cache` dictionary
        to be able to compute the backward pass.

        Parameters
        ----------
        X  input matrix of shape (batch_size, input_dim)

        Returns
        -------
        a matrix of shape (batch_size, output_dim)
        """
        # initialize layer parameters if they have not been initialized
        if self.n_in is None:
            self._init_parameters(X.shape)

        ### BEGIN YOUR CODE ###
        W = self.parameters["W"]
        b = self.parameters["b"]
        Z = X @ W + b

        # perform an affine transformation and activation
        out = self.activation(Z)

        # store information necessary for backprop in `self.cache`
        self.cache["Z"] = Z
        self.cache["X"] = X
        ### END YOUR CODE ###

        return out

    def backward(self, dLdY: np.ndarray) -> np.ndarray:
        """Backward pass for fully connected layer.
        Compute the gradients of the loss with respect to:
            1. the weights of this layer (mutate the `gradients` dictionary)
            2. the bias of this layer (mutate the `gradients` dictionary)
            3. the input of this layer (return this)

        Parameters
        ----------
        dLdY  derivative of the loss with respect to the output of this layer
              shape (batch_size, output_dim)

        Returns
        -------
        derivative of the loss with respect to the input of this layer
        shape (batch_size, input_dim)
```

```
        """
        ### BEGIN YOUR CODE ###
        W = self.parameters["W"]
        b = self.parameters["b"]

        # unpack the cache
        Z = self.cache["Z"]
        X = self.cache["X"]

        # compute the gradients of the loss w.r.t. all parameters as well as the
        # input of the layer
        dZ = self.activation.backward(Z, dLdY)
        dX = dZ @ W.T
        dW = X.T @ dZ
        db = np.sum(dZ, axis=0, keepdims=True)

        # store the gradients in `self.gradients`
        # the gradient for self.parameters["W"] should be stored in
        # self.gradients["W"], etc.
        self.gradients["W"] = dW
        self.gradients["b"] = db
        ### END YOUR CODE ###

        return dX
```

Implementation of `layers.Pool2D`:

```
class Pool2D(Layer):
    """Pooling layer, implements max and average pooling."""

    def __init__(
        self,
        kernel_shape: Tuple[int, int],
        mode: str = "max",
        stride: int = 1,
        pad: Union[int, Literal["same"], Literal["valid"]] = 0,
    ) -> None:

        if type(kernel_shape) == int:
            kernel_shape = (kernel_shape, kernel_shape)

        self.kernel_shape = kernel_shape
        self.stride = stride

        if pad == "same":
            self.pad = ((kernel_shape[0] - 1) // 2, (kernel_shape[1] - 1) // 2)
        elif pad == "valid":
            self.pad = (0, 0)
        elif isinstance(pad, int):
            self.pad = (pad, pad)
        else:
            raise ValueError("Invalid Pad mode found in self.pad.")

        self.mode = mode

        if mode == "max":
            self.pool_fn = np.max
            self.arg_pool_fn = np.argmax
        elif mode == "average":
            self.pool_fn = np.mean

        self.cache = {
            "out_rows": [],
            "out_cols": [],
            "X_pad": [],
```

```
            "p": [],
            "pool_shape": [],
        }
        self.parameters = {}
        self.gradients = {}

    def forward(self, X: np.ndarray) -> np.ndarray:
        """Forward pass: use the pooling function to aggregate local information
        in the input. This layer typically reduces the spatial dimensionality of
        the input while keeping the number of feature maps the same.

        As with all other layers, please make sure to cache the appropriate
        information for the backward pass.

        Parameters
        ----------
        X  input array of shape (batch_size, in_rows, in_cols, channels)

        Returns
        -------
        pooled array of shape (batch_size, out_rows, out_cols, channels)
        """
        ### BEGIN YOUR CODE ###
        n_examples, in_rows, in_cols, in_channels = X.shape
        kernel_height, kernel_width = self.kernel_shape

        out_rows = int(((in_rows + 2*self.pad[0] - kernel_height) / self.stride) + 1)
        out_cols = int(((in_cols + 2*self.pad[1] - kernel_width) / self.stride) + 1)
        X_padded = np.pad(X, ((0, 0), (self.pad[0], self.pad[0]), (self.pad[1], self.pad[1]), (0, 0)), mode='constant')

