1. Details about Identifying Pattern Imbalance

**Representation Visualization.** We first extract the output features before the full-connected layer of samples belonging to certain classes (e.g., class 0-airplane, 3-cat, 5-dog). Then TSNE is applied to reduce the dimension of these features to 2. We separately evaluate the model of epoch 1 and epoch 100 as not fully trained model and fully trained model. Classification loss is re-scaled to facilitate visualization, as $l \leftarrow \log(l + 1)$. Similar settings hold for the visualization of the activation path. As observed in the experimental results, both representations and activation paths show a clear imbalance.

**Clustering.** The classification loss of each sample is a scalar, and cannot be directly clustered. We repeat each scalar to a bivariate vector to facilitate clustering. For clustering, we employ the K-means algorithm \[1\] with the random state set as 1 and cluster number fixed as 10.

**Why not choose OOD datasets instead of CIFAR10?** Considering that there are multiple domains in one category in OOD datasets, it will naturally show an obvious imbalance, which will also make it difficult for us to find a suitable comparison to evaluate the clustering consistency criterion. Therefore, we tend to choose simpler settings and conduct experiments on the CIFAR10 dataset to verify that even if the model performs well on average, there still exists some modes that have not been fully optimized.

### Theorem 1 (Upper bound of seed category)

For a dataset with $N$ samples and threshold $\xi$, then the upper bound of seed categories satisfies,

$$|S| \leq \frac{N[1 + \ln(1 + F(\xi))]}{1 + F(\xi)}$$

where $|S|$ denotes the number of seed category.

**Proof:**

For the graph where the samples represent the vertices of the graph and the adjacent relationship between samples represents the edges between vertices, we can know this graph possesses $N$ vertices with minimum degree $\delta$.

We define all the vertices set as $V$, and then we construct a random subset $X$ of $V$ ($X \subset V$). Each sample in $X$ is taken from $V$ with a probability of $p$. Then the expectation scale of $X$ is,

$$E(|X|) = Np$$

We regard the subset $X$ as the candidate for seed set $S$. We can thus define the random set $Y_X$, which represents the samples in $V - X$ that do not have an adjacent sample in $X$, that is, for sample $v \in Y_X$, we can not find a sample $x \in X$ that $v$ is subordinate to $x$. This can also be interpreted as for $v \in Y_X$, any adjacent samples of $v$ not in $X$, so

$$P(v \in Y_X) = P(v \text{ and its adjacent samples not in } X)$$

$$\leq (1 - p)^{1+\delta(v)}$$

Then we can obtain,

$$E(|Y_X|) \leq N(1 - p)^{1+\delta}$$

It is apparent that $X \cup Y_X$ can be served as a seed set, and the number of seed categories can be represented as,

$$E(|X \cup Y_X|) \leq E(|X| + |Y_X|) \leq E(|X|) + E(|Y_X|)$$

$$= Np + N(1 - p)^{1+\delta} \leq Np + Ne^{-p(1+\delta)}$$

2. Theoretical Analysis

2.1. Upper Bound of the Number of Seed Categories

Once we determine a threshold $\xi$, the adjacent between samples is hence determined. It is apparent that the threshold dominates the density of edges in the graph, thus affecting the seed category number. We assume that the number of minimum adjacent points for a sample in the dataset is $\delta$, which satisfies $\delta = F(\xi)$, where $F$ is a non-increasing function. Then according to Arnautov-Payan theorem [2], we can get the following theorem.

\[\{\text{Supplementary Material}\}\]
Since we want to find the minimal seed set, which means we want to find the minimal value of \( NP + Ne^{-p(1+\delta)} \). We can then obtain that when \( p = \frac{\ln(\delta+1)}{\delta+1} \), the expectation get the minimum value,

\[
N[1 + \ln(\delta + 1)] \\
\delta + 1
\]

Therefore we can get

\[
|S| \leq \frac{N[1 + \ln(\delta + 1)]}{\delta + 1}, \quad \text{where} \quad \delta = F(N).
\]

2.2. Scale of Seed Category

We add a strong constraint to our assumption that the samples between different seed categories are not adjacent. In this way, we can get the following theorem,

**Theorem 2** (Scale of seed category) For a dataset with \( N \) samples and the samples between different seed categories are not adjacent. The scale of the seed category satisfies \( b < |S| < a \) if and only if the number of adjacent sample pairs satisfies,

\[
b^2 \left[ \frac{N}{b} \right] - N < E(N) \leq (a-1) + \frac{1}{2} N^2 - \frac{1}{2} + (a-1)N.
\]

We can also deduce a similar upper bound with a more concise form,

\[
E(N) \leq (1 - \frac{1}{a-1}) \frac{N^2}{2}
\]

, where \( |S| \) represent the scale of a certain seed category \( S \).

