Adversarial Training for RNN & U-Net in Audio Watermarking

Overview:

This document outlines the integration of RNN and U-Net for adversarial training in audio watermarking.

The RNN encodes metadata, while the U-Net generates watermarked audio.

A separate CNN acts as the discriminator, evaluating the generated audio in an adversarial setup to improve watermark robustness.

Key Points:

1. **RNN (Recurrent Neural Network) for Metadata Encoding**:

Role: The RNN encodes metadata that will be embedded into the audio.

Function: This metadata is passed to the U-Net for watermark generation.

2. **U-Net (Generator Network) for Watermarked Audio**:

Role: U-Net acts as the generator, creating watermarked audio from the metadata provided by the RNN.

Training: It is trained with adversarial loss to improve the robustness and imperceptibility of the watermarks.

3. **Adversarial Loss Setup**:

Function: The adversarial loss is applied to the U-Net during training. This loss encourages the U-Net to generate watermarked audio that is indistinguishable from the original (real) audio.

Discriminator (CNN): A separate CNN-based discriminator network is trained in parallel to distinguish between real and generated audio. This pushes the U-Net to generate more convincing watermarks.

4. **Discriminator (CNN) Network**:

Role: The CNN-based discriminator evaluates the output of the U-Net (generated watermarked audio) and tries to classify it as real or fake.

Training: The discriminator provides feedback to the U-Net, helping it generate better watermarks over time.

5. **Parallel Training**:

The RNN and U-Net are **trained in parallel**, not sequentially like a Feed-Forward Network (FFN). The RNN encodes metadata, and the U-Net uses it to generate the watermarked audio.

Adversarial Training: The U-Net (generator) and the CNN (discriminator) are trained simultaneously in a min-max setup, where the generator improves by trying to fool the discriminator, and the discriminator becomes better at distinguishing real from fake audio.

6. **Training Flow**:

   - **Step 1**: RNN encodes metadata.
   - **Step 2**: U-Net uses the encoded metadata to generate watermarked audio.
   - **Step 3**: The CNN discriminator distinguishes between real and generated audio.
   - **Step 4**: Both the U-Net and the discriminator are updated via adversarial loss, improving watermark robustness.

Goal:

The goal is to train the RNN, U-Net, and CNN in parallel to improve the robustness of the generated watermarks, making them resistant to attacks and more difficult to distinguish from real audio. The adversarial training process enables continuous improvement of both the generator (U-Net) and discriminator networks.