        X_pool = np.zeros((n_examples, out_rows, out_cols, in_channels))

        # implement the forward pass
        for r in range(out_rows):
            for c in range(out_cols):
                r_start, r_end = r * self.stride, (r * self.stride) + kernel_height
                c_start, c_end = c * self.stride, (c * self.stride) + kernel_width
                X_pool[:, r, c, :] = self.pool_fn(X_padded[:, r_start:r_end, c_start:c_end, :], axis=(1, 2))

        # cache any values required for backprop
        self.cache["X_pad"] = X_padded

        ### END YOUR CODE ###

        return X_pool

    def backward(self, dLdY: np.ndarray) -> np.ndarray:
        """Backward pass for pooling layer.

        Parameters
        ----------
        dLdY  gradient of loss with respect to the output of this layer
              shape (batch_size, out_rows, out_cols, channels)

        Returns
        -------
        gradient of loss with respect to the input of this layer
        shape (batch_size, in_rows, in_cols, channels)
        """
        ### BEGIN YOUR CODE ###
        X_padded = self.cache["X_pad"]
        n_examples, in_rows, in_cols, in_channels = X_padded.shape
        kernel_height, kernel_width = self.kernel_shape

        out_rows = int(((in_rows - kernel_height) / self.stride) + 1)
```

```
        out_cols = int(((in_cols - kernel_width) / self.stride) + 1)

        dX = np.zeros_like(X_padded)

        # perform a backward pass
        for r in range(out_rows):
            for c in range(out_cols):
                r_start, r_end = r * self.stride, (r * self.stride) + kernel_height
                c_start, c_end = c * self.stride, (c * self.stride) + kernel_width
                if self.mode == "max":
                    temp = X_padded[:, r_start:r_end, c_start:c_end, :]
                    tempMax = self.pool_fn(temp, axis=(1,2)).reshape(temp.shape[0], 1, 1, temp.shape[-1])
                    mask = np.equal(temp, tempMax).astype(int)
                    dX[:, r_start:r_end, c_start:c_end, :] += mask * dLdY[:, r, c, :].reshape(dLdY.shape[0], 1, 1, dLdY.shape[-1])
                else:
                    # print(self.pool_fn(dLdY[:, r, c, :]).shape)
                    # dX[:, r_start:r_end, c_start:c_end, :] = dX[:, r_start:r_end, c_start:c_end, :] + \
                        # (np.ones((1, kernel_height, kernel_width, 1)) / (kernel_width * kernel_height)) \
                        # * dLdY[:, r, c, :].reshape(dLdY.shape[0], 1, 1, dLdY.shape[-1])
                    for n in range(n_examples):
                        for ch in range(in_channels):
                            dLdY_avg = dLdY[n, r, c, ch] / (kernel_height * kernel_width)
                            dX[n, r_start:r_end, c_start:c_end, ch] = np.ones((kernel_height, kernel_width)) * dLdY_avg

        ### END YOUR CODE ###

        return dX
```

Implementation of `layers.Conv2D.__init__`:

```
    def __init__(
        self,
        n_out: int,
        kernel_shape: Tuple[int, int],
        activation: str,
        stride: int = 1,
        pad: str = "same",
        weight_init: str = "xavier_uniform",
    ) -> None:

        super().__init__()
        self.n_in = None
        self.n_out = n_out
        self.kernel_shape = kernel_shape
        self.stride = stride
        self.pad = pad

        self.activation = initialize_activation(activation)
        self.init_weights = initialize_weights(weight_init, activation=activation)
```

Implementation of `layers.Conv2D._init_parameters`:

```python
    def _init_parameters(self, X_shape: Tuple[int, int, int, int]) -> None:
        """Initialize all layer parameters and determine padding."""
        self.n_in = X_shape[3]

        W_shape = self.kernel_shape + (self.n_in,) + (self.n_out,)
        W = self.init_weights(W_shape)
        b = np.zeros((1, self.n_out))

        self.parameters = OrderedDict({"W": W, "b": b})
        self.cache = OrderedDict({"Z": [], "X": []})
        self.gradients = OrderedDict({"W": np.zeros_like(W), "b": np.zeros_like(b)})

        if self.pad == "same":
            self.pad = ((W_shape[0] - 1) // 2, (W_shape[1] - 1) // 2)
        elif self.pad == "valid":
            self.pad = (0, 0)
        elif isinstance(self.pad, int):
            self.pad = (self.pad, self.pad)
        else:
            raise ValueError("Invalid Pad mode found in self.pad.")
```

Implementation of `layers.Conv2D.forward`:

```python
    def forward(self, X: np.ndarray) -> np.ndarray:
        """Forward pass for convolutional layer. This layer convolves the input
        `X` with a filter of weights, adds a bias term, and applies an activation
        function to compute the output. This layer also supports padding and
        integer strides. Intermediates necessary for the backward pass are stored
        in the cache.

