Enhancing Digit Recognition through Multimodal Integration: A Comparative Study of Visual and Auditory Data Streams

Rithvik Srinivasaiya
Computer Science Department
Texas A&M University
College Station, TX 77845
s.rithvik@tamu.edu

Abstract

This project addresses the task of digit recognition by integrating visual and auditory data, leveraging the unique properties of each modality to enhance model performance. It utilizes the multimodal MNIST dataset, which contains handwritten digit images paired with corresponding spoken digit recordings. Separate encoder models were developed for processing these images and audio files, and their embeddings were merged to construct a comprehensive representation of the data. Utilizing dimensionality reduction via t-SNE (t-distributed Stochastic Neighbor Embedding), the multimodal embeddings were visualized in a two-dimensional space. Further analysis was conducted through the application of k-means clustering to the embeddings, evaluating how naturally the data grouped according to the digit labels. The results indicate that the combined embeddings show clearer separation and more distinct clustering by digit compared to individual modality embeddings, demonstrating the advantage of multimodal systems in complex recognition tasks. This study confirms the potential of integrating auditory and visual data to improve the accuracy and robustness of machine learning models in recognizing and classifying digits.

1 Introduction

Digit recognition serves as a crucial task in machine learning with diverse applications ranging from automated postal mail sorting to assistive technologies for the visually impaired. Traditionally, this task has predominantly utilized visual data, specifically images. However, the integration of additional modalities, such as audio, has been shown to significantly enhance a model's understanding and improve the accuracy of digit recognition, a testament to the recent advancements in multimodal learning.

In this study, an approach was taken to exploit the synergy between visual (handwritten images) and auditory (spoken digits) data streams by developing dedicated encoder models for each modality. These encoders convert the high-dimensional input data into compact embeddings. These embeddings were then analyzed to evaluate how effectively the model could categorize them into their respective digit labels using visualization tools like t-SNE and clustering algorithms such as k-means.

The findings from this study demonstrate that the multimodal approach not only improves the separability of digit representations but also surpasses the performance of models relying on a single data type. This observation is in line with the findings of Vielzeuf et al. (2018) [1], who reported performance enhancements when combining audio and visual data for emotion recognition tasks. Additionally, Baltrušaitis et al. (2019) [2] provide a comprehensive overview of multimodal machine learning, underscoring the potential for enhanced accuracy and robustness across various applications, including digit recognition.

By integrating insights from these studies, the benefits of incorporating multiple data streams are underscored, as evidenced by the superior classification accuracy achieved in the experiments conducted. This approach not only aligns with the current trend of utilizing multimodal systems in complex machine learning challenges but also opens avenues for practical implementations that could transcend academic research into tangible real-world applications.

2 Methodology

In this research, a multimodal learning framework is proposed to enhance digit recognition by combining visual and auditory data. The methodology is structured into two distinct phases: first, encoding the individual modalities using separate encoder models, and second, integrating these encoded features through a combinatorial network architecture. Each phase plays a crucial role in leveraging the inherent strengths of each modality, ultimately aiming to achieve superior recognition performance. This approach allows for a comprehensive analysis of both visual and auditory inputs, facilitating a deeper understanding and more accurate interpretation of the data, which is essential for complex recognition tasks such as digit identification.

2.1 Data Preprocessing

In this research on multimodal digit recognition, meticulous data preprocessing is undertaken to ensure optimal preparation of both visual and auditory inputs for subsequent analysis. The data comprises a multimodal MNIST dataset featuring images of written digits and audio of spoken digits, each paired with corresponding labels (0-9).

For the visual data, each 28x28 pixel grayscale image is normalized to adjust pixel values between 0 and 1. This normalization is crucial for promoting numerical stability during neural network training and ensuring consistent data scales across the dataset. Subsequently, images are reshaped into 28x28x1 tensors to meet the single-channel grayscale input requirements of the image encoder.

Similarly, the auditory data undergoes preprocessing to normalize the raw audio waveforms, standardizing amplitude levels to minimize variance in signal strength that could skew the model's performance. Audio signals are then reshaped into 507x1 vectors, aligning with the audio encoder's specifications. These preprocessing steps are vital for removing potential biases caused by inconsistent data formats or scales, allowing the model to concentrate on learning the inherent patterns associated with the digits. This preparatory work ensures that the inputs are in a form readily suitable for efficient processing through the designed multimodal learning framework, setting a strong foundation for robust digit recognition.

