Zero-Shot Image Classification

Chandrabhushan, Sai Ram Mohan, Manideepak, Samuel and Pramodh

***Abstract*— Zero-shot learning (ZSL) aims to predict unseen classes whose samples have never appeared during training. As annotations for class-level visual characteristics, attributes are among the most effective and widely used semantic information for zero-shot image classification. However, the current methods often fail to discriminate those subtle visual distinctions between images due to not only the lack of fine-grained annotations, but also the issues of attribute imbalance and co-occurrence. Generalized zero-shot learning (GZSL) has achieved significant progress, with many efforts dedicated to overcoming the problems of visual-semantic domain gap and seen-unseen bias. However, most existing methods directly use feature extraction models trained on ImageNet alone, ignoring the cross-dataset bias between ImageNet and GZSL benchmarks. Such a bias inevitably results in poor-quality visual features for GZSL tasks, which potentially limits the recognition performance on both seen and unseen classes. Zero-shot learning (ZSL) tackles the novel class recognition problem by transferring semantic knowledge from seen classes to unseen ones. Semantic knowledge is typically represented by attribute descriptions shared between different classes, which act as strong priors for localizing object attributes that represent discriminative region features, enabling significant and sufficient visual-semantic interaction for advancing ZSL. Existing attention-based models have struggled to learn inferior region features in a single image by solely using unidirectional attention, which ignore the transferable and discriminative attribute localization of visual features for representing the key semantic knowledge for effective knowledge transfer in ZSL.**

I. INTRODUCTION

Generalized Zero-Shot Learning (GZSL) identifies unseen categories by knowledge transferred from the seen domain, relying on the intrinsic interactions between visual and semantic information. Prior works mainly localize regions corresponding to the sharing attributes. When various visual appearances correspond to the same attribute, the sharing attributes inevitably introduce semantic ambiguity, hampering the exploration of accurate semantic-visual interactions. Zero-shot learning (ZSL) aims to recognize classes that do not have samples in the training set. One representative solution is to directly learn an embedding function associating visual features with corresponding class semantics for recognizing new classes. Many methods extend upon this solution, and recent ones are especially keen on extracting rich features from images, e.g., attribute features. These attribute features are normally extracted within each individual image; however, the common traits for features across images yet belonging to the same attribute are not emphasized.

II. RELATED WORK

A. FREE: Feature Refinement for Generalized Zero-Shot Learning

This paper proposes a method called Feature Refinement for Generalized Zero-Shot Learning (FREE). This method addresses the challenges of generalized zero-shot learning (GZSL), which aims to overcome the visual-semantic domain gap and seen-unseen bias in machine learning tasks1.

FREE incorporates a feature refinement (FR) module into a unified generative model. This module refines the visual features of both seen and unseen class samples by incorporating semantic-visual mapping. To guide the FR module, the authors propose a self-adaptive margin center loss (SAMC-loss) and a semantic cycle-consistency loss. These losses help FR learn class- and semantically-relevant representations. The refined features are then concatenated to extract fully refined features

B. DUET: Cross-modal Semantic Grounding for Contrastive Zero-shot Learning

This paper presents a transformer-based end-to-end zero-shot learning (ZSL) method named DUET. The goal of ZSL is to predict unseen classes that have never appeared during training. One of the most effective and widely used semantic information for zero-shot image classification are attributes, which are annotations for class-level visual characteristics.

DUET integrates latent semantic knowledge from pre-trained language models (PLMs) using a self-supervised multi-modal learning paradigm. It incorporates a cross-modal semantic grounding network to disentangle semantic attributes from images and applies an attribute-level contrastive learning strategy to enhance the model's discrimination on fine-grained visual characteristics. A multi-task learning policy is also proposed to consider multi-model objectives.

C. TransZero++: Cross Attribute-Guided Transformer for Zero-Shot Learning

This paper proposes a method called TransZero++. This method aims to refine visual features and learn accurate attribute localization for semantic-augmented visual embedding representations in zero-shot learning (ZSL).