**Proof:**

The lower bound is obvious. Here we only prove the upper bound.

We can model the adjacency relationship between samples as a bipartite graph with \( N \) points on both sides. The problem can be transformed into a matrix of \( M \in \mathbb{R}^{N \times N} \), where there is no all 1 sub-matrix of \( M_0 \in \mathbb{R}^{A \times A} \), and at this time, how many element 1 can there be in the matrix at most.

We count the following structures. We define that the left and right point sets of the bipartite graph are \( V_1 \) and \( V_2 \), and then assume that the structure \( p \) is selecting a point \( u \) from \( V_1 \) with \( a \) adjacent samples in \( V_2 \). Let’s start with point \( u \) in \( V_1 \), and the selection methods of \( a \) samples in \( V_2 \) is \( C_a^n \), then the total selections are \( \sum_{u \in V_1} C_a^n (d(u)) \). We can also start with \( a \) samples in \( V_2 \). And once we determine a point in \( V_2 \), there are at most \( a - 1 \) \( u \) in \( V_1 \), otherwise there will be an all 1 sub-matrix of \( M_0 \). We can thus obtain,

\[
\sum_{u \in V_1} C_a^n (d(u)) \leq C_N^n (a - 1)
\]

Following Jensen Inequality, \( f(x) = C_x^n \) is a convex function, then,

\[
\sum_{u \in V_1} \frac{1}{N} C_a^n (d(u)) \geq C_N^n \left( \frac{1}{N} \sum_{u \in V_1} d(u) \right) = C_N^n \frac{a}{N}
\]

So,

\[
NC_a^{\delta E} \leq \sum_{u \in V_1} C_a^n (d(u)) \leq C_N^n (a - 1)
\]

\[
= \frac{N(N - 1) \ldots (N - a + 1)}{a!} (a - 1)
\]

Retraction is conducted at both sides, and then,

\[
E(N) \leq (a - 1) + \frac{1}{2} N^2 - \frac{1}{2} + (a - 1)N
\]

After simplification, we can obtain,

\[
E(N) \leq (a - 1) \frac{1}{2} N^2 - \frac{1}{2} + (a - 1)N
\]

The proof of another upper bound is presented as follows.

Let the number of points be \( N \) and for every point \( x_i \), the number of neighboring points is \( d(x_i) \). Suppose a initial set \( C_\pi = \emptyset \), for all points, we introduce a random permutation \( x_1, x_2, x_3, ..., x_n \). For a certain permutation, if all points in front of \( x_i \) are \( x_i \)'s neighboring points, we put \( x_i \) into \( C_\pi \). Finally, all point pairs in \( C_\pi \) are neighboring points.

The probability of a certain point in \( C_\pi \) is \( p = \frac{1}{N-d(x_i)} \), then the mathematical expectation of the size of \( C \) is,

\[
|C_\pi| = \sum_{x_i} \frac{1}{N - d(x_i)}
\]

Suppose the size of maximal cluster in dataset is \( \omega(D) \), we apply Pigeonhole Principle [5] and get:

\[
\omega(D) \geq \sum_{x_i} \frac{1}{N - d(x_i)}
\]

What we need to satisfy is,

\[
a \geq \omega(G) \geq \sum_{v_i} \frac{1}{N - d(v_i)}
\]

According to Cauchy Inequality,

\[
a \sum_{v_i} (N - d(v_i)) \geq \sum_{v_i} \frac{1}{N - d(v_i)} \sum_{v_i} (N - d(v_i)) \geq N^2
\]

So,

\[
a(N^2 - 2|E|) \geq N^2
\]

We can thus obtain,

\[
|E| \leq \frac{N^2}{2} \left( 1 - \frac{1}{a-1} \right)
\]
2.3. Convergence Analysis

Since we have a similar training process with GroupDRO [4], we have similar conclusions about the convergence of the algorithm.

