Reference: http://math.stackexchange.com/questions/945871/derivative-of-softmax-loss-function
Inputs: x_i, where i = 1, 2 ...785
Hidden Layer: output is h_i and input is a_i, where i = 1, 2 ... 201
Output Layer: output is g_i and input is b_i, where i = 1, 2 ... 26
V is 200 x 785 weight matrix, V_{ij} is the weight from x_j to h_i
W is 26 x 201 weight matrix, W_{ij} is the weight from h_j to g_i
$h_i = t(a_i) = t(V_i x) = t(\Sigma_{j=1}^{785} V_{ij} x_j)$ where t is tranh function
$g_i = s(b_i) = s(W_i h) = s(\Sigma_{j=1}^{201} W_{ij} h_j)$ where s is softmax function
$\frac{dt(a)}{da} = 1 - t^2(a)$
$\frac{ds(b)}{db} = s(b) \cdot (1 - s(b))$
$\frac{dL(y,s(b))}{db} = b - y$

W:

$$\frac{dL}{dw_{ij}} = \frac{dL}{db_i} \cdot \frac{db_i}{dw_{ij}}$$
$$= (s(W_i h) - y_i) \cdot h_j$$
$$= (s(W_i h) - y_i) \cdot t(V_j x)$$

matrix form:

$$\frac{dL}{dW} = (b - y)^T \cdot h$$
$$= (Wh - y)^T \cdot h$$
$$= (Wt(Vx) - y)^T \cdot t(Vx)$$

V:

$$\frac{dL}{dv_{ij}} = \frac{dL}{dh_i} \cdot \frac{dh_i}{dv_{ij}} = \frac{dh_i}{dv_{ij}} \cdot \Sigma_{j=1}^{26} (\frac{dL}{db_j} \cdot \frac{b_j}{h_i})$$
$$= x_j \cdot (1 - t^2(a_i)) \cdot \Sigma_{j=1}^{26} ((s(b_j) - y_j) \cdot w_{ji})$$

matrix form:

$$\frac{dL}{dV} = W^T(g - y)(1 - h^2)x$$
$$= W^T(s(Wh) - y)(1 - h^2)x$$
$$= W^T(s(Wt(Vx)) - y)(1 - t(Vx)^2)x$$

Stochastic Gradient Descent:

$$w_{ij} = w_{ij} - \epsilon \cdot \frac{dL}{dw_{ij}}$$
$$v_{ij} = v_{ij} - \epsilon \cdot \frac{dL}{dv_{ij}}$$