Deep Imbalanced Regression via Hierarchical Classification Adjustment –Supplementary Materials

Haipeng Xiong Angela Yao
National University of Singapore
haipeng,ayao@comp.nus.edu.sg

1. Proof of the Propositions

1.1. Proof of the Proposition 1 (Range-Preserving Alignment)

Consider a classifier T which predicts $\hat{p} \in R^d$ with a mapping function G_T:

$$\hat{p}^T = \text{Softmax}(G_T(f)), \tag{1}$$

$\hat{p}^T$ then distills hierarchical information of $\hat{p}^h$, which typically adopts a Kullback–Leibler divergence loss between softmax normalized logits $\hat{p}^T$ and $\hat{p}^h$ [2, 5]. However, $\hat{p}_i^T \in R^{C_H}$ and $\hat{p}_i^h \in R^{C_h}$ have different resolutions, an alignment process is required before hierarchical distillation.

$$\ddot{p}^{T,h}[j] = \max_{k=1,\dots,C_H} T_{h,H}[j,k] \times \hat{p}^T[k]. \tag{2}$$

where "$\max$" denotes compute the maximum value, and then $\dot{p}^{T,h}$ is normalized to get $\overline{p}^{T,h} \in R^{C_h}$

$$\overline{p}^{T,h}[j] = \frac{\ddot{p}^{T,h}[j]}{\sum_{l=1}^{C_h} \ddot{p}^{T,h}[l]}. \tag{3}$$

After aligning $\hat{p}^T$ and $\hat{p}^h$, we can apply the Kullback–Leibler (KL) divergence between $\hat{p}_i^h$ and $\overline{p}_i^{T,h}$ as the hierarchical distillation loss functions L_{hd}^h as

$$L_{\text{hd}}^h = \text{KL}\{\hat{p}^h || \overline{p}^{T,h}\}, \tag{4}$$

and the overall hierarchical distillation loss is

$$L_{\text{hd}} = \sum_{h=1}^{H} L_{\text{hd}}^h. \tag{5}$$

Proposition 1 (Range-Preserving Alignment). *Let $v = argmax_j \overline{p}^{T,h}[j]$, $u = argmax_k \hat{p}^T[k]$. If $\overline{p}^{T,h}$ is computed by eqs. (2) and (3), then $T_{h,H}[v,u] = 1$, which indicates the class predicted by $\hat{p}^T$ is within the range of that predicted by $\overline{p}^{T,h}$.*

Proof: Let $v^{'}$ denotes the corresponding coarse class of u, which means

$$T_{h,H}[v^{'}, u] = 1 \tag{6}$$

$i)$ **Proving** $\ddot{p}^{T,h}[v^{'}] \geq \ddot{p}^{T,h}[v]$
Considering that Class transition matrix $T_{h,H}[j,k]$ can be 0 or 1, we have

$$\max_{k=1,\dots,C_H; T_{h,H}[j,k]=1} T_{h,H}[j,k] \times \hat{p}^T[k] \geq 0, \tag{7}$$

$$\max_{k=1,\dots,C_H; T_{h,H}[j,k]=0} T_{h,H}[j,k] \times \hat{p}^T[k] = 0. \tag{8}$$

Combining eqs. (2), (7) and eq. (8),

$$\begin{aligned} \ddot{p}^{T,h}[j] &= \max_{k=1,\dots,C_H} T_{h,H}[j,k] \times \hat{p}^T[k] \\ &= \max_{k=1,\dots,C_H; T_{h,H}[j,k]=1} \hat{p}^T[k]. \end{aligned} \tag{9}$$

Choosing $j = v^{'}$ in eq. (9) and combining eq. (6),

$$\ddot{p}^{T,h}[v^{'}] = \max_{k=1,\dots,C_H; T_{h,H}[v^{'},k]=1} \hat{p}^T[k] \geq \hat{p}^T[u]. \tag{10}$$

Since $u = \text{argmax}_k \hat{p}^T[k]$, we can derive eq. (11) as:

$$\ddot{p}^{T,h}[v^{'}] = \hat{p}^T[u] = \max_{k=1,\dots,C_H} \hat{p}^T[k]. \tag{11}$$

Choosing $j = v$ in eq. (12)

$$\ddot{p}^{T,h}[v] = \max_{k=1,\dots,C_H; T_{h,H}[v,k]=1} \hat{p}^T[k] \leq \hat{p}^T[u]. \tag{12}$$

From eqs. (11) and (12), we get the results of part i):

$$\ddot{p}^{T,h}[v^{'}] = \hat{p}^T[u] \geq \ddot{p}^{T,h}[v]. \tag{13}$$

ii) **Proving** $\ddot{p}^{T,h}[v^{'}] \leq \ddot{p}^{T,h}[v]$
Since $v = \text{argmax}_j \overline{p}^{T,h}[j]$, we have

$$\overline{p}^{T,h}[v^{'}] \leq \overline{p}^{T,h}[v]. \tag{14}$$

Multiplying $\sum_{l=1}^{C_h} \ddot{p}^{T,h}[l]$ in both sides of eq. (15), we can reach the conclusion:

$$\ddot{p}^{T,h}[v^{'}] \leq \ddot{p}^{T,h}[v]. \tag{15}$$

Combining i) **and** ii), we have:

$$\ddot{p}^{T,h}[v^{'}] = \ddot{p}^{T,h}[v]. \tag{16}$$

There are two situations for eq. (16):
- $\ddot{p}^{T,h}$ has single maximum value, and thus $v = v^{'}$. From eq. (6), we can get $T_{h,H}[v,u] = T_{h,H}[v^{'},u] = 1$;
- $\ddot{p}^{T,h}$ has multiple maximum values, and thus there exist $v^{'}$ satisfying $T_{h,H}[v^{'},u] = 1$ and $\ddot{p}^{T,h}[v^{'}] = \ddot{p}^{T,h}[v] = \max_j \ddot{p}^{T,h}[j]$.

