Multi-Modal Neural Network Model for Digit Classification: Write-up Report

Malik Rawashdeh
Computer Science Department
Texas A&M University
College Station, TX 77845
malikrawa@tamu.edu

Abstract

This paper addresses the challenge of multi-modal digit classification by developing and evaluating a machine learning model that leverages both image and audio data streams. I use a provided dataset consisting of 60,000 samples of 28x28 pixel grayscale images and corresponding audio clips. The project explores the effectiveness of a multi-modal fusion approach. The methodology involves training separate convolutional neural network (CNN) encoders for the image and audio data, followed by concatenating these encodings for classification through a neural network. Results from the Kaggle competition dataset revealed that the fusion model achieved an F1 Score of 0.978, significantly outperforming the individual modalities, which recorded maximum accuracies of approximately 83% for audio and 95% for images. The success of the fusion model highlights how combining different types of data can lead to more accurate classifications. To make the model even better, future work could explore using more advanced model architectures and fine-tuning the settings more precisely.

1 Introduction

This report presents the development and evaluation of a multi-modal machine learning model designed to classify numerical digits using two input streams; images and audio. The objective is to harness the distinct characteristics of image and audio inputs through two separate encoder models, enabling robust digit recognition. This project focuses on demonstrating a multi-modal fusion method which will help make more robust models that predict outputs more accurately.

Song et. al[4] explores various application of audio-visual multi-modal deep learning and cites that the first motivations to use multi-modal audio-visual fusion of data was for speech recognition. Similar to how neural networks resemble natural human neural structure, natural human phenomena such as the Mcgurk effect have spurred researches to take into account visual data to increase the robustness of the speech detection models. The McGurk effect illustrates how human perception of sound can change when accompanied by visual cues [5].

This assignment draws inspiration from these sorts of phenomena. In this assignment, I am tasked with creating a model to accurately predict the digit labels for data samples which contain 28x28 images of handwritten digits and audios of the spoken digits.

Other studies have utilized multimodal fusion to enhance speech recognition in noisy environments, such as the work by Kuniaki Noda et al., who employed an encoder-decoder model that integrates data streams into a multi-stream hidden Markov model [3]. In this project, I use an encoder model to process the two data streams, after which I concatenate the encodings and feed them into a neural network for classification. Many studies concur that concatenation is the most effective method for fusing these encodings [1]. I default to using a straightforward neural network for classification in this project since the audio clips are brief and the data does not contain significant noise that requires filtering, and from the previous homework, I have already achieved a high image classification accuracy using this sort of simpler architecture.

2 Your Method

The methodology involves training 2 separate neural network encoder models for the audio and image data streams separately. Then these encodings are fused and fed to another neural network which classifies the correct digit labels for each of the samples. The effectiveness of this method is demonstrated by embedding the encodings using t-SNE into 2D space and visualizing the formation of clusters corresponding to the ten digit labels. The effectiveness of this multi-modal fusion is also shown through k-means clustering, with a comparative analysis of the image and audio embeddings providing insights into the different contributions of each modality. This analysis draws on previous studies that have highlighted the benefits and challenges of multi-modal learning, offering a benchmark for evaluating our model's performance.

2.1 Data Preprocessing

The data is available on the course Kaggle page here The dataset is a multimodal MNIST dataset. It contains images of written digits as well as the audio of spoken digits, paired with the corresponding label (0-9).

The dataset provided consisted of numpy array files. The training dataset included two arrays: one for audio samples and another for images, accompanied by corresponding labels. This dataset comprised a total of 60,000 samples. A test dataset was also provided, structured similarly but without labels, intended for model predictions.

The image samples in the dataset were each a 28x28 pixel grayscale image, and the audio samples were arrays of 507 values in length. An inspection of the samples revealed that they were normalized, with all values scaled between 0 and 1. To facilitate processing the data in the 2D convolutional layers of the model, the samples were reshaped into a 2D matrix. Subsequently, they were batched into groups of 64 samples each and loaded into a dataloader. This preprocessing step was critical for ensuring efficient data handling and feeding into the machine learning model during training and testing phases.

2.2 Model Design

The model architecture for this project was inspired by the convolutional neural network (CNN) classifier I developed in Homework 4, where I successfully applied it to classify digit images from the same training dataset. Drawing on this previous experience, I adapted the CNN architecture to serve as the foundation for the encoder model used in this project. This decision was driven by the need to assess the CNN's capability for feature extraction from both audio and image datasets separately, evaluating its performance in a different modality.

2.2.1 Encoder Architecture Details

The architecture of the encoder model is essentially the same as the CNN classifier from Homework 4, with modifications to make it suitable for encoding rather than classification. Specifically, the linear layers, which included blocks of linear transformations typically used for classification, were removed. This adjustment focused the model on feature extraction, necessary for the encoding process rather than direct classification.

