June 5, 2017

Notation:

- $a \in \{\text{cloudy}, \text{clear}, \text{partial}, \text{haze}\}$ represents the possible atmospheric conditions.
- Vector $\boldsymbol{l} \in \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ 0 \\ \vdots \end{pmatrix}, \cdots \right\}$ represents whether the other labels are present, such that $l_i \in \{0, 1\}$ represents whether label i is present.

We are interested in obtaining the joint posterior distribution, $P\left(a, \boldsymbol{l} | \boldsymbol{o}_t\right)$. With this posterior, we can do training using either CE,

$$\mathcal{F}_{\text{CE}} = -\sum_t \log P\left(a_t^*, \boldsymbol{l}_t^* | \boldsymbol{o}_t\right),$$

or maximum expected F-score,

$$\mathcal{F}_{\text{F}} = -\sum_t \sum_{a,\boldsymbol{l}} F\left(a, \boldsymbol{l}, a_t^*, \boldsymbol{l}_t^*\right) P\left(a, \boldsymbol{l} | \boldsymbol{o}_t\right).$$

With this posterior, we can also do decoding either MAP,

$$a^*, \boldsymbol{l}^* = \arg\max_{a,\boldsymbol{l}} P\left(a, \boldsymbol{l} | \boldsymbol{o}_t\right),$$

or maximum expected F-score,

$$a^*, \boldsymbol{l}^* = \arg\max_{a,\boldsymbol{l}} \sum_{a',\boldsymbol{l}'} F\left(a', \boldsymbol{l}', a, \boldsymbol{l}\right) P\left(a', \boldsymbol{l}' | \boldsymbol{o}_t\right).$$

1 Obtaining the joint posterior distribution

There are several possible ways that we can choose to formulate the posterior:

1. Sigmoid for $c \Rightarrow$ cloudy, softmax for $b \Rightarrow$ {clear, partial, haze}, and separate sigmoids for $\boldsymbol{l} \Rightarrow$ {other labels}. b and $\boldsymbol{l}$ are conditionally independent given c.

$$P\left(c, b, \boldsymbol{l} | \boldsymbol{o}_t\right) = P\left(c | \boldsymbol{o}_t\right) P\left(b | c, \boldsymbol{o}_t\right) P\left(\boldsymbol{l} | c, \boldsymbol{o}_t\right)$$

2. Softmax for $a \Rightarrow$ {cloudy, clear, partial, haze} and separate sigmoids for $\boldsymbol{l} \Rightarrow$ {other labels}.

$$P\left(a, \boldsymbol{l} | \boldsymbol{o}_t\right) = P\left(a | \boldsymbol{o}_t\right) P\left(\boldsymbol{l} | a, \boldsymbol{o}_t\right)$$

3. Other labels are conditionally independent of the atmospheric condition.

$$P\left(a, \boldsymbol{l} | \boldsymbol{o}_t\right) = P\left(a | \boldsymbol{o}_t\right) P\left(\boldsymbol{l} | \boldsymbol{o}_t\right)$$

1.1 Group outputs into cloudy, {clear, partial, haze}, {other labels}

We need to formulate the joint label probability, $P\left(c, b, \boldsymbol{l} | \boldsymbol{o}_t\right)$. We can factorise the distribution into

$$P\left(c, b, \boldsymbol{l} | \boldsymbol{o}_t\right) = P\left(c | \boldsymbol{o}_t\right) P\left(b | c, \boldsymbol{o}_t\right) P\left(\boldsymbol{l} | c, \boldsymbol{o}_t\right).$$

Here, we have assumed that b and $\boldsymbol{l}$ are conditionally independent.

We can then model $P\left(c | \boldsymbol{o}_t\right)$ as a sigmoid output, $f\left(\boldsymbol{o}_t\right)$,

$$P\left(c | \boldsymbol{o}_t\right) = f^c\left(\boldsymbol{o}_t\right)\left[1 - f\left(\boldsymbol{o}_t\right)\right]^{1-c}.$$

We can model $P\left(a | c = 1, \boldsymbol{o}_t\right)$ as a softmax output, $\boldsymbol{g}\left(\boldsymbol{o}_t\right)$, then we have

$$P\left(b | c, \boldsymbol{o}_t\right) = \left[\delta\left(b, 0\right)\right]^c g_b^{1-c}\left(\boldsymbol{o}_t\right).$$

It can be seen that there is a problem, because 0 is not in the set of values that b can take.

