JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Effectiveness of Diffusion Models in Zero-Shot Learning

P Lokeshwar Reddy and Vandana Bhattacharjee
Birla Institute of Technology, Mesra

Abstract

Zero-Shot Learning (ZSL) is a transformative machine learning approach that enables models to classify unseen classes without needing labeled data for those categories, making it highly valuable in data-scarce environments. This paper compares two significant ZSL methods: ZeroDiff and the Diffusion Classifier, exploring their strengths, weaknesses, and appropriate use cases. ZeroDiff is a Diffusion-based Generative ZSL model that excels at creating features for unseen classes, enhancing generalization through its dual-branch architecture, which integrates feature generation with semantic alignment. While effective, it requires substantial computational resources, making it less suitable for environments with limited hardware. In contrast, the Diffusion Classifier uses pretrained models to perform zero-shot classification by estimating class-conditional densities. Its ability to classify unseen data with minimal retraining makes it computationally efficient, but its performance heavily depends on the quality of the pretrained model, which can limit its versatility in handling a wide variety of unseen classes. The paper discusses the trade-offs between flexibility in feature generation and efficiency in classification and proposes the potential of hybrid models that combine both strengths. It also highlights key challenges, such as domain shifts and data scarcity, and suggests future research directions to improve model robustness, scalability, and adaptability to different domains.

1 INTRODUCTION

Zero-Shot Learning (ZSL) represents a breakthrough in machine learning, enabling models to recognize objects, concepts, or data categories they have not encountered during training. Unlike traditional methods that require extensive labeled data for each class, ZSL relies on auxiliary information such as textual descriptions, semantic embeddings, or class attributes to bridge the knowledge gap between seen and unseen classes. The model learns to map both visual data and semantic representations into a shared space, where similarities and relationships can be inferred.

This approach is transformative for domains where gathering labeled data is impractical or impossible. In healthcare, for instance, diseases with rare cases lack sufficient training examples, yet ZSL can help classify these conditions by leveraging knowledge from related diseases. Similarly, in environmental monitoring or autonomous systems, it allows machines to identify novel species or objects without explicit training. A more advanced form of ZSL, Generalized Zero-Shot Learning (GZSL), expands the challenge by requiring models to generalize across both seen and unseen classes, reflecting real-world scenarios more accurately.

1.1 Role of Diffusion Models in Computer Vision

Diffusion models have recently become a cornerstone of generative AI, particularly in computer vision[1-2]. These models work by incrementally corrupting data with noise in a forward process and then learning to denoise it step-by-step in reverse. This probabilistic, iterative process allows diffusion models to capture intricate data distributions, making them exceptional at generating high-quality, realistic outputs. Their success spans applications like image synthesis, restoration, and inpainting, where the ability to reconstruct or generate content with fine details is crucial.

In the context of ZSL, diffusion models offer unique capabilities. By generating synthetic data or features, they address one of ZSL's biggest challenges: the absence of training data for unseen classes. Diffusion models can create rich, diverse representations that reflect the characteristics of unseen classes, enabling classifiers to perform better in these scenarios. Moreover, their ability to learn complex relationships and structures within data aligns well with the requirements of ZSL, where semantic and visual information must be intricately linked. Beyond generation, diffusion models are now being adapted for discriminative tasks like classification, further broadening their impact in computer vision.

1.2 Importance of Data Efficiency and Classification

Data efficiency is a critical challenge in machine learning, particularly in scenarios like ZSL, where labeled examples for unseen classes are nonexistent. Traditional supervised approaches, which depend on large, annotated datasets, fall short in such cases. Efficient use of data, whether by generating synthetic features or optimizing learning processes, is vital for enabling robust generalization in ZSL.

Diffusion models play a pivotal role in this aspect by synthesizing high-quality data that bridges the gap between seen and unseen classes. For instance, generating feature representations for unseen classes not only enhances the training set but also ensures that classifiers have a richer understanding of the unseen domain. This reduces dependency on manual data annotation and enables the model to perform well even in resource-constrained settings.

Equally important is the model's ability to maintain classification accuracy across both seen and unseen classes, particularly in GZSL settings. Models must balance generalization to unseen classes without compromising performance on seen ones. Achieving this balance is essential for deploying ZSL in real-world applications where both accuracy and reliability are paramount. Effective classification strategies paired with data-efficient methods ensure that ZSL can be applied in critical fields like healthcare diagnostics, autonomous vehicles, and beyond.

