SAMSUNG

AI Course

Capstone Project Final Report

For students (instructor review required)

Deteccion de voces falsas

Fecha (20/11/2024)

Los Nerds

Alejandro pimiento

Ethan Astorga

Content

1. Introduction

1.1. Background Information

El rápido avance de las tecnologías de inteligencia artificial ha permitido el desarrollo de voces artificiales casi indistinguibles de las humanas. Estas herramientas, como Vall-E 2, han abierto nuevas oportunidades pero también nuevos riesgos, como el uso malintencionado de estas voces en estafas telefónicas y suplantaciones de identidad. Una aplicación crítica de este proyecto es abordar desafíos reales como el "problema del CEO", en el cual se utilizan voces falsas para engañar a empresas y bancos, comprometiendo su seguridad.

1.2. Motivation and Objective

El objetivo principal de DetectVoice es desarrollar un sistema de detección avanzado que permita identificar de manera precisa y eficiente voces generadas artificialmente. Esto incluye la detección de voces falsas utilizadas en actividades fraudulentas. El proyecto busca proteger tanto a individuos como a organizaciones de ser víctimas de estafas que comprometen información sensible, fomentando un entorno más seguro para las comunicaciones digitales.

1.3. Members and Role Assignments

Ethan Astorga: Diseño del modelo e integración de redes CNN y RNN+CNN, implementación del modelo en problemas reales y análisis del impacto.

Alejandro Pimiento: Documentación del proyecto, diseño del flujo de trabajo, integracion de API para pag web del proyecto .

1.4. Schedule and Milestones

- **Paso 1: Investigación inicial y configuración**
- Identificar y recopilar bases de datos de audio relevantes para voces humanas y voces generadas artificialmente (Ethan).
- Evaluar las metodologías existentes para la detección de voces falsas.(Ethan y Alejandro)
- Preprocesamiento de datos: limpieza y normalización de audios para garantizar la calidad de las entradas (Alejandro).

- **Paso 2: Desarrollo del modelo**
- Selección de arquitecturas: CNN y combinación RNN+CNN como modelos principales (Ethan).
- Entrenamiento y validación del modelo utilizando bases de datos representativas. (Ethan y alejandro)
- Iteraciones de prueba para mejorar la precisión y minimizar los falsos positivos y negativos.(Alejandro)

- **Paso 3: Evaluación y pruebas finales**
- Implementación del modelo en escenarios simulados de estafas y suplantaciones (Alejandro).
- Comparación con tecnologías existentes, como Vall-E 2, para medir la efectividad del sistema (Ethan).

2. Project Execution

2.1. Data Acquisition

La adquisición de datos se enfocó en dos tipos principales de fuentes:

- **Voces reales**: Se recopilaron grabaciones de voces humanas provenientes de bases de datos públicas como VoxCeleb y LibriSpeech, garantizando diversidad en género, acentos y tonalidades.
- **Voces generadas artificialmente**: Se utilizaron ejemplos de herramientas avanzadas como Vall-E 2, Google Text-to-Speech y otros sistemas de síntesis de voz de código abierto. Estas grabaciones incluyen voces generadas en diferentes idiomas y niveles de realismo.

Además, los datos fueron organizados en formatos estándar (MP3, WAV, FLAC) para facilitar su preprocesamiento y asegurarse de que fueran representativos del entorno real.

2.2. Training Methodology

El entrenamiento del modelo siguió estos pasos:

1. **Preprocesamiento**:
    - Conversión de audios en espectrogramas utilizando herramientas como librosa, reduciendo el ruido y normalizando la amplitud.
    - Segmentación de los datos para trabajar con clips de duración uniforme, mejorando la coherencia del entrenamiento.
2. **Modelos utilizados**:
    - **Redes neuronales convolucionales (CNN)**: Para extraer patrones en los espectrogramas que distinguen voces humanas de generadas.
    - **Combinación RNN+CNN**: Para capturar patrones temporales y secuenciales en los audios, aumentando la precisión en escenarios complejos.
3. **Evaluación iterativa**:
    - Uso de técnicas como validación cruzada y métricas como precisión, sensibilidad y especificidad para ajustar los hiperparámetros.
4. **Aumento de datos**:
    - Aplicación de transformaciones como cambios de tono y velocidad para enriquecer el conjunto de entrenamiento y evitar sobreajuste.