We assume that all other labels are independent of each other,

$$P\left(\boldsymbol{l} | c, \boldsymbol{o}_t\right) = \prod_i P\left(l_i | c, \boldsymbol{o}_t\right).$$

We can model $P\left(l_i | c = 1, \boldsymbol{o}_t\right)$ as a sigmoid, $h_i\left(\boldsymbol{o}_t\right)$, then we have

$$P\left(l_i | c, \boldsymbol{o}_t\right) = \left[\delta\left(l_i, 0\right)\right]^c \left[h_i^{l_i}\left(\boldsymbol{o}_t\right)\left\{1 - h_i\left(\boldsymbol{o}_t\right)\right\}^{1-l_i}\right]^{1-c}.$$

The joint distribution is

$$P\left(c, b, \boldsymbol{l} | \boldsymbol{o}_t\right) = f^c\left(\boldsymbol{o}_t\right)\left[1 - f\left(\boldsymbol{o}_t\right)\right]^{1-c}\left[\delta\left(b, 0\right)\right]^c g_b^{1-c}\left(\boldsymbol{o}_t\right)\prod_i \left[\delta\left(l_i, 0\right)\right]^c \left[h_i^{l_i}\left(\boldsymbol{o}_t\right)\left\{1 - h_i\left(\boldsymbol{o}_t\right)\right\}^{1-l_i}\right]^{1-c}.$$

The CE criterion for frame t is

$$\begin{aligned}
\mathcal{F}_{\text{CE}} &= -\log P\left(c_t^*, b_t^*, \boldsymbol{l}_t^* | \boldsymbol{o}_t\right) \\
&= -c_t^* \log f\left(\boldsymbol{o}_t\right) - \left(1 - c_t^*\right)\log\left[1 - f\left(\boldsymbol{o}_t\right)\right] - c_t^* \log\delta\left(b_t^*, 0\right) - \left(1 - c_t^*\right) g_{b_t^*}\left(\boldsymbol{o}_t\right) \\
&\quad - \sum_i \left\{c_t^* \log\delta\left(l_{it}^*, 0\right) + \left(1 - c_t^*\right)\left[l_{it}^* \log h_i\left(\boldsymbol{o}_t\right) + \left(1 - l_{it}^*\right)\log\left(1 - h_i\left(\boldsymbol{o}_t\right)\right)\right]\right\}
\end{aligned}$$

There is a problem, because b cannot take the value 0. So, $\log\delta\left(b, 0\right)$ is always infinite. However, the gradient is still finite, as $\log\delta\left(b, 0\right)$ is not involved in the gradient. When we discard all the terms that do not depend on the model parameters, the criterion reduces to

$$\mathcal{F}_{\text{CE}} \Rightarrow -c_t^* \log f\left(\boldsymbol{o}_t\right) - \left(1 - c_t^*\right)\log\left[1 - f\left(\boldsymbol{o}_t\right)\right] - \left(1 - c_t^*\right)\sum_i \left[l_{it}^* \log h_i\left(\boldsymbol{o}_t\right) + \left(1 - l_{it}^*\right)\log\left[1 - h_i\left(\boldsymbol{o}_t\right)\right]\right].$$

1.2 Group outputs into {atmophere}, {other labels}

To avoid the infinite CE cost, we can instead group together all four atmospheric conditions into a single softmax output. The joint posterior can then be factorised as

$$P\left(a, \boldsymbol{l} | \boldsymbol{o}_t\right) = P\left(a | \boldsymbol{o}_t\right) P\left(\boldsymbol{l} | a, \boldsymbol{o}_t\right).$$

We can model $P\left(a | \boldsymbol{o}_t\right)$ as a softmax, $\boldsymbol{g}\left(\boldsymbol{o}_t\right)$,

$$P\left(a | \boldsymbol{o}_t\right) = g_a\left(\boldsymbol{o}_t\right).$$

We can again model $P\left(\boldsymbol{l} | a, \boldsymbol{o}_t\right) = \prod_i P\left(l_i | a, \boldsymbol{o}_t\right)$ as separate sigmoids, $h_i\left(\boldsymbol{o}_t\right)$, then we have

$$P\left(l_i | a, \boldsymbol{o}_t\right) = \left[\delta\left(l_i, 0\right)\right]^{\delta(a,\text{cloudy})} \left[h_i^{l_i}\left(\boldsymbol{o}_t\right)\left\{1 - h_i\left(\boldsymbol{o}_t\right)\right\}^{1-l_i}\right]^{1-\delta(a,\text{cloudy})}.$$

