CSC411 HW4 2019

1. a)

	#units	#weights	#connections
Conv Layer 1	290400	34 848	105 415 200
2	186624	307 200	223,948,800
3	64896	884736	149 520 384
4	64896	663552	112 140 288
5	43264	442 368	74 760 192
Fully Connected L1	4096	177,209,344	177 209 344
2	4096	16,777,216	16 777 216
Output Layer	1000	4,096,000	4,096,000

Example Calculations for Conv Layer 1:

First layer input: $224 \times 224 \times 3$, 96 Kernals size $11 \times 11 \times 3$. Stride 4.

Suppose input: $W_1 \times L_1 \times D_1$. Filter size $f_1 \times f_1 \times f_2$. Say # Kernals = K

output $W_2 = \frac{W_1 - F}{S} + 1$

$\therefore$ W_2 = First layer output width: $\frac{224-11}{4} + 1 = 54.25 = L_2$. Actually, paper does zero padding so output is 55×55

D_2 = # Kernals/Filters = $K = 96$

units $= W_2 \times L_2 \times D_2 = 55 \times 55 \times 96 = 290\,400$

Weights are shared. Each filter has $f_1 \times f_1 \times f_2$ weights.

$\therefore$ #weights $= (f_1 \times f_1 \times f_2) \times K = 11 \times 11 \times 3 \times 96 = 34\,848$

connections: Each output unit is connected via same # of weights, that is, by the filter size $f_1 \times f_1 \times f_2$.

There are $W_2 \times L_2 \times D_2$ output units.

$(W_2 \times L_2 \times D_2) \times (f_1 \times f_1 \times f_2)$ connections.

$= (55 \times 55 \times 96) \times (11 \times 11 \times 3) = 105\,415\,200$

For fully connected layer:

#weights = #connections = (#units layer)(#units previous layer)

connections for layer 2: $(186,624) \times (5 \times 5 \times 48) = 223\,948\,800$

NOT 96 because layers get split b/w two GPUs.

b) i) A large number of weights are from the fully connected layers, particularly after conv layer 5.

- Add another pooling layer(s) after conv layer 5
- Reduce the size of the fully connected layer to say around 3000.

ii) The source of the huge number of connections is the earlier conv layers.

The equation is:

output nodes $\times$ # weights.

$$= (W_2 \times L_2 \times K) \times (f_1 \times f_1 \times 3)$$

$$\approx \left(\frac{W_1 - f_1}{s} \times \frac{L_1 - f_1}{s} \times K\right) \times (f_1 \times f_1 \times 3)$$

$$= \frac{3f_1^2}{s^2} K \left(W_1 - f_1\right)^2 \quad \text{if } W_1 = L_1$$

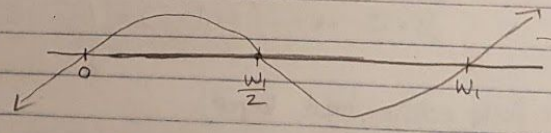
Clearly we can increase stride to radically lower # connections b/w conv layers 1, 2, 3, 4, 5. We can also decrease # filters K.

What about filter-size f_1?

$$\frac{3f_1^2 K}{s^2} 2(W_1 - f_1)(-1) + (W_1 - f_1)^2 \frac{6 f_1 K}{s^2} \quad \rightarrow \partial/\partial f_1$$

$$\frac{6(W_1 - f_1) f_1 K}{s^2} \left(-(W_1 - f_1) - f_1\right) = \frac{6(W_1 - f_1) f_1 K}{s^2} (W_1 - 2f_1)$$

$f_1 = W_1$ or $\frac{W_1}{2}$ or 0



b/c coeff of f_1^3 is positive.

I think it is clear that changing f_1 won't have a huge impact.

2. a) $P(x|y,\mu,\sigma) = \frac{P(y|x,\mu,\sigma)P(x|\mu,\sigma)}{P(y|\mu,\sigma)}$

$\therefore P(y|x,\mu,\sigma) = \frac{P(x|y,\mu,\sigma)P(y|\mu,\sigma)}{P(x|\mu,\sigma)}$

$P(x|\mu,\sigma) = \sum_{i=1}^{K} P(x|y=i,\mu,\sigma)P(y=i|\mu,\sigma)$

$\therefore P(Y=i|x,\mu,\sigma) = \frac{P(x|y=i,\mu,\sigma)\alpha_i}{\sum_{j=1}^{K} P(x|y=j,\mu,\sigma)\alpha_j} = \frac{1}{1+\sum_{j\neq i}\frac{P(x|y=j,\mu,\sigma)\alpha_j}{P(x|y=i,\mu,\sigma)\alpha_i}}$

$\frac{P(x|y=j,\mu,\sigma)}{P(x|y=i,\mu,\sigma)} = \frac{\left(\prod_{d=1}^{D} 2\pi\sigma_d^2\right)^{-1/2}\exp\left\{-\sum_{d=1}^{D}\frac{1}{2\sigma_d^2}(x_d-\mu_{jd})^2\right\}}{\left(\prod_{d=1}^{D} 2\pi\sigma_d^2\right)^{-1/2}\exp\left\{-\sum_{d=1}^{D}\frac{1}{2\sigma_d^2}(x_d-\mu_{id})^2\right\}}$

