FAKESOUND:
DEEPFAKE GENERAL AUDIO DETECTION

Emma Müller, Sofia Schreiner, Alexander Umansky

Content

- Paper
  - *Überblick*
  - *Experiment*
  - *Ergebnisse*
- Experiment
  - *Idee und Aufbau*
  - *Planung*

Paper – Überblick
Allgemeines

- Fortschritt in der Audiogenerierung
  - Generierung äußerst realistischer Audios (Deepfake)
  - Verbreitung von Deepfake-Audios hat negative Folgen

- Vorschlag der Autoren:
  - Nutzung des Datensatzes „FakeSound“ zur Erkennung von Deepfakes allgemeiner Art

- Beispiele einsehbar unter: https://FakeSoundData.github.io

Paper – Überblick
Allgemeines

- Arten von Deepfake-Audios:
  - gesamte Audio wurde neu generiert
  - einzelne Segmente mit anderem Material zusammengeschnitten
  - bestimmte Segmente durch ein generatives Modell ergänzt

- größtes Problem: dritte Art
  - Grund: enthält sowohl echte als auch generierte Elemente

Paper – Überblick
Allgemeines

- Grundidee:
  - man nimmt echte Audios
  - man verändert gezielt bestimmte Teile
  - so entstehen Deepfakes

- passiert durch eine Pipeline (mehrere Schritte hintereinander)

Paper – Überblick
FakeSound: Deepfake General Audio Detection

- Ground & Mask:
  - man sucht nach bestimmten Stellen, die man verändern möchte
    - Ergebnis: mehrere Schlüsselstellen (tragen Hauptinformation)
  - Grund: verändert man diese Stellen --> großer Effekt auf die Audio
  - eine der Stellen wird ausgewählt und gelöscht (auf Null gesetzt)
    - Effekt: im Original fehlt ein wichtiger Teil

Paper – Überblick
FakeSound: Deepfake General Audio Detection

- Regenerate & Replace:
  - KI-Modell generiert Audio neu
  - Grundlage: ausgeschnittene Stelle und lautliche Umgebung der eigentlichen Audio
    - nennt man **Inpainting**
  - weiteres Modell macht Sound besser
  - Einsetzen in die eigentliche Audio
  - Ergebnis: Audio, die größtenteils echt ist und teilweise verändert ist.

Paper – Überblick
FakeSound: Deepfake General Audio Detection

- Datensatz & Einstellungen:
  - Wie wird der Datensatz gebaut?
    - verwendeter Datensatz: AudioCaps
    - jeder Audioclip hat eine Textbeschreibung
    - benutzt zum Finden der Segmente und Generieren neuer Inhalte
  - Länge der Manipulation: 1 – 4 Sekunden
    - Grund:
      - zu kurz --> kaum merkbarer Effekt
      - zu lang --> schlechte Qualität

Paper – Überblick
FakeSound: Deepfake General Audio Detection

- Qualitätskontrolle: wenn keine passenden Segmente gefunden werden…
  - … wird Audio nicht verwendet
- 3 Schwierigkeitsstufen:
  - **Test-Easy:** einfachste Stufe
    - misst: Hat das Modell das Gelernte Verstanden?
  - **Test-Hard:** schwieriger
    - beliebige Segmentlängen, unterschiedliche Qualität, komplexe Szenarien
  - **Test-Zeroshot:** noch schwieriger
    - andere KI wird verwendet; Modell hat das noch nie gesehen

Paper – Überblick
Evaluation Metric

- das Modell muss 2 Dinge gleichzeitig können:
  - **erkennen (Classification)**
    - Ist das echt oder ist das fake?
  - **lokalisieren (Localization)**
    - Wo genau im Audio ist der Fake?

- viele Berechnungen (True / False Positives / Negatives, …)

Deepfake Detection Model tut genau das!

Paper – Experiment

- Deepfake-Erkennungsmodell wurde trainiert und getestet
- Daten & Verarbeitung:
  - Audio: 10 Sekunden
  - aufgeteilt in 500 kleine Abschnitte (je 20ms)
- Modell entscheidet für jeden Abschnitt, ob fake oder echt
- Vergleich mit Modell auf dem aktuellen Stand der Sprach-Deepfake-Erkennung

Paper – Experiment

- Evaluation durch den Menschen:
  - 10 VPn
  - jeweils 10 Audioclips aus 3 Datensätzen (nach den Schwierigkeitsstufen)
  - Aufgabe: Ist es Deepfake oder nicht? Falls ja – wo?
  - Bewertung mit gleichen Metriken
  - Ergebnisse gemittelt

Paper – Ergebnisse

- Vergleich: Menschen vs. Modelle
  - Ergebnisse verdeutlichen Schwierigkeit für Menschen
  - Durchschnittliche Genauigkeit im 3. Datensatz bei 0,51
    - das heißt: könnte genauso gut geraten sein

- Detektor übertrifft Vergleichsmodell in allen Aufgaben

Paper – Ergebnisse

- **Test-Easy-Datensatz:** nahezu perfekt
- **Test-Hard-Datensatz:** schwieriger aber besser als Vergleichsmodell und Menschen
- **Test-Zeroshot-Datensatz:** Leistungsabfall

- Modell performt besser wenn Daten ähnlich wie Trainingsdaten
- Domänenanpassung zukünftig wichtige Forschungsrichtung

Experiment – Idee und Aufbau

- **Idee:** Vergleich zwischen Mensch und frei zugänglichen Detektoren

- **Aufbau:**
  - *2 KI-generierte Audios (1 männliche und 1 weibliche Stimme)*
  - *4 aufgenommene Audios (2 männliche und 2 weibliche Stimmen)*
  - *20 VPn (10 männliche, 10 weibliche)*

Experiment – Idee und Aufbau

- **Aufbau:** Teil 1
  - *4 Versuchsgruppen (G1 – G4):*
    - G1 & G2 je 5 männliche VPn
    - G3 & G4 je 5 weibliche VPn
  - *den Gruppen werden Audios vorgespielt*
    - G1 & G3: 1x Deepfake und 2x echte Audio → nur männliche Stimmen
    - G2 & G4: 1x Deepfake und 2x echte Audio → nur weibliche Stimmen
    - Aufgabe: Identifiziere das Deepfake!
  - *Ziel: Überblick, wie gut Menschen (männlich vs. weiblich) bei Identifizierung verschiedener Stimmen (fake vs. echt) abschneiden*

Experiment – Idee und Aufbau

- **Aufbau:** Teil 2
  - *alle Audios werden 2 Deepfake-Detektoren gegeben*
    - sollen kostenlos zur Verfügung stehen
  - *Ziel: Überblick darüber, wie gut frei erhältliche Detektoren abschneiden*
    - Kann man sich auf solche (kostenlosen) Tools verlassen?...

- **Aufbau:** Teil 3
  - *Zusammenführung der Ergebnisse*

Experiment – Planung

- bis Freitag, 08.05.:
  - *alle Aufnahmen*
  - *Einigung auf die Detektoren*

- bis Freitag, 12.06.:
  - *Rekrutierung der VPn*
  - *Durchführung des Experiments*

Experiment – Planung

- bis Freitag, 26.06.:
  - *Evaluierung der Ergebnisse*
- 2 Wochen Puffer-Zone
- 14.07./17.07 (?):
  - *Minikonferenz*