Proof ends.

1.2. Proof of the Proposition 3 (Comparison of Error Bounds)

We provide a theoretical analysis of a simple case with hierarchical classifiers G_1 and G_2. Specifically, classifier G_1 has C_1 balanced classes, with $n_{1,i} = \frac{N}{C_1}$ samples for i-th class ; G_2 has $C_2 = 2C_1$ classes, with $n_{2,j}$ samples for j-th class. Note that i-th class of G_1 correspond to $(2i-1)$-th and $2i$-th classes in G_2.

Definition 1. *Following [3], the margin of i-th class of G_h is defined as $\gamma_i^h = \min_{y^h=i} \max_{l \neq y} \hat{p}^h[y^h] - \hat{p}^h[l]$, where y^h is the ground-truth for G_h.*

Definition 2. *Let $Pr(\hat{y}^h = j | y^h = i)$ denote the probability of i-th class in h-th classifier being mis-classified as j-th class by G_h. The classification error of G_h on the i-th class is defined as $L_i^h = \sum_{j \neq i} Pr(\hat{y}^h = j | y^h = i)$. The transformed classification error of G_{h+1} on the i-th class of G_h is defined as $L_i^{(h+1)\to h} = \sum_{j \neq 2i-1, 2i} Pr(\hat{y}^{h+1} = j | y^h = i)$.*

Proposition 2 (Generalization Error Bound [3])**.** *With probability $1 - \frac{1}{N^5}$, L_i^h is upper bounded by Δ_i^h:*

$$L_i^h \lesssim \Delta_i^h = \frac{1}{\gamma_i^h} \sqrt{\frac{C(G_h)}{n_{h,i}}} + \frac{\log(N)}{\sqrt{n_{h,i}}}, \tag{17}$$

where $C(G_h)$ is some proper complexity measure of function G_h, such as [1, 6], and we use $\lesssim$ to hide some constant factors.

Proposition 3 (Comparison of Error Bounds)**.** *Suppose the error of classifiers is uniformly distributed, with probability* $1-\frac{1}{N^5}$*, for* $i = 1, ..., C_1$,

$$L_i^{2\to1} \lesssim \Delta_i^{2\to1} = (1-\frac{1}{C_2-1})(\Delta_{2i-1}^2+\Delta_{2i}^2) \tag{18}$$

$$\frac{\Delta_i^{2\to1}}{\Delta_i^1} > \omega\cdot\eta_i > 1, \tag{19}$$

$$\frac{\Delta_{2i-1}^2+\Delta_{2i}^2}{\Delta_i^1} > \eta_i > 1, \tag{20}$$

where $\eta_i = \sqrt{1+r_i}+\sqrt{1+\frac{1}{r_i}}$, $r_i = \frac{n_{2,2i-1}}{n_{2,2i}}$ *and* $\omega = 1-\frac{1}{C_2-1}$.

Proof: According to Prop. 2, L_i^1 and L_j^2 have bounds as:

$$L_i^1 \lesssim \Delta_i^1 = \frac{1}{\gamma_i^1}\sqrt{\frac{C(G_1)}{n_{1,i}}}+\frac{\log(N)}{\sqrt{n_{1,i}}}, \qquad (i=1,...,C_1); \tag{21}$$

$$L_j^2 \lesssim \Delta_j^2 = \frac{1}{\gamma_j^2}\sqrt{\frac{C(G_2)}{n_{2,j}}}+\frac{\log(N)}{\sqrt{n_{2,j}}}, \qquad (j=1,...,C_2); \tag{22}$$

The error of classifiers is uniformly distributed means that for $j \neq i, j = 1, ..., C_h$,

$$Pr(\hat{y}^h = j|y^h = i) = \frac{1-Pr(\hat{y}^h = i|y^h = i)}{C_h-1}, \tag{23}$$

As per definition 2 and eq. (23),

$$\begin{aligned} L_i^{2\to1} &= \sum_{j\neq 2i-1,2i} Pr(\hat{y}^2 = j|y^1 = i) \\ &= \sum_{j\neq 2i-1,2i} Pr(\hat{y}^2 = j|y^2 = 2i-1) + \sum_{j\neq 2i-1,2i} Pr(\hat{y}^2 = j|y^2 = 2i) \\ &= L_{2i-1}^2 - Pr(\hat{y}^2 = 2i|y^2 = 2i-1) + L_{2i}^2 - Pr(\hat{y}^2 = 2i-1|y^2 = 2i) \\ &= (1-\frac{1}{C_2-1})L_{2i-1}^2 + (1-\frac{1}{C_2-1})L_{2i}^2, \end{aligned} \tag{24}$$

Substitute eq. (22) into eq. (24),

$$\begin{aligned} \Delta_i^{2\to1} &= (1-\frac{1}{C_2-1})(\Delta_{2i-1}^2+\Delta_{2i}^2) \\ &= (1-\frac{1}{C_2-1})(\frac{1}{\gamma_{2i-1}^2}\sqrt{\frac{C(G_2)}{n_{2,2i-1}}}+\frac{\log(N)}{\sqrt{n_{2,2i-1}}}+\frac{1}{\gamma_{2i}^2}\sqrt{\frac{C(G_2)}{n_{2,2i}}}+\frac{\log(N)}{\sqrt{n_{2,2i}}}) \\ &> (1-\frac{1}{C_2-1})(\frac{1}{\gamma_i^1}\sqrt{\frac{C(G_2)}{n_{2,2i-1}}}+\frac{\log(N)}{\sqrt{n_{2,2i-1}}}+\frac{1}{\gamma_i^1}\sqrt{\frac{C(G_2)}{n_{2,2i}}}+\frac{\log(N)}{\sqrt{n_{2,2i}}}) \\ &> (1-\frac{1}{C_2-1})(\frac{1}{\gamma_i^1}\sqrt{\frac{C(G_1)}{n_{2,2i-1}}}+\frac{\log(N)}{\sqrt{n_{2,2i-1}}}+\frac{1}{\gamma_i^1}\sqrt{\frac{C(G_1)}{n_{2,2i}}}+\frac{\log(N)}{\sqrt{n_{2,2i}}}) \\ &= (1-\frac{1}{C_2-1})(\frac{1}{\gamma_i^1}\sqrt{C(G_1)}+\log(N))(\frac{1}{\sqrt{n_{2,2i-1}}}+\frac{1}{\sqrt{n_{2,2i}}}) \\ &= (1-\frac{1}{C_2-1})(\frac{1}{\gamma_i^1}\sqrt{\frac{C(G_1)}{n_{1,i}}}+\frac{\log(N)}{\sqrt{n_{1,i}}})(\sqrt{1+r_i}+\sqrt{1+\frac{1}{r_i}}) \\ &= (1-\frac{1}{C_2-1})\Delta_i^1(\sqrt{1+r_i}+\sqrt{1+\frac{1}{r_i}}), \end{aligned} \tag{25}$$

