Eric Engelhart
Machine Learning

Final Project Report

Cross-Lingual Audio Retrieval with Neural Networks using the CVSS English-German subset

Abstract:
This project explores the feasibility of using a neural network model to facilitate cross-lingual audio retrieval between English and German speech. The model employs a contrastive loss function to learn embeddings from short audio clips, which facilitates the identification of semantically similar utterances across the two languages. Evaluation metrics such as recall @ k are used to assess performance, with the final iteration achieving 7% for recall @ 1 and almost 17% for recall @ 4, with recall @ 8 reaching 26%. The approach demonstrated potential for applications in language translation and language learning, although the model's effectiveness was constrained by hardware limitations, neural network size, and training duration.

Machine Learning Program Description:

The program created for this project is a neural network that generates embeddings from audio clips of speech. Translation between English and German audio data was approached by using contrastive loss through the PyTorch metric learning library. The Supervised Contrastive Loss implemented allowed for the precise comparison of semantically similar clips in differing languages. The code will be attached in a zip file, and as the various python files employed. Additionally, a text log of the terminal from the final training run will be attached.

Output of the Program:

The model outputs vector embeddings for given audio clips. These embeddings can be used to perform similarity searches through vector operations, such as the dot product which reveals cosine similarity. This mechanism empowers retrieval of the most similar English audio clip for a given German clip within the dataset, and vice-versa.

Analysis of the Output:

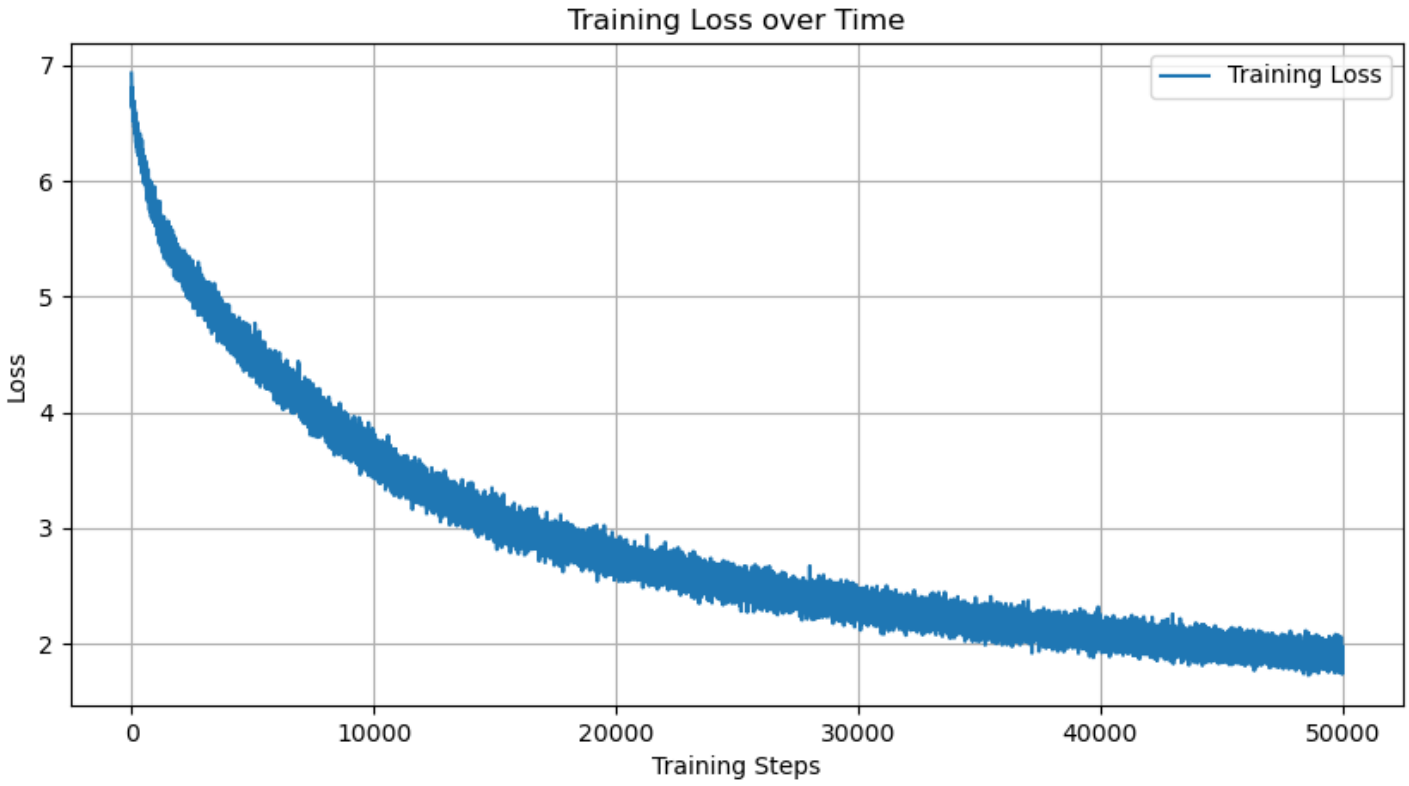
The analysis of the output indicates that while the model correctly identified the paired English audio clip 7% of the time at recall @ 1, expanding the search to the top four results increased the rate to 17%, and recall @ 8 increased to 26%. This suggests that