2.3. Workflow

El flujo de trabajo incluye las siguientes etapas:

1. **Entrada de datos**: Carga de audios en formatos MP3, WAV y FLAC desde las carpetas predefinidas en Google Drive o desde su computador.
2. **Preprocesamiento**:
    - Limpieza y transformación en espectrogramas.

    - Almacenamiento en Google Drive o computador para facilitar su reutilización.
3. **Entrenamiento**:
    - Entrenamiento del modelo CNN y RNN+CNN con los espectrogramas generados.
4. **Validación y pruebas**: Evaluación en escenarios simulados, incluyendo detección de voces falsas en grabaciones reales.
5. **Salida del sistema**: Predicción final con etiquetas **"Voz real"** o **"Voz generada"** y un índice de confianza.

2.4. System Design

El diseño del sistema sigue una arquitectura modular:

1. **Módulo de adquisición de datos**: Responsable de cargar y estructurar las bases de datos de voces reales y generadas.
2. **Módulo de preprocesamiento**: Convierte los audios en espectrogramas y realiza el almacenamiento organizado.
3. **Módulo de entrenamiento**:
    - Incluye el desarrollo y ajuste de modelos CNN y RNN+CNN.
    - Asegura la integración fluida entre las redes y sus respectivas salidas.
4. **Módulo de evaluación**: Pruebas exhaustivas en escenarios reales y simulados, con métricas detalladas para evaluar la precisión y el rendimiento.
5. **Interfaz de usuario (opcional)**: Diseño de un prototipo para que las personas o empresas carguen audios y obtengan resultados rápidamente.

3. Results

3.1. Data Preprocessing

Durante la etapa de preprocesamiento, se lograron los siguientes resultados:

- **Limpieza y conversión**: Se procesaron más de 1,000 audios en formatos MP3 y FLAC, convirtiéndolos en espectrogramas utilizando librosa. Esto permitió reducir la complejidad computacional al trabajar con imágenes en lugar de datos de audio en bruto.
- **Segmentación uniforme**: Los clips de audio fueron ajustados a una duración estándar de 10 a 50 segundos, facilitando y agregandole variedad a la consistencia en el entrenamiento del modelo.
- **Almacenamiento optimizado**: Los espectrogramas generados se almacenaron en Google Drive, asegurando acceso rápido y reducción del uso local de memoria.

3.2. Exploratory Data Analysis (EDA)

Se realizó un análisis exploratorio para entender las características principales de las voces reales y generadas:

- **Distribución de frecuencias**: Las voces reales mostraron patrones más variados y dinámicos en el dominio de frecuencia en comparación con las voces generadas, que tienden a ser más uniformes.

- **Análisis de espectrogramas**: Identificación de características específicas, como picos en ciertas frecuencias en voces generadas, lo que facilitó la selección de características relevantes para el modelo.
- **Comparativa temporal**: Se encontraron patrones recurrentes en los audios generados por IA, como una falta de naturalidad en la variación temporal.

3.3. Modeling

Se desarrollaron y evaluaron dos modelos principales:

1. **Modelo CNN**:
    - Logró una precisión del 95% en la clasificación de voces reales y generadas.
    - Fue efectivo para identificar características visuales únicas en los espectrogramas.
2. **Modelo RNN+CNN**:
    - Logro la precisión al 90%, gracias a su capacidad de analizar tanto patrones visuales como temporales.
    - Se destacó en la identificación de voces generadas más avanzadas, como las producidas por Vall-E 2.

3.4. User Interface

Se diseñó un prototipo básico de interfaz para pruebas iniciales:

- **Cargador de audios**: Permite al usuario cargar un archivo en formato MP3, WAV o FLAC.
- **Salida visual**: Muestra el espectrograma del audio analizado junto con el resultado de la predicción (**"Voz real"** o **"Voz generada"**) y un índice de confianza.