2 BACKGROUND

Zero-Shot Learning (ZSL) offers a transformative approach to machine learning by enabling models to recognize and classify data from unseen classes, without needing direct training examples from those classes. However, while ZSL provides significant benefits, it faces several inherent challenges. One of the most significant hurdles is the semantic gap—the difficulty of aligning visual features, such as pixels or shapes in images, with semantic descriptors like textual descriptions or class attributes. These visual and semantic domains are fundamentally different, making it difficult for models to accurately map them to each other, particularly when subtle visual distinctions do not align well with textual labels. Furthermore, domain shift presents a substantial challenge, especially in Generalized Zero-Shot Learning (GZSL), where models must perform well not only on seen classes but also on previously unseen ones. The distribution of seen data is often different from unseen classes, making it hard for the model to generalize effectively. ZSL also suffers from incomplete semantic representations, where the predefined attributes or embeddings may not fully capture the characteristics of unseen classes. For example, attributes like "large size" or "long tail" may not be enough to capture the full range of a class, limiting the model's ability to accurately classify unseen data. Additionally, models are often prone to overfitting to seen classes, leading to a bias where they perform well on seen classes but poorly on unseen ones. This imbalance in performance is particularly problematic in GZSL, where the model must balance predictions across both seen and unseen categories. Lastly, data imbalance and scalability remain critical issues, as many ZSL datasets suffer from underrepresented classes, complicating the task of generalizing to new data. As models scale to larger and more complex datasets, these challenges become even more pronounced.

To overcome these challenges, researchers have developed several approaches, each with its strengths and weaknesses. Embedding-based methods have been a traditional solution, aiming to map visual features and semantic descriptions into a shared latent space where relationships between classes can be inferred. These methods, however, often struggle with domain shifts and fail to generalize well to unseen classes. In contrast, generative models like

Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have been used to synthesize features for unseen classes by leveraging semantic attributes as input [3-8]. These synthetic features are then used to train classifiers. While these generative models are more flexible, they still face issues like mode collapse in GANs and low-quality sample generation in VAEs. Hybrid models have also emerged to combine the strengths of both embedding and generative methods, helping improve the synthesis of features while maintaining strong semantic alignment. Another growing area is contrastive learning, where models are trained to maximize similarity between positive pairs (e.g., an image and its corresponding text) while minimizing the distance between negative pairs. This method improves the alignment between visual and semantic features but often requires careful tuning and substantially labeled data to perform well. Some recent research has also explored Few-Shot Learning (FSL) in conjunction with ZSL, aiming to leverage a small number of labeled examples from unseen classes to improve generalization. This hybrid approach, though effective, moves away from the traditional definition of zero-shot learning.

In recent years, diffusion models have emerged as a promising tool to address these challenges in ZSL[9-14]. These models generate data by first corrupting it with noise through a forward diffusion process and then learning to reverse this corruption to reconstruct the original data. This process allows diffusion models to capture complex, high-dimensional data distributions and generate realistic, diverse samples. When applied to ZSL, diffusion models can create feature representations for unseen classes by using semantic information, thus helping bridge the semantic gap. By synthesizing realistic data for unseen classes, diffusion models also help improve data efficiency—one of the major limitations of traditional ZSL methods. Furthermore, by generating diverse features, these models can reduce the bias toward seen classes, making them effective in generalized ZSL tasks. Additionally, since diffusion models are inherently generative, they allow for the creation of high-quality features even from small amounts of data, which is crucial in domains where labeled data for unseen classes is limited. Through their flexibility and ability to model complex relationships within data, diffusion models are becoming an essential tool for tackling the challenges of zero-shot learning.

3 METHODOLOGY

3.1 Architectural Approaches

In Zero-Shot Learning (ZSL), the use of diffusion models has introduced novel architectural approaches that address the critical challenges of feature generation, data efficiency, and classification. **ZeroDiff** and the **Diffusion Classifier** are two such models that utilize diffusion processes, but they do so with differing architectures and strategies.

ZeroDiff's Two-Branch Structure

The ZeroDiff model is structured around a two-branch architecture, a design that distinguishes it from traditional generative models used in ZSL. This approach aims to enhance data efficiency and improve both instance-level and class-level performance, which are crucial when dealing with limited labeled data for unseen classes.

The first branch of this architecture, the Diffusion-based Feature Generator (DFG), focuses on instance-level efficiency by generating realistic feature representations for unseen classes. The DFG works by taking noisy input data (as generated through a forward diffusion process) and progressively refining it to produce features that resemble those of the unseen classes. The core idea here is that by applying a diffusion process, the model can create a wide variety of plausible instances that align with the distribution of the unseen classes, even when no direct data from those classes are available. This process essentially enables the model to expand the training data for unseen classes without requiring additional labeled data.

