Q1a)

	# unit	# weights	# connections
Convolution Layer 1	290400	34,848	105,415,200
Convolution Layer 2	186,624	307,200	111,974,400
Convolution Layer 3	64,896	884,736	149,520,384
Convolution Layer 4	64,869	663,552	112,140,288
Convolution Layer 5	43,264	442,368	74,760,192
Fully Connected Layer 1	4096	37,748,736	37,748,736
Fully Connected Layer 2	4096	16,777,216	16,777,216
Output Layer	1000	4,096,000	4,096,000

b) 1. Majority of the parameters are from fully connected layers. We can reduce the number of parameter by reducing the size of fully connected layers and the last convolutional layer.

2. Convolutional layers have a lot of connections. We can reduce the number of connections by reducing the size of convolutional layers and reducing the number of kernels.

a)

$$P(y=k) = \alpha_k$$

$$P(x|y=k, u, \sigma) = \left(\prod_{i=1}^{D} 2\pi\sigma^2\right)^{-\frac{1}{2}} \times e^{\left(-\sum_{i=1}^{D} \frac{1}{\sigma^2}(x-u_{ki})^2\right)}$$

Law of Total Probability: $P(x|u,\sigma) = \sum_{i=1}^{k} \left(p(x|y=i, u, \sigma)\, p(y=i|u,\sigma)\right)$

Baye's Rule: $P(y=k|x,u,\sigma) = \dfrac{P(x|y=k,u,\sigma)\, P(y=k,u,\sigma)}{P(x|u,\sigma)}$

$$P(y=k|x,u,\sigma) = \frac{\left(\prod_{i=1}^{D} 2\pi\sigma^2\right)^{-\frac{1}{2}} e^{\left(-\sum_{i=1}^{D}\frac{1}{\sigma^2}(x_i-u_{ki})^2\right)}}{P(x|u,\sigma)}$$

$$= \frac{\left(\left(\prod_{i=1}^{D} 2\pi\sigma^2\right)^{-\frac{1}{2}} e^{\left(-\sum_{i=1}^{D}\sigma^2(x_i-u_{ki})^2\right)} \times \alpha_k\right)}{\sum_{j=1}^{k}\left(P(x|y=j,u,\sigma)\, P(y=j|u,\sigma)\right)}$$

$$= \frac{\left(\left(\prod_{i=1}^{D} 2\pi\sigma^2\right)^{\frac{1}{2}} e^{\left(-\sum_{i=1}^{D}\sigma^2(x_i-u_{ki})^2\right)} \times \alpha_k\right)}{\sum_{n}^{k}\left(\left(\prod_{i=1}^{D} 2\pi\sigma_i^2\right)^{-\frac{1}{2}} e^{\left(-\sum_{j}^{D}\frac{1}{2\sigma_j^2}(x_j-u_{nj})^2\right)} \alpha_n\right)}$$

b)

$$P(x^i, y^i|\theta) = p(x^i|y^i\theta)\, p(y).$$

$$= ((2\pi \prod_{j=1}^{D} \sigma_j^2)^{-\frac{1}{2}} \exp(-\sum (2\sigma_j^2)^{-1} (x_j^i - \mu_{kj})^2))(\alpha_{k^i}).$$

$$-\log p(y^{(1)}, x^{(1)}, \ldots, y^{(N)}, x^{(N)}|\theta) = \sum_{i=1}^{N} \left(\log p(x^i|y^i, \theta) + \log p(y^i|\theta)\right).$$

$$= -\log[(\prod_{j=1}^{D} 2\pi\sigma^2)^{-\frac{N}{2}}] + \sum_{i=1}^{N}(\log \alpha_{k_i} - \sum_{n=1}^{N}\sum_{m=1}^{D}(2\sigma_m^2)^{-1}(x_m - \mu_{k_m}^n)^2)$$

$$= \frac{N}{2}\sum_{j=1}^{D}\log 2\pi\sigma^2 - \sum_{i=1}^{N}\log \alpha_{k_i} + \sum_{n=1}^{N}\sum_{m=1}^{D}(2\sigma_m^2)^{-1}(x_j - \mu_{k_m}^n)^2)$$

c). $\frac{\partial(\log l(\theta))}{\partial u_{ki}} = \frac{\partial \sum_{m=1}^{N}\left(\frac{1}{2}\log\left(\prod_{i=1}^{D} 2\pi\theta_i^2\right) + \sum_{i=1}^{D}\frac{1}{2\theta_i}(x_i^m - u_{y(m)i})^2 - \log\alpha_{y_i}\right)}{\partial u_{ki}}$

$$= -\sum_{i=1}^{N} \mathbb{1}[y=k](x_{ij} - u_{kj})\frac{1}{\sigma^2}$$

$$\frac{\partial(-\log l(\theta))}{\partial \sigma_j^2} = \frac{\partial \sum_{m=1}^{N}\left(\frac{1}{2}\log\left(\prod_{i=1}^{D} 2\pi\theta_i^2\right) + \sum \frac{1}{2\theta_i^2}(x_i^m - u_{y(m)i})^2 - \log\alpha_{y_i}\right)}{\partial \sigma_j^2}$$

$$= \frac{1}{2}\sum_{m}^{N}\sum_{n=1}^{D} \mathbb{1}(n=i)\left[(\sigma_n^2)^{-1} - (\sigma_n^2)^{-2}(x_n^m - u_{y^m n})^2\right]$$

$$= \frac{N}{2\sigma_j^2} - \sum_{i=1}^{N}(x_{ij} - u_{kj})^2\frac{1}{2\sigma_j^4}$$