From eq. (25), we have:

$$\Delta_i^{2\to 1} > (1 - \frac{1}{C_2 - 1})(\sqrt{1+r_i} + \sqrt{1+\frac{1}{r_i}})\Delta_i^1, \tag{26}$$

and

$$\begin{aligned}
\frac{\Delta_i^{2\to 1}}{\Delta_i^1} &> (1 - \frac{1}{C_2 - 1})(\sqrt{1+r_i} + \sqrt{1+\frac{1}{r_i}}) \\
&\geq (1 - \frac{1}{C_2 - 1})(\sqrt{1+1} + \sqrt{1+\frac{1}{1}}) \\
&\geq (1 - \frac{1}{4 - 1})(\sqrt{1+1} + \sqrt{1+\frac{1}{1}}) \\
&= \frac{4\sqrt{2}}{3} \\
&> 1,
\end{aligned} \tag{27}$$

Using the same derivation as eq. (24) and (25), we have

$$\frac{\Delta_{2i-1}^2 + \Delta_{2i}^2}{\Delta_i^1} > (\sqrt{1+r_i} + \sqrt{1+\frac{1}{r_i}}) \geq 2\sqrt{2} > 1. \tag{28}$$

Proof ends.

1.3. Proof of the Proposition 4 (MAE of HCA)

From the hierarchical predictions $\hat{p}^h$, we can estimate an adjusted prediction through a summation operation

$$\hat{p}^a = \hat{p}^H + \sum_{h=1}^{H-1} T_{h,H}^T \cdot \hat{p}^h, \tag{29}$$

or a multiplication operation:

$$\hat{p}^m = log(\hat{p}^H) + \sum_{h=1}^{H-1} T_{h,H}^T \cdot log(\hat{p}^h). \tag{30}$$

Proposition 4 (MAE of HCA)**.** *Let E_2 and E_{HCA} denote the mean absolute error (MAE) of G_2 and HCA (by eq. (29) or (30)). Suppose the error of classifiers is uniformly distributed, we have*

$$E_2 \lesssim U_2 = \sum_{i=1}^{C_2}\sum_{j=1}^{C_2} |j-i| \cdot \frac{\Delta_i^2}{C_2 - 1}, \tag{31}$$

$$E_{HCA} \lesssim U_{HCA} = \sum_{i=1}^{C_2}\{\nu_i - \rho_i + \sum_{j=1}^{C_2} |j-i| \cdot \rho_i\}, \tag{32}$$

$$U_2 - U_{HCA} \propto \sum_{i=1}^{C_1}(\Delta_{2i-1}^2 + \Delta_{2i}^2 - 2\Delta_i^1) > 0, \tag{33}$$

$$\frac{\Delta_{2i-1}^2 + \Delta_{2i}^2}{2\Delta_i^1} > \frac{\eta_i}{2} \geq \sqrt{2}, \tag{34}$$

where $\eta_i = \sqrt{1+r_i} + \sqrt{1+\frac{1}{r_i}}$, $r_i = \frac{n_{2,2i-1}}{n_{2,2i}}$, $\rho_i = \frac{1-\mu_i-\nu_i}{C_2-2}$,$\mu_i = (1 - \frac{C_1-2}{C_1-1}\Delta_i^1) \cdot \frac{\alpha}{\alpha+\beta}$, $\nu_i = (1 - \frac{C_1-2}{C_1-1}\Delta_i^1) \cdot \frac{\beta}{\alpha+\beta}$, $\alpha_i = (1 - \Delta_i^2)$ *and* $\beta_i = \frac{\Delta_i^2}{C_2-1}$.

Proof: Let $e_{2,i}$,$e_{a,i}$ denotes the mean absolute error (MAE) of i-th class samples predicted by $\hat{p}^2$ and $\hat{p}^a$ (or $\hat{p}^m$), respectively. We can compute $e_{2,i}$ and $e_{a,i}$ from classification errors as:

$$e_{2,i} = \sum_{j=1}^{C_2} |j-i| \cdot Pr(\hat{y}^2 = j | y^2 = i), \tag{35}$$

and

$$e_{a,i} = \sum_{j=1}^{C_2} |j-i| \cdot Pr(\hat{y}^a = j | y^2 = i), \tag{36}$$

where $Pr(\hat{y}^2 = j|y^2 = i)$ denotes the probability of i-th class in classifier G_2 being mis-classified as j-th class by $\hat{p}^2$, and $Pr(\hat{y}^a = j|y^2 = i)$ denotes the probability of i-th class in classifier G_2 being mis-classified as j-th class by $\hat{p}^a$.

