Revealing the Proximate Long-Tail Distribution in Compositional Zero-Shot Learning

Chenyi Jiang, Haofeng Zhang *

School of Computer Science and Engineering, Nanjing University of Science and Technology, China
{jiangchenyi, zhanghf}@njust.edu.cn,

Abstract

Compositional Zero-Shot Learning (CZSL) aims to transfer knowledge from seen state-object pairs to novel unseen pairs. In this process, visual bias caused by the diverse interrelationship of state-object combinations blurs their visual features, hindering the learning of distinguishable class prototypes. Prevailing methods concentrate on disentangling states and objects directly from visual features, disregarding potential enhancements that could arise from a data viewpoint. Experimentally, we unveil the results caused by the above problem closely approximate the long-tailed distribution. As a solution, we transform CZSL into a proximate class imbalance problem. We mathematically deduce the role of class prior within the long-tailed distribution in CZSL. Building upon this insight, we incorporate visual bias caused by compositions into the classifier's training and inference by estimating it as a proximate class prior. This enhancement encourages the classifier to acquire more discernible class prototypes for each composition, thereby achieving more balanced predictions. Experimental results demonstrate that our approach elevates the model's performance to the state-of-the-art level, without introducing additional parameters.

Introduction

Objects in the world often exhibit diverse states of existence; an *apple* can be *sliced* or *unripe*, while a *building* can be *ancient* or *huge*. Humans have the ability to recognize the composition of the unseen based on their knowledge of seen elements. Even if people have never seen a *green apple* before, they can infer the characteristics of a *green apple* from a *red apple* and a *green lemon*. To empower the machine with this capability, previous work (Misra, Gupta, and Hebert 2017; Purushwalkam et al. 2019) propose Compositional Zero-Shot Learning (CZSL), a task aims to identify unseen compositions from seen state-object compositions.

However, the combination of state-objects creates a visual bias for a attribute (state or object) in it, hindering the learning of distinguishable class prototypes. In the face of above challenge, early approaches in the domain of CZSL can be categorized into two distinct methods. The first method utilized two independent classifiers to categorize states and objects (Misra, Gupta, and Hebert 2017; Li et al. 2020; Purushwalkam et al. 2019; Li et al. 2022). The second method involved training a common embedding space where semantic and visual features could be projected to reduce the distance between them (Naeem et al. 2021; Mancini et al. 2021, 2022). Commonly, these studies concentrate on improving the structure of classifiers and investigating alternative architectures. However, minimal research has been conducted considering the problem in terms of data distribution.

We analyze the prior and posterior probabilities associated with attributes (states or objects) and compositions to determine a more suitable solution. Fig. 1 illustrates that the class prior follows a distinct trend differing from the posterior probabilities. For instance, even though the model is trained on a comparable number of samples, it demonstrates a low probability of predicting the object labeled as *O5*. This issue also extends to making inferences about compositions, which reminds us of the long-tail distribution or class imbalance (Menon et al. 2020; Tang et al. 2020; Kim et al. 2020).

We consider that certain samples are infected by the intricate interplay between objects and states within compositions (Atzmon et al. 2020), leading to significant bias from the ideal class prototype. Consequently, these samples with large visual bias make it difficult for the classifier to fit their intrinsic patterns, results in the inability to form effective classification boundaries. In contrast to class imbalance, we refer to this phenomenon as *'attribute imbalance'* below. The recent methods for CZSL (Saini, Pham, and Shrivastava 2022; Wang et al. 2023) synchronize the prediction of visual features to states and objects with the prediction of compositions in the common embedding space, which works as a model ensemble approach. While this design addresses the capability to categorize some classes, the non-interaction among the independent classifiers may lead to incomplete mutual compensation due to potential information gaps.

The identified shortcomings prompted a redesign of the model using the model ensemble approach. Building on the success of logit adjustment in addressing long-tail learning (Menon et al. 2020), this study treats attribute imbalance information as special prior knowledge (In the following we denote by *'attribute prior'*) that approximates the class prior. This attribute prior is derived from the estimation of available samples by two independent classifiers for states and objects. In other words, we construct this prior by modelling

*Corresponding author

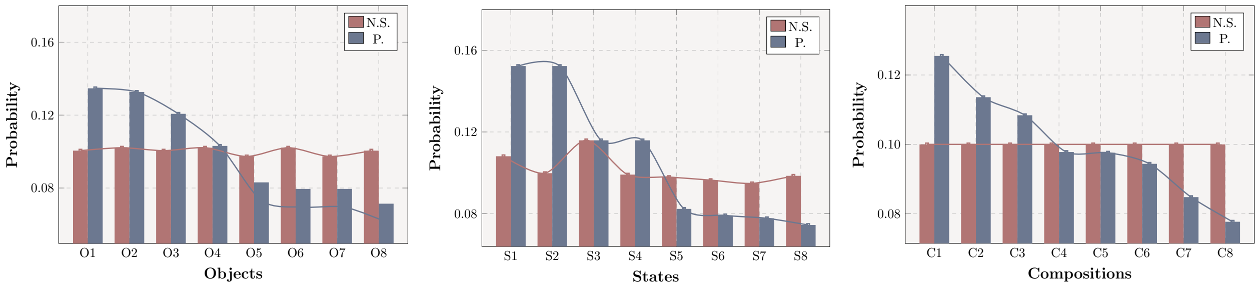

Figure 1: An example of prior and posterior probabilities (predicted by model) of same classes in MIT-States (Isola, Lim, and Adelson 2015). N.S. represents the prior probability calculated from the number of samples in each class. P. denotes the average value of posterior probabilities indicating the likelihood of the sample belonging to its class, which is predicted by a MLP with ResNet-18 (He et al. 2016) as backbone (same settings as $\mathcal{C}_o, \mathcal{C}_s$ and $\mathcal{C}_y$ in Implementation Details and trained via vanilla cross-entropy loss). The classes are selected on the basis of the closest sample size, and shows the results for the object (left), state (middle), and composition (right) classes. All data is simply normalized for ease of presentation.

the visual bias of states and objects from samples. During the training phase, we incorporate it through logit adjustment into the common embedding space. This approach enables the production of balanced posterior probabilities regarding the poorly-classified classes in Fig. 1, thereby preventing each independent classifier from ineffectively reinforcing the ability to classify the well-classified classes.

