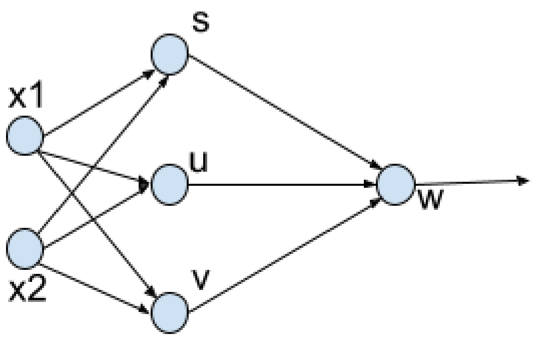

The loss of the above network is given by $f = ((w_1, w_2, w_3)^T(relu(s^T x), relu(u^T x), relu(v^T x)) - y)^2$ where $relu(x) = max(0, x)$ is the relu activation function. We need the first derivatives to optimize the network parameters.

In order to calculate the update equations let $z_1 = relu(s^T x) = relu(s_1 x_1 + s_2 x_2)$. This means I can write f as $f = ((w_1, w_2, w_3)^T(z_1, z_2, z_3) - y)^2$. Then

$df/dw_1 = 2\sqrt{(f)}z_1$ => same as $df/dw_1 = 2((w_1, w_2, w_3)^T(z_1, z_2, z_3) - y)\, z_1$

Thus we can write df/dw as

$df/dw \ = (2((w_1, w_2, w_3)^T(z_1, z_2, z_3) - y))(\, z_1, z_2, z_3)$

Now we calculate df/ds by doing the first coordinate df/ds1.

$df/ds_1 = (df/dz_1)(dz_1/ds_1)$

where $df/dz_1 = 2\sqrt{(f)}w_1$ and $dz_1/ds_1 = drelu(s^T x)/ds_1 = 0\ if\ relu(s^T x)\ <= 0\ and\ x_1\ if\ relu(s^T x)\ > 0$

since $drelu(f(x))/df(x) = 0\ if\ f(x) <= 0\ and\ df/dx\ if\ f(x)\ > 0$. Note that since relu is discontinuous at 0 and has a max we use the sub-gradient.

$df/ds_2 = (df/dz_1)(dz_1/ds_2)$

where
$df/dz_1 = 2\sqrt{(f)}w_1$ and

$dz_1/ds_2 = 0\ if\ relu\ <= 0\ and\ x_2\ if\ relu\ > 0$

This means $df/ds = (df/ds_1, df/ds_2) = df/dz_1(dz_1/ds_1, dz_2, ds_2) = df/dz_1(x_1, x_2)$ (assuming the relu outputs are positive)

From the above it is not too hard to calculate df/du and df/dv.

Compare the above to the sigmoid update given below:
$df/ds = (df/ds_1, df/ds_2) = df/dz_1 \sigma(s^T x)(1 - \sigma(s^T x))(x_1, x_2)$