The second component, the Diffusion-based Representation Generator (DRG), enhances the class-level efficiency. The DRG's primary role is to generate representations that are aligned with semantic information, which may not be fully captured by the raw feature generation of the DFG. By using supervised contrastive learning, the DRG aims to capture richer, more nuanced representations of classes, which allows for a better understanding of the

relationships between seen and unseen categories. This is particularly important in ZSL, where semantic information can bridge the gap between the visual data of seen classes and the unobserved features of unseen classes.

To improve the generative capabilities, ZeroDiff also integrates discriminators that evaluate the generated features. These discriminators assess the generated data from multiple perspectives: they evaluate how well the features match predefined semantics (e.g., labels or attributes), the effectiveness of the diffusion process (i.e., how well the data has been denoised), and the alignment of the generated features with contrastive representations. These discriminators help refine the generated features by providing valuable feedback, improving the overall quality and diversity of the synthesized data.

Through this two-branch structure, ZeroDiff effectively combines generative and discriminative components, ensuring that it can generate robust features for unseen classes and enhance the semantic richness of these features, which is essential for successful zero-shot learning.

Diffusion Classifier's Density Estimation Approach

The Diffusion Classifier, in contrast, adopts a fundamentally different approach by focusing on density estimation for zero-shot classification. Rather than generating new data for unseen classes, the Diffusion Classifier uses a pretrained diffusion model to estimate the class-conditional densities for unseen categories.

The process begins with the reverse denoising phase of the diffusion model, which reconstructs clean data from noisy inputs. This reverse process is used in the context of zero-shot classification to estimate the likelihood of a given class for an unseen sample. In the standard diffusion model, the goal is to model the data distribution through a series of probabilistic transformations. In the case of the Diffusion Classifier, these transformations are used to generate a probability distribution over the possible classes, with the model estimating the likelihood that a given sample belongs to a particular class.

The Diffusion Classifier does this by using a Monte Carlo sampling technique to compute class-conditional probabilities. By sampling multiple instances of the reverse diffusion process for each class, the classifier estimates the likelihood that an input data point belongs to each class. The model relies on the variational lower bound (ELBO), a technique commonly used in generative models to optimize the likelihood of the observed data, to guide the classification process. This approach does not require retraining for unseen classes, as the classifier leverages the pretrained generative model to perform zero-shot classification.

One of the key advantages of the Diffusion Classifier's approach is that it does not require additional training on unseen classes. Instead, it uses the pretrained diffusion model's knowledge of the data distribution to perform classification. This results in significant computational savings and allows the model to quickly adapt to unseen classes with minimal overhead. Additionally, this approach ensures that the model can handle distribution shifts—a common problem in ZSL—by utilizing the generative power of diffusion models to adapt to various data distributions.

The density estimation approach offers a powerful solution to the challenge of zero-shot classification, particularly for tasks where large amounts of labeled data for unseen classes are unavailable. By focusing on estimating class densities rather than generating new features, the Diffusion Classifier provides an efficient and robust framework for zero-shot learning.

In summary, both ZeroDiff and the Diffusion Classifier make use of diffusion models but take distinct architectural approaches to tackle the challenges of ZSL. ZeroDiff's two-branch structure focuses on generating high-quality features for unseen classes while leveraging contrastive representations to improve class-level understanding. In contrast, the Diffusion Classifier's density estimation approach bypasses feature generation, instead using the probabilistic framework of pretrained diffusion models to directly classify unseen data. Both architectures offer

promising solutions to the data scarcity and generalization challenges in ZSL, each bringing a unique set of advantages to the field.

3.2 Training Requirements

ZeroDiff's Training Procedure

The training process for ZeroDiff is specifically designed to address the challenges of zero-shot learning, especially when there is limited labeled data for unseen classes. ZeroDiff uses a two-branch generative architecture, which incorporates both a Diffusion-based Feature Generator (DFG) and a Diffusion-based Representation Generator (DRG). Each branch plays a critical role in improving data efficiency and enhancing the quality of the generated features for unseen classes.

Initially, the model begins by fine-tuning feature extractors on the labeled data from seen classes. The feature extractor is typically based on well-established neural networks like ResNet or EfficientNet. During this phase, the model applies cross-entropy loss for feature extraction through DFG, and supervised contrastive loss for semantic representation learning through DRG. This allows the model to develop strong foundational knowledge of how to extract and map visual and semantic features, which is crucial for generating meaningful data for unseen classes.