Specifically, we reconstructed the CZSL problem from the perspective of mutual information and adjusted the posterior values predicted by the model from the perspective of maximizing mutual information. In addition, we generalize the above attribute prior to the unseen class in order to optimize the lower bound of seen-unseen balanced accuracy (Xian, Schiele, and Akata 2017) obtained by Chen et al. (2022). We refer to this method as the logit adjustment for **Pro**ximate **L**ong-**T**ail Distribution (ProLT) in CZSL. Unlike previous methods, ProLT does not necessitate introducing additional parameters, yet it significantly enhances the overall CZSL model performance. Our contributions are summarized as follows:

- In our study, we conduct an analysis of the data distribution in CZSL. We translate the visual bias in compositions into an attribute imbalance and thereby generalize CZSL to a proximate long-tail learning problem.
- Our analysis involves a mathematical examination of both the training and inference phases of the model. This enables us to adapt the model's posterior probability based on the attribute prior.
- Our model enhances the prediction of relationships in compositions without the need for introducing additional parameters. Experimental results on three benchmark datasets demonstrate the effectiveness of our approach.

Related Work

Compositional Zero-Shot Learning (CZSL): Zero-Shot Learning (ZSL) transfers knowledge from seen classes to unseen ones by leveraging attributes (Akata et al. 2013; Lampert, Nickisch, and Harmeling 2013; Parikh and Grauman 2011; Frome et al. 2013; Akata et al. 2015). CZSL (Atzmon et al. 2020; Yang et al. 2022; Nagarajan and Grauman 2018; Wang et al. 2019) builds upon this foundation by incorporating the notion of composition learning (Hoffman and Richards 1984), with its extension primarily relying on the shared semantics of state and object within the composition of both seen and unseen classes.

Initial CZSL methodologies directly classify states and objects, effectively converting the task into a conventional supervised assignment (Misra, Gupta, and Hebert 2017; Chen and Grauman 2014; Yang et al. 2020; Lu et al. 2016; Li et al. 2022). However, the fusion of state-object pairs led to visual bias in both elements, impeding the acquisition of discernible class prototypes. Numerous subsequent strategies utilize visual-semantic alignment within a common embedding space (Naeem et al. 2021; Mancini et al. 2022, 2021) to grasp the entwined nature of objects and states within compositions. However, this technique is susceptible to domain shift challenges. Recent methodologies typically amalgamate these two models, creating a framework of model ensembles. For instance, Saini, Pham, and Shrivastava (2022) enhances the model's adaptability to unseen classes by disentangling visual features and subsequently reconstituting them for novel classes. Meanwhile, Wang et al. (2023) introduces conditional state generation to address visual alterations arising from object-state combination. ProLT aligns closely with this paradigm, although with a greater emphasis on direct inquiries into visual bias attributes.

Long-Tailed Classification: Numerous studies address the issue of imbalanced class distributions, with one prominent approach being posterior modification methods (Bunkhumpornpat, Sinapiromsaran, and Lursinsap 2012; Liu et al. 2019; Hou et al. 2021; Menon et al. 2020; Lin et al. 2017). Within ZSL, Chen et al. (2022) regards it as an imbalanced challenge involving seen and unseen classes, and then applies regulatory techniques based on logit adjustment. However, this approach does not readily extend to the issue of attribute imbalance in our context. Jiang et al. (2022) considers the presence of visual bias in samples re-weighting within the optimization process, but its localization-based weighting strategy ignores the differences between classes.

In this study, we introduce advanced logit adjustment strategies theoretically, aiming to enhance the equilibrium of predictions between various classes.

Methodology

Task Definition

Considering the two disjoint sets $\mathcal{Y}^S$ and $\mathcal{Y}^U$, *i.e.*, $\mathcal{Y}^S \cap \mathcal{Y}^U = \emptyset$. CZSL aims to classify sample $\mathbf{x} \in \mathcal{X}$ into a composition $y = (s, o) \in \mathcal{Y}$, where $\mathcal{Y} = \mathcal{Y}^S \cup \mathcal{Y}^U$, and samples from $\mathcal{Y}^U$ are unseen during training. y is composed by state $s \in \mathcal{S}$ and object $o \in \mathcal{O}$, $\mathcal{S}$ and $\mathcal{O}$ are sets of states and objects. Samples from $\mathcal{Y}^S$ and $\mathcal{Y}^U$ share the same objects o and states s, but their compositions (s, o) are different. Define the visual space $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and d_x is the dimension of the space, $\mathcal{X}$ can be divided into $\mathcal{X}^S$ and $\mathcal{X}^U$ based on whether their samples belong to seen classes. We can define the train set as $\mathcal{D}_{seen} = \{(\mathbf{x}, y)|\mathbf{x} \in \mathcal{X}^S, y \in \mathcal{Y}^S\}$ and an unseen set for evaluation of methods which is $\mathcal{D}_{unseen} = \{(\mathbf{x}, y)|\mathbf{x} \in \mathcal{X}^U, y \in \mathcal{Y}^U\}$. We employ the Generalized ZSL setup defined in Xian, Schiele, and Akata (2017), which requires both seen and unseen classes involves in testing.

Empirical Analysis on Model Ensemble

Method	M.E.	Sta.	Obj.	S.	U.	HM
MLP (Mancini et al. 2021)	F	27.9	31.8	25.3	24.6	16.4
	T	27.9	32.0	29.8	24.5	17.9
GCN (Naeem et al. 2021)	F	27.9	32.0	28.7	25.3	17.2
	T	28.3	33.4	28.9	26.0	18.8
I.C.	-	25.3	24.8	19.3	19.0	12.0

Table 1: The results of methods that incorporate a composition classifier, along with the addition of two classifiers for states and objects on top of them (*i.e.*, model ensemble), are presented. I.C. indicates only two independent classifiers are used. M.E. indicates the utilization of model ensemble in the methods, where F denotes false and T denotes true. The metrics in the table are defined in Evaluation Protocol.

