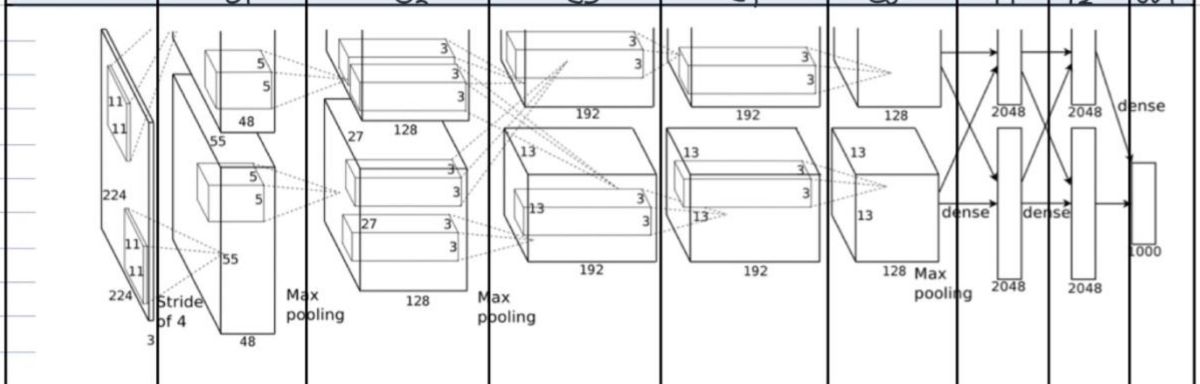
Name: Yuhan Shao
UTORid: shaoyuha
Student Number: 1002281327

CSC411 Assignment 4

1. (a)

	C1	C2	C3	C4	C5	F1	F2	OUT
# Units	$(55^2 \times 48) \times 2$ = 290,400	$(27^2 \times 128) \times 2$ = 186,624	$(13^2 \times 192) \times 2$ = 64,896	$(13^2 \times 192) \times 2$ = 64,896	$(13^2 \times 128) \times 2$ = 43,264	2048×2 = 4096	2048×2 = 4096	1000
# Weights	$(11^2 \times 3) \times (48 \times 2)$ = 34,848	$(5^2 \times 48) \times (128 \times 2)$ = 307,200	$(3^2 \times 128 \times 2) \times (192 \times 2)$ = 884,736	$(3^2 \times 192) \times (192 \times 2)$ = 663,552	$(3^2 \times 192) \times (128 \times 2)$ = 442,368	$(13^2 \times 192 \times 2) \times (2048 \times 2)$ = 177,209,344	$(2048 \times 2) \times (2048 \times 2)$ = 16,777,216	$(2048 \times 2) \times 1000$ = 4,096,000
# Connections	$(11^3 \times 3) \times (55^2 \times 48 \times 2)$ = 105,415,200	$(5^2 \times 48) \times (27^2 \times 128 \times 2)$ = 223,948,800	$(3^2 \times 128 \times 2) \times (13^2 \times 192 \times 2)$ = 149,520,384	$(3^2 \times 192) \times (13^2 \times 192 \times 2)$ = 112,140,288	$(3^2 \times 192) \times (13^2 \times 128 \times 2)$ = 74,760,192	$(13^2 \times 128 \times 2) \times (2048 \times 2)$ = 177,209,344	$(2048 \times 2) \times (2048 \times 2)$ = 16,777,216	$(2048 \times 2) \times 1000$ = 4,096,000



	# Units	# Weights	# Connections
Convolution Layer 1	290,400	34,848	105,415,200
Convolution Layer 2	186,624	307,200	223,948,800
Convolution Layer 3	64,896	884,736	149,520,384
Convolution Layer 4	64,896	663,552	112,140,288
Convolution Layer 5	43,264	442,368	74,760,192
Fully Connected Layer 1	4096	177,209,344	177,209,344
Fully Connected Layer 2	4096	16,777,216	16,777,216
Output Layer	1000	4,096,000	4,096,000

(b)

(i) To reduce the number of parameters for the network ⟶ reduce the number of weights

① Consider reducing the number of units generated by increasing the pooling size in the convolution layers.

