Weight-based Modality Fusion for Multimodal Digit Recognition

Chonglin Zhang
Computer Science Department
Texas A&M University
College Station, TX 77845
czhang71@tamu.edu

April 24, 2024

Abstract

Multimodal Digit Recognition involves the task of recognizing digits using both handwritten digit images and speech signals. This task explores the fusion of multiple sources of information to enhance the accuracy and robustness of digit recognition systems. In this paper, we propose a weight-based modality fusion method. This method learns a unique weight for each modality and then fuses the modalities based on these weights to obtain fused features, thereby improving the model's performance. Our approach achieved a macro F1 score of 0.992 on multimodal MNIST, surpassing previous benchmarks. Furthermore, a series of ablation experiments demonstrate that our method outperforms single-modal approaches.

1 Introduction

Digit recognition systems have wide-ranging applications in real-world scenarios such as human-computer interaction and biometric security. Traditional methods rely on a single modality to recognize digits, such as identifying handwritten digits from images. However, this approach encounters difficulties when the handwritten digit images are unclear and lack additional information for context. Multimodal digit recognition addresses this challenge by leveraging additional modalities. For instance, when the handwritten image is unclear, associated speech signals can provide supplementary context for accurate recognition. The challenge of multimodal digit recognition lies in how to integrate features from different modalities to produce more accurate and robust results.

In this paper, we propose a weight-based fusion method to address this challenge. Specifically, we construct a simple model for each modality to encode its features, resulting in feature representations for each modality. Additionally, we learn a separate weight for each modality. Subsequently, we combine each modality based on their respective weights to obtain the final fused features, which are then used for classification. We tested our method on the multimodal MNIST dataset and achieved a macro F1 score of 0.992.

2 Method

The overall structure of our method is illustrated in Figure 1. We first extract features from each modality and then fuse them using a weight-based method to obtain fused features. Finally, we use these fused features to derive the final classification results.

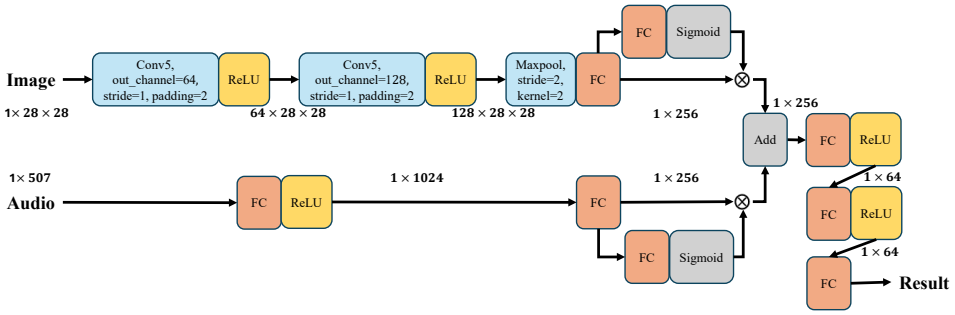

Figure 1: The overall structure of our proposed method, where $\otimes$ denotes element-wise multiplication, and *Add* denotes element-wise addition.

2.1 Data Preprocessing

We performed preprocessing on the image data. For each sample, we resized it to 28×28 dimensions and then divided all pixel values by 255 to normalize them between 0 and 1. This was done for ease of computation during backpropagation and to prevent gradients from becoming too large. As for the speech signal data, since it was already in the range of 0 to 1, we did not perform any further preprocessing on it.

We divided all the data into 80% for the training set and the remaining 20% for the validation set.

2.2 Model Design

The specific details of our model are illustrated in Figure 1. For the image data, we employed a convolutional neural network (CNN) to extract features. We experimented with various CNN model designs in HW4, including changing kernel sizes, activation functions, adding or removing dropout layers, etc. Consequently, we directly adopted the best-performing model from HW4 as the feature extraction network for images. We removed the final classification layer from the model, resulting in each sample being encoded into a 256-dimensional feature vector.

As for the speech data, since it is a one-dimensional signal, we employed fully connected (FC) layers to extract features. Similarly, we used 2 layers of FC networks to encode each sample into a 256-dimensional feature vector.