For the problem of approximate long-tailed distributions caused by visual bias in CZSL, ensemble-based methods have demonstrated exceptional performance in CZSL (Saini, Pham, and Shrivastava 2022; Wang et al. 2023). Typically, this approach combines the predictions of two models to produce the final prediction. The first model consists of two independent classifiers C_o and C_s for objects and states. The second model is a composition classifier C_y. The process of the model can be viewed as inputting the samples into three classifiers to estimate the posterior probabilities:

$$
\begin{aligned}
&p(s|\mathbf{x}) = \mathrm{softmax}[C_s(\mathbf{x})], p(o|\mathbf{x}) = \mathrm{softmax}[C_o(\mathbf{x})],\\
&p(y|\mathbf{x}) = \mathrm{softmax}[C_y(\mathbf{x})],\\
&\hat{p}(y|\mathbf{x}) = \delta p(y|\mathbf{x}) + (1-\delta)\left[p(s|\mathbf{x}) + p(o|\mathbf{x})\right],
\end{aligned} \tag{1}
$$

where $p(s|\mathbf{x}), p(o|\mathbf{x})$ and $p(y|\mathbf{x})$ are posterior probability from classifiers, $\hat{p}(y|\mathbf{x})$ is the final posterior probabilities. δ is a weight factor. $\mathcal{C}_y(\mathbf{x})$ denotes the logits for class y based on sample $\mathbf{x}$, and $\mathcal{C}_s(\mathbf{x})$, $\mathcal{C}_o(\mathbf{x})$ are similarly defined.

As demonstrated in Tab. 1, augmenting two additional posterior estimates $p(o|\mathbf{x})$ and $p(s|\mathbf{x})$ to $p(y|\mathbf{x})$ can significantly enhance CZSL results. However, only relying solely on $p(o|\mathbf{x})$ and $p(s|\mathbf{x})$ does not enable accurate estimation, this suggests the improvement in results is not due to the introduction of superior classifiers. Consequently, we can deduce the subsequent conjectures: The effectiveness of ensemble-based methods emanates from incorporating $\mathcal{C}_s$ and $\mathcal{C}_o$, aiding in the classification of compositions that encounter a relative disadvantage within $\mathcal{C}_y$. While attribute imbalances vary across states, objects, and compositions, all three elements might not simultaneously experience large visual bias for a particular class. Based on these preliminary studies, we can posit that effective classification of classes with large visual bias within common embedding spaces requires information compensation. In our study, we directly estimate visual bias as compensation described above. Considering that the visual bias generated by state-object combination is difficult to eliminate directly, we try to introduce it as an attribute prior into the training process from the classifier. In the following, we detail this process.

From the Perspective of Mutual Information

Let us first consider the problem from a simple CZSL approach based on common embedding spaces like (Mancini et al. 2021; Naeem et al. 2021). The optimization objective of these methods can be viewed as the maximum likelihood:

$$
\mathrm{argmin}_\theta \, \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{seen}}[-logp(y|\mathbf{x})], \tag{2}
$$

where $p(y|\mathbf{x})$ is defined in Eq. 1, which denotes the distribution of compositions predicted by the model. θ denotes the model parameters.

Given the characteristics of CZSL, where each sample is associated with two labels, s and o, there is a conditionality between the two in the setup of the dataset, *i.e.*,

$$
p(y) = p(s, o) = p(o)p(s|o), \tag{3}
$$

$p(y)$ and $p(o)$ denotes class prior of class y and object o, and $p(s|o)$ is conditional class prior of s and o.

Inspired by Su (2020), we look at the above issues through the perspective of mutual information (Kraskov, Stögbauer, and Grassberger 2004), we have:

$$
\begin{aligned}
I(Y;X) &\approx \mathbb{E}_y D_{KL}[p(y|\mathbf{x})\|p(y)]\\
&= \sum_{\mathbf{x},y} p(\mathbf{x}, y) log \frac{p(y|\mathbf{x})}{p(y)}\\
&= \sum_{\mathbf{x},y} p(\mathbf{x}, y) log \frac{p(y|\mathbf{x})}{p(o)p(s|o)},
\end{aligned} \tag{4}
$$

where X and Y are discrete random variables corresponding to $\mathbf{x}$ and y, respectively, and S and O are similarly defined. D_{KL} represents the Kullback-Leibler divergence, while $p(\mathbf{x}, y)$ represents the joint probability of the class y and the visual feature $\mathbf{x}$. Due to the real posterior probability between y and $\mathbf{x}$ is unknown, we use $p(y|\mathbf{x})$ as an approximation. We can interpret the optimization of maximum likelihood as follows, based on the posterior term in Eq. 4,

$$
log \frac{p(y|\mathbf{x})}{p(o)p(s|o)} \sim \mathcal{C}_y(\mathbf{x}), \tag{5}
$$

which can be transfer to:

$$logp(y|\mathbf{x}) \sim \mathcal{C}_y(\mathbf{x}) + logp(o)p(s|o), \tag{6}$$

here, $\mathcal{C}_y(\mathbf{x})$ represents the logits for class y, defined in Eq. 1, $\sim$ denotes approximately equal. The expression on the right-hand side is re-normalized using the $\mathrm{softmax}$ function, *i.e.*,

$$\begin{aligned} &- logp(y|\mathbf{x}) \\ &\sim log\left[1 + \sum_{o_i \neq o}\sum_{s_j \neq s} \frac{p(o_i)p(s_j|o_i)}{p(o)p(s|o)} e^{\mathcal{C}_{\hat{y}}(\mathbf{x}) - \mathcal{C}_y(\mathbf{x})}\right] \\ &\sim log\left[1 + \sum_{o_i \neq o}\sum_{s_j \neq s} \left(\frac{p(o_i)p(s_j|o_i)}{p(o)p(s|o)}\right)^{\eta} e^{\mathcal{C}_{\hat{y}}(\mathbf{x}) - \mathcal{C}_y(\mathbf{x})}\right], \end{aligned} \tag{7}$$

where $\hat{y} = (s_i, o_i)$, and η is an adjustment factor. Eq. 7 demonstrates that by incorporating the class prior $p(s|o)$ and $p(o)$ for state s and object o, we can optimize the model's mutual information. Consequently, we approach the CZSL problem from the perspective of mutual information.