The encoder architecture is outlined as follows:

- **CNN Layers:** Each CNN layer comprises convolutional blocks designed to process spatial features from the inputs. These blocks use a combination of convolutional layers, batch normalization, ReLU activation functions, and dropout layers to enhance the model's generalization capabilities.

- **Downsampling:** After processing through the CNN layers, the output is passed through a downsampling layer, which reduces the dimensionality of the feature maps, preparing them for encoding.

The key parameters of the CNN blocks, such as the number of hidden channels, kernel size, and stride, were retained from the CNN classifier model. These parameters were optimized through cross-validation during the initial classifier design, ensuring that the encoder benefited from the most effective configuration for feature extraction.

2.2.2 Fusion and Classification Model

After training the encoders and encoding all the images and audio samples in the training dataset, the encodings were flattened. Each image sample encoding was then concatenated with the corresponding audio sample encoding, effectively fusing the data representations from both modalities.

The fused data encodings were subsequently fed into a neural network consisting of two fully connected linear layers with a ReLU activation function between them. The output layer of this network produces a probability distribution over the 10 classes, with the highest probability indicating the predicted class.

This detailed description of the model design and architecture not only clarifies the modifications made to adapt the CNN classifier for encoding tasks but also provides a clear understanding of how the models were implemented and functioned together to achieve the project's objectives.

2.3 Model Training

The training of the models involved distinct steps for the encoder and the fusion model, with specific focus on optimizing performance for feature extraction and classification.

2.3.1 Encoder Training

The encoder models, both for images and audio, were trained using a similar approach but were treated separately to accommodate their respective data types. The training process can be described as follows:

- **Loss Function and Optimizer:** The encoder was trained using the Mean Squared Error (MSE) loss function. This choice is appropriate as the objective was to minimize the reconstruction error between the encoder's outputs and its inputs, thereby improving the model's ability to capture and reconstruct key features. The Adam optimizer was used with a learning rate of 1×10^{-3}, which provides an effective balance between speed and accuracy in convergence.
- **Training Procedure:** The model was set to training mode, initiating the process with the specified number of epochs. During each epoch, the training data was loaded in batches. Depending on whether the model was training on image or audio data, the appropriate input type was selected and forwarded to the device for processing.
- **Backpropagation:** For each batch, the gradients were initialized to zero to prevent accumulation from previous iterations. The model then performed a forward pass followed by the computation of the loss. The loss was then backpropagated to update the model parameters in an effort to minimize the reconstruction error.
- **Loss Monitoring:** The total and average losses were calculated and printed at the end of each epoch to monitor the training progress and ensure stability in the learning process.

This systematic approach to training ensured that each encoder effectively learned to compress and reconstruct its respective input data.

2.3.2 Fusion Model Training

Following the encoding phase, the fusion model was trained. This model took the encoded representations from both the image and audio encoders, concatenated them, and processed the combined features through fully connected layers. The model aimed to classify these features into one of the 10 predefined classes. The training for the fusion model followed standard supervised learning procedures, similar to those used for the encoders but tailored to classification tasks with appropriate loss functions and metrics for performance evaluation. A categorical cross-entropy loss function was employed to optimize the model. . Performance metrics such as accuracy and F1 were used to evaluate the model.

2.4 Hyperparameter Tuning

Hyperparameter tuning was a critical step in optimizing the performance of my CNN classification models. To achieve the best results, I employed cross-validation, systematically testing different combinations of hyperparameters to identify those that maximized classification accuracy. This was done in HW4 so I just reused the results from that. The hyperparameters I explored included:

- **Number of Layers:** I experimented with 2 and 3 layers to determine the impact of model depth on performance.
- **Activation Functions:** Both ReLU and Sigmoid functions were tested to evaluate their effectiveness in my network architecture.
- **Dropout Rates:** Dropout rates of 0.2 and 0.5 were considered to prevent overfitting and ensure generalization across unseen data.
- **Kernel Sizes:** I used kernel sizes of (3, 3) and (5, 5) to understand how the size of the filter affects feature extraction.
- **Strides:** A stride of 1 was maintained across all models to maintain a consistent step size across the input.

Using the best combination of these hyperparameters, I achieved a maximum accuracy of approximately 83% for audio classification and around 95% for image classification.

3 Results

This section details the performance of the developed models, assessed through a series of evaluations using a held out dataset available from the class Kaggle competition SP24-TAMU-CSCE-633-600 Machine Learning. Models were evaluated based on their F1 Score with macro averaging, focusing on their ability to classify multi-modal data accurately.

