CS482/682 Final Project Report Group 9

Autoencoder for Reconstructing Written Digits from Sound

Chenxiao Fan, Rongkun Zhou, Shiyang Lin, Ziheng Dai

1 Introduction

Background Cross-modal representation learning is an emerging field in machine learning aimed at bridging different data modalities. This project addresses the challenge of mapping audio recordings of spoken digits to their corresponding visual representations. Using the Audio-MNIST dataset, our goal was to develop a generative model that reconstructs MNIST-like images from audio inputs.

Related Work Research in cross-modal learning and spoken digit recognition has explored various approaches.

Sharan (reference: Spoken Digit Recognition Using Wavelet Scalogram and Convolutional Neural Networks) proposed using convolutional neural networks (CNNs) for spoken digit recognition by transforming audio signals into scalograms (time-frequency representations). This method leveraged image-like representations to achieve high accuracy.

Wan et al.(reference: Towards Audio to Scene Image Synthesis using Generative Adversarial Network) explored conditional generative adversarial networks (GANs) for audio-to-image synthesis, generating scene images from sounds. They demonstrated that advanced GAN techniques could improve image quality and model scalability based on audio variations.

Zelaszczyk and Mandziuk (reference: Audio-to-Image Cross-Modal Generation) investigated the use of variational autoencoders (VAEs) in an adversarial framework for reconstructing image archetypes from audio data. Their study highlighted a trade-off between consistency and diversity in generated images, emphasizing the importance of preserving critical features for classification.

2 Methods

Dataset The Audio-MNIST dataset (reference: Audio-MNIST dataset) is a collection of recordings of spoken digits (0-9) in WAV format, designed for tasks like speech recognition and audio classification. It contains 30,000 audio samples recorded by 60 different speakers, with each digit spoken multiple times under varying conditions.

Each sample was preprocessed using MFCC (Mel-Frequency Cepstral Coefficients) to extract 40-dimensional feature vectors by averaging across time frames. The dataset was split into training (70%), validation (15%), and test (15%) sets.

Setup, Training and Evaluation In order to generate the images that we want, we build two models: one for generating images and one for evaluate the generated images. For image-generating model, we use Variational AutoEncoder (VAE) as the base architecture. We use MLP as the encoder in the VAE to encode the preprocessed sound data into the latent space as the distributions, because MLP can work well on classifying preprocessed audios (reference: Audio MNIST Classification 0.98 Accuracy). Then we use transposed convolution layers to upsample the feature representation sampled from the latent distributions into the images, which is a prevailing architecture for generating images. For evaluation model, we use a existing trivial architecture that works very well on MNIST dataset (reference: Github mnist/example) and trained using MNIST training data with cross entropy loss. After training, the model reached 99% accuracy on MNIST test data.

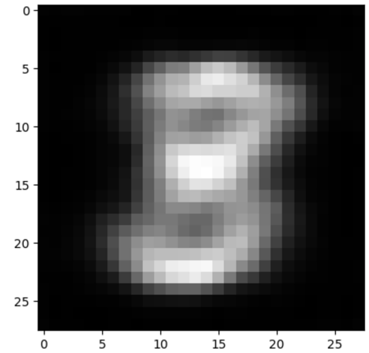
In the training process, we adopt the Evidence Lower BOund (ELBO) as our loss function in training the image-generating model. When we get an image generated from the model from a sound part with a class label, we randomly pick an image with the same class label in MNIST dataset as the baseline image in ELBO. This strategy can help increase the variation of the images generated by the model.

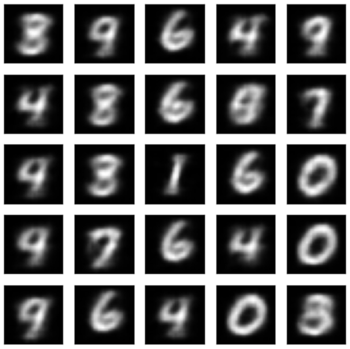
In order to evaluate the performance of the image-generating model during and after the training, we need to use the evaluation model. We use the images generated by the image-generating model with the input sounds as the input of the evaluation model, and calculate the accuracy of the output of the evaluation model in according to the class label of the input sound. This accuracy will be the way we measure the performance of the image-generating model.

But there is a problem: although the generated images can be identified by human, the evaluation model trained on MNIST can't identify them. That's because the images generated by our model are not as clear as these in MNIST dataset. In order to enhance the identification ability of the model, we generate a small sample of images ($n = 100$) and annotate them manually. Then we fine-tune the last two fully-connected layers using the small sample of annotated images. After fine-tuning, the evaluation model performs well: the evaluation accuracy gradually increases during the training process, and finally reaches 82.29% on the test dataset.

3 Results

The model successfully generated visually recognizable MNIST-like digit images from audio features, and achieved high accuracy of 82.29% evaluated by the fine-tuned AccuracyNet. In below figures, the first one displays a single batch of reconstructed digit images, while the second one shows a grid visualization of multiple reconstructed outputs.





All codes and results of this project can be viewed here at GitHub.

4 Discussion

Our cross-modal generation model demonstrates both achievements and limitations in audio-to-visual conversion, with the key challenge being the reliance on manual annotation for evaluation. We could have explored alternative architectures or pretraining strategies to achieve better image clarity without human intervention. While the VAE with ELBO loss effectively maintained digit consistency, investigating more sophisticated loss functions could have better balanced quality and variation. Looking ahead, focusing on end-to-end training methods and improved architecture design could eliminate the manual annotation step while enhancing the model's overall performance.