Estimating the Attribute Prior

The above idea comes from the logits adjustment (Menon et al. 2020) introduced to address class imbalance (Johnson and Khoshgoftaar 2019; Japkowicz and Stephen 2002), which demonstrate that the inclusion of a class prior enhances the maximization of mutual information, and we generalize it to CZSL task.

As stated in Introduction, we undertake the transformation of CZSL into an approximate long-tailed distribution issue caused by visual bias from state-object combinations. Our argument centers on the proposition that attribute imbalance within CZSL contributes to an approximate form of class imbalance, since visual bias hinders reduces the distinguishability of some of the samples. Therefore, exclusive reliance on the class prior is inadequate. Building upon this rationale, we propose to use the attribute prior to assume the function of the class prior within the long-tailed distribution, serving as an approximation.

We propose incorporating the model's conditional posterior probabilities as an approximation for this scenario. We continue to denote it as the 'prior' due to its function as a prior probability during the training process, despite being computed using posterior probability. Since attribute imbalance cannot be directly quantified from the dataset, we simulate it by utilizing the posterior probability of the additional classifiers, for $\mathbf{x}$ and its corresponding s, o, we have:

$$\hat{p}(s) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[p(s|\mathbf{x})], \hat{p}(o) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[p(o|\mathbf{x})], \tag{8}$$

where $\mathbf{x} \in \mathcal{D}_{seen}$, $p(s|\mathbf{x})$ and $p(o|\mathbf{x})$ are defined in Eq. 1, which are posterior probabilities from $\mathcal{C}_s$ and $\mathcal{C}_o$, we use their predicted expectations for all training samples as a special attribute prior. From this we can replace the class prior in Eq. 7 with following item:

$$k(s,o) = \mathrm{softmax}\left[\sigma(s,o)\hat{p}(s)\hat{p}(o)\right], \tag{9}$$

where $\sigma(s,o)$ is a function used to model the conditional nature of the composition, *i.e.*,

$$\sigma(s,o) = \begin{cases} 1 & (s,o) \in \mathcal{Y}^S \cup \mathcal{Y}^U, \\ 0 & else. \end{cases} \tag{10}$$

From this we obtain the final objective function according to Eq. 7:

$$\mathcal{L}_{cls} = log\left[1 + \sum_{o_i \neq o}\sum_{s_j \neq s}\left(\frac{k(s_j,o_i)}{k(s,o)}\right)^{\eta} e^{\mathcal{C}_{\hat{y}}(\mathbf{x}) - \mathcal{C}_y(\mathbf{x})}\right]. \tag{11}$$

Logit Adjustment for Inference

Due to the introduction of unseen classes in the inference phase we need to make additional adjustments. CZSL usually measures model performance in terms of $\mathcal{A}^H$ which denotes Harmonic Mean (HM) accuracy:

$$\mathcal{A}^H = 2/(\frac{1}{\mathcal{A}^S} + \frac{1}{\mathcal{A}^U}), \tag{12}$$

where $\mathcal{A}^S$, $\mathcal{A}^U$ denote seen and unseen accuracy. Chen et al. (2022) provides a lower bound of HM, below we briefly describe its conclusions. For HM's lower bound we have:

$$\mathcal{A}^H \geq 1/\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\frac{|\mathcal{Y}|p(\mathcal{Y})p(y|y\in\mathcal{Y})}{q(\mathcal{C}_{out}=y|\mathbf{x})p(y|\mathbf{x})}, \tag{13}$$

where $q(\mathcal{C}_{out} = y|\mathbf{x})$ represents the probability of predicting class y using our model. The set $\mathcal{Y}$ can be either $\mathcal{Y}^S$ or $\mathcal{Y}^U$, $p(y|y \in \mathcal{Y})$ represents the conditional class prior, and $|\mathcal{Y}|p(\mathcal{Y})$ can be seen as a hyper-parameter that quantifies the differences between seen and unseen classes. Considering that the gap between the domains of seen and unseen classes in CZSL is not significant, we can simply treat $|\mathcal{Y}|p(\mathcal{Y})$ as an ignorable constant in the following process.

Finding the Bayesian optimum for $\mathcal{A}^H$ is difficult. However, it is possible to maximize its lower bound, which is equal to minimizing the upper bound of its inverse, *i.e.*, the denominator term of Eq. 13 is minimized if:

$$\tilde{y} = \mathrm{argmax}_y\left[\mathcal{C}_y(\mathbf{x}) + \eta logp(y|y\in\mathcal{Y})\right], \tag{14}$$

where η is from Eq. 11, $\tilde{y}$ is the predicted label for sample $\mathbf{x}$. For conditional class prior $p(y|y \in \mathcal{Y}^S)$, which represents the true class frequency when y belong to seen classes. Following Eq. 11, we similarly replace the prior with the attribute prior estimate in Eq. 9 here, which is:

$$p(y|(s,o)\in\mathcal{Y}^S) := k(s,o), \quad (s,o)\in\mathcal{Y}^S, \tag{15}$$

and the attribute prior of unseen classes are not available to the model, we model it here using a combination of the estimation from Importance Sampling (Neal 2001) with the attribute prior from seen samples, which can be denoted as:

$$p(y|y\in\mathcal{Y}^U) := k(s,o) + \frac{\hat{k}_{\mathbf{x}}(s,o)}{\lambda k(s,o)}, \quad (s,o)\in\mathcal{Y}^U, \tag{16}$$

where $\frac{1}{\lambda}$ is a hyper-parameter denotes the distribution of $\mathbf{x}$. The above results are re-transformed into probability distributions in the actual calculation. And $\hat{k}_{\mathbf{x}}(s,o)$ is instance-based conditional posterior probability:

$$\hat{k}_{\mathbf{x}}(s,o) = \mathrm{softmax}\left[\sigma(s,o)p(s|\mathbf{x})p(o|\mathbf{x})\right], \tag{17}$$