**Theorem 3** (Convergence analysis) Suppose the classification loss is convex and nonnegative, $B_2$-Lipschitz continuous and bounded by $B_1$, and model parameter $||\theta||_2 \leq B_0$. The expectation error $\epsilon_t$ at step $t$ satisfies,

$$\mathbb{E}[\epsilon_t] \leq 2|S|\sqrt{\frac{10(B_2^2B_3^2 + B_1^2 \log |S|)}{t}}$$  \hspace{1cm} (21)

**Proof:**
We follow the convergence analysis in GroupDRO [4] that,

$$\mathbb{E}[\epsilon_t] \leq 2m\sqrt{\frac{10(B_2^2B_3^2 + B_1^2 \log m)}{t}}$$  \hspace{1cm} (22)

, where $m$ represents the number of all domains. Here the optimization unit is the seed category. So we replace the domain number $m$ with the number of seed category $S$, thus obtaining the convergence analysis of our method,

$$\mathbb{E}[\epsilon_t] \leq 2|S|\sqrt{\frac{10(B_2^2B_3^2 + B_1^2 \log |S|)}{t}}$$  \hspace{1cm} (23)

3. Proof of Calculating Seed Category by Loss Ranking

Supposing $P$ and $Q$ are two unit distributions,

$$JS(P||Q) = \frac{1}{2} KL(P||\frac{P+Q}{2}) + \frac{1}{2} KL(Q||\frac{P+Q}{2})$$  \hspace{1cm} (24)

Following f-GAN [3], we define $f$ as the generator function $f: \mathbb{R}_+ \rightarrow \mathbb{R}$ of JS divergence, and $f^*$ as its conjugate function, which is defined as $f^*(t) = \sup_{u \in \text{dom} f} \{ut - f(u)\}$.

For JS divergence, let $f(u) = -(u+1) \log 1 + u \log u$ and correspondingly $f^*(t) = -\log(2 - e^t)$, then following f-GAN [3],

$$JS(P||Q) = \int q(x)f\left(\frac{p(x)}{q(x)}\right)dx$$
$$= \int q(x) \sup_{t \in \text{dom} f^*} \left\{ t\frac{p(x)}{q(x)} - f^*(t) \right\}$$
$$\geq \sup_{t \in T} \left( \int p(x)T(x)dx - \int q(x)f^*(T(x))dx \right)$$
$$= \sup_{t \in T} \mathbb{E}_{x \sim P}[T(x)] - \mathbb{E}_{x \sim Q}[f^*(T(x))]$$

where $T$ represents a arbitrary mapping from data point $x$ to conjugate variable $t$ and $T$ represents qualified mapping groups.

We assume that $T$ is an optimized model with learnable parameters $\theta$ following f-GAN [3], and its optimal solution determines a classification for distribution $P$ and $Q$. We rewrite $T$ as a combination of feature extraction function $V_w$ and activation function $g_f$, i.e., $T_\theta(x) = g_f(V_w(x))$. Let the activation function be,

$$g_f(v) = \log 2 - \log(1 + e^{-v})$$  \hspace{1cm} (26)

and $D(v) = 1/(1 + e^{-v})$. Then the overlapping of distribution $P$ and $Q$, which is denoted as $D_{PQ}$, can be represented as,

$$D_{PQ} = -\mathbb{E}_{x \sim P}[g_f(V_w(x))] - \mathbb{E}_{x \sim Q}[-f^*(g_f(V_w(x)))]$$  \hspace{1cm} (27)

Then,

$$D_{PQ} = -\mathbb{E}_{x \sim P} \left[ \log \frac{2}{1 + e^{-V_w(x)}} \right]$$
$$- \mathbb{E}_{x \sim Q} \left[ \log \frac{2}{1 + e^{-V_w(x)}} \right]$$
$$= -\mathbb{E}_{x \sim P} \left[ \log \frac{1}{1 + e^{-V_w(x)}} \right]$$
$$- \mathbb{E}_{x \sim Q} \left[ \log \frac{1}{1 + e^{-V_w(x)}} \right] + \log 4$$

$$= -\mathbb{E}_{x \sim P} \left[ \log D(W_w(x)) \right]$$
$$- \mathbb{E}_{x \sim Q} \left[ \log (1 - D(W_w(x))) \right] + \log 4$$

We can omit the constant $\log 4$. If $P$ and $Q$ are balanced, the overlapping metric is equivalent to cross-entropy loss,

$$D_{PQ} = -\mathbb{E}_x[ y_x \log(D(W_w(x)))]$$
$$+ (1 - y_x) \log(1 - D(W_w(x)))$$  \hspace{1cm} (29)

where $y_x = 1(x \in P)$.

Due to the symmetry of JS divergence, it can be verified that the above indicator also holds symmetry property,

$$D_{PQ} = \mathbb{E}_{x \in P}[\log D_p(V_w(x))] + \mathbb{E}_{x \in Q}[\log (1 - D_p(V_w(x)))]$$
$$= \mathbb{E}_{x \in P}[\log (1 - D_q(V_w(x)))] + \mathbb{E}_{x \in Q}[\log D_q(V_w(x))]$$
$$= D_{QP}$$  \hspace{1cm} (30)

**References**


[2] SOLUTION MANUAL. Introduction to graph theory. 1
