Zero-Shot Learning with Feature Generation

Bogazici University, Fall 24-25, CMPE537 Computer Vision Term Project
Lami Kaan Kosesoy

Introduction

Zero-Shot Learning (ZSL) is a machine learning approach where a model is trained to recognize classes it has never seen before, using only semantic information (like attributes or descriptions) about those unseen classes.

Generalized Zero-Shot Learning (GZSL) extends ZSL by allowing the model to classify both seen and unseen classes at test time, making it a more realistic and challenging scenario.

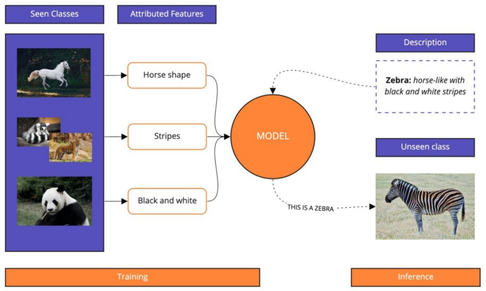

Figure 1. Zero-Shot Learning

Literature in the Field

The field of zero-shot learning (ZSL) has two main approaches: **embedding-based methods** and **generative-based methods**.

Embedding-based methods project image features and semantic attributes (e.g., class labels) into a shared embedding space. The model learns to match visual features with compatible semantic attributes, though it struggles with unseen class biases.

Generative-based methods use models like GANs or VAEs to synthesize image features for unseen classes by conditioning on their semantic attributes. These synthesized features enable training on both seen and unseen classes, improving generalization and reducing bias. Combined, these approaches represent key innovations in ZSL research.

Embedding-Based Methods	Generative-Based Methods
Out-Of-Distribution Detection-Based Methods	Generative Adversarial Networks (GANs)
Graph-Based Methods	Variational Autoencoders
Meta Learning-Based Methods	Combined GANs and VAEs
Attention-Based Methods	
Bidirectional Learning	
Autoencoder-Based Methods	

TF-VAEGAN Architecture

The TF-VAEGAN framework integrates a Variational Autoencoder (VAE), Generative Adversarial Network (GAN), and semantic embedding modules to address Zero-Shot Learning (ZSL) and Generalized ZSL (GZSL). The goal is to improve the quality and semantic consistency of synthesized features for unseen classes.

Feature Extraction and Feedback Module

- The encoder extracts visual features from input images, while the generator synthesizes initial features from semantic embeddings and noise.
- A feedback module refines synthesized features iteratively using reconstructed embeddings, improving alignment with real features.
- The semantic embedding decoder reconstructs embeddings during feedback, regularizing the generator.

Training Strategy

- Training involves iterative sub-steps: initial feature synthesis, embedding reconstruction, and feature refinement via feedback.
- Loss functions include reconstruction loss, adversarial loss, KL divergence, and cross-entropy loss.

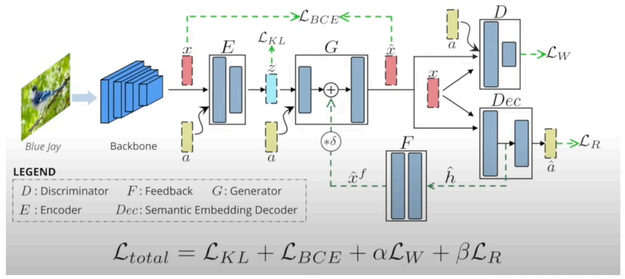

Figure 2. TF-VAEGAN Architecture

Classification with Discriminative Transformation

- The encoder extracts visual features from input images, A discriminative feature transformation leverages the decoder's embeddings during classification, enhancing separability between features.
- This transformation reduces ambiguities among class features, improving performance in ZSL and GZSL.

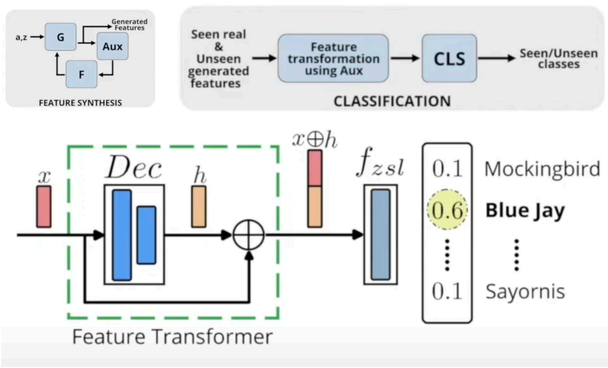

Figure 3. Classification with Discriminative Feature Transformation

Methodology & Experiments

The TF-VAEGAN framework integrates a Variational This project aims to evaluate the architecture proposed in the TF-VAEGAN paper under different settings, despite constraints in training the full model due to computational limitations. Instead of training the model, the focus is on analyzing how changes in image features and class attribute embeddings might affect its performance.