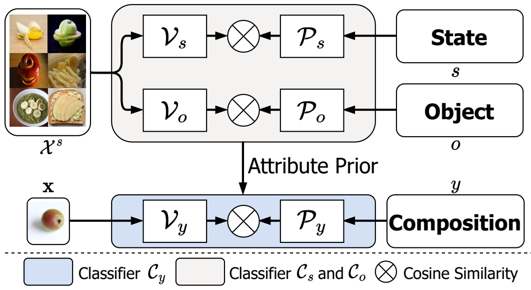

Figure 2: A brief demonstration of ProLT's training stage (detailed in Method Overview). $\mathcal{X}^s$ is the set of seen visual features, we obtain the attribute prior according to Eq. 8.

the aforementioned setup arises because during testing, we are unable to provide posterior probabilities $\hat{k}_{\mathbf{x}}(s, o)$ from multiple samples simultaneously. Furthermore, Importance Sampling results in significant variance when the number of samples is insufficient. To address this, we attempt to augment it by leveraging seen attribute prior.

ProLT makes inferences during testing phase based on Eq. 14, our aim is to integrate local information during testing with the prior derived from seen classes, to address the disparities between seen and unseen classes. With Eq. 14 ProLT theoretically achieves the best overall accuracy.

Method Overview

This section provides a concise summary of the aforementioned methods. Our approach, illustrated in Fig. 2, involves training two independent classifiers denoted as $\mathcal{C}_s$ and $\mathcal{C}_o$. These classifiers are implemented using prototype learners, namely $\mathcal{P}_s$ and $\mathcal{P}_o$, and visual embedders $\mathcal{V}_o$, $\mathcal{V}_s$, to determine the prototypes of states and objects, *i.e.*,

$$\mathcal{C}_s(\mathbf{x}) := \frac{cos(\mathcal{V}_s(\mathbf{x}), \mathcal{P}_s(s))}{\tau}, \mathcal{C}_o(\mathbf{x}) := \frac{cos(\mathcal{V}_o(\mathbf{x}), \mathcal{P}_o(o))}{\tau}, \tag{18}$$

where τ is the temperature. These classifiers are trained with vanilla cross-entropy loss:

$$\mathcal{L}_{ic} = log[1 + \sum_{s' \neq s} e^{\mathcal{C}_{s'}(\mathbf{x}) - \mathcal{C}_s(\mathbf{x})}][1 + \sum_{o' \neq o} e^{\mathcal{C}_{o'}(\mathbf{x}) - \mathcal{C}_o(\mathbf{x})}]. \tag{19}$$

Once the classifiers reach a specific training stage, we calculate the attribute prior using Eq. 8, and employ the loss function $\mathcal{L}_{cls}$ from Eq. 11 for training the classifier $\mathcal{C}_y$ for compositions:

$$\mathcal{C}_y(\mathbf{x}) := \frac{cos(\mathcal{V}_y(\mathbf{x}), \mathcal{P}_y(y))}{\tau}, \tag{20}$$

where $\mathcal{P}_y$ is the prototype learner for compositions and $\mathcal{V}_y$ is a visual embedder. After training, the model uses Eq. 14 for inference.

Experiments

Datas

There are numerous recent approaches to compositionality research, and three datasets have been primarily employed for evaluation: MIT-States (Isola, Lim, and Adelson 2015), UT-Zappos (Yu and Grauman 2014), and C-GQA (Naeem et al. 2021). We utilized a standardized evaluation dataset for a reasonable comparison with previous methods.

MIT-States presents a considerable challenge, consists of 53,753 images. It comprises 115 state classes, 245 object classes, and 1,962 compositions. In the total compositions, there are 1,262 seen compositions, and 700 compositions remain unseen. **UT-Zappos** is a collection of 50,025 images that focuses on various forms of footwear. It consists of 12 object classes and 16 state classes which is a fine-grained dataset, yielding 116 compositions, of which 83 are seen. **C-GQA** is introduced by Naeem et al. (2021), which encompasses a wide variety of real-world common objects. It comprises 413 states, 674 objects, and over 27,000 images, along with more than 9,000 compositions, consisting of 5,592 seen and 1,932 unseen compositions.

Evaluation Protocol

The setting of GZSL (Xian, Schiele, and Akata 2017) requires both seen and unseen compositions during testing. We report the **best accuracy** of **seen classes (best seen)**, the **unseen class (best unseen)**, and its **harmonic accuracy (HM)**. In order to measure the performance on attribute learning, we report the **best accuracy** of **states (best sta)** and **objects (best obj)**. Building upon the research of (Naeem et al. 2021) and (Wang et al. 2019), we calculate the **Area Under the Curve (AUC)** by comparing the accuracy on seen and unseen compositions with various bias terms.

Implementation Details

Below we present the details of the implementation of ProLT on ResNet-18 (He et al. 2016).

Visual Representations and Semantic: In line with prior methods, we employed ResNet-18 pre-trained on ImageNet (Deng et al. 2009) to extract 512-dimensional visual features from the images. For semantic information, we utilized GloVe (Pennington, Socher, and Manning 2014) to extract attribute names as 300-dimensional word vectors.