$= \exp\left\{-\left[\sum_{d=1}^{D}\frac{1}{2\sigma_d^2}\left((x_d-\mu_{jd})^2-(x_d-\mu_{id})^2\right)\right]\right\}$

$\therefore P(Y=i|x,\mu,\sigma) = \left(1+\sum_{j\neq i}\frac{\alpha_j}{\alpha_i}\exp\left\{-\sum_{d=1}^{D}\frac{1}{2\sigma_d^2}\left((x_d-\mu_{jd})^2-(x_d-\mu_{id})^2\right)\right\}\right)^{-1}$

b) Assuming indep. of data:

$P(y^1,x^1,\ldots,y^N,x^N) = P(y^1,x^1)\cdots P(y^N,x^N)$

$= \prod_{i=1}^{N} P(x^i|y^i)P(y^i)$

$= \prod_{i=1}^{N}\left(\prod_{d=1}^{D} 2\pi\sigma_d^2\right)^{-1/2}\exp\left\{-\sum_{d=1}^{D}\frac{1}{2\sigma_d^2}(x_d^i-\mu_{y^i d})^2\right\}\alpha_{y^i}$

$\log P = N\sum_{d=1}^{D}\log\left((2\pi\sigma_d^2)^{-1/2}\right) + \sum_{i=1}^{N}\left(-\sum_{d=1}^{D}\frac{1}{2\sigma_d^2}(x_d^i-\mu_{y^i d})^2\right) + \sum_{i=1}^{N}\log(\alpha_{y^i})$

$\therefore \ell(\theta,D) = +N\sum_{d=1}^{D}\log(2\pi\sigma_d^2) + \sum_{i=1}^{N}\sum_{d=1}^{D}\frac{1}{2\sigma_d^2}(x_d^i-\mu_{y^i d})^2 - \sum_{i=1}^{N}\log(\alpha_{y^i})$

c) $\frac{d\ell(\theta;D)}{d\mu_{kd}} = 0$

$$\Leftrightarrow \sum_{i=1}^{N} \delta(y^i - k) \frac{1}{\sigma_d^2} (x_d^i - \mu_{kd})(-1) = 0$$

$$\sum_{i=1}^{N} x_d^i \delta(y^i - k) = \sum_{i=1}^{N} \delta(y^i - k) \mu_{kd}$$

Let there be $N_k < N$ examples of the k^{th} class

$$\sum_{i=1}^{N} x_d^i \delta(y^i - k) = \sum_{i=1}^{N_k} \mu_{kd} = N_k \mu_{kd}$$

$$\therefore \boxed{\mu_{k,d} = \frac{1}{N_k} \sum_{i=1}^{N} x_d^i \delta(y^i - k)} \text{ as expected.}$$

- For the mean of the d^{th} component for class K, average the d^{th} component for all examples of class K.

$$\frac{d\ell}{d\sigma_d} = 0$$

$$\frac{N}{2} \cdot \frac{1}{2\pi\sigma_d^2} 4\pi\sigma_d + \sum_{i=1}^{N} (x_d^i - \mu_{y^i d})^2 \cdot \left(\frac{-4\sigma_d}{4\sigma_d^4}\right) = 0$$

$$\frac{2N}{2\sigma_d} - \sum_{i=1}^{N} (x_d^i - \mu_{y^i d})^2 \frac{1}{\sigma_d^3} = 0$$

$$\frac{1}{\sigma_d} \left(2N - \frac{1}{\sigma_d^2} \sum_{i=1}^{N} (x_d^i - \mu_{y^i d})^2 \right) = 0$$

$$\boxed{\sigma_d^2 = \frac{1}{N} \sum_{i=1}^{N} (x_d^i - \mu_{y^i d})^2}$$

again, as expected.

d) $\frac{dl}{d\alpha_k} = 0$

$$-\sum_{i=1}^{N} \frac{1}{\alpha_k}\delta(y^i - k) = 0$$

→ In some ways this makes sense.

If $\alpha_k = \infty$, then prob is maximized.

But then probability laws are violated.

We require: $\sum_{j=1}^{K} \alpha_j = 1$ or $1 - \sum_{j=1}^{K} \alpha_j = 0 \equiv g$

Minimize l subject to $\sum_{j=1}^{K} \alpha_j = 1$.

Define $\mathcal{L} = l - \lambda g = 0$

Solve:

$\nabla \nabla \mathcal{L} = 0$

$\nabla g = 0$

$$\nabla_{\alpha_k} \mathcal{L} = -\sum_{i=1}^{N} \frac{1}{\alpha_k}\delta(y^i - k) + \lambda = 0$$

$$-\frac{N_k}{\alpha_k} + \lambda = 0 \quad \forall k$$

$$\alpha_k = \frac{N_k}{\lambda}$$

$$\sum_{j=1}^{K} \alpha_j = 1 = \sum_{j=1}^{K} \frac{N_j}{\lambda} = \frac{1}{\lambda}\sum_{j=1}^{K} N_j = \frac{N}{\lambda}$$

$$\lambda = N$$

$$\boxed{\alpha_j = \frac{N_k}{N}}$$ as required.