GENERAL

- PSET 5 Posted.
- One day extension (now due on the Friday two weeks from now)
- You get to implement SLIC, and your own conv layer!
  - and work on the “world famous” MNIST data set.
- Has extra credit question—but prioritize your projects.
- Look through the entire code!
- Regular Office Hours Tomorrow.
- Recitation Monday (9/20) Evening at 5:30pm
  - In Jolley 420
- Wed Morning Office Hours as usual

AUTOGRAD

Code from pset5/mnist.py

```python
# Inputs and parameters
inp = edf.Value()
lab = edf.Value()
W1 = edf.Param()
B1 = edf.Param()
W2 = edf.Param()
B2 = edf.Param()

# Model
y = edf.matmul(inp, W1)
y = edf.add(y, B1)
y = edf.RELU(y)
y = edf.matmul(y, W2)
y = edf.add(y, B2) # This is our final prediction
```

When you construct an object, it just does book-keeping!
# Inputs and parameters
inp = edf.Value()
lab = edf.Value()
W1 = edf.Param()
B1 = edf.Param()
W2 = edf.Param()
B2 = edf.Param()

# Model
y = edf.matmul(inp,W1)
y = edf.add(y,B1)
y = edf.RELU(y)
y = edf.matmul(y,W2)
y = edf.add(y,B2) # This is our final prediction

# Inputs and parameters
inp = edf.Value()
lab = edf.Value()
W1 = edf.Param()
B1 = edf.Param()
W2 = edf.Param()
B2 = edf.Param()

# Model
y = edf.matmul(inp,W1)
y = edf.add(y,B1)
y = edf.RELU(y)
y = edf.matmul(y,W2)
y = edf.add(y,B2) # This is our final prediction

# Inputs and parameters
inp = edf.Value()
lab = edf.Value()
W1 = edf.Param()
B1 = edf.Param()
W2 = edf.Param()
B2 = edf.Param()

# Model
y = edf.matmul(inp,W1)
y = edf.add(y,B1)
y = edf.RELU(y)
y = edf.matmul(y,W2)
y = edf.add(y,B2) # This is our final prediction
# Inputs and parameters

```python
# Model
y = edf.add(y, B1)
y = edf.relu(y)
y = edf.add(y, B2) # This is our final prediction
```
# Inputs and parameters

- `inp = edf.Value()`
- `lab = edf.Value()`
- `W1 = edf.Param()`
- `B1 = edf.Param()`
- `W2 = edf.Param()`
- `B2 = edf.Param()`

```python
y = edf.matmul(inp, W1)
y = edf.add(y, B1)
y = edf.RELU(y)
y = edf.matmul(y, W2)
y = edf.add(y, B2)
```

```python
loss = edf.smaxloss(y, lab)
loss = edf.mean(loss)
```

```python
acc = edf.accuracy(y, lab)
acc = edf.mean(acc)
```

**BRIEF DETOUR**

- **What are RELUs?**
  - Element-wise non-linear activations
  - Previous activations would be sigmoid-like: $\sigma(x) = \frac{\exp(x)}{1 + \exp(x)}$
  - Great when you want a probability, bad when you want to learn by gradient descent.
  - For both high and low-values of $x$, $\frac{\partial \sigma(x)}{\partial x} = 0$
  - So if you weren't careful, you would end up with high-magnitude activations, and the network stops learning.
  - Gradient descent is very fragile!

- RELU($x$) = $\max(0, x)$

- But this can saturate too!

  What is $\frac{\partial \text{RELU}(x)}{\partial x}$?

  0 if $x < 0$, 1 otherwise.
**BRIEF DETOUR**

- What are RELUs?

\[ \text{RELU}(x) = \max(0, x) \]

But this can saturate too!

If all your \(x\)'s are negative.

- Will talk about batch-norm later, but we try to make sure all activations are zero-mean.
- Normalize inputs to be zero mean unit variance.
- Initialize each layer's weights randomly, such that if its inputs are zero mean unit variance, so are its outputs in expectation.
- And then, learn very very slowly.

**BRIEF DETOUR**

- What is \(-\text{max}\)?

We are now doing multi-class prediction. Instead of binary, our answers are one of \(K\) classes.

- Binary case: produce a single number and checking if it is \(>0\) or \(<0\)
- \(K\) Class Case: produce a vector of \(K\) numbers, and do \(\arg\max\). We say \(k\) is the right answer, if the \(k\)th element of this vector is the highest. (soft-max is a soft version of \(\arg\max\), not max)
- Binary Case: Predicted log-odds
- Multi-class Case: Predict un-normalized log probabilities

**BRIEF DETOUR**

- What is \(-\text{max}\)?

