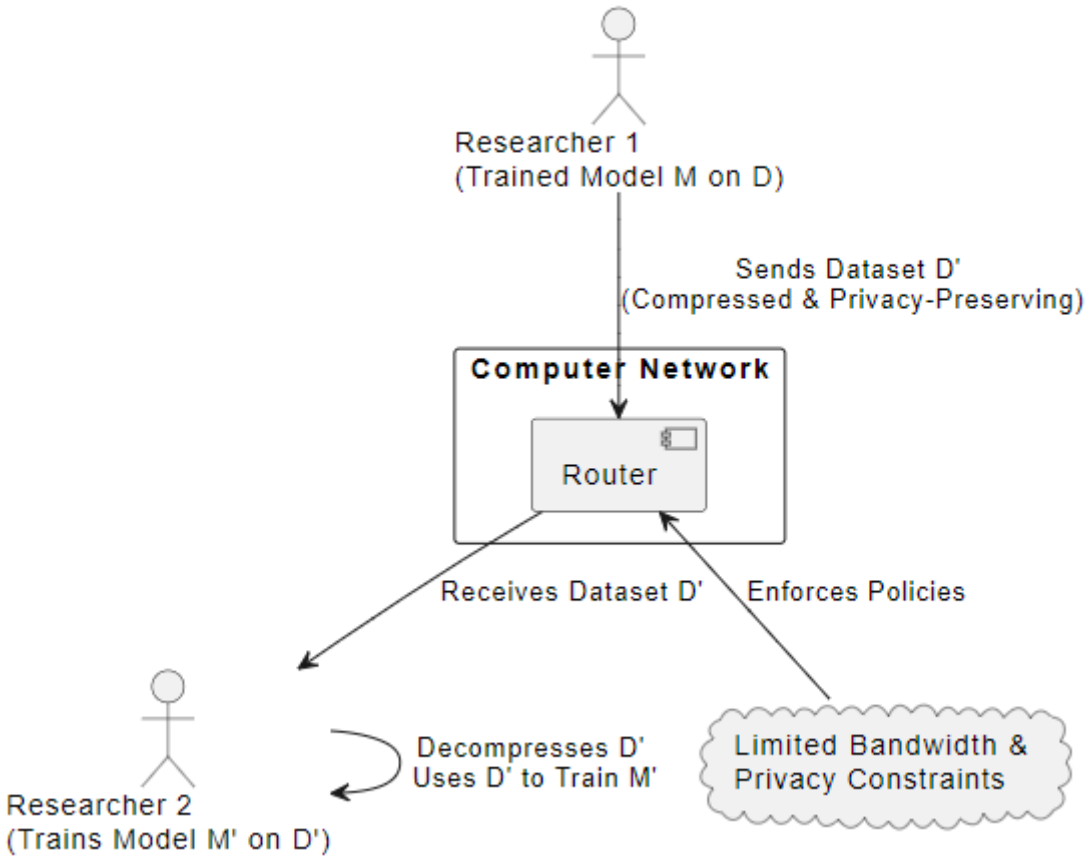
Latent Gaussian Compression: Dataset Compression and Reconstruction with Autoencoders and Gaussian Mixture Models

Abstract

This report explores a machine learning-based approach for compressing and reconstructing large-scale image datasets. We combine autoencoders and Gaussian Mixture Models (GMMs) to create a compact, efficient representation that preserves key features for classification. We evaluate the effectiveness of this approach (Latent Gaussian Compression - LGC) by comparing the performance of a classifier trained on reconstructed data to the original dataset. To establish a strong baseline, we compared our approach to coreset selection with k-medoids. Additionally, we experimented with various autoencoder architectures (vanilla autoencoder, VAE, conditional VAE, contrastive VAE, and AE) to optimize performance. We further validated our method on the SpuCo (spurious correlation) dataset to assess its robustness to spurious correlations.

