Few-shot and Zero-shot Learning for Cross-modal Data Generation

クロスモーダルデータ生成のための少数ショット学習およびゼロショット学習

Abstract

Advancing AI integrates text and images using cross-modal generative models. This project evaluates these models, focusing on generalization, consistency, and realism. Building on prior work, it explores using GANs, VAEs, and contrastive learning in few-shot and zero-shot learning for cross-modal data. The study looks at encoding, retention, and retrieval in these models to boost robustness and interpretability. Goals include developing new methods for better performance with limited data and proposing enhancements for more effective AI. This has implications for image generation, segmentation, and other cross-modal tasks.

概要:

急速に発展している人工知能の分野では、クロスモーダル生成モデルが前端に入り、異なるデータ類型（例えば、テキストと画像）が組み合わせられ、統一されたマルチモーダルが作られています。私の研究計画は、これらのモデルの一般性、一貫性、真実性に重点を置き、分析解釈を行うことです。卒論の設計のもとで、私は、GANs、VAEs その他の新しい生成モデルなどの生成モデルを考察したいと思います。サンプル学習とゼロサンプル学習シナリオにおけるクロスモーダルデータ生成の応用を行います。そして、これらのモデルにおけるコード、保持、検索プロセスを研究することにより、本論文はそのロバスト性と説明性を強化することを目的としています。限られたデータを扱う際のモデルの性能を向上させるために、テストや改善手法を使います。この研究は、画像生成のような他のクロスモーダル課題にも重要な影響を与えています。

Research Background

Cross-modal generative models are at the forefront in AI field, integrating different data types, such as text and images, to create unified multimodal outputs. I plan to analyze these models, concentrating on their generalization, consistency, and realism in my final year project. Following, this research proposal builds on that foundation by investigating the use of various generative models, including but not limited to GANs, in few-shot learning scenarios for cross-modal data generation. By extending my initial work, I aim to determine how effectively these models can produce high-quality cross-modal data from limited examples, thereby enhancing their applicability and robustness. This will require an in-depth examination of their encoding, retention, and retrieval processes, aligning with the research group's focus on machine learning.

Previous Research

Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders [1]: This study introduces a model where a shared latent space of image features and class embeddings is learned using modality-specific aligned variational autoencoders. This approach captures essential multi-modal information associated with unseen classes and establishes a new state of the art on generalized zero-shot and few-shot learning across multiple benchmark datasets, including CUB, SUN, AWA1, AWA2, and ImageNet.

GanOrCon: Are Generative Models Useful for Few-shot Segmentation? [2]: This paper compares the effectiveness of generative models and contrastive learning methods for few-shot segmentation

tasks. It finds that while GANs can compactly represent datasets for few-shot discrimination tasks, they may lose information during the learning process, whereas contrastive learning methods excel in learning invariances from unlabelled data.

Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models [3]: This work demonstrates that leveraging cross-modal information can improve few-shot learning performance. By using multimodal foundation models like CLIP, the authors propose a cross-modal adaptation approach that achieves state-of-the-art results in vision-language tasks and benefits other methods like prefix tuning and adapters. They also introduce an audiovisual few-shot benchmark to enhance both image and audio classification.

A Systematic Evaluation and Benchmark for Embedding-Aware Generative Models: Features, Models, and Any-shot Scenarios [4]: This study evaluates the performance of Embedding-Aware Generative Models in generalized zero-shot learning scenarios, highlighting the importance of feature selection in zero-shot learning benchmarks and discussing the application of these models in few-shot learning.

These studies provide a comprehensive foundation for further research into enhancing the robustness, interpretability, and applicability of generative models in few-shot and zero-shot cross-modal learning tasks.

Research Objectives

The primary objectives of this research are to evaluate the performance of cross-modal generative models in few-shot and zero-shot learning scenarios, and to develop new methods for enhancing the robustness and interpretability of these models. Additionally, this study aims to compare the efficacy of various generative models, including GANs, VAEs, contrastive learning models, and potentially latest architectures in these tasks. Another focus is to investigate the encoding, retention, and retrieval processes within these models, examining their contributions to generalization and robustness. The ultimate goal is to propose improvements to existing models or architectures that excel in few-shot and zero-shot cross-modal generation.

Research Methods

For the project, I will start with data collection and preprocessing using existing datasets and possibly creating new benchmark datasets for few-shot and zero-shot learning tasks. Preprocessing ensures compatibility with various generative models.

For model development and training, I will implement and fine-tune advanced GANs, VAEs, and other generative models. This might involve creating new architectures or modifying existing ones for better performance in few-shot and zero-shot scenarios.

To assess the models, I will define evaluation metrics focusing on generalization, consistency, realism, and interpretability. I will conduct systematic benchmarking across different models for a comprehensive evaluation.

Additionally, I will perform comparative analysis between GANs, VAEs, and other models like contrastive learning methods. This analysis will identify strengths and weaknesses in performance, robustness, and information retention.

Furthermore, I will develop methods to enhance the robustness of generative models to data perturbations and improve the interpretability of model outputs, this can be done by understanding the latent space and feature representations. These enhancements should significantly contribute to

the models' overall effectiveness and reliability in practical applications.

Expected Outcome

I expect to achieve a comprehensive evaluation of cross-modal generative models in few-shot and zero-shot learning scenarios, identifying their strengths and weaknesses, as well as other notable features. This research aims to offer valuable insights into enhancing the robustness, interpretability, and usability of these models. Additionally, I anticipate developing new methods and benchmarks that will contribute to future advancements in cross-modal generative modeling. These outcomes should help in understanding the encoding, retention, and retrieval processes of generative models, leading to improved performance in handling diverse and limited data scenarios. The findings could also provide a solid foundation for further research and practical applications in various fields such as image generation, segmentation, and beyond.

Significance

This research targets improving the use of generative models for few-shot and zero-shot learning. By evaluating and enhancing these models, the study aims to help develop more robust and interpretable AI systems for diverse, limited data scenarios. The findings may impact applications like image generation, segmentation, and cross-modal tasks.

References

[1] E. Schönfeld, S. Ebrahimi, S. Sinha, T. Darrell, and Z. Akata, "Generalized Zero- and Few-Shot Learning via Aligned Variational Autoencoders," arXiv preprint arXiv:1812.01784, 2018.

[2] O. Saha, Z. Cheng, and S. Maji, "GANORCON: Are Generative Models Useful for Few-shot Segmentation?," in *2022 IEEE/CVF CVPR*, Jun. 2022, doi: 10.1109/CVPR52688.2022.00975.

[3] Z. Lin, S. Yu, Z. Kuang, D. Pathak, and D. Ramanan, "Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning with Multimodal Models," in *2023 IEEE/CVF CVPR*, Jun. 2023, doi: 10.1109/CVPR52729.2023.01852.

[4] L. Feng, J. Zhao, and C. Zhao, "A Systematic Evaluation and Benchmark for Embedding-Aware Generative Models: Features, Models, and Any-shot Scenarios," arXiv preprint arXiv:2302.04060, Feb. 2023.