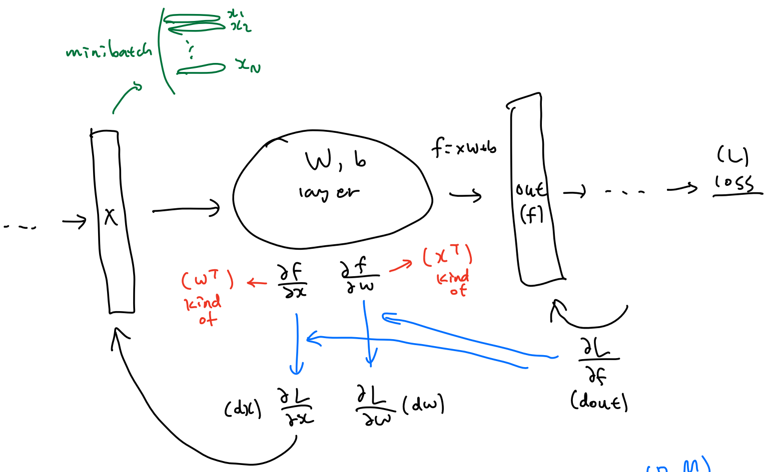

forward method : $\left.\begin{matrix} X \\ w \end{matrix}\right\} \to$ out, cache

cache $(X, w, (z), out)$

(z): some intermediate value

backward method : $\left.\begin{matrix} \frac{\partial(\text{loss})}{\partial(\text{out})} \ (\text{dout}) \\ \text{cache} \end{matrix}\right\} \to \left(\begin{matrix} \frac{\partial(\text{loss})}{\partial x} \\ \frac{\partial(\text{loss})}{\partial w} \end{matrix}\right.$

affine_forward (x, w, b)   [w: (D, M), b: (M,)]

→ (flatten x into 2d (N, D) → X)

→ out = X.dot(w) + b
  cache = (x, w, b)

affine_backward (dout, cache)   [dout: (N, M), cache: x, w, b]

→ (flatten x into 2d (N, D) → X)

→ dx = dout.dot(w^T) → (unflatten)
  dw = X^T.dot(dout)
  db = dout.sum(axis=0)

relu_forward (x)

→ out = relu(x)
  cache = x

relu_backward (dout, cache)   [dout: same dims as x; cache: x]

→ dx = dout * (x > 0)

affine-relu_forward (x, w, b)

$\left.\begin{matrix} x \\ w \\ b \end{matrix}\right\} \to$ affine forward → (a → relu forward → $\boxed{\text{out}}$; (x, w, b))

$\boxed{\text{cache}}$ = ((x, w, b), a)

affine_relu_backward (dout, cache)

cache → (a, (x, w, b)); dout → relu back; a → relu back → da → affine back; (x, w, b) → affine back → $\boxed{dx, dw, db}$

svm_loss (x, y)

x: (N, C) (minibatch) stores for each of the C classes

y: (N,) correct class

→ loss = average loss for N vectors (scalar)
  dx = $\frac{\partial(\text{loss})}{\partial x}$ (size (N, C))

softmax_loss (x, y)

(same but softmax loss)

batchnorm_backward_alt

$$\hat{X} = \frac{X - \mu}{\sqrt{v + \varepsilon}}$$

mean (over each feature)

variance

$$y = \gamma \cdot \hat{X} + \beta$$

$* \ \sigma = \sqrt{v+\varepsilon}$

$$\frac{\partial \mu}{\partial x} = \frac{1}{N} \quad (N, D)$$

$$\frac{\partial v}{\partial x} = \frac{2}{N} x - 2 \cdot \frac{1}{N} \cdot \mu = \frac{2}{N}(x - \mu) \quad (N, D)$$

$$\frac{\partial \sigma}{\partial x} = \frac{\partial}{\partial x}\sqrt{v+\varepsilon} = \frac{\frac{\partial v}{\partial x}}{2\sqrt{v+\varepsilon}} = \frac{1}{N\sigma}(x - \mu) \quad (N, D)$$

$$\frac{N-1}{N\sigma} - (\quad)$$

$$\frac{\partial \hat{X}}{\partial x} = \frac{1}{\sigma}$$

$$\frac{\partial \hat{X}}{\partial \mu} = -\frac{1}{\sigma} \quad \rightarrow \quad -\frac{1}{N\sigma}$$

$$\frac{\partial \hat{X}}{\partial \sigma} = (x - \mu)\left(-\frac{1}{\sigma^2}\right) \quad \rightarrow \quad -\frac{\cancel{(x-\mu)^2}}{\cancel{N}\sigma^3}$$

$$\sigma^2 = \frac{1}{N}(x - \mu)^2$$