Exploring and Utilizing Pattern Imbalance Supplementary Material

Shibin Mei, Chenglong Zhao, Shengchao Yuan, Bingbing Ni*
Shanghai Jiao Tong University, Shanghai 200240, China
{adair327, cl-zhao, sc_yuan, nibingbing}@sjtu.edu.cn

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 10 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 can not 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.

2. Theoretical Analysis

2.1. Upper Bound of the Number of Seed Categories

Once we determine a threshold ξ, 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 δ, which satisfies $\delta = \mathcal{F}(\xi)$, where $\mathcal{F}$ is a non-increasing function. Then according to Arnautov-Payan theorem [2], we can get the following theorem,

Theorem 1 *(Upper bound of seed category) For a dataset with N samples and threshold ξ, then the upper bound of seed categories satisfies,*

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

, where $|\mathcal{S}|$ 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 δ.

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,

$$\mathbb{E}(|X|) = Np \tag{2}$$

We regard the subset X as the candidate for seed set $\mathcal{S}$. 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

$$\begin{aligned} P(v \in Y_X) &= P(\text{v and its adjacent samples not in X}) \\ &= (1-p)^{1+d(v)} \\ &\leq (1-p)^{1+\delta} \end{aligned} \tag{3}$$

Then we can obtain,

$$\mathbb{E}(|Y_X|) \leq N(1-p)^{1+\delta} \tag{4}$$

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,

$$\begin{aligned} \mathbb{E}(|X \cup Y_X|) &\leq \mathbb{E}(|X| + |Y_X|) \leq \mathbb{E}(|X|) + \mathbb{E}(|Y_X|) \\ &= Np + N(1-p)^{1+\delta} \leq Np + Ne^{-p(1+\delta)} \end{aligned} \tag{5}$$

*Corresponding author.

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,

$$\frac{N[1+\ln(\delta+1)]}{\delta+1} \tag{6}$$

Therefore we can get

$$|\mathcal{S}| \leq \frac{N[1+\ln(\delta+1)]}{\delta+1}, \quad \text{where} \quad \delta = \mathcal{F}(\xi). \tag{7}$$

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 $b < |\mathcal{S}_i| < a$ if and only if the number of adjacent sample pairs satisfies,*

$$b^2\lfloor\frac{N}{b}\rfloor - N < E(N) \leq (a-1)^{\frac{1}{a}}N^{2-\frac{1}{a}} + (a-1)N. \tag{8}$$

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

$$E(N) \leq (1-\frac{1}{a-1})\frac{N^2}{2} \tag{9}$$

, where $|\mathcal{S}_i|$ represent the scale of a certain seed category $\mathcal{S}_i$.

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 $0-1$ matrix of $\mathcal{M} \in \mathbb{R}^{N\times N}$, where there is no all 1 sub-matrix of $\mathcal{M}_0 \in \mathbb{R}^{A\times A}$, 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 $C^a_{d(u)}$, then the total selections are $\sum_{u\in V_1} C^a_{d(u)} = |S|$. 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 $\mathcal{M}_0 \in \mathbb{R}^{A\times A}$. We can thus obtain,

$$\sum_{u\in V_1} C^a_{d(u)} \leq C^a_N(a-1) \tag{10}$$

Following Jensen Inequality, $f(x) = C^a_x$ is a convex function, then,

$$\sum_{u\in V_1}\frac{1}{N}C^a_{d(u)} \geq C^a_{\frac{1}{N}\sum_{u\in V_1} d(u)} = C^a_{\frac{|E|}{a}} \tag{11}$$

So,

$$\begin{aligned} NC^a_{\frac{|E|}{a}} &\leq \sum_{u\in V_1} C^a_{d(u)} \leq C^a_N(a-1) \\ &= \frac{N(N-1)...(N-a+1)}{a!}(a-1) \end{aligned} \tag{12}$$

Retraction is conducted at both sides, and then,

$$N\frac{(\frac{|E|}{N}-a+1)^a}{a!} < NC^a_{\frac{|E|}{a}} < \frac{N^a}{a!}(a-1) \tag{13}$$

After simplification, we can obtain,

$$E(N) \leq (a-1)^{\frac{1}{a}}N^{2-\frac{1}{a}} + (a-1)N \tag{14}$$

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 $C_\pi = \varnothing$, for all points, we introduce a random permutation $\mathcal{O} : x_1, x_2, x_3, ..., x_n$. For a certain permutation, if all points in front of x_i are x_i's neighboring points, we put x_i into C_π. Finally, all point pairs in C_π are neighboring points.