We analyze the case that i is an odd number using the Prop. 2 (the analysis is the same when i is an even number). According to eq (35), (22) and (23),

$$\begin{aligned}
e_{2,i} &= \sum_{j=1}^{C_2} |j-i| \cdot Pr(\hat{y}^2 = j|y^2 = i) \\
&= \sum_{k=1}^{C_1} |2k-1-i| \cdot Pr(\hat{y}^2 = 2k-1|y^2 = i) + |2k-i| \cdot Pr(\hat{y}^2 = 2k|y^2 = i) \\
&\lesssim 0 * (1-\Delta_i^2) + 1 * \frac{\Delta_i^2}{C_2 - 1} + \sum_{k=1, k \neq \frac{2i+1}{2}}^{C_1} (|2k-1-i| + |2k-i|) \cdot \frac{\Delta_i^2}{C_2-1} \\
&= 0 * \alpha_i + 1 * \beta_i + \sum_{k=1, k \neq \frac{2i+1}{2}}^{C_1} (|2k-1-i| + |2k-i|) \cdot \frac{1-\alpha_i - \beta_i}{C_2 - 2} \triangleq u_{2,i} \\
&= \sum_{j=1}^{C_2} |j-i| \cdot \frac{\Delta_i^2}{C_2 - 1},
\end{aligned} \tag{37}$$

where $\alpha_i = (1 - \Delta_i^2)$, $\beta_i = \frac{\Delta_i^2}{C_2 - 1}$.

$$\begin{aligned}
e_{a,i} &= \sum_{j=1}^{C_2} |j-i| \cdot Pr(\hat{y}^a = j|y^2 = i) \\
&= \sum_{k=1}^{C_1} |2k-1-i| \cdot Pr(\hat{y}^a = 2k-1|y^2 = i) + |2k-i| \cdot Pr(\hat{y}^a = 2k|y^2 = i) \\
&= \sum_{k=1}^{C_1} Pr(\hat{y}^1 = k|y^1 = \frac{i+1}{2}) \cdot \{|2k-1-i| \cdot Pr(\hat{y}^2 = 2k-1|\hat{y}^2 = 2k-1, 2k) + |2k-i| \cdot Pr(\hat{y}^2 = 2k|\hat{y}^2 = 2k-1, 2k)\} \\
&\lesssim 0 * \mu_i + 1 * \nu_i + \sum_{k=1, k \neq \frac{2i+1}{2}}^{C_1} (|2k-1-i| + |2k-i|) \cdot \frac{1-\mu_i - \nu_i}{C_2 - 2} \triangleq u_{a,i} \\
&= \nu_i - \frac{1-\mu_i - \nu_i}{C_2 - 2} + \sum_{j=1}^{C_2} |j-i| \cdot \frac{1-\mu_i-\nu_i}{C_2-2},
\end{aligned} \tag{38}$$

where $\mu_i = (1 - \Delta^1_{\lceil i/2 \rceil}) \cdot \frac{\alpha}{\alpha+\beta}$ and $\nu_i = (1 - \Delta^1_{\lceil i/2 \rceil}) \cdot \frac{\beta}{\alpha+\beta}$. "$\lceil x \rceil$" denotes rounding up to the nearest integer greater than or equal to x. According to Prop. 3, we have:

$$\alpha_i < \mu_i, \beta_i < \nu_i, \tag{39}$$

Combing eq. (37)∼(39), it can be derived that

$$u_{a,i} < u_{2,i}, \tag{40}$$

Finally, the overall MAE E_2 and E_a for $\hat{p}_2$ and $\hat{p}_a$ can be computed as:

$$E_2 = \sum_{i=1}^{C_2} e_{2,i} < \sum_{i=1}^{C_2} u_{2,i} = \sum_{i=1}^{C_2}\sum_{j=1}^{C_2} |j-i| \cdot \frac{\Delta_i^2}{C_2 - 1} \triangleq U_2, \tag{41}$$

and

$$E_a = \sum_{i=1}^{C_2} e_{a,i} < \sum_{i=1}^{C_2} u_{a,i} = \sum_{i=1}^{C_2} \{\nu_i - \frac{1-\mu_i-\nu_i}{C_2-2} + \sum_{j=1}^{C_2} |j-i| \cdot \frac{1-\mu_i-\nu_i}{C_2-2}\} \triangleq U_{HCA}. \tag{42}$$

Combining eq. (40)∼(42), it can be derived that

$$U_2 - U_{HCA} \propto \sum_{i=1}^{C_1} (\Delta_{2i-1}^2 + \Delta_{2i-1}^2 - \Delta_i^1) > 0, \tag{43}$$

where "$\propto$" denotes being propositional to. Eq. (34) has already been proved in eq. (28).
Proof ends.
Remarks on Data Sufficiency: i) When the data is sufficient ($n_{h,i} \to \infty$), the upper bounds Δ_i^h for a given classifier, as given in Eq. (17) approaches zero. Therefore, each of the Δ_i^h terms on the RHS of Eq. (33) will progressively shrink, *i.e.* $(\Delta_{2i-1}^2 + \Delta_{2i}^2) - 2\Delta_i^1$ becomes smaller, resulting in a limited gap between U_2 and U_{HCA} (eq. (33)).
ii) The converse is true for Δ_i^h when the data is limited and the gap between U_2 and U_{HCA} will become more prominent, as eq. (34) suggests that RHS of eq. (33) is larger than $\sum_{i=1}^{C_1} 2(\sqrt{2}-1)\Delta_i^1$. Moreover, as per eq. (33) and (34), the more imbalanced the data (the larger r_i), the larger the difference between U_2 and U_{HCA}.

2. Experiment and Discussion

2.1. More Ablation Studies

IMDB-WIKI-DIR [23] and SHTech Part A (SHA) [24] data are chosen for ablation studies. Mean absolute error (MAE) and its balanced version bMAE [13] are adopted as evaluation metrics for SHA and IMDB-WIKI-DIR, respectively. Lower MAE and bMAE denote better performance.
i) *Influence of Class Number* C_H In [21], C_H is chosen as 100. We further explored the influence of larger C_H in the SHA dataset. Specifically, $1 \sim (H-1)$-th classifiers are kept the same while the class number of H-th classifier is increased. Fig. 1 visualize the results of various C_H ranging from 100 to 3200. As shown in Fig. 1, increasing C_H will increase the MAE of a single classifier due to the fewer sample per class, which is consistent with the results in [21]. Moreover, HCA shows consistent improvement over all the C_H, especially when $C_H \geq 1600$.