        Parameters
        ----------
        X  input with shape (batch_size, in_rows, in_cols, in_channels)

        Returns
        -------
        output feature maps with shape (batch_size, out_rows, out_cols, out_channels)
        """
        if self.n_in is None:
            self._init_parameters(X.shape)

        W = self.parameters["W"]
        b = self.parameters["b"]

        kernel_height, kernel_width, in_channels, out_channels = W.shape
        n_examples, in_rows, in_cols, in_channels = X.shape
        kernel_shape = (kernel_height, kernel_width)

        ### BEGIN YOUR CODE ###
        X_col, p = im2col(X, kernel_shape, self.stride, self.pad)
        W_col = W.transpose(3, 2, 0, 1).reshape(out_channels, -1)
        out_rows = int((in_rows + p[0] + p[1] - kernel_height) / (self.stride) + 1)
        out_cols = int((in_cols + p[2] + p[3] - kernel_width) / (self.stride) + 1)

        # implement a convolutional forward pass
        Z = ((W_col @ X_col).reshape(out_channels, out_rows, out_cols, n_examples).transpose(3, 1, 2, 0)) + b
        out = self.activation(Z)

        # cache any values required for backprop
        self.cache["Z"] = Z
        self.cache["X"] = X
        ### END YOUR CODE ###

        return out
```

Implementation of `layers.Conv2D.backward`:

```
def backward(self, dLdY: np.ndarray) -> np.ndarray:
    """Backward pass for conv layer. Computes the gradients of the output
    with respect to the input feature maps as well as the filter weights and
    biases.

    Parameters
    ----------
    dLdY  derivative of loss with respect to output of this layer
          shape (batch_size, out_rows, out_cols, out_channels)

    Returns
    -------
    derivative of the loss with respect to the input of this layer
    shape (batch_size, in_rows, in_cols, in_channels)
    """
    ### BEGIN YOUR CODE ###
    Z = self.cache["Z"]
    X = self.cache["X"]
    W = self.parameters["W"]
    b = self.parameters["b"]

    kernel_height, kernel_width, in_channels, out_channels = W.shape
    n_examples, in_rows, in_cols, in_channels = X.shape
    kernel_shape = (kernel_height, kernel_width)

    # perform a backward pass
    dZ = self.activation.backward(Z, dLdY)
    dZ_col = dZ.transpose(3, 1, 2, 0).reshape(out_channels, -1)
    W_col = W.transpose(3, 2, 0, 1).reshape(out_channels, -1).T
    X_col, p = im2col(X, kernel_shape, self.stride, self.pad)
    print(dZ_col.shape, X_col.shape)
    dW = ((dZ_col @ X_col.T).reshape(out_channels, in_channels, kernel_height, kernel_width).transpose(2, 3, 1, 0))
    dB = np.sum(dZ_col, axis=1).reshape(1, -1)
    dX_col = W_col @ dZ_col
    dX = col2im(dX_col, X, W.shape, self.stride, p).transpose(0, 2, 3, 1)

    self.gradients["W"] = dW
    self.gradients["b"] = dB
    ### END YOUR CODE ###

    return dX
```

Loss Function Implementations:

Implementation of `losses.CrossEntropy`:

```
class CrossEntropy(Loss):
    """Cross entropy loss function."""

    def __init__(self, name: str) -> None:
        self.name = name

    def __call__(self, Y: np.ndarray, Y_hat: np.ndarray) -> float:
        return self.forward(Y, Y_hat)

    def forward(self, Y: np.ndarray, Y_hat: np.ndarray) -> float:
        """Computes the loss for predictions `Y_hat` given one-hot encoded labels
        `Y`.

