Project Proposal: Boost Model Performance from Small Amount of Multi-Modal Data

Yibin Wang, Jiaxi Xie, Yuyuan Yang, Kina Huang, Zijin Hu

1. Introduction and Motivation

Humans efficiently integrate multi-modal information like sounds and vision. Can machine learning models replicate this with multi-modal inputs? Previous work successfully combined images and text in vision encoders, now used in LLMs like GPT (Raford et al., 2021). Animal classification has been done using either images or sounds separately (J.Deng et al, 2009; Hagiwara et al., 2023). This project will comparatively study encoder architectures for multi-modal classification, focusing on small datasets of animal acoustics and images. We'll also analyze generative and contrastive learning approaches, explore label noise impact in pre-training, and investigate bidirectional image-audio translation, all in limited data scenarios.

2. Objectives

This study aims to compare generative and contrastive learning in multi-modal classification using small animal acoustic and image datasets. We will evaluate the impact of label noise injection during pre-training on encoder robustness and generalization, and assess the scalability and performance of different models on classification and bidirectional modality translation tasks with varying training set sizes (20%, 50%, and 80% of the total dataset).

3. Methodology

Our approach utilizes CNN-based encoders for image and audio modalities, employing a traditional CNN for images and a 1D-CNN for audio. This captures unique modal features while maintaining architectural consistency for comparison. We'll investigate both generative and contrastive learning approaches.

To study the effect of label noise, we will introduce label noise during pre-training at various levels (e.g., 5%, 10%, 20%). This will allow us to compare model robustness and generalization with and without label noise, providing insights into the impact of noisy labels on small datasets.

For modality translation, we will implement both image→audio and audio→image translation models to explore the bidirectional relationship between modalities and its impact on classification performance in the context of animal acoustic and image data.

To analyze model behavior, we'll employ feature map visualization using dimensionality reduction (e.g., t-SNE), and feature space structure examination (e.g., K-means). Additionally,

we'll evaluate the models' training size scalability by assessing performance with 20%, 50%, and 80% of the total dataset, simulating different scenarios of data availability. Also, we plan to compare the performance of a classification model trained on a large number of images with our model trained with images and audios to check the benefits of adding one more modality to data.

4. Datasets

We plan to use several datasets focusing on animal acoustics and paired images:

1. Animal Sound Dataset (https://github.com/YashNita/Animal-Sound-Dataset)
2. BEANS Dataset (https://github.com/earthspecies/beans)
3. Bioacoustics Datasets (https://bioacousticsdatasets.weebly.com/)
4. Animal Image Dataset (https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals)

To simulate learning from small data, we will create subsets of these datasets at 20%, 50%, and 80% of the total data. This approach will allow us to systematically study the impact of dataset size on model performance and generalization.

5. Expected Outcomes

Our study aims to comprehensively compare generative and contrastive learning approaches for multi-modal classification using small animal acoustic and image datasets. We'll also evaluate the effect of label noise on model's performance and generalization. The potential benefits of cross-modal information in small data scenarios will be evaluated against single-modality models trained on larger datasets. Visualizations of feature maps and clustering results will be provided to understand how different methods structure feature spaces, elucidating their effectiveness in small data contexts.

6. Planned Timeline

October 2 - October 8: Dataset preparation and subset creation (20%, 50%, 80%).
October 9 - October 30: Implement baseline models. Develop models (CNN, generative, contrastive), set up pre-training with label noise, and develop multi-modality integration and translation approaches.
October 31 - November 24: Train models, run experiments, and tune hyperparameters.
November 25 - December 4: Evaluate model performance, analyze results, and visualize features (t-SNE, etc.).
December 5 - December 12: Finalize results, write the report, and prepare the presentation.

Reference:

Banerjee, Sourav. “Animal Image Dataset (90 Different Animals).” Kaggle, July 17, 2022. https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals.

“Bioacoustics Datasets.” BIOACOUSTICS DATASETS. Accessed October 2, 2024. https://bioacousticsdatasets.weebly.com/.

Hagiwara, Masato, Benjamin Hoffman, Jen-Yu Liu, Maddie Cusimano, Felix Effenberger, and Katie Zacarian. "BEANS: The benchmark of animal sounds." In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5. IEEE, 2023.

J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. "Learning transferable visual models from natural language supervision." In *International conference on machine learning*, pp. 8748-8763. PMLR, 2021.

YashNita. “Yashnita/Animal-Sound-Dataset.” GitHub. Accessed October 2, 2024. https://github.com/YashNita/Animal-Sound-Dataset.