Unsupervised Learning Project: VAE for Hybrid Language Music Clustering

Course: Neural Networks
Prepared By: Moin Mostakim
Submission Due: January $10^{th}, 2026$

Project Overview

In this project, you will implement an unsupervised learning pipeline inspired by Variational Autoencoders (VAE) for clustering hybrid language music tracks. The goal is to extract latent representations from audio and/or lyrics and perform clustering. You will explore different VAE architectures, analyze clustering results, and compare with baseline methods (e.g., K-Means, PCA).

Deliverables:

1. GitHub Repository: Code implementation, dataset processing scripts, and results.
2. NeurIPS-like Paper Report: PDF report in a scientific format describing method, experiments, results, and discussion.

Project Tasks

Easy Task

- Implement a basic VAE for feature extraction from music data.
- Use a small hybrid language music dataset (English + Bangla songs).
- Perform clustering using K-Means on latent features.
- Visualize clusters using t-SNE or UMAP.
- Compare with baseline (PCA + K-Means) using Silhouette Score and Calinski-Harabasz Index.

Medium Task

- Enhance VAE with convolutional architecture for spectrograms or MFCC features.
- Include hybrid feature representation: audio + lyrics embeddings.
- Experiment with clustering algorithms: K-Means, Agglomerative Clustering, DBSCAN.

- Evaluate clustering quality using metrics: Silhouette Score, Davies-Bouldin Index, Adjusted Rand Index (if partial labels are available).

- Compare results across methods and analyze why VAE representations perform better/worse than baselines.

Hard Task

- Implement Conditional VAE (CVAE) or Beta-VAE for disentangled latent representations.

- Perform multi-modal clustering combining audio, lyrics, and genre information.

- Quantitatively evaluate using metrics: Silhouette Score, Normalized Mutual Information (NMI), Adjusted Rand Index (ARI), Cluster Purity.

- Provide detailed visualizations: latent space plots, cluster distribution over languages/genres, reconstruction examples from VAE latent space.

- Compare VAE-based clustering with PCA + K-Means, Autoencoder + K-Means, and direct spectral feature clustering.

Sample Datasets

- **Million Song Dataset (MSD)**: `http://millionsongdataset.com/` Large-scale dataset with audio features, metadata, and lyrics (partial). Useful for clustering music by audio features and genres.

- **GTZAN Genre Collection**: `http://marsyas.info/downloads/datasets.html` 1000 audio tracks categorized by 10 genres. Can extract MFCC/spectrogram features for VAE input.

- **Jamendo Dataset**: `https://www.kaggle.com/datasets/andradaolteanu/jamendo-music-da` Songs with audio previews, metadata, and lyrics. Great for hybrid (audio + lyrics) embedding experiments.

- **MIR-1K Dataset**: `https://sites.google.com/site/unvoicedsoundseparation/mir-1k` 1000 song clips in Mandarin and English. Useful for multi-language experiments.

- **Lakh MIDI Dataset (LMD)**: `https://colinraffel.com/projects/lmd/` MIDI dataset for music modeling. Can be converted to audio features and combined with lyrics from other sources.

- **Kaggle Lyrics Datasets**: `https://www.kaggle.com/datasets?search=lyrics` Multiple datasets containing song lyrics in English, Hindi, Bangla, and other languages. Can combine with audio features for hybrid VAE experiments.

Metrics and Comparison

Metric	Description	Higher Better?	Use Case
Silhouette Score	Measures how similar an object is to its own cluster compared to other clusters	Yes	Cluster quality evaluation
Calinski-Harabasz Index	Ratio of between-cluster variance to within-cluster variance	Yes	Cluster evaluation
Davies-Bouldin Index	Average similarity of each cluster with its most similar cluster	No	Cluster evaluation
Adjusted Rand Index (ARI)	Measures agreement of clustering with ground truth labels	Yes	Partial label evaluation
Normalized Mutual Information (NMI)	Quantifies the mutual information between clusters and labels	Yes	Cluster alignment with labels
Cluster Purity	Fraction of dominant class in cluster, useful if labels are available	Yes	Label-based cluster quality evaluation

Metrics and Mathematical Formulations

For each clustering metric, we provide a description and the mathematical formula to help you understand and compute them clearly.

1. Silhouette Score (S)

Measures how similar an object is to its own cluster compared to other clusters.

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where:

- $a(i)$ = average distance between i and all other points in the same cluster.
- $b(i)$ = minimum average distance between i and all points in other clusters.

The score ranges from -1 to 1. Higher is better.