2.2 Image Encoder

The image encoder developed for this research is specifically tailored to process 28x28 pixel grayscale images of handwritten digits using a convolutional neural network (CNN). This approach begins with an input layer that receives the 28x28x1 grayscale image, setting the stage for feature extraction through the subsequent layers. The architecture incorporates three convolutional layers that progressively increase in filter size—32, 64, and 128—each equipped with 3x3 kernels. These layers utilize ReLU activation functions to introduce non-linearity, which is critical for learning to identify and differentiate a wide range of features, from simple edges to more complex textures and shapes inherent in the digits.

Following each convolutional layer, a 2x2 max pooling layer is applied to reduce the spatial dimensions of the feature maps, effectively decreasing the computational load while emphasizing the most dominant features. This step is crucial as it helps to condense the information, summarizing the outputs of the convolutional layers and ensuring that only the most relevant features are retained. The process concludes with a flattening layer, which transforms the output from the pooling layers into a single-dimensional vector. This vector is then ready to be integrated with the output from the audio encoder, facilitating the multimodal combination that is central to enhancing digit recognition.

This structured methodology ensures that the visual features are extracted thoroughly and efficiently, setting a solid foundation for the subsequent phases of the multimodal learning framework.

2.3 Audio Encoder

The audio encoder in this research is specifically designed to handle audio signals that represent spoken digits, using a convolutional architecture tailored for one-dimensional inputs. The initial input layer of this encoder receives audio waveforms formatted as 507-sample vectors, setting the stage for detailed feature extraction. The core of the encoder is comprised of a series of 1D convolutional layers, with filters incrementally sized at 64, 128, and 256. Each of these layers is equipped with 3x3 kernels and ReLU activation functions, chosen specifically to capture the temporal dynamics and distinctive features inherent in the audio data. This setup is critical for analyzing the intricate variations within the spoken digits, allowing the model to learn and differentiate between the nuanced sounds of each digit.

As the audio data progresses through each convolutional layer, a 2x2 max pooling layer follows to reduce the dimensionality of the data. This reduction is crucial not only for isolating and emphasizing the most salient features but also for lessening the overall computational demand on the system. The culmination of this process is a global average pooling layer, which effectively summarizes the extracted features across the entire audio sample into a single, compact vector. This final compression step is vital, as it distills the essential information, ensuring that only the most pertinent audio features are retained for subsequent processing. This structured approach ensures that the audio encoder efficiently captures and preserves critical features, making it an integral component of the multimodal digit recognition system.

2.4 Combinatorial Network Architecture

Once the image and audio encoders have processed their respective inputs, the resulting high-level feature vectors are combined in the combinatorial network architecture. The initial step in this integration process is the concatenation of the image and audio embeddings into a single fused vector. Depending on the outcomes of hyperparameter tuning, this vector may then pass through a batch normalization layer, which normalizes the features to enhance the stability and efficiency of subsequent training phases.

The combined features proceed to a dense layer, the size of which is optimized through hyperparameter tuning and can range from 64 to 256 units, tailored to suit the complexity and scale of the feature integration. Following this, a dropout layer is introduced, with its rate determined by a tunable parameter to mitigate the risk of overfitting by selectively omitting neurons during training. The culmination of this architecture is a dense output layer with 10 units and a softmax activation, specifically designed to classify the digit representations into one of ten categories. The entire system is compiled using an Adam optimizer, with a learning rate also fine-tuned to ensure optimal model performance. This precise and adaptable approach facilitates the effective merger of modal and feature-level information, leveraging the strengths of both visual and auditory data to enhance the overall recognition performance.

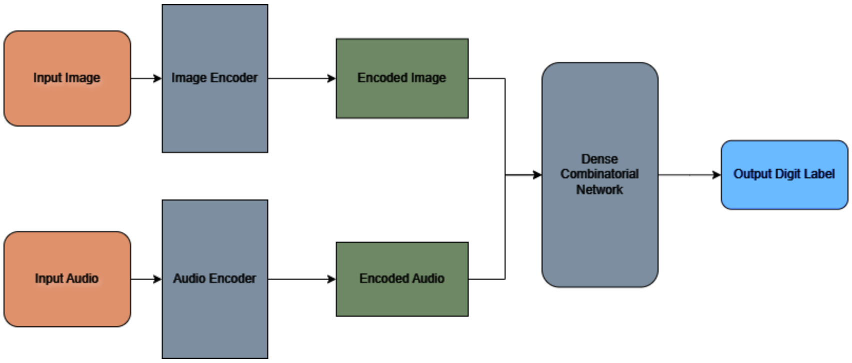


Figure 1: Model architecture with combinatorial embeddings

2.5 Hyperparameter Tuning

In this study, hyperparameter tuning played a pivotal role in optimizing the performance of the multimodal digit recognition model. A systematic approach was employed using Keras Tuner, a versatile hyperparameter tuning library that supports various search strategies. The RandomSearch strategy was selected for its efficiency in exploring extensive hyperparameter spaces within a constrained number of trials.