② Consider increasing the stride to reduce the output units.

(ii) To reduce the number of connections:

① Reduce the kernel size in the convolution layers.

② Add "bottle-neck" layer between fully-connected layers.

2.(a) $p(y=k \mid \vec{x}, \vec{\mu}, \vec{\sigma}) = \dfrac{P(\vec{x} \mid y=k, \vec{\mu}, \vec{\sigma})\, P(y=k \mid \vec{\mu}, \vec{\sigma})}{P(\vec{x} \mid \vec{\mu}, \vec{\sigma})}$ # By Bayes' Rule: $P(A|B) = \dfrac{P(B|A)\,P(A)}{P(B)}$

$$= \frac{P(\vec{x} \mid y=k, \vec{\mu}, \vec{\sigma})\, P(y=k \mid \vec{\mu}, \vec{\sigma})}{\sum_{j=1}^{k} P(\vec{x} \mid y=j, \vec{\mu}, \vec{\sigma})\, P(y=j \mid \vec{\mu}, \vec{\sigma})}$$

$$= \frac{\left(\prod_{d=1}^{D} 2\pi\sigma_d^2\right)^{-1/2} \exp\left\{-\sum_{d=1}^{D} \frac{1}{2\sigma_d^2}(x_d - \mu_{kd})^2\right\} \cdot \alpha_k}{\sum_{j=1}^{k}\left[\left(\prod_{d=1}^{D} 2\pi\sigma_d^2\right)^{-1/2} \exp\left\{-\sum_{d=1}^{D} \frac{1}{2\sigma_d^2}(x_d - \mu_{jd})^2\right\} \cdot \alpha_j\right]}$$ # By (1), (2).

(b) $\ell(\vec{\theta}; D) = -\log p(y^{(1)}, \vec{x}^{(1)}, y^{(2)}, \vec{x}^{(2)}, \dots, y^{(N)}, \vec{x}^{(N)} \mid \vec{\theta})$

$= -\log\left[p(y^{(1)}, \vec{x}^{(1)} \mid \vec{\theta}) \cdot p(y^{(2)}, \vec{x}^{(2)} \mid \vec{\theta}) \cdot \dots \cdot p(y^{(N)}, \vec{x}^{(N)} \mid \theta)\right]$ # By data are i.i.d.

$= -\left[\log p(y^{(1)}, \vec{x}^{(1)} \mid \vec{\theta}) + \log p(y^{(2)}, \vec{x}^{(2)} \mid \vec{\theta}) + \dots + \log p(y^{(N)}, \vec{x}^{(N)} \mid \theta)\right]$

$= -\sum_{i=1}^{N} \log p(y^{(i)}, \vec{x}^{(i)} \mid \vec{\theta})$

$= -\sum_{i=1}^{N}\left[\log\left(p(\vec{x}^{(i)} \mid y^{(i)}, \theta)\, p(y^{(i)} \mid \vec{\theta})\right)\right]$ # By Chain Rule.

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

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

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

(c) $\ell(\vec{\theta}; D) = \sum_{i=1}^{N}\left[\frac{1}{2}\sum_{d=1}^{D}\log(2\pi\sigma_d^2) + \sum_{d=1}^{D} \frac{1}{2\sigma_d^2}(x_d^{(i)} - \mu_{y^{(i)}d})^2 - \log \alpha_{y^{(i)}}\right]$

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

- $\dfrac{\partial \ell(\vec{\theta}; D)}{\partial \mu_{y^{(i)}d}} = 0 + \sum_{i=1}^{N}\sum_{d=1}^{D} \frac{1}{\sigma_d^2}(x_d^{(i)} - \mu_{y^{(i)}d}) \cdot (-1) - 0$

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

$\Rightarrow \dfrac{\partial \ell(\vec{\theta}; D)}{\partial \mu_{kd}} = \sum_{i=1}^{N} \frac{1}{\sigma_d^2}(\mu_{kd} - x_d^{(i)})\, \mathbb{I}[y^{(i)} = k]$.

