Relatório de Implementação Deepfake Detection

Evellyn Nicole Machado Rosa[1], **Guilherme Henrique dos Reis**[2], **Gustavo dos Reis Oliveira**[3], **Isadora Stéfany Rezende Remigio Mesquita**[4]

[1]Instituto de Informática – Universidade Federal de Goiás (UFG)
Caixa Postal 131 – 74001-970 – Goiânia – GO – Brazil

{nicole,guilherme_reis,gustavo.reis2,isadora.mesquita}@discente.ufg.br

Abstract. *For the study of classifying real and fake audio samples from a dataset provided in a Kaggle challenge within the Audio and Speech Processing discipline, machine learning algorithms were employed, incorporating both classical and self-supervised techniques discussed in the classroom.*

Resumo. *Para o estudo de classificação de áudios reais e fakes de um dataset fornecido em um desafio do Kaggle na disciplina de Processamento de Áudio e Voz, utilizou-se algoritmos de machine learning com técnicas clássicas e autosupervisionadas, as quais foram discutidas em sala de aula.*

1. Introdução

O presente trabalho tem como objetivo apresentar os resultados alcançados na competição de Deepfake Detection para a disciplina de Processamento de Áudio e Voz, ministrada pelo professor Arlindo Galvão, semestre 2024/1.

Para resolver o problema proposto de classificação de áudios reais e falsos, serão empregados métodos ensinados em sala de aula, a fim de obter uma experiência prática e aperfeiçoar os conhecimentos adquiridos.

O dataset de treino fornecido para competição contém 7277 arquivos de áudio catalogados como reais e falsos, em formato mp3, com um total de cerca de 390 minutos de dados. Em relação aos dados de teste, foram fornecidos áudios não catalogados para que o modelo seja capaz de avaliar e classificar cada áudio em real ou falso, retornando a probabilidade para cada um deles.

Este relatório está organizado em quatro seções, que abordam sobre as técnicas e algoritmos utilizados, discute e descreve os resultados obtidos e, finalmente, apresenta as conclusões a respeito dos desafios enfrentados, vantagens e limitações da melhor solução.

2. Metodologia

Para iniciar o projeto, foi necessário realizar um pré-processamento, onde todos os áudios foram definidos com o tamanho de 3 (três) segundos e convertidos para mono, se necessário, calculando a média dos canais. Nesse sentido, para áudios mais curtos foi aplicada a técnica de padding com zeros, já para áudios mais longos foi aplicada a técnica de truncamento. Ao analisar as taxas de amostragem, foi percebido diferentes taxas de amostragem dos áudios (24000, 48000, 8000, 16000 e 44100 hz). A taxa de amostragem definida foi de 16000 Hz. Por fim, foi extraído o Mel espectrograma dos áudios.

2.1. Baseline

Para a competição, foi fornecida a arquitetura de uma Rede Neural Convolucional (CNN) simples como baseline. O número de filtros convolucionais aplicados aos dados de entrada inicia em 16 e dobra até 128. Além disso, foi utilizada a função de ativação ReLu. Por fim, foi aplicado a técnica de Pooling, seguida por uma camada flatten, uma camada densa e uma softmax. O resultado da submissão dessa arquitetura foi 0.86833 para o score público e 0.87809 para o score privado, e é o ponto de partida do nosso projeto.

2.2. Algoritmos Clássicos

Ao realizar os testes com modelos clássicos, seguimos as seguintes abordagens:

- Pré-processamento, como relatado na seção Metodologia.
- Extração de características: As características extraídas foram os coeficientes cepstrais de frequência mel (MFCCs), energia do sinal e frequência fundamental média (pitch).
- Treinamento e avaliação: Utilizamos 5 modelos diferentes para realizar o treinamento: Regressão Logística, SVM, XGBoost, KNN e Random Forest. O modelo que mostrou as melhores métricas tanto nos dados de treinamento quanto na competição do Kaggle foi o SVM, atingindo os seguintes resultados:

Metrica	Valor
acurácia	0.90
precision	0.86 (real); 0.92 (fake)
recall	0.85 (real); 0.93 (fake)
f1-score:	0.85 (real), 0.92 (fake)
Score Kaggle	0.60 (publico); 0.66 (privado)

Table 1. Métricas SVM em dados de validação.

2.3. CNN

A arquitetura da CNN implementada se assemelha a rede fornecida no baseline, composta por camadas convolucionais, camadas de pooling e camadas lineares. Porém, a CNN implementada para o projeto, adota uma abordagem de redução gradual do número de canais nas camadas convolucionais, começando com 128 canais e reduzindo para 16. Foram definidas quatro camadas convolucionais, cada uma seguida por uma camada de ativação ReLU e uma camada de pooling.

Outra diferença a ser observada é o tratamento da saída das camadas convolucionais. Na CNN implementada, a saída da última camada convolucional é achatada para um vetor antes de passar por uma sequência de camadas lineares, cada uma seguida por uma ativação ReLU.

2.4. ResNet-18

Uma abordagem para problemas de aúdio onde a extração de features manuais pode ser árdua e difícil, é utlizar embeddings, que são representações do aúdio após passar por uma rede neural. A ideia central é a de que modelos que já aprenderam outras tasks também aprenderam a extrair características importantes da entrada ao longo de suas camadas.

Para validar essa ideia usamos a Resnet para extração de características do áudio, os áudios são convertidos em representações bidimensionais para serem adequados para entrada na ResNet, optamos por usar os Mel espectrogramas.

Metrica	Valor
acurácia	0.85
precision	0.87
recall	0.89
f1-score:	0.88
Score Kaggle	0.70 (público); 0.64 (privado)

Table 2. Resultados da CNN em dados de validação.