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

$$|C_\pi| = \sum_{x_i}\frac{1}{N-d(x_i)} \tag{15}$$

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

$$\omega(\mathcal{D}) \geq \sum_{x_i}\frac{1}{N-d(x_i)} \tag{16}$$

What we need to satisfy is,

$$a \geq \omega(G) \geq \sum_{v_i}\frac{1}{N-d(v_i)} \tag{17}$$

According to Cauchy Inequality,

$$a\sum_{v_i}(N-d(v_i)) \geq \sum_{v_i}\frac{1}{N-d(v_i)}\sum_{v_i}(N-d(v_i)) \geq N^2 \tag{18}$$

So,

$$a(N^2-2|E|) \geq N^2 \tag{19}$$

We can thus obtain,

$$|E| \leq \frac{N^2}{2}(1-\frac{1}{a-1}) \tag{20}$$

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_Δ-Lipschitz continuous and bounded by B_l, and model parameter $\|\theta\|_2 \leq B_\Theta$. The expectation error ϵ_t at step t satisfies,*

$$\mathbb{E}[\epsilon_t] \leq 2|\mathcal{S}|\sqrt{\frac{10(B_\Theta^2 B_\Delta^2 + B_l^2 \log|\mathcal{S}|)}{t}} \tag{21}$$

Proof:

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

$$\mathbb{E}[\epsilon_t] \leq 2m\sqrt{\frac{10(B_\Theta^2 B_\Delta^2 + B_l^2 \log m)}{t}} \tag{22}$$

, 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 $\mathcal{S}$, thus obtaining the convergence analysis of our method,

$$\mathbb{E}[\epsilon_t] \leq 2|\mathcal{S}|\sqrt{\frac{10(B_\Theta^2 B_\Delta^2 + B_l^2 \log|\mathcal{S}|)}{t}} \tag{23}$$

3. Proof of Calculating Seed Category by Loss Ranking

Supposing P and Q are two unit distributions,

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

Following f-GAN [3], we define f as the generator function $f : \mathbb{R}_+ \to \mathbb{R}$ of JS divergence, and f^* as its conjugate function, which is defined as $f^*(t) = \sup_{u \in dom\, f}\{ut - f(u)\}$. For JS divergence, let $f(u) = -(u+1)\log\frac{1+u}{2} + u\log u$ and correspondingly $f^*(t) = -\log(2 - e^t)$, then following f-GAN [3],

$$\begin{aligned} JS(P||Q) &= \int q(x) f(\frac{p(x)}{q(x)})dx \\ &= \int q(x) \sup_{t \in dom\, f^*}\left\{t\frac{p(x)}{q(x)} - f^*(t)\right\} \\ &\geq \sup_{T\in\mathcal{T}}\Big(\int p(x)T(x)dx - \int q(x)f^*(T(x))dx\Big) \\ &= \sup_{T\in\mathcal{T}} \mathbb{E}_{x\sim P}[T(x)] - \mathbb{E}_{x\sim Q}[f^*(T(x))] \end{aligned} \tag{25}$$

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

We assume that T is an optimized model with learnable parameters θ 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,

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

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,

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

Then,

$$\begin{aligned} D_{PQ} &= -\mathbb{E}_{x\sim P}[\log\frac{2}{1+e^{-V_w(x)}}] \\ &\quad - \mathbb{E}_{x\sim Q}[\log(2 - e^{\log\frac{2}{1+e^{-V_w(x)}}})] \\ &= -\mathbb{E}_{x\sim P}[\log\frac{1}{1+e^{-V_x(x)}}] \\ &\quad - \mathbb{E}_{x\sim Q}[\log\frac{e^{-V_w(x)}}{1+e^{-V_w(x)}}] + \log 4 \\ &= -\mathbb{E}_{x\sim P}[\log D(V_w(x))] \\ &\quad - \mathbb{E}_{x\sim Q}[\log(1 - D(V_w(x)))] + \log 4 \end{aligned} \tag{28}$$

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

$$\begin{aligned} D_{PQ} &= -\mathbb{E}_x[y_x \log(D(V_w(x))) \\ &\qquad + (1 - y_x)\log(1 - D(V_w(x)))] \\ &= \mathbb{E}_x[\mathbb{C}\mathbb{E}_{V_w(x)}(x)] \end{aligned} \tag{29}$$

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

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

$$\begin{aligned} D_{PQ} &= \mathbb{E}_{x\in P}[\log D_p(V_w(x))] + \mathbb{E}_{x\in Q}[\log(1 - D_p(V_w(x)))] \\ &= \mathbb{E}_{x\in P}[\log(1 - D_q(V_w(x)))] + \mathbb{E}_{x\in Q}[\log D_q(V_w(x))] \\ &= D_{QP} \end{aligned} \tag{30}$$

References

[1] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967. 1

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

[3] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 271–279, 2016. 3

[4] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 3

[5] Wojciech A Trybulec. Pigeon hole principle. *Journal of Formalized Mathematics*, 2(199):0, 1990. 2