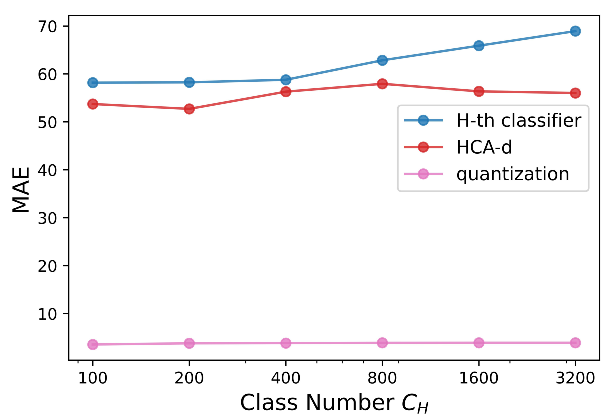

Figure 1. Comparison of varing class numbers C_H of H-th classfiers.

$ii)$ *Settings of Classifiers* For classifier G_h $(h = 1, ..., H)$, we adopt a single linear layer, which maps features $f \in R^d$ to outputs $\hat{p}^H \in R^{C_h}$ For classifiers G_T, we investigate linear and non-linear settings. In the linear setting, one fully connected layer (1-fc) is adopted, which maps features $f \in R^d$ to outputs $\hat{p}^T \in R^{C_H}$. For the non-linear setting, two fully connected layers (2-fc) with feature dimensions $\frac{d}{4}$ and C_H are adopted, and softplus is adopted as the activation function. Table 1 presents the quantitative results. It can be observed that: $i)$ both of the 1-fc and 2-fc G_Ts show significant improvement compared to the (1-fc and 2-fc) vanilla classifiers; $ii)$ 2-fc G_T shows slightly better performance than adopting 1-fc G_T, suggesting a linear G_T is not adequate for distilling hierarchical information from classifiers G_h $(h = 1, ..., H)$; $iii)$ the vanilla classifiers have the same class-splitting as classifier G_T, but 2-fc CLS does not show significant improvement to 1-fc CLS.

	G_h	G_T	IMDB-WIKI-DIR				SHA
			All	Many	Med.	Few	
CLS	1-fc		13.58	7.13	13.95	33.21	58.2
CLS	2-fc		13.51	7.43	13.95	31.97	58.1
HCA-d	1-fc	1-fc	13.06	7.00	13.17	31.72	54.5
HCA-d	1-fc	2-fc	12.70	7.00	13.18	29.94	53.7

Table 1. Comparing settings of Classifiers.

$iii)$ **One-hot or Gaussian-smoothed ground-truth labels** One-hot and Gaussian-smoothed [4] ground truths p^h are two common choices for cross-entropy losses. Compared to one-hot p^h, Gaussian-smoothed ground truths further encode the ordinal relationship among labels. We compare both of them in Table 2. From Table 2, we can observe that HCA shows improvements with both hard and soft ground truths, and HCA with soft ground truths delivers better performance. We use soft labels by default in all of the remaining experiments.

Method	GT	IMDB-WIKI-DIR				SHA
		All	Many	Med.	Few	
CLS	one-hot	13.48	7.25	13.65	32.57	58.8
HCA-d		12.93	7.20	12.81	30.71	55.0
CLS	soft [4]	13.58	7.13	13.95	33.21	58.2
HCA-d		12.70	7.00	13.18	29.94	53.7

Table 2. Soft vs. hard one-hot ground truth of classification.

$iv)$ **Can HCA be a regularizer to regression?** We combine HCA and regression in a single network to see the combination effect of them. Results are shown in Table 3. Training HCA and regression together will improve the regression performance (MAE from 65.4 to 58.7). However, the performance of HCA will be harmed by regression (MAE from 53.7 to 58.6), implying that learning imbalanced regression targets together is harmful to HCA.

$v)$ **Imbalanced Ratios** We do ablation studies on imbalanced ratios in Table 4. It can be observed that HCA outperforms both regression and vanilla classification in all imbalance ratios. The larger the imbalance ratio r, the greater the improvement from vanilla classification to HCA. Theoretically, as indicated in eq.16&17, an increasing r leads to larger η_i, thereby amplifying the improvement from vanilla classification to HCA.

	HCA or Regression		HCA+Regression	
	MAE↓	RMSE↓	MAE↓	RMSE↓
Regression	65.4	103.3	**58.7**	**101.8**
HCA-d	**53.7**	**87.8**	58.6	100.4

Table 3. Comparison on SHTech dataset Part A (SHA) [24]. (Left) Training HCA or Regression with L1 loss separately. (Right) Training HCA and Regression together in a network.

2.2. Comparison with SOTA on Regression Tasks

SHTech Dataset SHTech [24] is a crowd-counting dataset, which presents severe imbalanced distribution [9, 21, 22]. It has two subsets, part A and part B. Part A presents crowded scenes captured in arbitrary camera views, while part B presents

Configuration	$r=19$	$r=49$	$r=99$
	1900:100	1960:40	1980:20
Regression	6.78±0.04	8.07±0.07	8.07±0.13
CLS	6.78±0.03	7.64±0.13	7.65±0.08
HCA-d	6.72±0.04	7.57±0.01	7.54±0.04

Table 4. Comparison on subsampled subsets of IMDB-WIKI-DIR [23] with different imbalanced ratios. The sample number of each subset is the same. "$n_1 : n_2$" denotes the sample number of the major and the minor classes, and "r" denotes the imbalance ratio.

relatively sparse scenes captured by surveillance cameras. We follow the same network setting as [21], where 100 logarithm classes are adopted for C_H. Mean absolute error (MAE) and rooted mean square error are adopted as evaluation metrics. Both MAE and RMSE are the lower, the better. Quantitative results are presented in Table 5. It can be observed that Hierarchical classification shows the best performance and improves plain classification by a large margin.