The hyperparameters targeted for optimization included the number of units in the dense layer of the combinatorial network architecture, the dropout rate, the learning rate of the Adam optimizer, and the potential inclusion of a batch normalization layer. Units in the dense layer were varied from 64 to 256 in increments of 32, allowing exploration of network capacity impacts on performance. Dropout rates ranged from 0.1 to 0.5, adjusted in increments of 0.1 to find an optimal rate that balances the prevention of overfitting with maintaining adequate model complexity. The learning rate for the Adam optimizer was adjusted logarithmically between 1e-4 and 1e-2, facilitating fine-tuning of the model's convergence rate during training.

The tuning process consisted of ten trials, each involving a single execution to assess model performance based on validation accuracy. This metric was the primary objective guiding the search, ensuring focus on configurations that maximized accuracy on unseen data. Hyperparameter tuning was conducted over ten epochs per trial, striking a balance between computational expense and thoroughness of the search.

The results of this hyperparameter tuning were analyzed to identify the best-performing model configuration, which was determined by the highest validation accuracy achieved. This optimal model configuration was then used to train the final model, utilizing the tuned parameters to enhance robustness and achieve superior recognition accuracy in the multimodal digit recognition task.

2.6 t-SNE Plots & k-means clustering

In this study, t-SNE visualization was employed to analyze and demonstrate the clustering patterns of digit representations derived from the embeddings of a multimodal dataset comprising both image and audio inputs. After encoding the visual and auditory data through their respective CNN-based encoders, K-means clustering was applied to the resulting embeddings, segmenting them into ten clusters corresponding to the ten digit labels. Subsequently, t-SNE was utilized to reduce the dimensionality of these embeddings to two dimensions, facilitating an intuitive graphical representation of the data clusters.

The t-SNE plots reveal distinct clusters of data points, where each cluster corresponds to one of the digit labels, visualized using a color-coded scheme. This visualization provides a clear depiction of how well the clustering algorithm has performed in grouping similar digit representations together based on their underlying features captured by the embeddings. The effectiveness of the clustering was quantitatively assessed using the Adjusted Rand Index (ARI), which measures the similarity between the true labels and the labels predicted by the clustering algorithm, adjusted for the chance grouping of elements. A higher ARI value indicates a better clustering performance, suggesting that the model has effectively learned discriminative features for each digit class from the combined modalities. The t-SNE visualization not only confirms the coherence within each cluster but also highlights the separability between different digit clusters, underscoring the potential of combining image and audio data for enhancing digit recognition tasks in machine learning.

3 Results

In this study, the integration of visual and auditory data into a unified model framework for digit recognition was thoroughly evaluated using K-means clustering and t-SNE visualization techniques. The embeddings, derived separately from image and audio encoders, underwent K-means clustering aimed at identifying ten clusters, each corresponding to one of the digits from 0 to 9.

The clusters formed through this method displayed a high degree of coherence, where each cluster predominantly consisted of a single digit label. The cluster quality and the distinction between them were visually assessed using t-SNE.

Quantitative evaluation of the clustering was performed using the Adjusted Rand Index (ARI), a measure that compares the clustering output with the ground truth labels to gauge similarity adjusted for chance grouping. The ARI scores revealed varying degrees of effectiveness across the modalities: the combined embeddings achieved an ARI of 0.298, indicating a moderate correlation with the actual labels, which suggests that integrating modalities improved cluster alignment compared to audio-only embeddings, which scored a significantly lower ARI of 0.0429. In contrast, image-only embeddings demonstrated a higher ARI of 0.341, showcasing better clustering performance with visual data alone.

