Bonus 2 Soft Targets v/s Label Flipping

Murali Krishnan Rajasekharan Pillai

October 23, 2019

Let's assume that a model with parameters θ defines a function $f : \mathcal{X} \to y$, the maps from the input vectors $\boldsymbol{x}_i \forall\, i \in \{1, \cdots, N\}$ to the probability of the input vector being in class $\{j \;\forall\; j \in \{1, \cdots, K\}\}$.

Lets consider one data point which is of class 1, for our analysis. The softmax units in a neural network estimates the probability of input vector $\boldsymbol{x}$ in all the possible classes:

$$q_\theta(y|x_i) = \left(\frac{\exp(z_{y_1})}{\sum_{j=1}^{K} \exp(z_{y_j})}, \cdots, \frac{\exp(z_{y_K})}{\sum_{j=1}^{K} \exp(z_{y_j})} \right),$$

In a sense, the labeled dataset defines the true distribution of labels as their normalized values $p(y|x_i)$, and we want $q_\theta(.)$ to approximate the true probability. The true probability distribution is as follows:

$$p(y_j|x) = \begin{cases} 1, & \text{if } j = 1 \\ 0, & \text{otherwise} \end{cases}$$

It is important to note here that $\sum_{j=1}^{K} q_\theta(y_j|x_i) = 1$ and $\sum_{j=1}^{K} p(y_j|x_i) = 1$ Now, we can define the cross-entropy loss for $\boldsymbol{x}$ as:

$$\begin{aligned}
\mathcal{L} &= \sum_{j=1}^{K} H_j(p, q_\theta) \\
&= -\sum_{j=1}^{K} p(y_j|x) \log(q_\theta(y_j|x)) \\
&= -(1) \cdot \log(q_\theta(y_1|x)) && (\because \text{all other contributions vanish}) \\
&= -\left(z_{y_1} - \log\left(\sum_{j=1}^{K} \exp(z_{y_j}) \right) \right) && (\because \text{assuming } i \text{ is the correct label}) \\
&= -\Big(z_{y_1} - \log\big(\exp(z_{y_1}) + K - 1\big) \Big) \\
&\approx -\Big(z_{y_1} - \|z_y\|_\infty \Big)
\end{aligned}$$

Now let us see how this cost function varies when we do label smoothing.

Label Smoothing

Label smoothing refers to regularizing a model by flipping output labels with a certain probability (say ϵ). It encodes the prior belief that the provided labels in the training set may not be completely trust-worthy and that they are correct with the probability of $1 - \epsilon$ This can be achieved by two ways,

- **Soft Labels:** For softmax output units with k output units, it can be incorporated in the cost function by replacing the 0 or 1 value in the on-hot-encoded vectors with $\frac{\epsilon}{k-1}$ and $1 - \epsilon$, respectively.

Carrying forward the example from above, If the correct label of $\boldsymbol{x}$ is 1:

$$q_\theta(y|x) = \left(\frac{\exp((1-\epsilon)z_1)}{\sum_{j=1}^{K} \exp(z_{y_j})}, \cdots, \frac{\exp(\frac{\epsilon}{K-1} z_{y_K})}{\sum_{j=1}^{K} \exp(z_{y_j})} \right)$$

Similarly, we define soft labels as:

$$p(y_j|x) = \begin{cases} 1-\epsilon, & \text{if } j = 1 \\ \frac{\epsilon}{K-1}, & \text{otherwise} \end{cases}$$

Now, we can define the cross-entropy loss as:

$$\begin{aligned}
\mathcal{L} &= \sum_{l=1}^{N} H_i(p, q_\theta) \\
&= -\sum_{j=1}^{K} p(y_j|x) \log(q_\theta(y_j|x)) \\
&= -\left[(1-\epsilon) \cdot \log(q_\theta(y_1|x)) + \cdots + \frac{\epsilon}{K-1} \log(q_\theta(y_K|x)) \right] && (\because \text{assuming } i \text{ is the correct label}) \\
&\approx -\left[(1-\epsilon) \cdot \left((1-\epsilon) z_{y_1} - M \right) + \cdots + \frac{\epsilon}{K-1} \left(\frac{\epsilon}{K-1} \cdot z_{y_K} - M \right) \right] && (\because \text{assuming } \|z_y\|_\infty = M) \\
&\approx -\left[\left[(1-\epsilon)^2 z_{y_1} + \left(\frac{\epsilon}{K-1} \right)^2 \sum_{j \neq i} z_{y_j} \right] - M \right]
\end{aligned}$$

Now we can see that the final loss function has a contribution from even incorrect classification.

- **Label Flipping:** We can also achieve the former by sampling from a distribution such that

$$\Pr(y = 1|x) = 1 - \epsilon$$

This means that

$$\Pr(y \neq 1|x) = \epsilon$$

Now, since there are $K-1$ possibilities of where $y \neq 1$ (as there are K classes),

$$\Pr(y = \text{incorrect class}|x) = \frac{\epsilon}{K-1}$$

What we observe here is that as defined above, by flipping the labels with $1-\epsilon$ probability, we obtain the same $q_\theta(y|x)$ and same $p(y_j|x)$ as in label smoothing. Hence this would lead to similar loss functions. This shows that label flipping is equivalent to replacing hard labels with soft labels.