3.1 Model Performance Evaluation

3.1.1 Fusion Model Performance

The fusion model, which integrates the features extracted from both the image and audio encodings, achieved an F1 Score of 0.978 on the Kaggle test dataset. This score surpasses the third benchmark set by the competition, highlighting the effectiveness of the fusion approach in handling complex, multi-modal datasets.

3.1.2 Individual CNN Model Performance

In comparison to the fusion model, individual CNN models trained on single modalities showed lower performance:

- The audio-only CNN model reached a maximum accuracy of 83.4% on the validation dataset.
- The image-only CNN model achieved a higher accuracy of 95.1% on the validation data set.

These results illustrate the substantial improvement gained through the fusion of audio and image data over using single modality data for classification.

3.2 Visualizations of Model Encodings

To further analyze the model's capability in feature extraction and representation, 2D embeddings of the image and audio encodings were visualized. The following figures display these embeddings, color-coded by labels to show the clustering of the encodings:

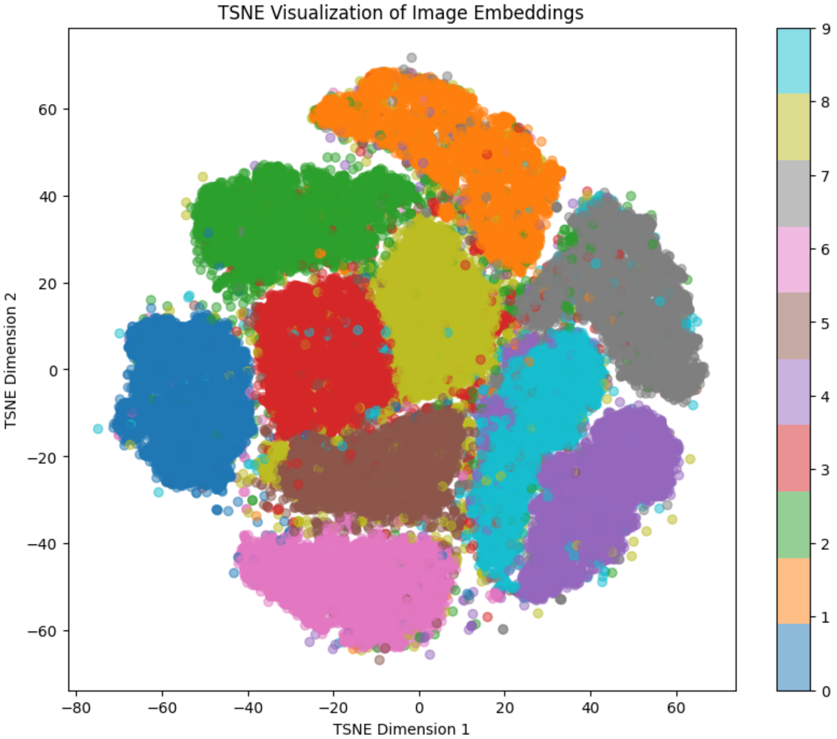

Figure 1: 2D Embeddings of Image Encodings, showing distinct clusters by labels.

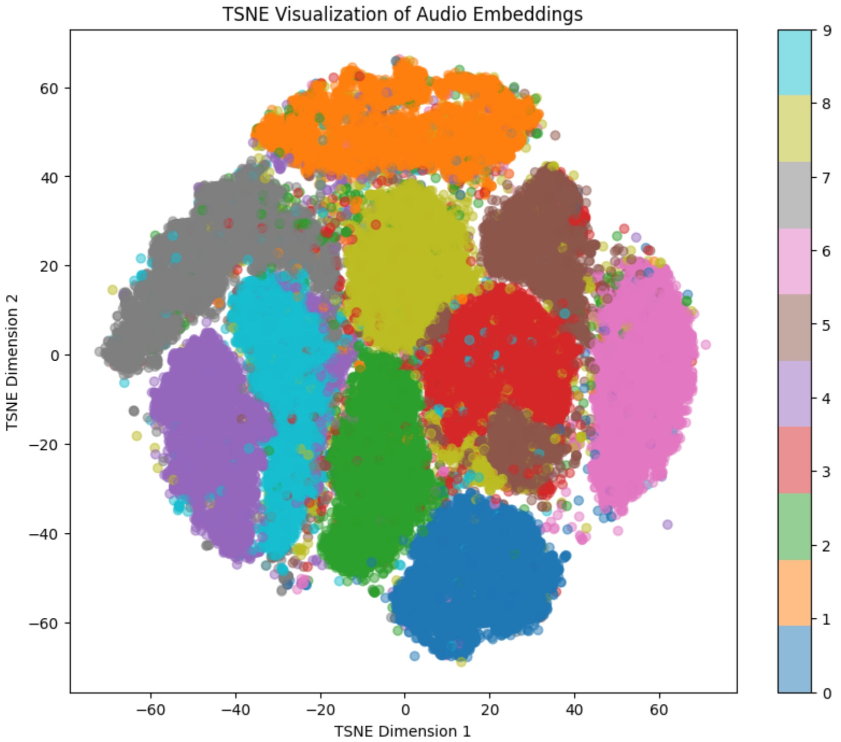

Figure 2: 2D Embeddings of Audio Encodings, illustrating clustering effectiveness.