Os Mel espectrogramas são representações gráficas do espectro de frequência de um sinal de áudio em função do tempo. Essa representação é altamente informativa e preserva as características importantes do sinal.

O espectograma é então passado pela ResNet até a penúltima camada, para obtenção dos embeddings, estes são o vetor de representações após passar pelas camadas da rede, passamos até a penúltima descartando a camada linear final de classificação da ResNet. Então esses embeddings são passados como features para um MLP classificador com 5 camadas lineares e ativação relu que irá predizer a probabilidade do áudio ser fake ou não.

O treinamento ocorreu por 10 épocas, e um batchsize igual a 100. Os resultados em dados de teste podem ser observados na Tabela 3.

Metrica	Valor
acurácia	0.87
precision	0.87
recall	0.93
f1-score	0.90
Score Kaggle	0.76374 (publico); 0.71698 (privado)

Table 3. Resultados do modelo ResNet-18 em dados de validação.

2.5. Wav2vec

Wav2vev2 é um modelo de aprendizado auto-supervisionado que busca aprender representações latentes do áudio bruto, para depois aprender tarefas específicas através de um fine tuning. No escopo do nosso trabalho usamos alguns modelos dessa arquiteturas pré treinados para realizar o fine tuning na tarefa de classificação de deepfake.

Os modelos ultilizados foram: wav2vec-base, wav2vec2-large-xlsr-53-gender-recognition-librispeechxls. Dito isso, todos esses modelos estao na plataforma HuggingFace

O pipeline de audio dessa arquitetura se diferencia das outras apresentadas até então por usar a forma bruta do áudio para o autoaprendizado de features, e usar uma duração de 4 segundos para os áudios, realizando padding ou cutting se necessário.

Para o pipeline do treinamento foram usadas 20 épocas e uma learning rate de 0.00003.

Foram implementadas técnicas de label smoothing, para suavizar as labels e reduzir a confiança do modelo para cada classe. O label smoothing padrão foi de 0.2. Esse valor foi escolhido empiricamente através do envio de submissões para a competição e vendo qual perfomava melhor, mantendo uma boa distribuição de previsões corretas para as labels.

Ademais, foi usado data augmentation de TimeMask, que mascara alguns trechos

aleatórios do áudio, e um filtro passa baixa para filtrar as frequências do áudio. O data augmentation foi feito com probabilidade de 0.5.

Técnicas Aplicadas

- **ls** - *Label Smoothing* 0.2
- **dg** - *Data Augmentation TimeMask* e *Low Pass Filter*

modelo	F1	KP	kPriv
wav2vec base	0,98	4,41	5,14
wav2vec base + ls 0.2	0,98	0,57	0,5
wav2vec2-large-xlsr-53-gender-recognition-librispeechxls	0,98	0,57	0,6
Fine tuning wav2vec2-large-xlsr-53-gender-recognition-librispeechxls + ls	0,96	0,38	0,45
Fine tuning wav2vec2-large-xlsr-53-gender-recognition-librispeechxls + ls + dg	0,94	0,36	0,28
wav2vec2-large-xlsr-53-gender-recognition-librispeechxls + ls + dg + 10 epochs	0,95	0,26	0,23

Table 4. Resultados fine tuning modelos wav2vec. KP - kaggle public ; Kpriv - kaggle private.

Wav2vec-base Escolhemos esse modelo para iniciar os testes e treinos do *Wav2vec* para validar a ideia do uso de modelos auto-supervisionados

wav2vec2-large-xlsr-53-gender-recognition-librispeechxls Escolhemos esse modelo para testar o *transfer learning* entre modelos. Esse modelo foi originalmente treinado para reconhecer gênero pela voz, o motivo pelo qual escolhemos ele é porque ele já foi treinado em uma classificação binária.

3. Resultados

Das técnicas de aprendizado de maquina aplicadas à detecção de deepfake em áudios, as que trouxeram melhores resultados foram devido ao fine tuning de modelos autosupervisionados pré treinados.

Abordagens clássicas, como o SVM, conseguiram performar melhor que abordagens mais comuns para classifcação, como CNNs, e modelos de classifcação em cima de modelos de extração de features como Resnet18+MLP. Contudo, ela não conseguiu bater os modelos autosupervisionados pré treinados.

Isso se dá porque esses modelos pré treinados tem um alto poder de extração de features de áudios que serão úteis na etapa de classificação. Ao usarmos um modelo já treinado em outra task para fazermos um transfer learning, potencializa-se esse poder de extração de features.

Apesar do bom resultado na competição do melhor modelo, 0.23 de score privado e métricas de avaliação, mais estudos são necessários sobre a viabilidade de um sistema de biometria usando esse modelo em wild-data.

4. Conclusão

O trabalho apresenta abordagens clássicas e avançadas para detecção de áudios deepfake, conseguindo bons resultados em conjuntos in-domain. Abordagens clássicas conseguem se manter competitivas frente a algumas abordagens mais recentes. Contundo, não conseguem desempenho frente a abordagens com maior poder de abstração de features auto-extraídas. Ademais, o trabalho mostra o poder da capacidade de aprendizado de modelos pré-treinados de maneira

autosupervisionada para classificação de áudios, aliando técnicas de label smoothing, data augmentation e transfer learning para melhorar os resultados e tornar o modelo mais robusto à variações do áudio.

5. References

BAEVSKI, Alexei et al. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, v. 33, p. 12449-12460, 2020.

BABU, Arun et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale. arXiv preprint arXiv:2111.09296, 2021.

https://huggingface.co/alefiury/wav2vec2-large-xlsr-53-gender-recognition-librispeech, Acesso em 13/05/2024