1. Introduction

Traditional compression methods may discard information crucial for machine learning tasks. This project leverages autoencoders and generative models to achieve a balance between compression factor and classification performance.



Motivation: In many real-world scenarios, researchers need to share large datasets for collaborative research or machine learning model training. However, sharing raw data can pose significant challenges due to privacy

concerns, bandwidth limitations, and storage costs. Our Latent Gaussian Compression (LGC) scheme provides a solution to these challenges by enabling efficient and privacy-preserving data sharing.

Consider a scenario where two researchers, Researcher 1 and Researcher 2, collaborate on a machine learning project. Researcher 1 has trained a model M on a large dataset D. Researcher 2 wants to train a similar model M' on a similar dataset D', but they are constrained by limited bandwidth and privacy concerns.

Using LGC, Researcher 1 can compress and anonymize their dataset D into a compressed representation D'. This compressed representation can be transmitted over the network to Researcher 2 with minimal bandwidth overhead. Researcher 2 can then decompress D' and use it to train their model M'.

By using LGC, researchers can benefit from reduced bandwidth consumption, enhanced privacy, efficient data sharing, and improved model performance. LGC addresses the challenges of data sharing, enabling collaboration and innovation in machine learning research.

Contribution: We investigate the effectiveness of LGC for dataset compression and reconstruction. We explore different autoencoder architectures (vanilla, VAE) and incorporate submodular maximization for data selection. The project investigates the trade-off between compression and performance and proposes theoretical guarantees for information retention.

2. Related Work

Autoencoders have been used for dimensionality reduction and feature learning in various applications [1]. Variational Autoencoders (VAE) introduce a probabilistic framework for learning latent representations [2]. GMMs are employed for density estimation and data modeling [3]. Submodular maximization techniques have been explored for efficient data selection [4].

3. Methodology

The project is divided into the following stages:

Autoencoder Compression

An autoencoder is trained to compress images into a lower-dimensional latent space.

Part 2: Latent Space Distribution Learning with GMMs

A GMM is fitted to the latent representations to capture the underlying distribution.

Selecting the Optimal Number of Components for GMMs

Determining the optimal number of components (K) in a Gaussian Mixture Model (GMM) is crucial for achieving accurate and robust modeling. In this work, we employed the Bayesian Information Criterion (BIC) to select the appropriate number of components for each class in our dataset.

The BIC score is a statistical criterion that balances model fit with model complexity. It is defined as:

$$ \text{BIC} = k \ln(n) - 2 \ln(\widehat{L}) $$

where:

- `log-likelihood`: The log-likelihood of the data given the model.

- k: The number of parameters in the model.
- n: The number of data points.