Once the extractors are fine-tuned, the generative models (DFG and DRG) are trained using a combination of adversarial learning and contrastive learning. The DFG generates visual features for unseen classes by leveraging the diffusion process to transform noisy data into clean, realistic samples. The DRG, meanwhile, enhances the quality of these features by generating richer, more semantically meaningful representations through contrastive learning. These two branches are trained together, optimizing for both realistic feature generation and semantic alignment with class descriptions.

ZeroDiff employs multiple discriminators to evaluate the generated data. These discriminators assess the quality of the features from different perspectives: one checks if the generated features match predefined class semantics (e.g., attribute labels), another ensures that the features adhere to the expected diffusion process, and the last one evaluates how well the features align with contrastive representations. To integrate these evaluations, ZeroDiff uses a Wasserstein-distance-based mutual learning loss, which helps improve the feature generation process by encouraging mutual learning among the discriminators, thereby refining the overall output. This multi-faceted training procedure allows ZeroDiff to create high-quality, realistic features for unseen classes and ensure these features are semantically aligned.

Diffusion Classifier's Zero-Shot Nature

The Diffusion Classifier, on the other hand, takes a different approach that is more streamlined for zero-shot tasks. The primary advantage of the Diffusion Classifier is its zero-shot nature, meaning it can perform classification tasks without requiring retraining on unseen classes. Instead of generating new features for unseen categories, the Diffusion Classifier estimates class-conditional densities directly using a pretrained diffusion model. This eliminates the need for additional labeled data or retraining for unseen classes, making it highly efficient.

The Diffusion Classifier begins with a pretrained diffusion model, which has been trained on a large dataset of seen classes. This model has already learned the structure and distribution of the data, and can now be leveraged to classify new, unseen classes. The classifier does not perform additional training on the unseen data, but instead uses the model's knowledge to estimate class probabilities. This is done through a process known as density estimation, where the reverse diffusion process is used to generate class-conditional distributions.

The core idea is to apply Monte Carlo sampling to estimate the class probabilities for an input image. In this approach, the model performs multiple iterations of the reverse denoising process, sampling different paths to

estimate the likelihood of a given class. By using the variational lower bound (ELBO), the Diffusion Classifier optimizes the likelihood of classifying the unseen classes correctly, without the need for additional training. This enables the model to perform zero-shot classification tasks effectively, leveraging the pretrained model's generative capabilities.

In addition to this, the Diffusion Classifier can incorporate semantic inputs (e.g., text descriptions or embeddings) to further guide the class-conditional density estimation process. These inputs help the model better understand the relationship between visual features and semantic categories, allowing it to classify unseen classes based on the semantic cues provided. This zero-shot capability is particularly useful in scenarios where labeled data for unseen classes is unavailable, making the Diffusion Classifier a highly efficient solution for zero-shot tasks.

ZeroDiff requires a more involved training procedure that focuses on feature generation and alignment with semantic representations through its dual-branch architecture, leveraging adversarial and contrastive learning. This training process equips the model to generate high-quality data for unseen classes. In contrast, the Diffusion Classifier operates by estimating class-conditional densities directly using a pretrained diffusion model, which allows it to classify unseen classes without additional retraining. While both approaches tackle the challenges of ZSL, they use fundamentally different strategies—ZeroDiff focuses on feature generation and enhancement, while the Diffusion Classifier uses pretrained models to perform efficient zero-shot classification.

3.3 Data Efficiency

Data Efficiency

In the context of Zero-Shot Learning (ZSL), achieving high data efficiency is a key challenge, particularly when dealing with unseen classes for which no labeled data is available. Both ZeroDiff and Diffusion Classifier tackle this problem in unique ways, using different strategies to maximize the utility of limited data and reduce the need for large, labeled datasets.

ZeroDiff's Instance-Level and Class-Level Strategies

ZeroDiff improves data efficiency by implementing strategies that optimize both the instance-level and class-level learning.

1. Instance-Level Efficiency: ZeroDiff's approach at the instance level focuses on generating diverse features for unseen classes using the Diffusion-based Feature Generator (DFG). This component takes noisy input and progressively denoises it, generating realistic features that mimic the characteristics of unseen classes. This is particularly useful because ZSL typically suffers from a lack of data for unseen categories. By creating synthetic data that resembles real instances from unseen classes, ZeroDiff can significantly increase the number of available examples for training. Even though the model does not have access to actual data from unseen classes, the generated features help fill this gap, allowing the model to perform well even with limited real data.