~~ta~~ let $\frac{\partial(-\log \ell(\theta))}{\partial \mu_{ki}} = 0.$

$$-\sum \mathbb{1}(y^m=k) \sum_{n=1}^{D} (n=i) \frac{1}{\sigma_n^2} (\mu_{y^m n} - x_i^m) = 0.$$

$$-\sum_{m=1}^{N} \mathbb{1}(y^m=k) \sum_{n=1}^{D} \mathbb{1}(n=i) \frac{1}{\sigma_n^2} \times \mu_{y^m_n} = -\sum \mathbb{1}(y^m=k) \sum \mathbb{1}(n=i) \frac{x_i^m}{\sigma_n^2}$$

$$\mu_{ki} = \frac{\sum \mathbb{1}(y^m=k) \sum \mathbb{1}(n=i)\, x_n^m}{\sum \mathbb{1}(y^m=k) \sum \mathbb{1}(n=i)}$$

$$= \frac{1}{N} \sum_{m=1}^{N} \mathbb{1}(y^m=k)\, x_n^m$$

let $\frac{\partial(-\log \ell(\theta))}{\partial(\sigma_i^2)} = 0.$

$$\frac{1}{2} \sum \sum \mathbb{1}(n=i) \left[(\sigma_n^2)^{-1} - (\sigma_n^2)^{-2} (x_n^m - \mu_{y^m n})^2 \right] = 0$$

$$\sum_{m=1}^{N} \sum_{n=1}^{D} \mathbb{1}(n=i) [(x_n^m - \mu_n^m)^2] (\sigma_n^2)^{-2} = \sum_{m=1}^{N} \sum_{n=1}^{D} \mathbb{1}(n=i) (\sigma_n^2)^{-1}$$

$$\sigma_i^2 = \frac{\sum_{m=1}^{N} \sum_{n=1}^{N} \mathbb{1}(n=i) [x_n^m - \mu_{y^m_n}]^2}{\sum_{m=1}^{N} \sum_{n=1}^{D} \mathbb{1}(n=i)}$$

$$\hat{\sigma}_i^2 = \frac{1}{N} \sum_{m=1}^{N} (x_i^m - \mu_{y^m i})^2$$

$$\sigma_i = \sqrt{\frac{1}{N} \sum (x_i^m - \mu_{y^m i})^2}.$$

d) $\arg\max(L(\theta, D)) = \underset{\alpha}{\arg\max}\left(\sum_{j=1}^{N}(-\log \alpha_{y^j})\right) = \underset{\alpha}{\arg\min}\left(\sum_{j=1}^{N}\log \alpha_{y^j}\right)$

$= \underset{\alpha}{\arg\min}\left(\sum_{j=1}^{N}\sum_{m=1}^{k} I(y^j = m)\log \alpha_m\right)$

let $f(\alpha_1, \dots, \alpha_k) = \sum_{j=1}^{N}\sum_{m=1}^{k} I(y^j = m)\log \alpha_m$

by lagrange thm,

$\exists \lambda \in \mathbb{R} \quad \forall k = 1, 2, \dots k.$

$$\frac{\partial f}{\partial \alpha_k} = \lambda \frac{\partial g}{\partial \alpha_k}.$$

$$\frac{\partial f}{\partial \alpha_k} = \sum_{j=1}^{N} \frac{\partial}{\partial \alpha_k}\left(I(y^j = k)\log(\alpha_k)\right) = \sum_{j=1}^{N} \frac{1}{\alpha_k} I(y^k = k)$$

$$= \frac{1}{\alpha_k}\sum_{j=1}^{N} I(y^j = k)$$

$$\frac{\partial g}{\partial \alpha_k} = 1$$

$$\frac{1}{\alpha_k}\sum_{j=1}^{N} I(y^j = k) = \lambda.$$

$$\sum_{j=1}^{N} I(y^j = k) = \lambda \alpha_k.$$

$$\sum_{k=1}^{k}\sum_{j=1}^{N} I(y^j = k) = \lambda \sum_{k=1}^{k} \alpha_k$$

$$\sum_{j=1}^{N}\sum_{k=1}^{k} I(y^j = k) = \lambda$$

$$\sum_{j=1}^{N} 1 = \lambda$$

$\therefore \lambda = N$

then $\frac{1}{\alpha_k}\sum^{N} I(y^j = k) = \lambda = N.$

$$\alpha_{k_{ML}} = \frac{1}{N}\sum_{j=1}^{N} I(y^j = k) = \frac{1}{N}\sum_{j=1}^{N} 1(y^k = k)$$