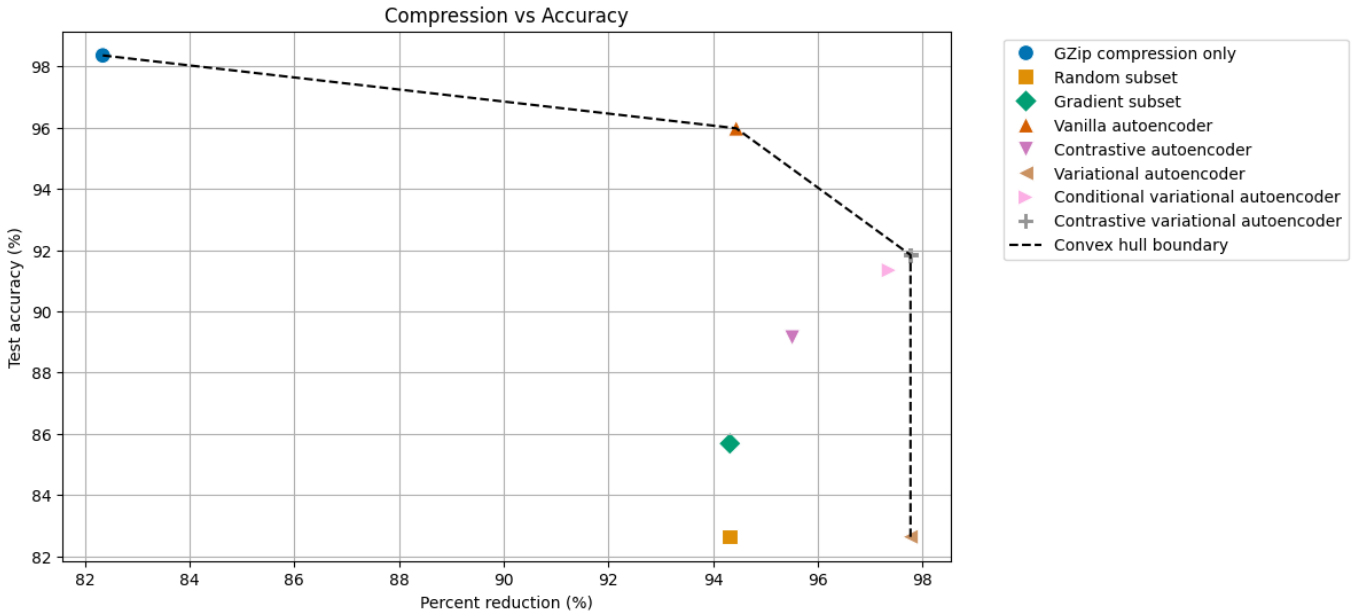
A lower BIC score indicates a better model fit. By comparing BIC scores for different values of K, we can identify the optimal number of components that balances model complexity and predictive accuracy.

For each class in our dataset, we trained GMMs with a varying number of components (K) and calculated their corresponding BIC scores. The BIC scores were then plotted against the number of components, resulting in the curves shown in Figure 1. The optimal number of components for each class was selected as the value of K that minimized the BIC score.

By carefully selecting the number of components using the BIC criterion, we ensured that our GMMs effectively captured the underlying data distribution without overfitting or underfitting. This optimal model selection is crucial for accurate data compression and reconstruction.

- **Part 3: Dataset Transportation and Reconstruction:** Samples are drawn from the learned GMM and decoded to reconstruct new images.
- **Part 4: Dataset Reconstruction and Training:** A classifier is trained on the reconstructed dataset and evaluated for performance comparison.
- **Part 5: Comparison to Submodular Maximization for Dataset Summarization:** Explores submodular maximization for data selection before applying LGC.

4. Experiments and Results



We evaluated the performance of our Latent Gaussian Compression (LGC) approach against several baseline methods, including GZIP compression, random subsetting, gradient-based subset selection, and coreset selection with k-medoids. We experimented with various autoencoder architectures (vanilla, VAE, conditional VAE, contrastive VAE, and contrastive VAE with convex hull boundary) within the LGC framework. Our results, illustrated in the figure above, demonstrate a compelling trade-off between compression ratio and test accuracy.

LGC consistently outperformed baseline methods, especially at higher compression ratios. While coreset selection with k-medoids offered reasonable performance, LGC generally provided superior accuracy,

particularly at higher compression levels.

However, it's important to note that all methods experienced a significant drop in accuracy at extremely high compression ratios. This highlights the inherent trade-off between compression and performance.

In conclusion, our LGC approach offers a promising solution for compressing large-scale datasets while preserving essential information for downstream tasks. Further research can explore more advanced autoencoder architectures and optimization techniques to further improve the performance of LGC.

5. Conclusion

This project investigates the feasibility of LGC for dataset compression and reconstruction. We explore various techniques and analyze the trade-off between compression rate and classification performance.

6. Future Work

Future work includes exploring:

- Integration of contrastive learning to enhance latent space structure.
- Incorporation of GMM structured priors for complex latent distributions.
- Derivation of theoretical guarantees for information retention as a function of the compression ratio.

7. References

- [1] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.
- [2] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational 1 inference. arXiv preprint arXiv:1312.6114.
- [3] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

8. Appendix

TODO