the embeddings capture enough linguistic and acoustic features to group similar utterances, even if they are not the exact pairs, hinting at broader semantic understanding.

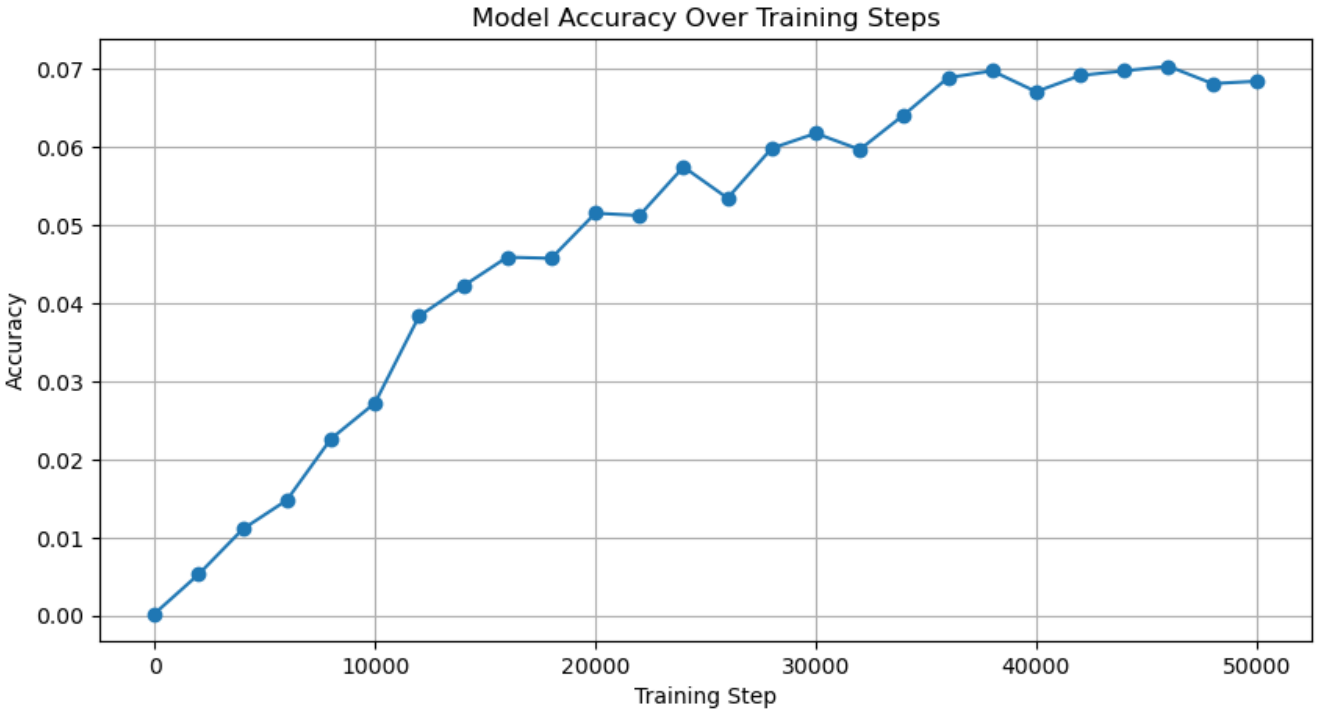
Description of the Learning Process:
The model learned to map audio clip embeddings into a multi-dimensional space where similar linguistic content, despite language differences, would be closer. This was achieved through a supervised contrastive learning approach, which strategically used paired English-German clips within training batches. The model's learning was evidenced not only by improved recall rates but also by how its performance evolved to capture broader semantic relationships beyond exact matches.

To illustrate the learning process, here is a plot of the loss and a plot of the accuracy of the final model.