        Parameters
        ----------
        Y      one-hot encoded labels of shape (batch_size, num_classes)
        Y_hat  model predictions in range (0, 1) of shape (batch_size, num_classes)

        Returns
        -------
        a single float representing the loss
        """
        ### YOUR CODE HERE ###
        return -np.sum(Y * np.log(Y_hat + np.finfo(np.float64).tiny)) / Y.shape[0]

    def backward(self, Y: np.ndarray, Y_hat: np.ndarray) -> np.ndarray:
        """Backward pass of cross-entropy loss.
        NOTE: This is correct ONLY when the loss function is SoftMax.

        Parameters
        ----------
        Y      one-hot encoded labels of shape (batch_size, num_classes)
        Y_hat  model predictions in range (0, 1) of shape (batch_size, num_classes)

        Returns
        -------
        the derivative of the cross-entropy loss with respect to the vector of
        predictions, `Y_hat`
        """
        ### YOUR CODE HERE ###
        return -Y / (Y_hat * Y.shape[0])
```

Implementation of `losses.L2`:

```
class L2(Loss):
    """Mean squared error loss."""

    def __init__(self, name: str) -> None:
        self.name = name

    def __call__(self, Y: np.ndarray, Y_hat: np.ndarray) -> float:
        return self.forward(Y, Y_hat)

    def forward(self, Y: np.ndarray, Y_hat: np.ndarray) -> float:
        """Compute the mean squared error loss for predictions `Y_hat` given
        regression targets `Y`.

        Parameters
        ----------
        Y      vector of regression targets of shape (batch_size, 1)
        Y_hat  vector of predictions of shape (batch_size, 1)

        Returns
        -------
        a single float representing the loss
        """
        ### YOUR CODE HERE ###
        return ...

    def backward(self, Y: np.ndarray, Y_hat: np.ndarray) -> np.ndarray:
        """Backward pass for mean squared error loss.

        Parameters
        ----------
        Y      vector of regression targets of shape (batch_size, 1)
        Y_hat  vector of predictions of shape (batch_size, 1)

        Returns
        -------
        the derivative of the mean squared error with respect to the last layer
        of the neural network
        """
        ### YOUR CODE HERE ###
        return ...
```

Model Implementations:

Implementation of `models.NeuralNetwork.forward`:

```
    def forward(self, X: np.ndarray) -> np.ndarray:
        """One forward pass through all the layers of the neural network.

        Parameters
        ----------
        X  design matrix whose must match the input shape required by the
           first layer

        Returns
        -------
        forward pass output, matches the shape of the output of the last layer
        """
        ### YOUR CODE HERE ###
        # Iterate through the network's layers.
        Y = X
        for layer in self.layers:
            Y = layer.forward(Y)
        return Y
```

Implementation of `models.NeuralNetwork.backward`:

```
def backward(self, target: np.ndarray, out: np.ndarray) -> float:
    """One backward pass through all the layers of the neural network.
    During this phase we calculate the gradients of the loss with respect to
    each of the parameters of the entire neural network. Most of the heavy
    lifting is done by the `backward` methods of the layers, so this method
    should be relatively simple. Also make sure to compute the loss in this
    method and NOT in `self.forward`.

    Note: Both input arrays have the same shape.

    Parameters
    ----------
    target  the targets we are trying to fit to (e.g., training labels)
    out     the predictions of the model on training data

    Returns
    -------
    the loss of the model given the training inputs and targets
    """
    ### YOUR CODE HERE ###
    # Compute the loss.
    # Backpropagate through the network's layers.
    L = self.loss.forward(target, out)
    dLdY = self.loss.backward(target, out)
    for layer in reversed(self.layers):
        dLdY = layer.backward(dLdY)
    return L
```

Implementation of `models.NeuralNetwork.predict`:

```
def predict(self, X: np.ndarray, Y: np.ndarray) -> Tuple[np.ndarray, float]:
    """Make a forward and backward pass to calculate the predictions and
    loss of the neural network on the given data.

    Parameters
    ----------
    X  input features
    Y  targets (same length as `X`)

    Returns
    -------
    a tuple of the prediction and loss
    """
    ### YOUR CODE HERE ###
    # Do a forward pass. Maybe use a function you already wrote?
    # Get the loss. Remember that the `backward` function returns the loss.
    Yhat = self.forward(X)
    L = self.backward(Y, Yhat)
    return (Yhat, L)
```