	SHA		SHB	
	MAE↓	RMSE↓	MAE↓	RMSE↓
CSRNet [8]	68.2	115.0	10.6	16.0
DRCN [16]	64.0	98.4	8.5	14.4
BL [10]	62.8	101.8	7.7	12.7
PaDNet [17]	59.2	98.1	8.1	12.2
MNA [18]	61.9	99.6	7.4	11.3
OT [20]	59.7	95.7	7.4	11.8
GL [19]	61.3	95.4	7.3	11.7
Regression [21]	65.4	103.3	10.7	19.5
DC-regression [21]	60.7	101.0	7.1	**11.0**
CLS	58.2	96.7	7.0	11.8
HCA-add	55.9	92.8	**6.7**	11.4
HCA-mul	54.7	91.6	6.8	11.4
HCA-d	**53.7**	**87.8**	6.8	11.8

Table 5. Comparison on SHTech dataset [24]. Methods are grouped as density map regression, local count regression and classification approaches.

IMDB-WIKI-DIR Dataset IMDB-WIKI-DIR [23] is a large age estimation dataset, which is an imbalanced subset sampled from IMDB-WIKI [14]. There are 191509 training samples, 11022 validation samples, and 11022 testing samples. Table 6 presents the quantitative results. It can be observed that hierarchical classification shows the best result on the whole range of the target space. We choose three baselines of classification, they are: i) vanilla classification, which is H-th classifier of HCA; ii) classification with label distribution smoothing (LDS) [23], which re-weight samples with inverse class frequency; iii) classification with label distribution smoothing (LDS) and ranksim [7] regularization, ranksim [7] regularizes feature space to have the same ordering as label space. Their HCA counterparts are also included.

From the results in Table 6, we can observe that: i) HCA shows clear improvement in bMAE over naive classification baselines. Specifically, HCA-d can improve all the shots for "CLS" and "CLS+LDS" baselines, while for strong baseline "CLS+LDS+ranksim", since the baseline results are already saturated for the many-shot, there is still a slight trade-off between many and few-shot (many-shot bMAE increases from 6.70 to 6.88). ii) HCA outperforms its regression baselines and other regression approaches. Noted that Balanced MSE [13] is a logit adjustment version for regression, it improves the few/medium-shot performances via significantly harming the many-shot (bMAE from 7.32 to 7.56), while for HCA-d, many-shot performance is roughly maintained or improved.

AgeDB-DIR Dataset AgeDB-DIR [23] is an imbalanced re-sampled version of AgeDB dataset [12]. It contains 12208 training samples, 2140 validation samples and 2140 testing samples, with ages ranging from 0 to 101. Table 7 presents the quantitative results. HCA approaches show consistent improvement over classification baselines and outperform regression approaches.

NYUDv2-DIR Dataset NYUDv2-DIR [23] is an imbalanced version sampled from the NYU Depth Dataset V2 [15]. The depth values range from 0 to 10 meters, which are divided into 100 logarithm classes for C_H. Mean absolute error (MAE), rooted mean square error (RMSE), relative absolute error (RelAbs), δ_1, δ_2 and δ_3 are adopted as evaluation metrics. Noted that all classes in NYUDv2-DIR have more than 10^7 samples, which should be all categorized as many-shot classes according

Methods	bMAE↓				MAE↓			
	All	Many	Med.	Few	All	Many	Med.	Few
Regression [23]	13.92	7.32	15.93	32.78	8.06	7.23	15.12	26.33
Regression+LDS [23]	13.37	7.55	13.96	30.92	8.11	7.47	13.41	23.50
Regression+LDS +ranksim [7]	12.83	7.00	13.28	30.51	7.56	6.94	12.61	23.43
Regression+FDS +ranksim [7]	12.39	6.91	12.82	29.01	7.35	6.81	11.50	22.75
Balanced MSE [13]	12.66	7.65	12.68	28.14	8.12	7.58	12.27	23.05
DC-regression [21]	14.18	7.30	16.04	34.00	8.05	7.18	15.40	26.48
DC-regression+LDS [21]	13.04	8.11	13.62	27.82	8.62	8.04	13.50	**22.04**
CLS	13.58	7.13	13.95	33.21	7.75	7.04	13.60	25.17
CLS+LA [11]	13.04	7.82	11.89	30.10	8.22	7.75	11.75	22.40
HCA-add	12.86	6.98	13.15	30.80	7.53	6.90	12.70	23.53
HCA-mul	12.89	7.00	13.36	30.74	7.57	6.92	12.91	23.52
HCA-d	12.70	7.00	13.18	29.94	7.54	6.91	12.69	22.96
CLS+LDS	12.85	7.31	13.40	29.54	7.84	7.25	12.53	23.56
HCA-add+LDS	12.64	7.15	12.83	29.47	7.66	7.09	12.20	23.31
HCA-mul+LDS	12.68	7.18	13.03	29.42	7.70	7.11	12.35	23.34
HCA-d+LDS	12.42	7.28	12.47	28.24	7.77	7.21	12.25	22.43
CLS+LDS+ranksim	12.33	**6.70**	13.16	29.10	**7.25**	**6.63**	12.26	22.77
HCA-add+LDS+ranksim	12.15	6.77	12.09	28.80	7.26	6.72	11.39	23.48
HCA-mul+LDS+ranksim	12.24	6.69	12.69	29.01	7.22	6.63	11.84	23.22
HCA-d+LDS+ranksim	**11.92**	6.88	**11.67**	**27.72**	7.31	6.82	**10.99**	**22.04**

Table 6. Comparison on IMDB-WIKI-DIR Dataset.

