Zero-Shot Learning of Image Classification through Text-to-Image Generative Model

Fanyi Meng
fanyimeng@link.cuhk.edu.cn

March 22, 2024

1 Introduction

Zero-Shot Learning (ZSL) is a critical technique in machine learning, enabling the recognition of unseen objects or categories beyond the scope of training data. It excels in scenarios where comprehensive data collection is challenging. This project proposes using generative models, like Stable Diffusion, to fill data gaps by generating features of unseen categories. The aim is to enhance ZSL's performance through experiments with these models, thereby testing their ability to improve data richness and the ablity of generalizing from known to unknown classes.

2 Related work

Zero-shot Learning Zero-shot learning operates on the assumption that while no images of a new class are in the training set, other information like attributes or textual descriptions is available. Based on the connections between new class and training class in semantic space. The model can still classify these new class even corresponding data is not available. (Yu and Aloimonos, 2010).Some work has already been done to integrate generative models with ZSL. Mishra et al. (2018) trains a Conditional Variational Autoencoder to understand the probability distribution of image features, thus improve the performance of ZSL. Wang et al. (2018) used similar intuition and adapted the model for semi-supervised and few-shot learning scenarios, outperforming existing methods in tests on various benchmark datasets.

Text-to-Image Generative Model Recently, with the advancement of large language models(LLM), there has been significant development in models capable of generating images from textual descriptions. This technology allows models to create visual content based on given text instructions. Generative text-to-image models based on denoising diffusion probabilistic models (Ho et al., 2020) such as Imagen (Saharia et al., 2022), Dalle-2 (Ramesh et al., 2022), and Stable Diffusion (Rombach et al., 2022) can generate realistic high-resolution images and generalize to diverse text prompts. Their strong performance shows the feasibility of using generative models to provide data for new classes

3 Research questions

Following (Mishra et al., 2018), our problem is described as follows: Given a set of train classes (seen classes) $\mathcal{Y}_s = \{y_s^1, y_s^2, ...y_s^n\}$ and a set of test classes (unseen classes) $\mathcal{Y}_u = \{y_u^1, y_u^2, ...y_u^m\}$. For each class y in $\mathcal{Y} = \mathcal{Y}_u \cup \mathcal{Y}_s$, we have a class semantic embedding vector A_y, that describes

the class. Given d-dimensional labelled training data from the seen classes $\mathcal{Y}_s$, i.e $\{X_s, Y_s\}$. We will generate some new training data for the unseen classes $\mathcal{Y}_u$ based on the Text-to-Image Generative Model as mentioned in last part, and then we will construct a model $f : \mathbb{R}^d \rightarrow \mathcal{Y}_s \cup \mathcal{Y}_u$. Finally we will test their performance and compare it with traditional zero-shot learning.

4 Methodology

We plan to utilize the same CNN architecture to perform three distinct tasks, demonstrating the effectiveness of our approach across different scenarios: Assume there are ten classes in the entire dataset, among which 8 classes are seen and two are unseen. The first task involves training the CNN only on seen dataset. And the model will be tested on the entire dataset. The second task involves training the CNN on the dataset which consists of seen data and generative data based on the label. And the model will be tested on the entire dataset. Based only on intuition, the second task should achieve better classification performance since some additional features have been provided to the dataset. We aim to verify the extent of this performance improvement through specific experiments.

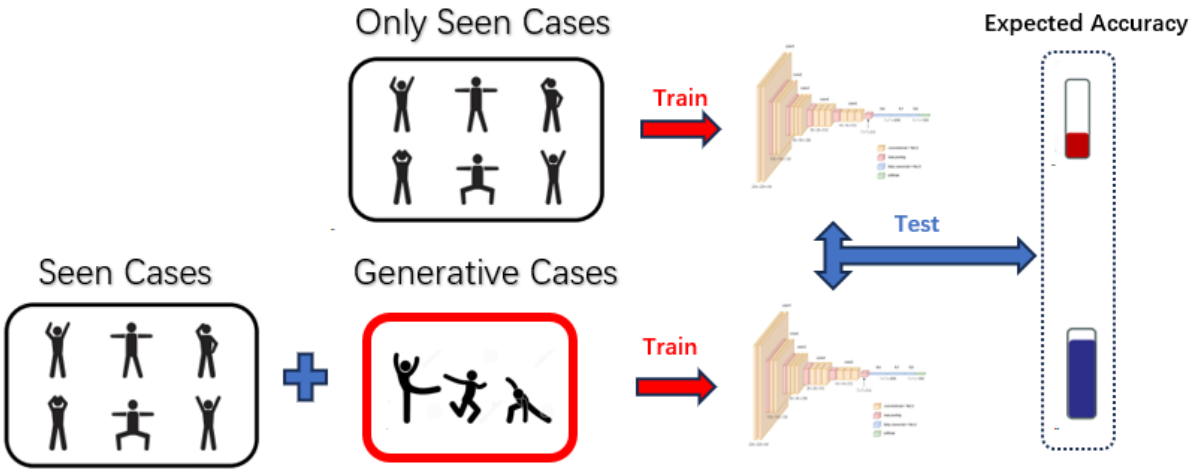

Figure 1: Workflow of Project

5 Tools and datasets used

We plan to use Dataset Yang et al. (2023) to complete this work. This dataset contains datas of figures when human performing various actions. Additionally, it encompasses many other modalities of data like lidar. If feasible, we consider extending this work to include these other modalities as well.

We plan to use Chen et al. (2023) to generate new data, which combines three multimodal skills of pre-built AI models without additional model training: visual chat of LLaVA , image segmentation from SEEM , and image generation and editing from GLIGEN . So the generated data will be more closed to real data.

References

W.-G. Chen, I. Spiridonova, J. Yang, J. Gao, and C. Li. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. arXiv preprint arXiv:2311.00571, 2023.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33:6840–6851, 2020.

A. Mishra, S. Krishna Reddy, A. Mittal, and H. A. Murthy. A generative model for zero shot learning using conditional variational autoencoders. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 2188–2196, 2018.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125, 2022.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10684–10695, 2022.

C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In ACM SIGGRAPH 2022 Conference Proceedings, pages 1–10, 2022.

W. Wang, Y. Pu, V. Verma, K. Fan, Y. Zhang, C. Chen, P. Rai, and L. Carin. Zero-shot learning via class-conditioned deep generative models. In Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.

J. Yang, H. Huang, Y. Zhou, X. Chen, Y. Xu, S. Yuan, H. Zou, C. X. Lu, and L. Xie. Mm-fi: Multi-modal non-intrusive 4d human dataset for versatile wireless sensing. In Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2023. URL https://openreview.net/forum?id=1uAsASS1th.

X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part V 11, pages 127–140. Springer, 2010.