2. Class-Level Efficiency: At the class level, ZeroDiff uses the Diffusion-based Representation Generator (DRG) to enhance the semantic understanding of each class. The DRG improves the richness of the class representations by incorporating semantic embeddings, which can include textual descriptions or attribute-based representations. By learning better class representations through contrastive learning, the model can align visual features with their semantic meanings more effectively. This allows ZeroDiff to generalize across both seen and unseen classes, improving its ability to handle a broad set of categories with limited labeled data. The DRG helps the model connect high-level attributes with specific visual features, making it more adept at recognizing unseen classes based on their semantic descriptors.

Through this combination of instance-level and class-level strategies, ZeroDiff maximizes data efficiency by generating realistic features and enriching class representations. This approach ensures that the model can operate effectively even with limited training data, which is a common limitation in ZSL scenarios.

Diffusion Classifier's Leverage of Pre-Trained Models

The Diffusion Classifier, by contrast, relies on a more direct approach to data efficiency by utilizing pretrained models. Instead of generating new data for unseen classes, the Diffusion Classifier leverages the generative capabilities of a pretrained diffusion model to perform zero-shot classification without requiring additional training on unseen categories.

Pretraining on Large Datasets: The Diffusion Classifier begins with a pretrained diffusion model, which has already been trained on a large dataset of seen classes. This model has learned the underlying data distribution, including how various classes are visually represented and how they relate to semantic descriptors. Because this model has already internalized a significant amount of knowledge about the data distribution, it does not need retraining for new classes, which significantly reduces the need for labeled data. The pretrained model is then applied to classify unseen classes by estimating the class-conditional densities using a process known as density estimation.

Generating Class-Conditional Densities: The classifier estimates the probability distribution of each class using the generative power of the pretrained diffusion model. By employing Monte Carlo sampling methods, the Diffusion Classifier approximates the likelihood of a given class for an unseen sample, without needing any data from those classes. This makes the approach highly efficient, as the model can directly predict class probabilities for unseen data using the knowledge it has already acquired. The ability to perform zero-shot classification with a pretrained model means that the Diffusion Classifier requires no additional labeled data or retraining on new classes, making it an extremely data-efficient solution for zero-shot tasks.

By relying on a pretrained model, the Diffusion Classifier avoids the need for generating new instances or retraining for unseen classes. This greatly reduces the computational overhead and data requirements, which is especially advantageous when labeled data for unseen classes is sparse or nonexistent. The model's ability to handle zero-shot classification tasks with minimal additional data is a significant strength, making it highly suitable for environments where data is limited.

Both ZeroDiff and the Diffusion Classifier tackle the problem of data efficiency in Zero-Shot Learning but do so in distinct ways. ZeroDiff generates synthetic features for unseen classes at the instance level while enhancing class representations through semantic embeddings at the class level. This two-pronged approach boosts data efficiency by augmenting available data and improving the model’s understanding of classes. On the other hand, the Diffusion Classifier maximizes data efficiency by leveraging a pretrained diffusion model, enabling zero-shot classification without the need for additional training or data from unseen classes. Both approaches significantly reduce the need for large, labeled datasets and enhance the ability to generalize to new, unseen categories.

4 PERFORMANCE ANALYSIS

Benchmark Comparisons

To assess the effectiveness of ZeroDiff and the Diffusion Classifier, it's essential to compare their performance across standard zero-shot learning benchmarks. These include well-known datasets like AWA2 (Animals with Attributes 2), CUB-200 (Caltech-UCSD Birds), and SUN Attribute datasets. These datasets provide a diverse range of categories and have been widely used in the ZSL community to test the ability of models to classify unseen classes based on semantic information.

ZeroDiff, with its dual-branch architecture, has shown excellent performance in Generalized Zero-Shot Learning (GZSL) tasks. On datasets like AWA2 and CUB-200, it significantly outperforms traditional models like SJE (Semantic Embedding) and DEVISE (Deep Visual-Semantic Embedding). ZeroDiff’s ability to generate high-quality synthetic features for unseen classes allows it to perform well across both seen and unseen categories, even when the available labeled data is limited. Its focus on generating both diverse instance-level features (via DFG) and rich class-level semantic representations (via DRG) makes it a powerful model for generalization in zero-shot learning settings.

On the other hand, the Diffusion Classifier leverages pretrained models and focuses on density estimation for zero-shot classification. By using a pretrained diffusion model, it estimates the probability distribution over classes for unseen data, without the need for feature generation. This model shows competitive performance on the same benchmarks, often performing on par with ZeroDiff but with significantly less computational overhead. The Diffusion Classifier's efficiency makes it an attractive option for real-world applications that require quick adaptation to new classes with minimal retraining.