The joint distribution is

$$P\left(a, \boldsymbol{l} | \boldsymbol{o}_t\right) = g_a\left(\boldsymbol{o}_t\right) \prod_i \left[\delta\left(l_i, 0\right)\right]^{\delta(a,\text{cloudy})} \left[h_i^{l_i}\left(\boldsymbol{o}_t\right)\left\{1 - h_i\left(\boldsymbol{o}_t\right)\right\}^{1-l_i}\right]^{1-\delta(a,\text{cloudy})}.$$

The CE criterion for frame t is

$$\mathcal{F}_{\text{CE}} = -\log g_{a_t^*}\left(\boldsymbol{o}_t\right) - \sum_i \left\{\delta\left(a_t^*, \text{cloudy}\right) \log \delta\left(l_{it}^*, 0\right) + \left[1 - \delta\left(a_t^*, \text{cloudy}\right)\right]\left[l_{it}^* \log h_i\left(\boldsymbol{o}_t\right) + \left(1 - l_{it}^*\right) \log\left(1 - h_i\left(\boldsymbol{o}_t\right)\right)\right]\right\}$$

By discarding all the terms that do not depend on the model parameters, the criterion reduces to

$$\mathcal{F}_{\text{CE}} \Rightarrow -\log g_{a_t^*}\left(\boldsymbol{o}_t\right) - \left[1 - \delta\left(a_t^*, \text{cloudy}\right)\right] \sum_i \left[l_{it}^* \log h_i\left(\boldsymbol{o}_t\right) + \left(1 - l_{it}^*\right) \log\left(1 - h_i\left(\boldsymbol{o}_t\right)\right)\right].$$

1.3 Conditionally independent a and $\boldsymbol{l}$

Although the task instructions mentions that images labelled with cloudy should not have any other labels, labelling errors do exists in the training set. It may be possible that labelling errors also exist in the test set. If this is the case it may be better not to force $P\left(\boldsymbol{l}=\mathbf{0}|a=\text{cloudy},\boldsymbol{o}_t\right)=1$. We can remove this constraint by assuming that a and $\boldsymbol{l}$ are conditionally independent, so that the joint posterior factorises into

$$P\left(a,\boldsymbol{l}|\boldsymbol{o}_t\right)=P\left(a|\boldsymbol{o}_t\right)P\left(\boldsymbol{l}|\boldsymbol{o}-t\right).$$

We can again model $P\left(a|\boldsymbol{o}_t\right)$ as a softmax,

$$P\left(a|\boldsymbol{o}_t\right)=g_a\left(\boldsymbol{o}_t\right).$$

We can model $P\left(\boldsymbol{l}|\boldsymbol{o}-t\right)=\prod\limits_i P\left(l_i|\boldsymbol{o}-t\right)$ as separate sigmoids, regardless of the atmospheric condition,

$$P\left(l_i|\boldsymbol{o}_t\right)=h_i^{l_i}\left(\boldsymbol{o}_t\right)\left\{1-h_i\left(\boldsymbol{o}_t\right)\right\}^{1-l_i}.$$

The joint posterior is then

$$P\left(a,\boldsymbol{l}|\boldsymbol{o}_t\right)=g_a\left(\boldsymbol{o}_t\right)\prod_i h_i^{l_i}\left(\boldsymbol{o}_t\right)\left\{1-h_i\left(\boldsymbol{o}_t\right)\right\}^{1-l_i}.$$

The CE criterion for frame t is

$$\mathcal{F}_{\text{CE}}=-\log g_{a_t^*}\left(\boldsymbol{o}_t\right)-\sum_i\left\{l_{it}^*\log h_i\left(\boldsymbol{o}_t\right)+\left(1-l_{it}^*\right)\log\left[1-h_i\left(\boldsymbol{o}_t\right)\right]\right\}.$$