\(y = f(x) \in \mathbb{R}^K\)

- Predict class \(k\) if \(y^k \geq y^{k'} \forall k'\)

Here \(y^k\) refers to the \(k\)th element of \(y\), not an exponent

\[ P(\text{class} = k) = p^k = \frac{\exp(y^k)}{\sum_{k'} \exp(y^{k'})} \]

- Cross Entropy Loss

\[ L(C, y) = -\log p^C = \log(\sum_{k'} \exp(y^{k'}) - \exp(y^C) \]  

for true \(C \in \{0, 2, \ldots, K - 1\}\)

**BRIEF DETOUR**

- What is \(-\text{max}\)?

\(y = f(x) \in \mathbb{R}^K\)

- Predict class \(k\) if \(y^k \geq y^{k'} \forall k'\)

\[ P(\text{class} = k) = p^k = \frac{\exp(y^k)}{\sum_{k'} \exp(y^{k'})} \]

- Cross Entropy Loss

\[ L(C, y) = -\log p^C = \log(\sum_{k'} \exp(y^{k'}) - \exp(y^C) \]

What is \(\frac{\partial L}{\partial y^k}\)? For \(k = c\) and \(k \neq c\)?

\[ \frac{\partial L}{\partial y^k} = p^k \text{ if } k \neq C \]

\[ \frac{\partial L}{\partial y^C} = p^C - 1 \]
**BRIEF DETOUR**

- What is soft-max?
  
  \[ y = f(x) \in \mathbb{R}^K \]

  Predict class \( k \) if \( y^k \geq y^{k'} \) \( \forall k' \)

  \[ L(C,y) = -\log p^C = \log(\sum_{k'} \exp y^{k'}) - y^C \]

  \[ \frac{\partial L}{\partial y^k} = p^k \text{ if } k \neq C \]

  \[ \frac{\partial L}{\partial y^C} = p^C - 1 \]

  Similar form to binary case: \( p^k - p^C \)

  Soft-max is not an element-wise non-linearity. True class \( C \) creates gradients for all classes.

  You'll want to implement soft-max + cross-entropy usually as a joint "layer" for numerical stability and efficiency.

**AUTOGRAD**

Now let's train this thing!

Beginning of training:

```python
nHidden = 1024; K = 10
W1.set(xavier((28*28,nHidden)))
B1.set(np.zeros((nHidden)))
W2.set(xavier((nHidden,K)))
B2.set(np.zeros((K)))
```

Initialize weights randomly.

In each iteration of training:

```python
for iters in range(...):
    inp.set(train_im[idx[b:b+BSZ],:])
    lab.set(train_lb[idx[b:b+BSZ]])

    loss = edf.smaxloss(y,lab)
    loss = edf.mean(loss)

    acc = edf.accuracy(y,lab)
    acc = edf.mean(acc)
```

**AUTOGRAD**

set is the only function that the classes param and value have.

```python
class Value:
    def __init__(self):
        values.append(self)

    def set(self,value):
        self.top = np.float32(value).copy()

class Param:
    def __init__(self):
        params.append(self)

    def set(self,value):
        self.top = np.float32(value).copy()
```

Sets a member called "top" to be an array that holds these values.
And this will work. It will print the loss and accuracy values for the set inputs, given the current value of the parameters.

What is this magical function `forward`?

```python
W1.set(xavier((28*28,nHidden)))
...
B2.set(np.zeros((K)))

for iters in range(...):
    inp.set(train_im[idx[b:b+BSZ],:])
    lab.set(train_lb[idx[b:b+BSZ]])
    edf.Forward()
    print(loss.top,acc.top)
```

From `edf.py`

```
# Global forward
def Forward():
    for c in ops: c.forward()
```

But the operation classes have their own forward function.

```
class matmul:
    def __init__(self,x,y):
        ops.append(self)
        self.x = x
        self.y = y

    def forward(self):
        self.top = np.matmul(self.x.top,self.y.top)
```

...
So the forward pass computes the loss. But we want to learn the parameters.

The SGD function is pretty simple. Requires `p.grad` (gradients with respect to loss) to be present.

```python
for iters in range(...):
    inp.set(train_im[idx[b:b+BSZ],:])
    lab.set(train_lb[idx[b:b+BSZ]])
    edf.Forward()
    print(loss.top, acc.top)
    edf.Backward(loss)
    edf.SGD(lr)
```

- The SGD function is pretty simple

```python
def SGD(lr):
    for p in params:
        p.top = p.top - lr*p.grad
```

That's what backward does!
**AUTOGRAD**

From edf.py

```python
# Global backward
def Backward(loss):
    for c in ops:
        c.grad = np.zeros_like(c.top)
    for c in params:
        c.grad = np.zeros_like(c.top)
    loss.grad = np.ones_like(loss.top)
    for c in ops[::-1]: c.backward()
```

- Called with an op object as the loss.
- Initializes all gradients to zero.
- x.grad is defined as the gradient of the loss wrt to x.top
- And so, initializes loss.grad to all ones.
- Each op has a backward function. Calls this in reverse order.