Strengths and Limitations of Each Approach

Both ZeroDiff and the Diffusion Classifier come with their own set of strengths and weaknesses, which influence their suitability for different tasks. ZeroDiff's main advantage is its ability to generate diverse, realistic features for unseen classes. This is achieved through its two-branch architecture that combines feature generation with semantic alignment. By generating new features and aligning them with class semantics, ZeroDiff can generalize well to unseen classes, making it effective in data-scarce environments. Additionally, its use of contrastive learning in the Diffusion-based Representation Generator (DRG) allows it to create more accurate class representations, further enhancing its performance across different ZSL tasks. Despite its strong performance, ZeroDiff's computational requirements can be a significant drawback. The training process is computationally demanding, as it involves training both generative and discriminative components. The model's reliance on adversarial training and multiple discriminators increases training time and the resources needed to train the model effectively. Moreover, although feature generation helps in generating unseen class features, ZeroDiff is not immune to overfitting on the seen classes, which can reduce its ability to generalize effectively to novel, unseen categories.

The Diffusion Classifier shines in its efficiency. By leveraging pretrained models, it avoids the need for additional training on unseen classes, making it highly computationally efficient. Its ability to perform zero-shot classification without retraining is a significant advantage in scenarios where labeled data for unseen classes is scarce. The Diffusion Classifier also benefits from its robustness to domain shifts, as pretrained models are generally exposed to a wide variety of data, making the model more adaptable to new data distributions.However, the Diffusion Classifier's performance is heavily dependent on the quality of the pretrained models. If the pretrained model was not trained on a sufficiently diverse or representative dataset, the model may struggle to generalize well to unseen classes. Additionally, although Monte Carlo sampling helps in density estimation, it can introduce variability in predictions, potentially affecting the consistency and accuracy of results, especially in challenging, out-of-distribution scenarios.

Computational Requirements

ZeroDiff requires substantial computational resources, particularly during the training phase. The model involves a complex dual-branch architecture, where both the DFG and DRG must be trained with multiple discriminators. This process, which integrates adversarial learning and contrastive loss, demands significant processing power and memory. Training ZeroDiff on large datasets or with complex architectures can be time-consuming, requiring considerable hardware resources, such as GPUs or TPUs, and longer training times. Additionally, the generation of synthetic features for unseen classes adds extra computational complexity.

In contrast, the Diffusion Classifier is more computationally efficient. Since it leverages pretrained models, the model does not need to generate synthetic features for unseen classes, and there is no need for retraining on those classes. Inference time is relatively fast compared to ZeroDiff, making the Diffusion Classifier a better choice in real-time applications or when computational resources are limited. However, the Monte Carlo sampling process used during classification can still be computationally demanding, particularly in scenarios where high accuracy is

required. Despite this, it remains more efficient overall, especially in comparison to the complex training processes of ZeroDiff.

Robustness Considerations

Both models show promising robustness in handling distribution shifts and generalizing to unseen data, but they exhibit different vulnerabilities. ZeroDiff is robust to domain shifts because of its ability to generate features that align well with unseen classes. The generation of features from a generative process makes the model less reliant on direct labeled examples and helps it adapt to new unseen classes. However, the model can still suffer from overfitting to seen classes, particularly if the synthetic feature generation process does not adequately represent the diversity of unseen categories.

The Diffusion Classifier also demonstrates strong robustness, particularly in zero-shot settings. By leveraging a pretrained diffusion model, it can generalize well to new unseen classes without retraining. However, the robustness of the Diffusion Classifier is directly tied to the quality of the pretrained model. If the pretrained model is not diverse enough or trained on data that doesn't generalize well to unseen classes, its performance can drop significantly. Additionally, Monte Carlo sampling introduces an element of variability in the results, which can sometimes affect consistency, especially in challenging classification tasks with complex or noisy data.

Both ZeroDiff and the Diffusion Classifier are effective models in zero-shot learning, but they differ significantly in terms of computational demands, efficiency, and robustness. ZeroDiff is highly effective in generating diverse features for unseen classes and performs well in generalized zero-shot learning, but it comes at a high computational cost and is prone to overfitting. Diffusion Classifier, on the other hand, is more computationally efficient, leveraging pretrained models for zero-shot classification with minimal overhead. While it shows strong robustness and efficiency, its performance is dependent on the quality of pretrained models and may struggle with complex or highly variable unseen classes. The choice between these models will largely depend on the specific needs of the application, including resource constraints, data availability, and the complexity of the unseen classes being classified.

5 DISCUSSION

Trade-offs Between Approaches

When comparing ZeroDiff and the Diffusion Classifier, there are notable trade-offs in terms of computational efficiency, the approach to feature generation and classification, and their robustness in handling different types of data.