Implementations and Hyper-Parameters: For three prototype learner $\mathcal{P}_s, \mathcal{P}_o$ and $\mathcal{P}_y$ are GloVe connects with two Fully Connected (FC) layers with ReLU (Nair and Hinton 2010) following the first layer. And the three visual embedders $\mathcal{V}_s, \mathcal{V}_o$, and $\mathcal{V}_y$ are also two FC layers with ReLU and Dropout (Srivastava et al. 2014). All FCs embed the input features in 512 dimensions and the hidden layer is 1024 dimensions. The overall model is trained using the Adam optimizer (Kingma and Ba 2014) on NVIDIA GTX 2080Ti GPU, and it is implemented with PyTorch (Paszke et al. 2019). We set the learning rate as 5×10^{-4} and the batchsize as 128. We train the $\mathcal{C}_s, \mathcal{C}_o$ and $\mathcal{C}_y$ with an early-stopping strategy, it needs about 400 epochs on MIT-States, 300 epochs on UT-Zappos and 400 epochs on C-GQA. For hyper-parameters, we set τ as $0.1, 0.1, 0.01$, η as $1.0, 1.0, 1.0$ and λ as $50, 10, 100$ for MIT-States, UT-Zappos, and C-GQA, respectively.

	Methods	MIT-States						UT-Zappos						C-GQA					
		AUC	HM	S.	U.	Sta.	Obj.	AUC	HM	S.	U.	Sta.	Obj.	AUC	HM	S.	U.	Sta.	Obj.
	LE+ (2017)	2.0	10.7	15.0	20.1	23.5	26.3	25.7	41.0	53.0	61.9	41.2	69.2	0.8	6.1	18.1	5.6	-	-
	AttOp (2018)	1.6	9.9	14.3	17.4	21.1	23.6	25.9	40.8	59.8	54.2	38.9	69.6	0.7	5.9	17.0	5.6	-	-
	TMN (2019)	2.9	13.0	20.2	20.1	23.3	26.5	29.3	45.0	58.7	60.0	40.8	69.9	1.1	7.5	23.1	6.5	-	-
	SymNet (2020)	3.0	16.1	24.4	25.2	26.3	28.3	23.9	39.2	53.3	57.9	40.5	71.2	2.1	11.0	26.8	10.3	-	-
†	CompCos (2021)	4.5	16.4	25.3	24.6	27.9	31.8	28.7	43.1	59.8	62.5	44.7	73.5	2.6	12.4	28.1	11.2	-	-
	CGE (2021)	5.1	17.2	28.7	25.3	27.9	32.0	26.4	41.2	56.8	63.6	45.0	73.9	2.3	11.4	28.1	10.1	-	-
	SCEN (2022)	5.3	18.4	29.9	25.2	28.2	32.2	32.0	47.8	63.5	63.1	47.3	75.6	2.9	12.4	28.9	12.1	13.6	27.9
	Co-CGE (2022)	5.1	17.5	27.8	25.2	-	-	29.1	44.1	58.2	63.3	-	-	2.8	12.7	29.3	11.9		
	OADis (2022)	5.9	18.9	31.1	25.6	28.4	33.2	30.0	44.4	59.5	65.5	46.5	75.5	-	-	-	-	-	-
	DECA (2022)	5.3	18.2	29.8	25.5	-	-	31.6	46.3	62.7	63.1	-	-	-	-	-	-	-	-
	CANet (2023)	5.4	17.9	29.0	26.2	30.2	32.6	33.1	47.3	61.0	66.3	48.4	72.6	**3.3**	**14.5**	30.0	13.2	17.5	22.3
	ProLT (Ours)	**6.0**	**19.3**	30.9	26.5	29.5	34.2	**33.4**	**49.3**	62.7	64.0	46.1	74.2	3.2	14.4	32.1	13.7	17.8	32.5
	CSP (2022)	19.4	36.3	46.6	49.9	-	-	33.0	46.6	66.2	64.2	-	-	6.2	20.5	26.8	28.8	-	-
‡	DFSP (2023)	20.8	37.7	52.8	47.1	-	-	36.0	47.2	66.7	71.7	-	-	10.5	27.1	38.2	32.0		
	ProLT (Ours)	**21.1**	**38.2**	49.1	51.0	49.8	59.0	**36.1**	**49.4**	66.0	70.1	52.6	79.4	**11.0**	**27.7**	39.5	32.9	24.9	50.1

Table 2: The SoTA comparisons on three datasets. We compare ProLT with others on AUC, best HM, best sta (Sta.), best obj (Obj.), best seen (S.) and best unseen (U.). † denotes ResNet-based methods and ‡ denotes CLIP-based methods. The best AUC and HM for ResNet-based methods and CLIP-based methods are shown in bold.

Compared with State-of-the-Arts

ProLT is mainly compared with recent methods using fixed ResNet-18 as backbone with the same settings. We also compared ProLT with the CLIP-based approaches (Lu et al. 2023; Nayak, Yu, and Bach 2022) after using CLIP (Radford et al. 2021) to learn visual and semantic embeddings. The comparison results are shown in Tab. 2.

The results demonstrate that ProLT achieves a new state-of-the-art performance when using ResNet-18 as backbone on the MIT-States, UT-Zappos, and C-GQA. Specifically, our method achieves the highest AUC of 6.0% on MIT-States, surpassing CANet by 0.6%. On the UT-Zappos, we achieve the highest HM of 49.3%, outperforming CANet by 2.0%. Although ProLT has a slight disadvantage on the C-GQA dataset, it remains competitive with the state-of-the-art methods, achieving an HM of 14.4%. As for the CLIP-based approaches, ProLT has produced remarkable outcomes. Unlike DFSP, our method avoids the incorporation of extra self-attention or cross-attention mechanisms. Despite this, we excel across all three datasets, attaining an HM of 38.2% on MIT-States and 49.4% on UT-Zappos. These results underscore the compatibility of ProLT when combined with CLIP.

Ablation Study

In this section, we verify that each of these modules plays an active role by ablating each of its parts on UT-Zappos with ResNet-18. The results are shown in Tab. 3 and Tab. 4.