3.5. Testing and Improvements

Pruebas en escenarios reales: El sistema fue probado con grabaciones de estafas telefónicas simuladas y voces generadas con diferentes herramientas. En estas pruebas, el modelo RNN+CNN logró detectar correctamente el 93% de las voces generadas.
Iteración y optimización: Se realizaron ajustes en los hiperparámetros y técnicas de regularización para reducir falsos positivos en un 15%.
Pruebas con nuevos datos: Se incluyeron voces generadas por herramientas más recientes para garantizar la adaptabilidad del modelo.

4. Projected Impact

4.1. Accomplishments and Benefits

Logros alcanzados:

- **Detección precisa de voces falsas:** El modelo RNN+CNN alcanzó una precisión del 90%, destacándose en la identificación de voces generadas por herramientas avanzadas como Vall-E 2.
- **Protección ante estafas:** DetectVoice demostró ser eficaz en la detección de fraudes basados en la suplantación de identidad vocal, contribuyendo a la protección de individuos y organizaciones contra pérdidas económicas y de información sensible.
- **Arquitectura modular:** La arquitectura del sistema facilita futuras integraciones y actualizaciones, permitiendo su adaptación a nuevas herramientas de síntesis de voz.
- **Accesibilidad:** El prototipo inicial de interfaz ofrece un entorno amigable para que usuarios no técnicos puedan analizar audios de manera sencilla y obtener resultados confiables.

Beneficios previstos:

- **Mayor seguridad:** Implementar DetectVoice en bancos, empresas y servicios de atención al cliente puede prevenir estafas telefónicas, especialmente aquellas dirigidas a personas mayores o sectores vulnerables.
- **Contribución tecnológica:** El proyecto sienta las bases para investigaciones futuras en detección de deepfakes y otras formas de falsificación digital.
- **Concienciación:** Promueve una mayor comprensión pública sobre los riesgos asociados con el uso malintencionado de la inteligencia artificial en la generación de voces.

4.2. Future Improvements

Ampliación de la base de datos:

- Incorporar más voces generadas por nuevas tecnologías y herramientas emergentes para mantener la relevancia del modelo frente a avances futuros.
- Incluir voces en múltiples idiomas y acentos para mejorar la generalización global del sistema.

Optimización del modelo:

- Reducir el tamaño del modelo sin comprometer su precisión, facilitando su implementación en dispositivos con recursos limitados.
- Experimentar con arquitecturas avanzadas como transformers para capturar mejor las dependencias temporales complejas.

Desarrollo de la interfaz:

- Crear una versión web o móvil del sistema para aumentar su accesibilidad.
- Agregar funcionalidades como reportes automáticos y recomendaciones basadas en los resultados obtenidos.

Integración con sistemas de seguridad existentes:

- Diseñar APIs para integrar DetectVoice con software de seguridad y plataformas de análisis de riesgos.
- Colaborar con empresas y gobiernos para implementar el sistema en entornos críticos.

Educación y sensibilización:

- Desarrollar talleres y materiales educativos para capacitar a las personas en la detección de voces falsas, ayudándoles a identificar posibles amenazas.

5. Team Member Review and Comment

<ATTACH A TEAM PICTURE HERE>

NAME	REVIEW and COMMENT
Ethan Astorga	Reconozco que fue un muy buen curso, a pesar de ser solamente 2 personas la pasamos bien dentro del proyecto y estamos más que contentos de poder hacer 2 modelos además de que reducimos el trabajo de 2 años a un mes y medio.
Alejandro Pimiento	La verdad me voy muy feliz de esta experiencia, pude conocer gente simpatica y trabajadora que quieren luchar por su futuro, ojalá poder participar en otra actividad o curso por parte de samsung

6. Instructor Review and Comment

CATEGORY	SCORE	REVIEW and COMMENT
IDEA	__/10	
APPLICATION	__/30	

RESULT	__/30	
PROJECT MANAGEMENT	__/10	
PRESENTATIO N & REPORT	__/20	
TOTAL	__/100	