Methods	bMAE↓				MAE↓			
	All	Many	Med.	Few	All	Many	Med.	Few
Regression [23]	9.72	6.62	8.80	16.66	7.57	6.61	8.73	13.48
Regression+LDS [23]	9.12	6.98	8.87	13.66	7.67	6.98	8.87	10.91
Regression+LDS +ranksim [7]	7.96	**6.34**	7.84	11.35	**6.91**	**6.34**	7.80	9.92
Balanced MSE [13]	8.97	7.65	7.43	12.65	7.78	7.65	7.45	9.99
DC-regression [21]	9.70	6.82	8.77	16.16	7.65	6.82	8.70	12.55
DC-regression+LDS [21]	9.48	7.36	9.14	14.04	8.03	7.36	9.13	11.26
CLS	9.14	6.89	8.62	14.08	7.58	6.89	8.51	11.60
CLS+LA [11]	8.86	7.80	8.82	11.03	8.20	7.80	8.87	10.06
HCA-add	8.95	6.91	8.26	13.53	7.49	6.91	8.17	11.05
HCA-mul	8.97	6.93	8.35	13.52	7.52	6.93	8.25	11.10
HCA-d	8.85	6.86	8.31	13.26	7.45	6.86	8.22	10.90
CLS+LDS	8.75	7.17	8.29	12.27	7.63	7.17	8.30	10.14
HCA-add+LDS	8.40	7.22	7.83	11.18	7.53	7.22	7.82	9.61
HCA-mul+LDS	8.54	7.25	8.02	11.49	7.60	7.25	8.02	9.70
HCA-d+LDS	8.46	7.11	7.80	11.64	7.47	7.11	7.77	10.06
CLS+LDS+ranksim	7.99	6.66	7.21	11.20	6.97	6.66	7.16	9.34
HCA-add+LDS+ranksim	**7.82**	6.67	**7.12**	**10.59**	6.94	6.67	**7.07**	**9.10**
HCA-mul+LDS+ranksim	7.85	6.68	7.14	10.71	6.95	6.68	7.10	9.17
HCA-d+LDS+ranksim	7.87	6.74	7.14	10.66	7.01	6.74	7.13	9.22

Table 7. Comparison on AgeDB-DIR Dataset.

to the criteria in IMDB-WIKI-DIR [23] ($>$ 100 samples). We report the overall results in Table 8 and detailed results for relatively many/medium/few shots can be found in Table 9. We can observe that HCA shows improvements to its naive classification baselines and it is also comparable to or better than other regression methods.

References

[1] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *NeurIPS*, 30, 2017. 2

Methods	MAE↓	RMSE↓	AbsRel↓	$\delta_1\uparrow$	$\delta_2\uparrow$	$\delta_3\uparrow$
Regression [23]	1.004	1.486	**0.179**	0.678	0.908	0.975
Regression+LDS [23]	0.968	1.387	0.188	0.672	0.907	**0.976**
Regression+LDS+ranksim [7]	0.931	1.389	0.183	0.699	0.905	0.969
Balanced MSE [13]	0.922	**1.279**	0.219	0.695	0.878	0.947
CLS	1.011	1.512	0.184	0.678	0.906	0.958
HCA-add	0.987	1.470	0.180	0.686	0.909	0.961
HCA-mul	0.991	1.478	0.181	0.685	0.909	0.960
HCA-d	0.987	1.475	0.181	0.689	0.915	0.961
CLS+LDS	0.924	1.383	0.181	0.711	0.909	0.965
HCA-add+LDS	0.919	1.375	0.180	0.710	0.910	0.965
HCA-mul+LDS	0.920	1.377	0.180	0.710	0.910	0.965
HCA-d+LDS	0.911	1.367	**0.179**	0.714	0.911	0.966
CLS+LDS+ranksim	0.904	1.335	0.182	**0.715**	0.916	0.972
HCA-add+LDS+ranksim	0.901	1.330	0.181	0.714	**0.919**	0.972
HCA-mul+LDS+ranksim	0.902	1.332	0.181	0.714	0.918	0.972
HCA-d+LDS+ranksim	**0.895**	1.321	0.180	**0.715**	**0.919**	0.972

Table 8. Comparison on NYUD2-DIR dataset. Methods are grouped as regression and classification approaches.

[2] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *CVPR*, pages 12506–12515, 2020. 1

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 32, 2019. 2

[4] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In *CVPR*, pages 4738–4747, 2019. 7

[5] Ashima Garg, Depanshu Sani, and Saket Anand. Learning hierarchy aware features for reducing mistake severity. In *ECCV*, pages 252–267. Springer, 2022. 1

[6] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *COLT*, pages 297–299. PMLR, 2018. 2

[7] Yu Gong, Greg Mori, and Fred Tung. RankSim: Ranking similarity regularization for deep imbalanced regression. In *ICML*, pages 7634–7649, 2022. 8, 9, 10, 11

[8] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, pages 1091–1100, 2018. 8

[9] Liang Liu, Hao Lu, Haipeng Xiong, Ke Xian, Zhiguo Cao, and Chunhua. Shen. Counting objects by blockwise classification. *IEEE TCSVT*, 30(10):3513–3527, 2019. 7

[10] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *ICCV*, pages 6142–6151, 2019. 8

[11] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021. 9

[12] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *CVPRW*, pages 51–59, 2017. 8

[13] Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. In *CVPR*, 2022. 6, 8, 9, 10, 11

[14] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *IJCV*, 126(2-4):144–157, 2018. 8

[15] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760, 2012. 8

[16] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *Technical Report*, 2020. 8

[17] Yukun Tian, Yiming Lei, Junping Zhang, and James Z Wang. Padnet: Pan-density crowd counting. *IEEE TIP*, 29:2714–2727, 2019. 8