Attribute Prior versus Class Prior: As mentioned above, we use the attribute prior in place of the class prior due to the attribute imbalance. To further validate this, we replaced Eq. 11 and Eq. 14 using class prior, shown in Tab. 3. To make the results more robust, we tested two different prototype learners, *i.e.*, the GCN from the CGE (Naeem et al. 2021) and the FC layers. The results in Tab. 3 indicate that incorporating a class prior yields improvements over the baseline. We attribute this enhancement mainly to Menon et al. (2020), the class sizes of datasets are not solely identical. However, ProLT exhibits a substantial advantage over the other methods, which demonstrates the more dominant influence of potential attribute imbalances in CZSL.

Method	Prior	AUC	HM	S.	U.	Sta.	Obj.
GCN	N.	28.4	45.0	58.9	60.0	43.4	70.0
	C.P.	29.6	46.2	58.9	61.3	43.9	71.8
	A.P.	32.3	48.4	61.7	62.4	45.9	73.2
FC	N.	32.3	46.8	61.1	64.8	44.1	72.3
	C.P.	32.0	47.9	62.2	62.1	44.2	73.7
	A.P.	33.4	49.3	62.7	64.0	46.1	74.2

Table 3: A comparison of different priors for Eq. 11 and Eq. 14 on UT-Zappos. GCN: GCN is used as the prototype learner, FC: FC layers are used as the prototype learner. N. represents no prior is introduced, using pure ensemble method, C.P. represents class prior from datasets is utilized, and A.P. denotes attribute prior is incorporated.

Effect of Components: We eliminate the effects of each component by adjusting the hyper-parameters η in Eq. 11 and the attribute prior in Eq. 16 to verify the role played by each component. In Tab. 4, we set η to 0 to convert Eq. 11 to a vanilla cross entropy loss and the inference phase is converted to same as CGE. For $p = 0$, we remove the attribute prior in inference phase. We also tested on both prototype learners. We can observe that each part of the ablation leads to a decrease in outcome, with $\eta = 0$ being the most significant. This reflects the effectiveness of our method.

Hyper-Parameter Analysis

Our method primarily comprises the subsequent hyper-parameters: 1) logit-adjusting factor (η), and 2) factor about

Figure 3: Qualitative results on MIT-States (first row), UT-Zappos (second row) and C-GQA (third row), where the left part contains the top-3 results contains correct predicts, and the rights contains the top-3 predicts do not contain correct predicts. The label is indicated in black above the image, with correctly predicted results indicated in blue and incorrect ones in red.

Method	$\eta=0$	$p=0$	AUC	HM	S.	U.	Sta.	Obj.
GCN.	✓	×	29.4	44.2	59.9	63.0	43.4	70.0
	×	✓	31.3	47.3	60.8	60.5	43.7	73.6
	×	×	32.3	48.4	61.7	62.4	45.9	73.2
FC.	✓	×	28.9	44.1	60.1	62.9	44.1	73.2
	×	✓	32.7	48.3	61.8	64.6	45.8	74.1
	×	×	33.4	49.3	62.7	64.0	46.1	74.2

Table 4: Ablation results for each component on UT-Zappos. $p=0$ deontes we remove the prior in Eq. 14, ✓ indicates setting p or η to 0, and × indicates the opposite.

the sample distribution (λ). We test on the UT-Zappos under various hyper-parameters based on ResNet-18, shown in Fig. 4. For η, the best AUC are observed when $\eta = 1.6$, and the gap between seen and unseen begins to decrease as η increases. Concerning λ, the outcomes are documented within the interval $\lambda \in [1.0, 50.0]$ with increments about 5.0. The pinnacle value for the seen class is observed at 20.0, and 35.0 for unseen class. Overall, these hyper-parameter settings yield results characterized by minimal fluctuations, thus underscoring the robustness of our methodology.

Qualitative Results

Qualitative results for unseen compositions, accompanied by the top-3 predictions when we use ResNet-18 as backbone, are displayed in Fig. 3. Concerning MIT-States, we argue that certain erroneous predictions as partially justifiable. For instance, the phrase *tiny dog*, for which the model's incorrect predictions involve *small dog* and *tiny animal*, exhibits a high degree of semantic similarity. A similar phenomenon can be observed for the *brown chair* in C-GQA. For UT-Zappos, ProLT's limitation in fine-grained classification persists. An illustrative example is the outcomes for *leather boot.M*, our approach encounters challenges in making nuanced differentiations within the category of boots.

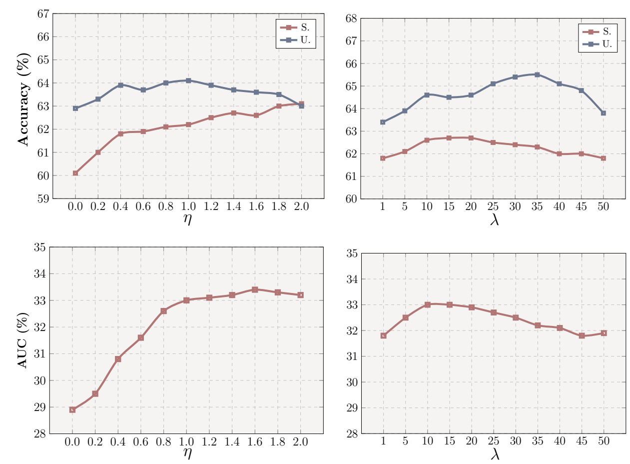

Figure 4: Influence of hyper-parameters on UT-Zappos about best seen (S.), best unseen (U.) and AUC.

Conclusion

This paper presents from an experimental analysis aimed at revealing the concealed proximate long-tail distribution issue within CZSL. In our work, CZSL is transformed into an underlying proximate class imbalance problem, and the logit adjustment technique is employed to refine the posterior probability for individual classes. Diverging from conventional methods for handling long-tailed distributions, the introduced attribute prior is derived from the model's sample estimation of visual bias. Experimental results demonstrate that our approach attains state-of-the-art outcomes without necessitating the introduction of supplementary parameters.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC) under the Grant No. 62371235.

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2013. Label-embedding for attribute-based classification. In *CVPR*, 819–826.