ZeroDiff is based on a generative model that creates realistic features for unseen classes, making it highly versatile for generating data in environments where labeled data for new categories is scarce. This feature generation process is beneficial for generalized zero-shot learning (GZSL) tasks where the model must handle both seen and unseen classes simultaneously. However, the downside is that the training procedure is computationally expensive due to the complex architecture, including dual branches and adversarial components. While this allows ZeroDiff to produce high-quality features, it also results in a higher resource demand for training and inference, limiting its use in real-time applications or environments with limited computational resources.

In contrast, the Diffusion Classifier utilizes a pretrained diffusion model, making it much more computationally efficient, especially in zero-shot classification tasks. Since it doesn’t require generating new features for unseen classes, the model can classify unseen data based on class-conditional densities learned from a large dataset. This pretrained model approach allows it to perform effectively with minimal retraining, making it a faster and more resource-efficient solution. However, this efficiency comes at the cost of flexibility. The Diffusion Classifier's performance is closely tied to the quality of the pretrained model, meaning that it can struggle if the pretrained model

was not trained on a sufficiently diverse dataset. It also cannot generate new features for unseen classes, which can limit its applicability in scenarios where such generation is needed for optimal performance.

Application Scenarios

The choice between ZeroDiff and the Diffusion Classifier often depends on the specific needs of the application. ZeroDiff excels in situations where feature generation for unseen classes is crucial. For example, in medical imaging, where detecting rare diseases or classifying unseen conditions is important, ZeroDiff's ability to generate plausible features for these rare classes would be highly beneficial. Similarly, in wildlife conservation, where new species are continually discovered, ZeroDiff could be used to generate synthetic data for species that haven't yet been observed, enhancing the model's ability to recognize them.

On the other hand, Diffusion Classifier is more suited to environments where zero-shot classification with minimal overhead is required. In real-time systems, such as autonomous vehicles or surveillance, where quick adaptation to new and unseen data is essential, the Diffusion Classifier's pretrained model allows for faster deployment. This makes it a good fit for scenarios that demand low latency and computational efficiency. Additionally, in applications where classification accuracy is important but generating synthetic data is unnecessary, such as in automated content tagging or text-to-image applications, the Diffusion Classifier proves to be both efficient and effective.

For applications that combine data efficiency with the need for feature generation and quick classification, a hybrid approach could work well. For instance, a system could leverage ZeroDiff's feature generation for unseen classes while using a Diffusion Classifier to handle the final classification task in real-time. Such a combination could offer the best of both worlds, balancing data generation with computational efficiency.

Future Research Directions

Both ZeroDiff and Diffusion Classifier offer promising solutions to challenges in zero-shot learning, but several areas could be explored to improve their performance and broaden their applicability.

1. Improving Generative Models: In ZeroDiff, future research could focus on optimizing the feature generation process to make it more efficient. This could involve using techniques like meta-learning or few-shot learning, allowing the model to generate features with less data while retaining high accuracy for unseen classes. Another promising direction could be improving the adversarial learning framework to reduce training time and computational demands, making the model more practical in real-world applications.

2. Enhancing Pretrained Models: For the Diffusion Classifier, improving the quality and diversity of pretrained models is crucial. Researchers could explore methods to adapt pretrained models to specific domains or applications through fine-tuning. This would allow the model to be more flexible in handling domain-specific tasks, such as medical or industrial applications, where unseen classes may differ significantly from those seen during pretraining.

3. Hybrid Models: There is significant potential in developing hybrid architectures that combine the generative capabilities of ZeroDiff with the computational efficiency of the Diffusion Classifier. A hybrid approach could generate features for unseen classes where necessary while maintaining the efficiency of pretrained models for real-time classification tasks. This would offer greater flexibility and robustness, especially in dynamic environments where both feature generation and quick classification are needed.

4. Domain-Specific Applications: Another exciting direction for future research is creating domain-specific models that focus on particular fields like medical imaging, wildlife conservation, or automated content recognition. By fine-tuning diffusion models and zero-shot learning techniques for specific tasks, researchers could improve the accuracy and efficiency of models in those fields, addressing challenges unique to those domains, such as class imbalance and semantic misalignment.

ZeroDiff and Diffusion Classifier offer substantial contributions to the field of zero-shot learning, each with its strengths and limitations. Future research will likely focus on optimizing their respective computational efficiency, improving generalization to unseen classes, and exploring hybrid solutions that combine the benefits of both

approaches. These advancements will further enhance the application of ZSL models in real-world scenarios, where adaptability, data efficiency, and accuracy are paramount.