Additionally, K-means clustering was applied to these embeddings to investigate potential patterns and separability:

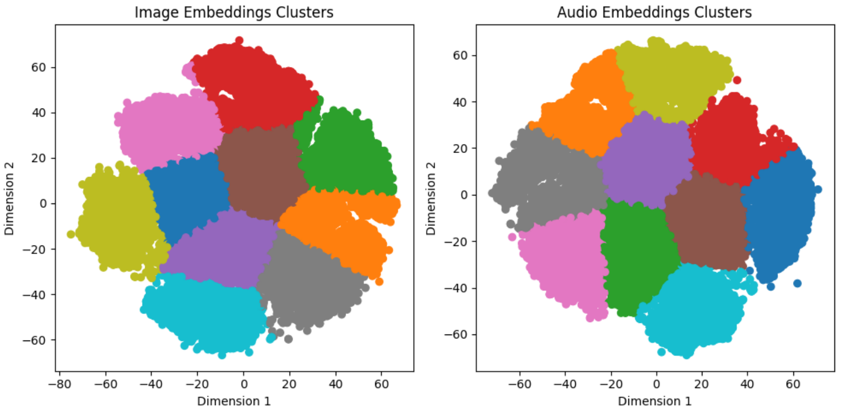

Figure 3: K-means Clustering of Image Embeddings and Audio Embeddings, highlighting the data separability.

Each figure provides insights into how well the encodings of different modalities are differentiated and grouped, which correlates with the classification performance observed.

4 Conclusion

This project's journey through the development and evaluation of a machine learning model for multi-modal data classification culminated in significant insights and achievements. The choice of employing a fusion model, integrating features from both image and audio data, was verified by the high performance on the Kaggle competition dataset, where an F1 Score of 0.978 was achieved. This score not only surpassed the third benchmark level set which shows that this model was reasonably and highly accurate when combining both data streams. Using only a single mode of data, yielded maximum accuracies of approximately 83% for audio and 95% for images.

The success of the fusion model can be attributed to several key factors. Firstly, the rigorous hyperparameter tuning and cross-validation ensured that each component model encoding the data streams was operating at optimally. Secondly, the decision to use CNN for both modalities was crucial due to its proven capability in extracting robust features from complex datasets.

Despite these successes, there is always room for improvement. Other students were able to achieve over .99 accuracy. To enhance the model's performance further, future work could explore several avenues:

- **Advanced Architectural Innovations:** Employing newer neural network architectures like Transformers, which have shown great promise in various domains, could potentially boost the classification accuracy. A relevant study in this context is by Naranchimeg et. al, which presents CNN-based multimodal learning models utilizing both visual (images) and audio (sounds) data for classifying bird species [2].
- **More Extensive Hyperparameter Optimization:** Utilizing automated hyperparameter optimization techniques such as Bayesian Optimization could lead to more refined model parameters and potentially better outcomes. I could have added more layers to the fusion model and overall experimented with the fusion model archetecture more.

The methods and decisions taken during this project have proven largely successful, setting a solid foundation for future explorations. The high performance of the fusion model illustrates the potential of machine learning techniques in multi-modal data analysis.

References

[1] Navonil Majumder, Devamanyu Hazarika, Alexander Gelbukh, E. Cambria, and Soujanya Poria. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl. Based Syst.*, 161:124–133, 2018.

[2] B. Naranchimeg, Chao Zhang, and T. Akashi. Cross-domain deep feature combination for bird species classification with audio-visual data. *IEICE Trans. Inf. Syst.*, 102-D:2033–2042, 2018.

[3] K. Noda, Yuki Yamaguchi, K. Nakadai, HIroshi G. Okuno, and T. Ogata. Audio-visual speech recognition using deep learning. *Applied Intelligence*, 42:722 – 737, 2014.

[4] Xiaoyu Song, Hong Chen, Qing Wang, Yunqiang Chen, Mengxiao Tian, and Hui Tang. A review of audio-visual fusion with machine learning. *Journal of Physics: Conference Series*, 1237, 2019.

[5] K. Tiippana. What is the mcgurk effect? *Frontiers in Psychology*, 5, 2014.