**AUTOGRAD**

```python
# Matrix multiply (fully-connected layer)
class matmul:
    def __init__(self,x,y):
        ops.append(self)
        self.x = x
        self.y = y
    def forward(self):
        self.top = np.matmul(self.x.top,self.y.top)
    def backward(self):
        if self.x in ops or self.x in params:
            self.x.grad = self.x.grad +
            np.matmul(self.y.top,self.grad.T).T
        if self.y in ops or self.y in params:
            self.y.grad = self.y.grad +
            np.matmul(self.x.top.T,self.grad)
```

- The backward function adds to the gradients of its inputs. Those inputs could've been used by other ops.
- Computes gradients of its inputs based on the value of its own gradients.
- Assumes that by the time backward is called on it, self.grad will exist and be final. Why?

**AUTOGRAD**

```python
# Matrix multiply (fully-connected layer)
class matmul:
    def __init__(self,x,y):
        ops.append(self)
        self.x = x
        self.y = y
    def forward(self):
        self.top = np.matmul(self.x.top,self.y.top)
    def backward(self):
        if self.x in ops or self.x in params:
            self.x.grad = self.x.grad +
            np.matmul(self.y.top,self.grad.T).T
        if self.y in ops or self.y in params:
            self.y.grad = self.y.grad +
            np.matmul(self.x.top.T,self.grad)
```

- Assumes that by the time backward is called on it, self.grad will exist and be final.
- Because anything that could add to its gradient would have been called before it.
- Only computes grads for params and ops (not values).
AUTOGRAD

• Allows us to specify a complex estimation function as a composition of simpler functions.
• Need to just provide an ‘implementation’ for each function so that we know how to:
  • Compute outputs given inputs
  • Compute contribution to gradients of inputs given gradient of output
• Given this model, can allow us to compute gradients of all parameters automatically!

CONVOLUTIONAL NETWORKS

y = edf.matmul(inp, W1)

for iters in range(...):
    inp.set(train_im[idx[b:b+BSZ],:])
    lab.set(train_lb[idx[b:b+BSZ]])

• So far, even though our inputs were images, we were treating them as vectors.
• inp is a matrix of size $B \times N$, where $B$ is the batch-size, and $N$ was the number of pixels ($28^2$ for MNIST).
• So each sample is represented by a row vector of size $N$.
• This means that parameters of the first layer $W_1 = N \times H$, where $H$ is the dimensionality of the encoding.
• That’s a lot of weights, especially as you go to “real” images.
• We can’t solve this by reducing $H$. Because then we would be assuming that there is a “linear” function that can reduce dimensionality and keep information required for classification.
• Need to do this slowly.

CONVOLUTIONAL NETWORKS

Instead, consider our input to be what it is, an image!

• Work with 4-D arrays of size $B \times H \times W \times C$ instead of matrices.
• Use convolutional layers, which produce multi-channel output images from multi-channel input images, except that each output pixel only depends on a small number of input pixels in its neighborhood.
• And this dependence is translation invariant.

$$g[b, y, x, c_2] = \sum_k \sum_{y} \sum_{x_1} f[b, y + k, x + k, c_1] k[y, k, c_1, c_2]$$

(Note that this is actually ‘correlation’ not convolution)

In the problem set, you’ll have to implement forward and backward.
CONVOLUTIONAL NETWORKS

```python
y = edf.conv2(inp,K1)
y = edf.down2(y); y = edf.down2(y);
y = edf.add(y,B1)
y = edf.RELU(y)
y = edf.flatten(y)
y = edf.matmul(y,W2)
y = edf.add(y,B2) # This is our final prediction
```

This is actually a huge waste of computation! We're computing values and then throwing them away.

CONVOLUTIONAL NETWORKS

```python
# Downsample by 2
class down2:
    def __init__(self,x):
        ops.append(self)
        self.x = x

    def forward(self):
        self.top = self.x.top[:,::2,::2,:]

    def backward(self):
        if self.x in ops or self.x in params:
            grd = np.zeros_like(self.x.top)
            grd[:,::2,::2,:] = self.grad
            self.x.grad = self.x.grad + grd
```

Still to come

- Actual full network models and tasks.
- Regularization: Data augmentation, early stopping, dropout
- Other optimization techniques: batchnorm, Adam, etc.
- Handling Vanishing Gradients: Residual Networks!