TransZero++ consists of an attribute $\rightarrow$ visual Transformer sub-net (AVT) and a visual $\rightarrow$ attribute Transformer sub-net (VAT). The AVT module employs a feature augmentation encoder to alleviate the cross-dataset problem and improve the transferability of visual features. It also employs an attribute $\rightarrow$ visual decoder to localize the image regions most relevant to each attribute in a given image for attribute-based visual feature representations. Similarly, the VAT module refines

visual features using a similar feature augmentation encoder and applies them in a visual → attribute decoder to learn visual-based attribute features. By introducing semantical collaborative losses, the two attribute-guided transformers teach each other to learn semantic-augmented visual embeddings via semantical collaborative learning.

D. Progressive Semantic-Visual Mutual Adaption for Generalized Zero-Shot Learning

This paper proposes a method called Progressive Semantic-Visual Mutual Adaption (PSVMA). This method addresses the challenges of Generalized Zero-Shot Learning (GZSL), which aims to identify unseen categories by transferring knowledge from the seen domain.

PSVMA introduces a dual semantic-visual transformer module (DSVTM) to progressively model the correspondences between attribute prototypes and visual features. This module constitutes a progressive semantic-visual mutual adaption network for semantic disambiguation and knowledge transferability improvement. The DSVTM includes an instance-motivated semantic encoder that learns instance-centric prototypes to adapt to different images, enabling the recast of unmatched semantic-visual pairs into matched ones. It also employs a semantic-motivated instance decoder to strengthen accurate cross-domain interactions between matched pairs for semantic-related instance adaption, encouraging the generation of unambiguous visual representations. Additionally, a debiasing loss is proposed to mitigate the bias towards seen classes in GZSL.

E. Boosting Zero-Shot Learning via Contrastive Optimization of Attribute Representations

The paper proposes a new framework to boost zero-shot learning (ZSL) by explicitly learning attribute prototypes beyond images and contrastively optimizing them with attribute-level features within images. The authors highlight two key elements for attribute representations: a prototype generation module that generates attribute prototypes from attribute semantics, and a hard example-based contrastive optimization scheme that reinforces attribute-level features in the embedding space. The proposed framework is built using two alternative backbones: CNN-based and transformer-based

III. LIMITATIONS OF EXISTING APPROACHES

A. FREE: Feature Refinement for Generalized Zero-Shot Learning

FREE is a generative model, and as such, it is susceptible to overfitting to the seen classes. This is especially true if the number of unseen classes is large. Also, the performance of FREE depends on the quality of the semantic embedding used. If the embedding is not well-learned, the feature refinement module may not be able to learn class- and semantically-relevant representations.

To mitigate overfitting to seen classes, we can use regularization techniques, such as data augmentation and dropout. we can also use a two-stage training approach, where the feature refinement module is trained separately from the rest of the model.

To improve the robustness to the quality of the semantic embedding, we can use a pre-trained semantic embedding that has been shown to perform well on other tasks. Fine-tune the semantic embedding during the training of FREE. We can also use a more robust semantic embedding method, such as a graph-based embedding.

B. DUET: Cross-modal Semantic Grounding for Contrastive Zero-shot Learning

There are ways to improve performance on CUB, we could consider involving 2D relative geometry relationships via adding relative position encoding. Meanwhile, due to the intra-class image differences (images of the sample bird class in CUB are different due to e.g., different image shooting angles and different bird ages), we think DUET could be further improved by e.g., applying instance-level feature similarity as an auxiliary sample filtering strategy on the basis of our attribute-level contrastive learning (ACL) method, or pre-grounding the object, to get higher quality positive / negative samples.

C. TransZero++: Cross Attribute-Guided Transformer for Zero-Shot Learning

TransZero++ is a more complex model than many other ZSL methods, and as such, it is more computationally expensive to train and test. To reduce the computational cost we can use a smaller and more efficient XAT-Net architecture. We can also use pre-trained visual and semantic encoders to reduce the computational complexity further.

D. Progressive Semantic-Visual Mutual Adaption for Generalized Zero-Shot Learning

The proposed PSVMA network relies heavily on pre-trained models like ViT and GloVe. The effectiveness of the model might be limited by the quality and relevance of these pre-trained representations. Changes or advancements in pre-trained models could impact the model's performance. Also PSVMA has a number of hyperparameters that need to be tuned carefully to achieve optimal performance. This can be time-consuming and challenging.