[18] Jia Wan and Antoni Chan. Modeling noisy annotations for crowd counting. *NeurIPS*, 33, 2020. 8

[19] Jia Wan, Ziquan Liu, and Antoni B Chan. A generalized loss function for crowd counting and localization. In *CVPR*, pages 1974–1983, 2021. 8

[20] Boyu Wang, Huidong Liu, Dimitris Samara, and Minh Hoai. Distribution matching for crowd counting. In *NeurIPS*, 2020. 8

	MAE↓				RMSE↓				AbsRel↓			
	All	Many	Med.	Few	All	Many	Med.	Few	All	Many	Med.	Few
Regression [23]	1.004	0.400	0.639	1.748	1.486	0.562	0.845	2.162	0.179	0.153	0.165	0.210
Regression+LDS [23]	0.968	0.485	0.716	1.548	1.387	0.671	0.913	1.954	0.188	0.188	0.189	0.187
Regression+LDS+ranksim [7]	0.931	0.452	0.708	1.495	1.389	0.639	0.922	1.967	0.183	0.180	0.195	0.181
Balanced MSE [13]	0.922	0.630	0.726	1.289	1.279	0.819	0.917	1.705	0.219	0.270	0.252	0.156
CLS	1.011	0.425	0.755	1.695	1.512	0.642	1.028	2.152	0.184	0.159	0.189	0.207
HCA-add	0.987	0.429	0.755	1.635	1.470	0.652	1.016	2.081	0.180	0.160	0.185	0.199
HCA-mul	0.991	0.428	0.754	1.645	1.478	0.650	1.021	2.094	0.181	0.159	0.187	0.201
HCA-d	0.987	0.427	0.755	1.637	1.475	0.648	1.021	2.090	0.181	0.159	0.187	0.200
CLS+LDS	0.924	0.483	0.860	1.388	1.383	0.766	1.196	1.851	0.181	0.179	0.205	0.173
HCA-add+LDS	0.919	0.481	0.852	1.382	1.375	0.754	1.181	1.845	0.180	0.178	0.205	0.173
HCA-mul+LDS	0.920	0.482	0.854	1.383	1.377	0.757	1.184	1.847	0.180	0.178	0.205	0.173
HCA-d+LDS	0.911	0.479	0.849	1.368	1.367	0.746	1.174	1.835	0.179	0.177	0.204	0.171
CLS+LDS+ranksim	0.904	0.507	0.747	1.361	1.335	0.770	1.014	1.807	0.182	0.193	0.193	0.167
HCA-add+LDS+ranksim	0.901	0.503	0.743	1.359	1.330	0.756	1.006	1.804	0.181	0.191	0.192	0.167
HCA-mul+LDS+ranksim	0.902	0.503	0.744	1.361	1.332	0.758	1.008	1.807	0.181	0.191	0.193	0.167
HCA-d+LDS+ranksim	0.895	0.503	0.741	1.347	1.321	0.754	1.006	1.791	0.180	0.191	0.193	0.165
	δ_1 ↑				δ_2 ↑				δ_3 ↑			
	All	Many	Med.	Few	All	Many	Med.	Few	All	Many	Med.	Few
Regression [23]	0.678	0.788	0.762	0.536	0.908	0.960	0.942	0.843	0.975	0.991	0.984	0.955
Regression+LDS [23]	0.672	0.701	0.706	0.630	0.907	0.932	0.929	0.875	0.976	0.984	0.982	0.964
Regression+LDS+ranksim [7]	0.699	0.734	0.713	0.658	0.905	0.933	0.911	0.870	0.969	0.985	0.973	0.953
Balanced MSE [13]	0.695	0.617	0.813	0.728	0.878	0.835	0.944	0.896	0.947	0.900	0.963	0.988
CLS	0.678	0.784	0.708	0.562	0.911	0.949	0.913	0.871	0.959	0.985	0.974	0.928
HCA-add	0.686	0.782	0.709	0.582	0.918	0.948	0.923	0.885	0.962	0.985	0.975	0.935
HCA-mul	0.685	0.783	0.707	0.579	0.917	0.948	0.921	0.883	0.961	0.985	0.974	0.933
HCA-d	0.689	0.783	0.707	0.588	0.915	0.949	0.921	0.879	0.961	0.985	0.974	0.933
CLS+LDS	0.711	0.736	0.678	0.697	0.909	0.931	0.894	0.892	0.965	0.977	0.974	0.949
HCA-add+LDS	0.710	0.736	0.679	0.696	0.910	0.933	0.897	0.892	0.965	0.979	0.974	0.949
HCA-mul+LDS	0.710	0.736	0.678	0.696	0.910	0.933	0.896	0.892	0.965	0.978	0.974	0.948
HCA-d+LDS	0.714	0.738	0.676	0.705	0.911	0.934	0.899	0.892	0.966	0.979	0.974	0.949
CLS+LDS+ranksim	0.715	0.725	0.726	0.701	0.916	0.929	0.929	0.897	0.972	0.976	0.978	0.966
HCA-add+LDS+ranksim	0.714	0.725	0.724	0.699	0.919	0.931	0.931	0.902	0.972	0.977	0.976	0.965
HCA-mul+LDS+ranksim	0.714	0.725	0.724	0.699	0.918	0.930	0.310	0.901	0.972	0.977	0.976	0.965
HCA-d+LDS+ranksim	0.715	0.725	0.721	0.703	0.919	0.931	0.931	0.902	0.972	0.978	0.975	0.965

Table 9. Detailed results on NYUD2-DIR dataset. Methods are grouped as regression and classification approaches.

[21] Haipeng Xiong and Angela Yao. Discrete-constrained regression for local counting models. In *ECCV*, pages 621–636. Springer, 2022. 6, 7, 8, 9

[22] Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *ICCV*, pages 8362–8371, 2019. 7

[23] Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *ICML*, 2021. 6, 8, 9, 10, 11

[24] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597, 2016. 6, 7, 8