Environmental Sound Classification

Saksham Singh Kushwaha, sxk230060

- **Dataset**:
  - Environmental Sound Classification (ESC-50) [2]
  - This is a labeled collection of 2000 audio recordings suitable for benchmarking methods of environmental sound.
- **Problem**: To analyze and compare the performance of several machine learning models in supervised and zero-shot settings on the ESC-50 dataset.
- **Exploratory Data Analysis**:
  - Analysing the data statistics like number of classes, number of folds, number of datapoints per class, and audio length distribution of datapoints.
  - Visualizing some examples from different classes by plotting waveforms and spectrograms.
- **Machine algorithms** :
  I will compare the following models on ESC-50 dataset :
  - **Supervised classification**
    - **Models:** Logistic Regression, SVM, DecisionTree, XGBoost, Neural networks, and Convolutional Neural networks (CNN)
    - **Features:** For CNN spectograms and melspectogram will be used as input features while for other methods audio-extracted features i.e. Mel-Frequency Cepstral Coefficients (MFCCs), Spectral Rolloff, Intensity, loudness etc. will be used.

    - **Hyper-parameters:** For all the models learning rate will be a parameter. For tree-based models(i.e. decision tree and xgboost) max-depth, max-leaf nodes etc will be the hyperparameters. We can tune these hyperparameters by comparing performance on validation data.
  - **Zero-shot approach** Pretrained audio and text encoders will be used extract audio and text embeddings and their cosine similarity will be used to predict the class label for test data points of ESC-50.
    - **Models:** AudioClip [1], Laion-clap [3] (Note that the training data of these models does not contain ESC-50 and hence it's a zero-shot classification)
    - **Features:** Raw audio or spectrograms
    - **Hyperparameters :** No hyperparameters
  - **Eval Criterion**: It is a single-label multiclass classification problem and hence I will use accuracy.
  - **Tools**: Python, scikit learn, Pytorch, matplotlib

References

[1] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[2] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

[3] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.