2. Calinski-Harabasz Index (CH)

Measures ratio of between-cluster variance to within-cluster variance.

$$\mathrm{CH} = \frac{\mathrm{tr}(B_k)/(k-1)}{\mathrm{tr}(W_k)/(n-k)}$$

where:

- k = number of clusters

- n = number of points
- B_k = between-cluster dispersion matrix
- W_k = within-cluster dispersion matrix

Higher values indicate better clustering.

3. Davies-Bouldin Index (DB)

Average similarity between each cluster and its most similar cluster.

$$\text{DB} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d_{ij}} \right)$$

where:

- σ_i = average distance of points in cluster i to its centroid
- d_{ij} = distance between centroids of clusters i and j

Lower DB indicates better clustering.

4. Adjusted Rand Index (ARI)

Measures similarity between predicted clusters and ground truth labels.

$$\text{ARI} = \frac{\text{RI} - \text{Expected RI}}{\max(\text{RI}) - \text{Expected RI}}$$

where RI (Rand Index) measures the fraction of pairs correctly clustered. ARI is adjusted for chance. Higher is better (1 = perfect match, 0 = random labeling).

5. Normalized Mutual Information (NMI)

Measures mutual information between predicted clusters and true labels, normalized to [0,1].

$$\text{NMI}(U, V) = \frac{2I(U; V)}{H(U) + H(V)}$$

where:

- $I(U; V)$ = mutual information between clustering U and labels V
- $H(U), H(V)$ = entropies of clusters and labels

Higher values indicate better agreement with labels.

6. Cluster Purity

Fraction of the dominant class in each cluster.

$$\text{Purity} = \frac{1}{n}\sum_{k}\max_{j}|c_k \cap t_j|$$

where:

- c_k = set of points in cluster k
- t_j = set of points in true class j
- n = total number of points

Higher purity indicates better clustering with respect to ground truth.

Marks Distribution

Task/Component	Max Marks	Comments
Easy Task Implementation	20	Implementation of a basic VAE and clustering with visualization and metric evaluation
Medium Task Implementation	25	Multi-modal feature use, convolutional enhancements, improved clustering, and analysis
Hard Task Implementation	25	Advanced VAE architectures (CVAE/Beta-VAE), multi-modal clustering, extensive evaluation
Evaluation Metrics	10	Correct computation of clustering metrics and thorough analysis
Visualization	10	Clear latent space visualizations, cluster plots, and reconstructions
Report Quality	10	Well-structured NeurIPS-like report with clarity and completeness
GitHub Repository	10	Organized, readable code, reproducibility, and clear instructions

Grading / Judging Criteria

1. Correctness of Implementation: Your VAE implementation runs without errors and produces meaningful latent features.
2. Feature Engineering and Multi-Modality: Proper extraction and use of audio and lyric features; advanced task includes CVAE or Beta-VAE.
3. Clustering Quality: High-quality clusters as measured by Silhouette Score, ARI, NMI, and other metrics.

4. Visualization: Latent space representations and cluster visualizations should be clear, informative, and interpretable.
5. Comparison with Baselines: Results should be compared with PCA, K-Means, or Autoencoder baselines and discussed properly.
6. Report Quality: Your paper should be well-structured, follow NeurIPS format, and clearly present methods, experiments, and conclusions.
7. Code Quality and Reproducibility: Your GitHub repo should include readable code, clear instructions, and scripts to reproduce results.
8. Creativity and Effort: Extra points for innovative approaches, hyperparameter tuning, disentangled latent space, or multi-modal fusion.

GitHub Repository Structure (Suggested)

```
project/
 data/
    audio/
    lyrics/
 notebooks/
    exploratory.ipynb
 src/
    vae.py
    dataset.py
    clustering.py
    evaluation.py
 results/
    latent_visualization/
    clustering_metrics.csv
 README.md
 requirements.txt
```

Sample Report Template

URL of Overleaf Latex template of NEURL-IPS(2024) https://www.overleaf.com/latex/templates/neurips-2024/tpsbbrdqcmsh

1 Abstract

Brief summary of project goal, method, experiments, and key results.

2 Introduction

Motivation, background, and hybrid music clustering context.

3 Related Work

Summary of VAE methods, clustering techniques, and music representation.

4 Method

- VAE architecture description
- Feature extraction (audio + lyrics)
- Clustering methods

5 Experiments

Dataset description, preprocessing steps, hyperparameters, and training details.

6 Results

- Clustering metrics: Silhouette, ARI, NMI, etc.
- Latent space visualizations
- Comparison with baseline methods

7 Discussion

Analysis of results, limitations, and interpretation of clusters.

8 Conclusion

Summary of findings and future work.

9 References

Include relevant references for VAE, clustering, and music representation.