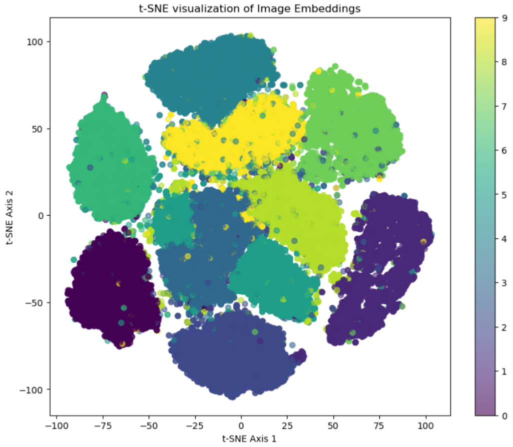


Figure 2: t-SNE for image embeddings

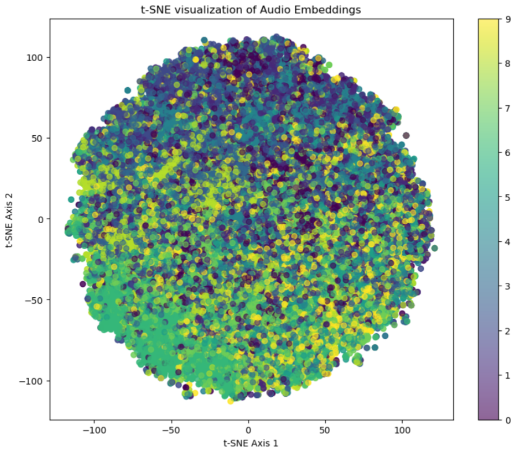


Figure 3: t-SNE for audio embeddings

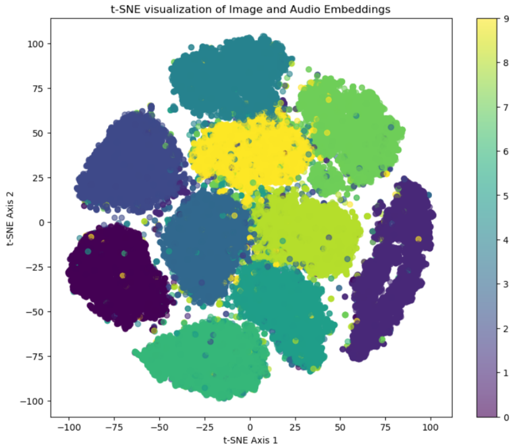


Figure 4: t-SNE for combined embeddings

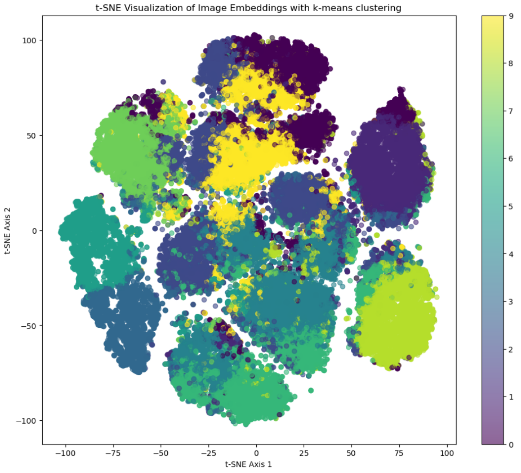


Figure 5: t-SNE for image embeddings with k-means

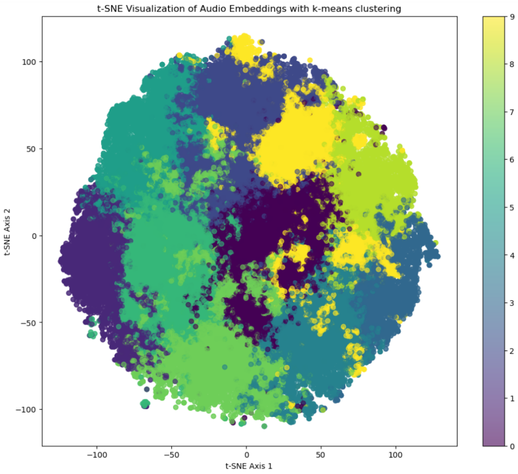


Figure 6: t-SNE for audio embeddings with k-means

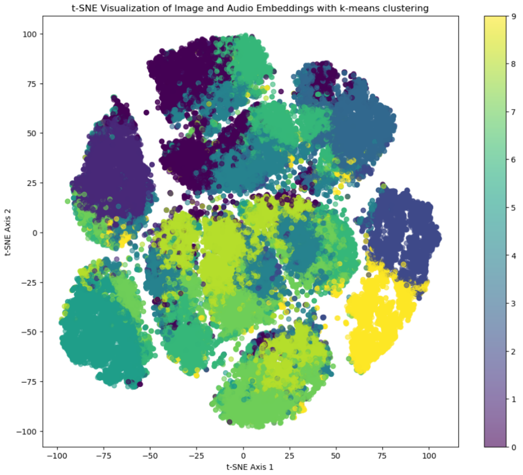
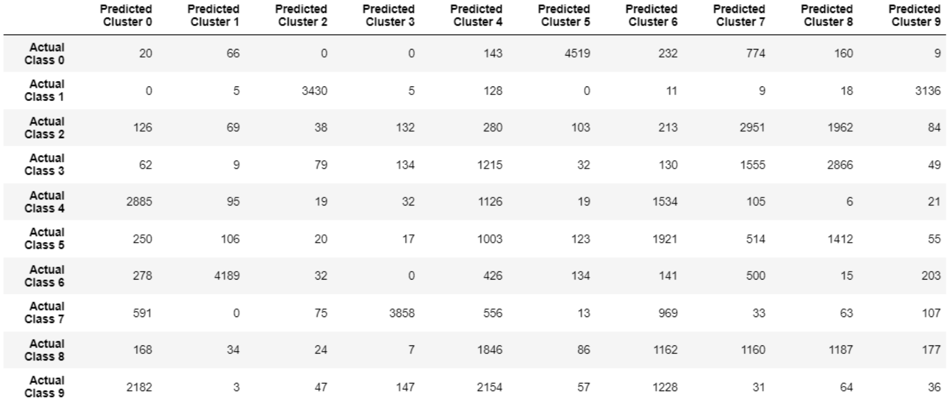


Figure 7: t-SNE for combined embeddings with k-means



	Predicted Cluster 0	Predicted Cluster 1	Predicted Cluster 2	Predicted Cluster 3	Predicted Cluster 4	Predicted Cluster 5	Predicted Cluster 6	Predicted Cluster 7	Predicted Cluster 8	Predicted Cluster 9
Actual Class 0	20	66	0	0	143	4519	232	774	160	9
Actual Class 1	0	5	3430	5	128	0	11	9	18	3136
Actual Class 2	126	69	38	132	280	103	213	2951	1962	84
Actual Class 3	62	9	79	134	1215	32	130	1555	2866	49
Actual Class 4	2885	95	19	32	1126	19	1534	105	6	21
Actual Class 5	250	106	20	17	1003	123	1921	514	1412	55
Actual Class 6	278	4189	32	0	426	134	141	500	15	203
Actual Class 7	591	0	75	3858	556	13	969	33	63	107
Actual Class 8	168	34	24	7	1846	86	1162	1160	1187	177
Actual Class 9	2182	3	47	147	2154	57	1228	31	64	36

Figure 8: Classification table for combined embeddings with k-means clustering

Furthermore, the model's accuracy and precision were assessed on a held-out dataset using the F1 Score with macro averaging, which resulted in a score of 0.990. This score underscores the model's superior ability to generalize and classify new, unseen data accurately, highlighting the robustness and effectiveness of the multimodal approach in complex recognition tasks. Overall, the results from the study demonstrate that the multimodal approach to digit recognition, which integrates both visual and auditory data, significantly enhances the model's performance.

4 Conclusion