Akata, Z.; Reed, S.; Walter, D.; Lee, H.; and Schiele, B. 2015. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2927–2936.

Atzmon, Y.; Kreuk, F.; Shalit, U.; and Chechik, G. 2020. A causal view of compositional zero-shot recognition. *NeurIPS*, 33: 1462–1473.

Bunkhumpornpat, C.; Sinapiromsaran, K.; and Lursinsap, C. 2012. DBSMOTE: density-based synthetic minority oversampling technique. *Applied Intelligence*, 36: 664–684.

Chen, C.-Y.; and Grauman, K. 2014. Inferring analogous attributes. In *CVPR*, 200–207.

Chen, D.; Shen, Y.; Zhang, H.; and Torr, P. H. 2022. Zero-shot logit adjustment. *arXiv preprint arXiv:2204.11822*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Frome, A.; Corrado, G.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M.; and Mikolov, T. 2013. Devise: A deep visual-semantic embedding model. In *NeurIPS*, 2121–2129.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hoffman, D. D.; and Richards, W. A. 1984. Parts of recognition. *Cognition*, 18(1-3): 65–96.

Hou, Z.; Yu, B.; Qiao, Y.; Peng, X.; and Tao, D. 2021. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 14646–14655.

Isola, P.; Lim, J. J.; and Adelson, E. H. 2015. Discovering states and transformations in image collections. In *CVPR*, 1383–1391.

Japkowicz, N.; and Stephen, S. 2002. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5): 429–449.

Jiang, C.; Chen, D.; Wang, S.; Shen, Y.; Zhang, H.; and Shao, L. 2022. Mutual Balancing in State-Object Components for Compositional Zero-Shot Learning. *arXiv preprint arXiv:2211.10647*.

Johnson, J. M.; and Khoshgoftaar, T. M. 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1): 1–54.

Kim, D.-J.; Sun, X.; Choi, J.; Lin, S.; and Kweon, I. S. 2020. Detecting human-object interactions with action co-occurrence priors. In *ECCV*, 718–736. Springer.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kraskov, A.; Stögbauer, H.; and Grassberger, P. 2004. Estimating mutual information. *Physical review E*, 69(6): 066138.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2013. Attribute-based classification for zero-shot visual object categorization. *PAMI*, 36(3): 453–465.

Li, X.; Yang, X.; Wei, K.; Deng, C.; and Yang, M. 2022. Siamese Contrastive Embedding Network for Compositional Zero-Shot Learning. In *CVPR*, 9326–9335.

Li, Y.-L.; Xu, Y.; Mao, X.; and Lu, C. 2020. Symmetry and group in attribute-object compositions. In *CVPR*, 11316–11325.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *CVPR*, 2980–2988.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-scale long-tailed recognition in an open world. In *CVPR*, 2537–2546.

Lu, C.; Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2016. Visual relationship detection with language priors. In *ECCV*, 852–869. Springer.

Lu, X.; Guo, S.; Liu, Z.; and Guo, J. 2023. Decomposed soft prompt guided fusion enhancing for compositional zero-shot learning. In *CVPR*, 23560–23569.

Mancini, M.; Naeem, M. F.; Xian, Y.; and Akata, Z. 2021. Open world compositional zero-shot learning. In *CVPR*, 5222–5230.

Mancini, M.; Naeem, M. F.; Xian, Y.; and Akata, Z. 2022. Learning graph embeddings for open world compositional zero-shot learning. *PAMI*.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2020. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

Misra, I.; Gupta, A.; and Hebert, M. 2017. From red wine to red tomato: Composition with context. In *CVPR*, 1792–1801.

Naeem, M. F.; Xian, Y.; Tombari, F.; and Akata, Z. 2021. Learning graph embeddings for compositional zero-shot learning. In *CVPR*, 953–962.

Nagarajan, T.; and Grauman, K. 2018. Attributes as operators: factorizing unseen attribute-object compositions. In *ECCV*, 169–185.

Nair, V.; and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*.

Nayak, N. V.; Yu, P.; and Bach, S. H. 2022. Learning to compose soft prompts for compositional zero-shot learning. *arXiv preprint arXiv:2204.03574*.

Neal, R. M. 2001. Annealed importance sampling. *Statistics and computing*, 11: 125–139.

Parikh, D.; and Grauman, K. 2011. Relative attributes. In *ICCV*, 503–510. IEEE.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Purushwalkam, S.; Nickel, M.; Gupta, A.; and Ranzato, M. 2019. Task-driven modular networks for zero-shot compositional learning. In *ICCV*, 3593–3602.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Saini, N.; Pham, K.; and Shrivastava, A. 2022. Disentangling Visual Embeddings for Attributes and Objects. In *CVPR*, 13658–13667.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 15(1): 1929–1958.

Su, J. 2020. Mitigating class imbalances through mutual information ideology. https://spaces.ac.cn/archives/7615. Accessed: 2020-07-19.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased scene graph generation from biased training. In *CVPR*, 3716–3725.

Wang, Q.; Liu, L.; Jing, C.; Chen, H.; Liang, G.; Wang, P.; and Shen, C. 2023. Learning Conditional Attributes for Compositional Zero-Shot Learning. In *CVPR*, 11197–11206.

Wang, X.; Yu, F.; Darrell, T.; and Gonzalez, J. E. 2019. Task-aware feature generation for zero-shot compositional learning. *arXiv preprint arXiv:1906.04854*.

Xian, Y.; Schiele, B.; and Akata, Z. 2017. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 4582–4591.

Yang, M.; Deng, C.; Yan, J.; Liu, X.; and Tao, D. 2020. Learning unseen concepts via hierarchical decomposition and composition. In *CVPR*, 10248–10256.

Yang, M.; Xu, C.; Wu, A.; and Deng, C. 2022. A decomposable causal view of compositional zero-shot learning. *IEEE Transactions on Multimedia*.

Yu, A.; and Grauman, K. 2014. Fine-grained visual comparisons with local learning. In *CVPR*, 192–199.