Deep Model Reference: Simple but Effective Confidence Estimation for Image Classification - Supplementary Material

Yuanhang Zheng[1,5], Yiqiao Qiu[3], Haoxuan Che[4], Hao Chen[4], Wei-Shi Zheng[1,5], and Ruixuan Wang[1,2,5⋆]

[1] School of Computer Science and Engineering, Sun Yat-sen Univerisity, Guangzhou, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Department of Computer Science and Engineering, University of California, San Diego, United States
[4] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[5] Key Laboratory of Machine Intelligence and Advanced Computing, MOE, Guangzhou, China

A Detailed proof of proposition 1

Given M models and K categories, denote $\hat{y}(\mathbf{x})$ as the predicted category and $\mathbf{p}_m = [p_{m,1}\ p_{m,2}\ \dots\ p_{m,K}]^\intercal \in \mathbb{R}^K$ as the output probability vector from the m-th model. The estimated confidence of DE and DMR are defined as

$$S_e(\mathbf{x}) = \max_{k\in\{1,2,\dots,K\}} \frac{1}{M}\sum_{m=1}^{M} p_{m,k}, \tag{1}$$

$$S_r(\mathbf{x}) = \frac{1}{M}\sum_{m=1}^{M} p_{m,k^*}\ , \tag{2}$$

respectively, where $k^* = \arg\max_{k\in\{1,2,\dots,K\}} p_{i,k}$ is the predicted category of the main model(i-th classifier). Define d as the difference between estimated and expected confidence. Formally,

$$d(\mathbf{x}) = \begin{cases} S(\mathbf{x}), & \text{if } \mathbf{x} \text{ is misclassified} \\ 1 - S(\mathbf{x}), & \text{if } \mathbf{x} \text{ is not misclassified} \end{cases} \tag{3}$$

Proof. The whole test dataset D can be divided into two cases as below,

1. When $\hat{y}(\mathbf{x})_r = \hat{y}(\mathbf{x})_e$, obviously, $S_r = S_e$ is established. So $d_r = d_e$, that is, $\mathbb{E}_1(d_r) = \mathbb{E}_1(d_e)$ always holds.

⋆ Corresponding author: wangruix5@mail.sysu.edu.cn

2. When $\hat{y}(\mathbf{x})_r \neq \hat{y}(\mathbf{x})_e$, based on Equation 1 and Equation 2, below inequality always holds.
$$S_e > S_r \tag{4}$$
Otherwise, $\hat{y}(\mathbf{x})_r = \hat{y}(\mathbf{x})_e$ holds, which is contradictory. And below inequality always holds because $\sum_k^K p_{m,k} = 1$.
$$S_e + S_r \leq 1 \tag{5}$$
Now let's divide it into 3 subcases. Let $y(\mathbf{x})$ be the ground truth category.

   (a) When both $\hat{y}(\mathbf{x})_e \neq y(\mathbf{x})$ and $\hat{y}(\mathbf{x})_r \neq y(\mathbf{x})$, we have $d_e > d_r$ always holds. This can be derived from Equation 3 and Equation 4,
$$d_e - d_r = |S_e - 0| - |S_r - 0| = S_e - S_r > 0. \tag{6}$$
So $\mathbb{E}_a(d_e) > \mathbb{E}_a(d_r)$ holds. Let N_a as the number of samples.

   (b) When $\hat{y}(\mathbf{x})_e = y(\mathbf{x})$ but $\hat{y}(\mathbf{x})_r \neq y(\mathbf{x})$, since $S(\mathbf{x})_e = \max(\mathbf{p}_e)$, we have $\min(S_e) = \min_p \max(\mathbf{p}_e) = \frac{1}{K}$. Since $\hat{y}_r \neq \hat{y}_e$, there exists a model m^*,
$$p_{m^*,\hat{y}_r} = \max(\mathbf{p}_{m^*}) > p_{m^*,\hat{y}_e}. \tag{7}$$
Hence, $\max(p_{m^*,\hat{y}_e}) = \frac{1}{2}$ and $\max(p_{m,\hat{y}_e}) = 1$, where $m \neq m^*$. So
$$\max(S_e) = \frac{(M-1)\max(p_{m,\hat{y}_e}) + \max(p_{m^*,\hat{y}_e})}{M} = \frac{2M-1}{2M}, \tag{8}$$
which leads to $S_e \in [\frac{1}{K}, \frac{2M-1}{2M}]$.
Similarly, since $\min(p_{m^*,\hat{y}_r}) = \dfrac{1}{K}$ and $\min(p_{m,\hat{y}_r}) = 0$, where $m \neq m^*$,
$$\min(S_r) = \frac{(M-1)\min(p_{m,\hat{y}_r}) + \min(p_{m^*,\hat{y}_r})}{M} = \frac{1}{KM}. \tag{9}$$
Additionally, from Equation 4 and Equation 5, we can get $\max(S_r) = \frac{1}{2}$, which leads to $S_r \in [\frac{1}{KM}, \frac{1}{2}]$.
According to Equation 3, we have $d_e \in [\frac{1}{2M}, \frac{K-1}{K}], d_r \in [\frac{1}{KM}, \frac{1}{2}]$. Now assume d follows a truncated Gaussian distribution, the expectation can be derived as below. Let N_b as the number of samples in this case.
$$\mathbb{E}_b(d_e) = (\frac{1}{2M} + \frac{K-1}{K}) \,/\, 2 = (K + 2MK - 2M) \,/\, 4KM$$
$$\mathbb{E}_b(d_r) = (\frac{1}{KM} + \frac{1}{2}) \,/\, 2 = (2 + KM) \,/\, 4KM$$

   (c) When $\hat{y}(\mathbf{x})_e \neq y(\mathbf{x})$ but $\hat{y}(\mathbf{x})_r = y(\mathbf{x})$, similar to Case 2b, we can derive the expectations below. Let N_c as number of samples in this case.
$$\mathbb{E}_c(d_e) = (2M + 2MK - K) \,/\, 4KM$$
$$\mathbb{E}_c(d_r) = (3KM - 2) \,/\, 4KM$$

Now let's combine Case 2b and 2c, we have

$$\begin{aligned}
\mathbb{E}_2(d_e) &= \frac{N_a \cdot \mathbb{E}_a(d_e) + N_b \cdot \mathbb{E}_b(d_e) + N_c \cdot \mathbb{E}_c(d_e)}{N_a + N_b + N_c} \\
\mathbb{E}_2(d_r) &= \frac{N_a \cdot \mathbb{E}_a(d_r) + N_b \cdot \mathbb{E}_b(d_r) + N_c \cdot \mathbb{E}_c(d_r)}{N_a + N_b + N_c} \\
\mathbb{E}_2(d_e) - \mathbb{E}_2(d_r) &= \frac{N_a[\mathbb{E}_a(d_e) - \mathbb{E}_a(d_r)] + (N_b - N_c)(K-2)(M+1)}{4KM(N_a + N_b + N_c)}
\end{aligned}$$

Since $N_a \geq 0, \mathbb{E}_a(d_e) > \mathbb{E}_a(d_r), K \geq 2, M \geq 2$, and $N_b \geq N_c$ according to Assumption 1, $\mathbb{E}_2(d_e) - \mathbb{E}_2(d_r) \geq 0$ holds, which leads to $\mathbb{E}_2(d_e) \geq \mathbb{E}_2(d_r)$.

To sum up Case 1 and Case 2, $\mathbb{E}(d_e) \geq \mathbb{E}(d_r)$ always holds when Assumption 1 satisfied. And based on Lemma 1, proposition 1 is proved.