To reduce the sensitivity to hyperparameters we can use a hyperparameter optimization algorithm to automatically search for the best hyperparameter values.

E. Boosting Zero-Shot Learning via Contrastive Optimization of Attribute Representations

One shortcoming of this method is that it relies on human-crafted attributes given in the dataset and cannot be applied to datasets without explicit attribute information. This limits the generalizability of this method. A possible solution may be directly learning shared prototypes across classes to simulate the attributes.

Besides above, there are also other places that can be improved: for instance, they currently use hard samples for the contrastive learning of attribute features. This may be

improved using the graph neural network where attribute relationships can be encoded into graph edges, so that attribute features can be optimized and refined via the message passing in the graph.

IV. Proposed Method

Currently the TransZero++ model is able to give SOTA results on benchmarks like CUB dataset but not on SUN dataset because the number of images in a single class are very less in the latter ones (approx. 16 training images per class). With this few number of training examples, the TransZero++ based ZSL does not give efficient semantic augmented visual embeddings. Therefore this heavily limits the performance of ZSL model.

Since per class training examples are very limited, we can improve the model performance by data augmentation using generative models like Variational Auto Encoders (VAEs), Generative Adversarial Nerworks (GANs) or Generative Flows.

Also currently the Attribute Regression Loss and Attribute based Cross Entropy Loss optimise the model on seen classes only. This causes the TransZero++ model to overfit on the seen classes.

Inspired from "Progressive Semantic-Visual Mutual Adaption for Generalized Zero-Shot Learning" research paper, we have introduced a "Debiasing Loss" to mitigate this seen-unseen bias we have. The debiasing loss (denoted by $\mathcal{L}_{deb}$) is formulated as follows

$$\mathcal{L}_{deb} = ||\alpha_s - \alpha_u||^2 + ||\beta_s - \beta_u||^2$$

Here α_s and α_u represent the mean and variance values of seen prediction scores and β_s and β_u represent the mean and variance values of unseen prediction scores.

We have introduced another loss function called "Correlation Loss". The final class which we get after the zero shot prediction, should be closer to the classes which are similar to the actual class and farther from the remaining classes. The similarity between the classes is calculated using cosine similarity.

So the final loss is given by
$\mathcal{L}_{total} = \mathcal{L}_{AVT} + \lambda_{VAT}\mathcal{L}_{VAT} + \lambda_{fsCL}\mathcal{L}_{fsCL} + \lambda_{psCL}\mathcal{L}_{psCL} + \mathcal{L}_{deb} + \mathcal{L}_{corel}$

Here,
$\mathcal{L}_{AVT}$ represents Attribute-Vision Transformer Loss
$\mathcal{L}_{VAT}$ represents Vision-Attribute Transformer Loss
$\mathcal{L}_{fsCL}$ represents Feature based Colloboration Loss
$\mathcal{L}_{psCL}$ represents Prediction based Colloboration Loss
$\mathcal{L}_{deb}$ represents Debiasing Loss
$\mathcal{L}_{corel}$ represents Correlation Loss

V. Results

Due to hardware limitations we were only able to run our model on the CUB dataset with batch size 10 and number of epochs 50. The final results are as follows:

Category	**Accuracy**
Seen	71.12%
Unseen	66.34%
Harmonic Mean	68.65%

References

[1] Shiming Chen. (2023). TransZero++: Cross Attribute-Guided Transformer for Zero-Shot Learning.
[2] Man Liu. (2022). Progressive Semantic-Visual Mutual Adaption for Generalized Zero-Shot Learning.
[3] Zhuo Chen. (2023). DUET: Cross-modal Semantic Grounding for Contrastive Zero-shot Learning.
[4] Yu Du. (2023). Boosting Zero-Shot Learning via Contrastive Optimization of Attribute Representations
[5] Shiming CHen. (2021). FREE: Feature Refinement for Generalized Zero-Shot Learning.