During the fusion stage, we learned a weight for each modality. Specifically, for each modality, we passed its features through an FC layer with a single output to obtain a weight. These weights were then mapped to the range of 0 to 1 using a sigmoid function. Subsequently, we multiplied each modality's features by its corresponding weight and summed them to obtain the fused feature. The fused feature was then passed through 3 layers of FC networks to obtain the final output.

2.3 Model Training

We constructed our model using PyTorch. For the loss function, we employed the commonly used cross-entropy loss for classification tasks. The model was trained using the SGD optimizer for a total of 40 epochs.

2.4 Hyperparameter Tuning

For the SGD optimizer, we set the momentum to 0.9 and weight decay to 0.0001. The initial learning rate was set to 0.01 and decreased by a factor of 10 every 10 epochs. During training, the batch size was set to 64.

3 Results

3.1 Dataset

We conducted experiments on the Multimodal MNIST dataset to validate our method. Each sample in the Multimodal MNIST dataset consists of both image and speech modalities. The dataset contains a total of 60,000 samples for training. We used 80% of the dataset for training and the remaining 20% for validation.We used the macro F1 score as the evaluation metric to assess the models.

3.2 Ablation Study

Table 1: Comparison of Different Feature Fusion Methods on Multimodal MNIST.

Fusion Method	F1 socre
Image	0.991
Audio	0.752
Add w/o weights	0.991
Add w/ weights	0.992

We conducted experiments on various feature fusion methods, and the results are shown in Table 1. Multimodal fusion yielded better results compared to single modality approaches. Furthermore, the weight-based addition method outperformed the direct addition fusion method.

3.3 Comparison with Other Methods

Table 2: Comparison with Other Methods on Multimodal MNIST.

Method	F1 socre
benchmark_01	0.606
benchmark_02	0.886
benchmark_03	0.952
Ours	**0.992**

The experimental results are shown in Table 2. Our method achieved a macro F1 score of 0.992, surpassing all previous benchmarks, thus validating the effectiveness of our approach.

3.4 Visualization

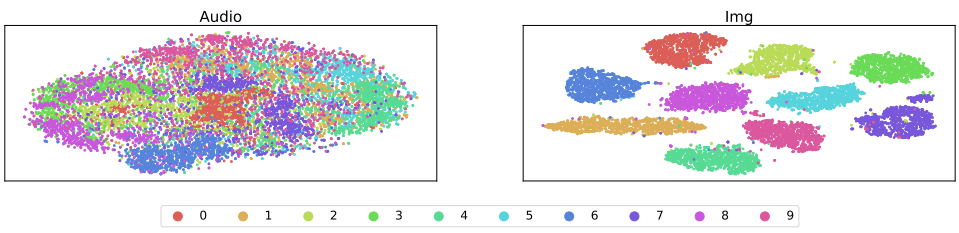

Figure 2: Visualization of Speech and Image Features.

To further evaluate the contribution of different modalities to the model, we visualized the features of the image and speech modalities separately. We used t-SNE to visualize the features extracted by the model on the validation set for both image and speech modalities. The results are shown in Figure 2.

In the visualization of image features, the clusters for each category are distinct and well-separated. In contrast, the visualization of speech features shows less distinct separation between clusters for each category, leading to potential confusion.

The visualizations further support the results in Table 1 regarding the performance of individual modalities. We hypothesize that this outcome may be due to handwritten image data being more easily recognizable compared to speech signals. Additionally, another reason could be the simplicity of our speech feature extraction model.

4 Conclusion

In this study, we proposed a weight-based fusion method for multimodal digit recognition. Our experimental results on the Multimodal MNIST dataset demonstrate that our approach outperforms single-modal methods and other fusion techniques, achieving a macro F1 score of 0.992.

However, our method still has some limitations, notably in the simplicity of the speech feature extraction model, which may not sufficiently extract high-quality features. In future work, we plan to design more sophisticated speech feature extraction models to improve feature extraction capabilities, thus enhancing the accuracy and robustness of the model.