- $\dfrac{\partial \ell(\vec{\theta}; D)}{\partial \sigma_d^2} = \frac{1}{2}\sum_{i=1}^{N} \frac{1}{\sigma_d^2} + \sum_{i=1}^{N} \frac{1}{-2\sigma_d^4}(x_d^{(i)} - \mu_{y^{(i)}d})^2 - 0$

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

- Set $\dfrac{\partial \ell(\vec{\theta}; D)}{\partial \mu_{kd}} = 0 \Rightarrow \sum_{i=1}^{N} \frac{1}{\sigma_d^2}(\mu_{kd} - x_d^{(i)})\, \mathbb{I}[y^{(i)} = k] = 0$

$\Rightarrow \sum_{i=1}^{N} \mu_{kd}\, \mathbb{I}[y^{(i)} = k] = \sum_{i=1}^{N} x_d^{(i)}\, \mathbb{I}[y^{(i)} = k]$

$\Rightarrow \mu_{kd} = \dfrac{\sum_{i=1}^{N} x_d^{(i)}\, \mathbb{I}[y^{(i)} = k]}{\sum_{i=1}^{N} \mathbb{I}[y^{(i)} = k]}$ # Let $N_k = \sum_{i=1}^{N} \mathbb{I}[y^{(i)} = k]$

$\Rightarrow \mu_{kd} = \dfrac{\sum_{i=1}^{N} x_d^{(i)}\, \mathbb{I}[y^{(i)} = k]}{N_k}$ # maximum likelihood estimate for μ_{kd} where $1 \le d \le D$

- Set $\dfrac{\partial \ell(\vec{\theta}; D)}{\partial \sigma_d^2} = 0 \Rightarrow \frac{N}{2} \cdot \frac{1}{\sigma_d^2} - \sum_{i=1}^{N} \frac{1}{2\sigma_d^4}(x_d^{(i)} - \mu_{y^{(i)}d})^2 = 0$

$\Rightarrow \frac{N}{2} \cdot \frac{1}{\sigma_d^2} = \sum_{i=1}^{N} \frac{1}{2\sigma_d^4}(x_d^{(i)} - \mu_{y^{(i)}d})^2$

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

$\Rightarrow \sigma_d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_d^{(i)} - \mu_{y^{(i)}d})^2}$ # maximum likelihood estimate for σ_d where $1 \le d \le D$

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

$\Rightarrow \dfrac{\partial \ell(\vec{\theta}; D)}{\partial \alpha_k} = 0 + 0 - \sum_{i=1}^{N} \frac{1}{\alpha_k}\, \mathbb{I}[y^{(i)} = k]$.

$= -\frac{1}{\alpha_k}\sum_{i=1}^{N} \mathbb{I}[y^{(i)} = k] = -\frac{N_k}{\alpha_k}$ # Let $N_k = \sum_{i=1}^{N} \mathbb{I}[y^{(i)} = k]$

- Let $g(\alpha_1, \alpha_2, \dots, \alpha_k) = \sum_{i=1}^{k} \alpha_i = 1 \Rightarrow \dfrac{\partial g}{\partial \alpha_i} = 1$ (*)

→ By Langrange Multipliers: ($f' = \lambda g'$ with constraint $g(\alpha_1, \dots, \alpha_k) = \sum_{i=1}^{k} \alpha_i = 1$)

$\left(-\frac{N_1}{\alpha_1}, -\frac{N_2}{\alpha_2}, \dots, -\frac{N_k}{\alpha_k}\right) = \lambda(1, 1, \dots, 1)$ # By (*)

$\Rightarrow \alpha_i = -\frac{N_i}{\lambda}$ where $1 \le i \le k$

→ Use constraint $\sum_{i=1}^{k} \alpha_i = 1$, we have:

$-\frac{1}{\lambda}\sum_{i=1}^{k} N_i = 1 \Rightarrow -\frac{N}{\lambda} = 1 \Rightarrow \lambda = -N$

- Therefore, $\alpha_i = -\frac{N_i}{\lambda} = -\frac{N_i}{-N} = \frac{N_i}{N}$

$\Rightarrow \alpha_k = \frac{N_k}{N} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}[y^{(n)} = k]$