The loss:

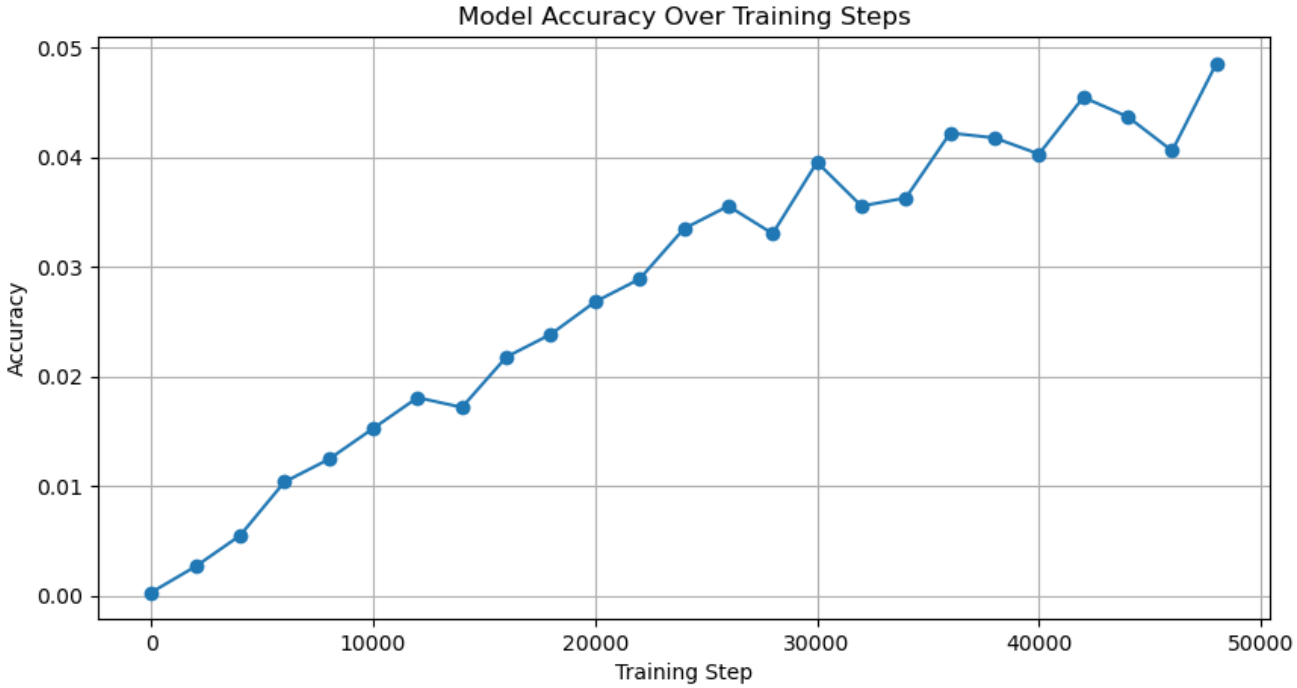


The accuracy (Recall @ 1):



What is interesting is that a previous training run with a bigger model (more layers), but a smaller batch size, had a noticeably worse result in terms of accuracy.

Earlier training run accuracy plot (notice the lower maximum on the y axis):



Unfortunately, these results used all the memory on the GPU used to train the model, so the model had to be made smaller in order to increase the batch size, which was done for the final (and highest performing) model. However, this result is not unexpected, since the contrastive loss compares each batch element against the rest of the training batch, so a larger batch should result in a better performance. Unfortunately, due to hardware constraints, how the models performance would change with batch size and model size was not explored deeply.

Model Limitations and Future Work:
Despite promising findings, the model's potential was limited by the neural network complexity, limited number of training steps due to time restrictions, and batch size restrictions due to hardware constraints. Future work could involve scaling up the network, increasing the batch size, extending training time, and utilizing a larger dataset to refine and improve upon these initial promising results.

Summary of Findings:
This work successfully demonstrates a prototype model for cross-lingual audio retrieval between English and German. It suggests the viability of using neural network-based embeddings for translation and language-learning applications. However, the study also highlights the critical role of training duration, network size, and computational resources in achieving higher accuracy and effectiveness.

Sources:
- CVSS Dataset
- PyTorch Metric Learning Library Documentation
- Einops Library Documentation
- PyTorch Documentation
- Supervised Contrastive Loss literature: https://arxiv.org/abs/2004.11362
- Using ChatGPT for software development, general ideas, and help writing this report.