6 CONCLUSION

In the realm of Zero-Shot Learning (ZSL), both ZeroDiff and the Diffusion Classifier offer valuable contributions, each with its strengths and trade-offs. ZeroDiff is distinguished by its ability to generate high-quality features for unseen classes through its two-branch architecture. This makes it particularly effective in generalized zero-shot learning (GZSL) scenarios, where a model must classify both seen and unseen classes simultaneously. Its strength lies in generating diverse, synthetic features for categories that lack labeled data, which is a significant advantage when facing data scarcity. However, the trade-off is that its training process is computationally expensive, requiring a substantial amount of resources, which could limit its application in real-time scenarios or situations with limited computational power.

The Diffusion Classifier, on the other hand, shines in computational efficiency and is particularly suited for zero-shot classification tasks. By leveraging pretrained models, it can classify unseen data without needing additional training on those classes. This eliminates the need for data generation, making it faster and more resource-efficient compared to ZeroDiff. However, this approach's main drawback is its reliance on the quality of pretrained models. If these models are not sufficiently diverse or have limitations, the Diffusion Classifier might struggle to generalize effectively to new, unseen categories. Still, its efficiency and minimal need for retraining make it an attractive choice for quick deployment in environments where speed and scalability are crucial.

References

1. Z. Ye, S. N. Goda et al. "Exploring Data Efficiency in Zero-Shot Learning with Diffusion Models", arXiv:2406.02929
2. A. C. Li, M. Prabhudesai et al. "Your Diffusion Model is Secretly a Zero-Shot Classifier", arXiv:2303.16203
3. Z. Akata, S. Reed, D. Walter, Honglak Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2015, doi: 10.1109/cvpr.2015.7298911.
4. W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An Empirical Study and Analysis of Generalized Zero-Shot Learning for Object Recognition in the Wild," Computer Vision – ECCV 2016, pp. 52–68, 2016, doi: 10.1007/978-3-319-46475-6_4.
5. S. Seo and J. Kim, "Hierarchical Semantic Loss and Confidence Estimator for Visual-Semantic Embedding-Based Zero-Shot Learning," Applied Sciences, vol. 9, no. 15, p. 3133, Aug. 2019, doi: 10.3390/app9153133.
6. Z. Zhang and G. Yang, "Deep Representation of Hierarchical Semantic Attributes for Zero-shot Learning," 2020 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, Jul. 2020, doi: 10.1109/ijcnn48605.2020.9206924.
7. Y. Xie, T. Song, and W. Li, "Semantic-aware visual attributes learning for zero-shot recognition," Journal of Visual Communication and Image Representation, vol. 74, p. 103010, Jan. 2021, doi: 10.1016/j.jvcir.2020.103010.
8. Y. Xie, T. Song, and J. Yuan, "Learning discriminative visual semantic embedding for zero-shot recognition," Signal Processing: Image Communication, vol. 115, p. 116955, Jul. 2023, doi: 10.1016/j.image.2023.116955.
9. L. Khachatryan et al., "Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators," 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15908–15918, Oct. 2023, doi: 10.1109/iccv51070.2023.01462.
10. G. Couairon, M. Careil, M. Cord, S. Lathuilière, and J. Verbeek, "Zero-shot spatial layout conditioning for text-to-image diffusion models," 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2174–2183, Oct. 2023, doi: 10.1109/iccv51070.2023.00207.

11. X. Liu, S. Huang, Y. Kang, H. Chen, and D. Wang, "VGDIFFZERO: Text-To-Image Diffusion Models Can Be Zero-Shot Visual Grounders," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2765–2769, Apr. 2024, doi: 10.1109/icassp48485.2024.10445945.
12. D. Jiang, A. Shah, S. Yeung, J. Zhu, K. Singh, and G. Goldenberg, "Multi-Label Classification for Fashion Data: Zero-Shot Classifiers via Few-Shot Learning on Large Language Models," Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, pp. 250–257, 2024, doi: 10.5220/0012899100003838.
13. E. Frakes, U. Khalid, and C. Chen, "Efficient and consistent zero-shot video generation with diffusion models," Real-Time Image Processing and Deep Learning 2024, p. 8, Jun. 2024, doi: 10.1117/12.3013575.
14. H. Kong, D. Lian, M. B. Mi, and X. Wang, "DreamDrone: Text-to-Image Diffusion Models Are Zero-Shot Perpetual View Generators," Computer Vision – ECCV 2024, pp. 324–341, Oct. 2024, doi: 10.1007/978-3-031-72624-8_19.