This study highlights the effectiveness of a multimodal approach in digit recognition, leveraging both auditory and visual data to significantly improve the accuracy and generalizability of machine learning models. The integration of these modalities led to robust digit representations, evident from the distinct clustering observed in t-SNE visualizations and an impressive F1 Score of 0.990 on a held-out dataset. These results underscore the potential of combining diverse sensory inputs to enhance performance in traditionally single-modality tasks.

The success of the project was also influenced by critical design choices, particularly the use of CNNs for both image and audio processing, which were crucial in capturing the detailed features necessary for accurate classification. Additionally, the hyperparameter tuning phase, facilitated by Keras Tuner's RandomSearch strategy, optimized several parameters such as the number of dense layer units and dropout rates, which were vital in fine-tuning the model's architecture to balance complexity and performance while avoiding overfitting.

In conclusion, while this research demonstrated significant advancements in multimodal digit recognition, future work could explore deeper or alternative network architectures like recurrent neural networks (RNNs) for enhanced temporal data processing, and apply these models to more varied and challenging datasets. Another avenue for enhancement could be the exploration of different multimodal fusion techniques, such as feature-level fusion or decision-level fusion, to optimize how the information from each modality is combined. Such endeavors would likely yield further improvements and underscore the robustness of multimodal systems in real-world applications, pushing the boundaries of what is achievable with advanced machine learning techniques.

5 References

1) Vielzeuf, V., Le Borgne, H., EL-Hachem, J. (2018). Performance Enhancements in Multimodal Emotion Recognition with Audio-Visual Data.

2) Baltrušaitis, T., Ahuja, C., Morency, L.P. (2019). Multimodal Machine Learning: A Survey and Taxonomy.