For image features, representations were extracted from the CUB dataset using three different backbone architectures: ResNet-101, ConvNeXt (Tiny), and ViT (B_16). These extracted features were visualized using t-SNE, highlighting the top 10 most frequent classes in the dataset. The spatial clustering of these image features was examined to determine how well images of the same class are represented. Additionally, the most similar and dissimilar class pairs were identified based on the average feature embeddings.

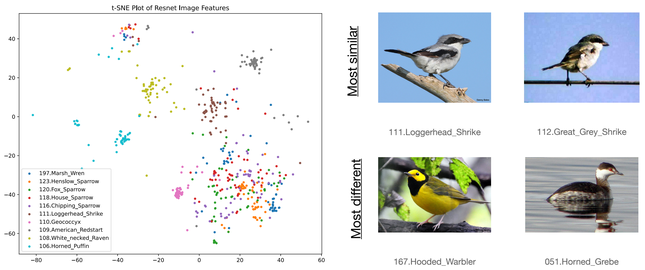

Figure 4. Results for ResNet features

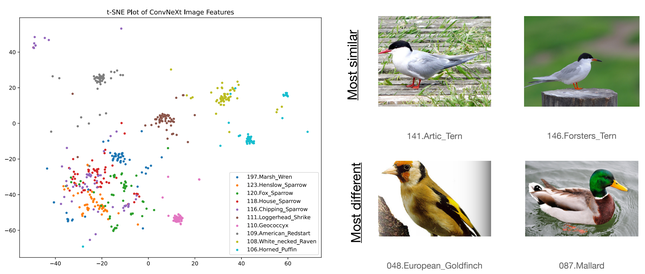

Figure 5. Results for ConvNeXt features

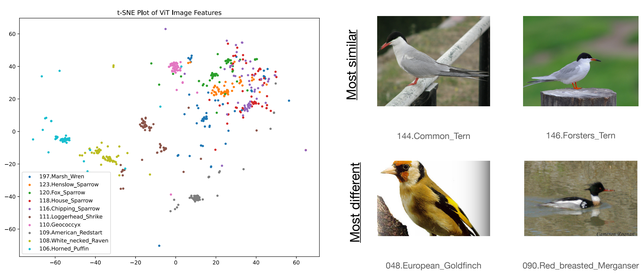

Figure 6. Results for ViT features

For class attribute embeddings, the sentence descriptions in the CUB dataset were utilized. Using Sentence Transformers (SBERT), two approaches were followed to create new class attribute vectors:

1. All 10 descriptions per image were aggregated into a single text, an embedding was extracted for each image, and the embeddings for all images in a class were averaged.
2. Embeddings of individual sentences for each image were averaged, followed by computing the class-level average from the resulting image embeddings.

The generated class attributes were visualized using UMAP due to their sparse structure. Clustering behavior and relationships between classes were explored for both the original (not shown here due to limited space) and sentence-derived attributes.

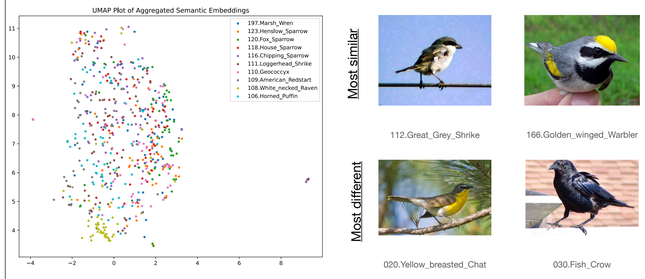

Figure 7. Results for aggregated sentence embeddings

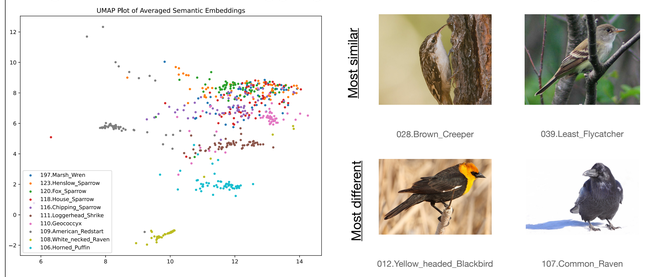

Figure 8. Results for averaged sentence embeddings

Conclusion & Feature Work

This study highlights the potential of using advanced image features and sentence-derived semantic embeddings to enhance zero-shot classification. Future work will focus on fully training the TF-VAEGAN model with these modified features, evaluating its performance, and exploring additional backbone architectures and semantic embedding techniques for further improvements.

References

1. Pourpanah, F., Abdar, M., Luo, Y., Zhou, X., Wang, R., Lim, C. P., & Wu, Q. J. (2022). A review of generalized zero-shot learning methods. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4), 4051–4070.
2. Xian, Y., Lorenz, T., Schiele, B., & Akata, Z. (2018). Feature generating networks for zero-shot learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5542–5551).
3. Narayan, S., Gupta, A., Khan, F. S., Snoek, C. G., & Shao, L. (2020). Latent embedding feedback and discriminative features for zero-shot classification. In Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII (pp. 479–495). Springer International Publishing.