Projeto

Date @November 5, 2025

Urban Sound Classification: Deep Learning Approaches for Audio Recognition

Aprendizagem Computacional II

1. **Escolha dos classificadores que vamos usar**

   Pela pesquisa que fiz, o CNN e o RNN (os que a prima da Bea usou) são aqueles mais complexos e que tornam o trabalho mais completo porque têm em conta uma série de coisas fundamentais para a análise dos sons.

    Portanto, **por mim usaríamos esses: CNN e RNN.**

2. **Organização das pastas**

   Gostei mais da organização dos ficheiros da mentora da Bea. Acho que tem em geral um trabalho mais completo e melhor estruturado.

📙 Portanto, **por mim seguíamos a estrutura da mentora da Bea:**

```
├— README.md
├— research
   ├— summary.md
   ├— artigo1.pdf
   ├— artigo2.pdf
   ├— ...
├— augmented_datasets
   ├— augmented_dataset_fold1.csv
   ├— ...
├— datasets
   ├— dataset_fold1.csv
   ├— ...
├— images (se usarmos)
   ├— imagem1.png
   ├— ...
├— Assignment.ipynb
├— CNN.ipynb
├— RNN.ipynb
├— DeepFold.ipynb
├— slides
   ├— Apresentação.pdf
├— ...
```

Notebook “principal” - basicamente de exploração de dados

- Introdução
- Bibliotecas importadas
- Librosa VS Scipy.io (**Prima Bea**) — se acharem útil/necessário

- Exploração dos Dados
  - Data Understanding (**Francisca**)
  - Data Reading (**Francisca**)
    - Extração de features
    - Limpeza dos dados
    - Leitura dos CSV
    - Limpeza dos CSV
- Análise exploratória dos Dados (**Prima Bea**)
  - Pré-processamento dos Dados

Podemos fazer um mix dos dois. Há coisas que gosto mais dum e outras coisas do outro.

Notebook MODELO 1

Notebook MODELO 2

Notebook Deep Fold

Pesquisa

Já fiz uma pesquisa de artigos e um resumo de tudo Summary.md

---

Classificação de Sons Urbanos com Deep Learning

1. Introdução e Motivação

A **classificação de sons urbanos** tem-se tornado uma área de investigação central para aplicações em informática urbana, vigilância acústica, reconhecimento ambiental e recuperação multimédia.

O objetivo fundamental é permitir que sistemas automáticos identifiquem sons como buzinas, sirenes, perfurações ou música de rua — elementos cruciais na perceção sonora das cidades.

Durante muitos anos, este campo enfrentou duas barreiras principais:

1. **Ausência de uma taxonomia comum**, que dificultava a comparação entre estudos.
2. **Escassez de datasets anotados de grande escala**, com sons reais e variados do ambiente urbano.

Essas limitações foram superadas a partir de 2014, quando Salamon, Jacoby e Bello (Salamon et al., 2014) criaram o UrbanSound8K, um dataset amplamente adotado que se tornou o padrão de referência para avaliação de modelos de deep learning aplicados a som urbano.

2. Dataset e Taxonomia Urbana

O UrbanSound8K é um subconjunto anotado do UrbanSound Dataset, composto por cerca de **8.732 amostras** (**8,75 horas** de áudio), agrupadas em **10 classes principais** de sons urbanos:

- air conditioner
- car horn
- children playing
- dog bark
- drilling
- engine idling
- gun shot

- jackhammer
- siren
- street music

Cada gravação tem no máximo **4 segundos** e foi cuidadosamente segmentada e anotada quanto à ocorrência sonora e à saliência acústica (sons de fundo vs. sons em primeiro plano).

As classes foram selecionadas com base na frequência de queixas registadas no sistema **NYC 311**, tornando o dataset representativo dos problemas acústicos mais comuns nas cidades modernas.

A metodologia de Salamon incluiu também um protocolo de validação em **10 folds**, garantindo a reprodutibilidade dos resultados e evitando sobreposição entre treino e teste (slices do mesmo áudio permanecem no mesmo fold).

3. Extração de Características

Escolha das Características - MFCCs

Grande parte da literatura posterior (Piczak, 2015; Massoudi et al., 2021; Tyagi et al., 2023) baseia-se em coeficientes cepstrais na escala Mel (MFCCs), que representam o conteúdo espectral de forma compacta e perceptualmente significativa.

Processo de Extração

- O sinal é dividido em **janelas de 23,2 ms**, com **50% de sobreposição**.
- Para cada janela, são calculados **40 filtros Mel** cobrindo o intervalo de 0–22.050 Hz.
- Os **25 primeiros coeficientes MFCC** são extraídos por frame.
- Em seguida, são calculadas **estatísticas agregadas** (média, variância, assimetria, curtose, derivadas) para descrever o som completo.

Esse processo resulta em vetores de cerca de **225 atributos por amostra**, permitindo representar o som como uma "imagem" tempo-frequência, que serve de entrada para redes CNN.

4. Modelos Baseados em CNN (Convolutional Neural Networks)

CNN 2D - Espectrogramas e MFCCs

Os primeiros modelos de deep learning aplicados a sons urbanos trataram os espectrogramas como imagens. Piczak (2015) foi pioneiro ao utilizar **CNNs 2D** sobre espectrogramas Mel, demonstrando que este tipo de rede supera métodos clássicos como SVM ou Random Forest.

Mais tarde, Massoudi, Verma e Jain (2021) reforçaram a eficácia desta abordagem, usando Mel-spectrogramas derivados de MFCCs como entrada de uma CNN. O modelo alcançou **91% de acurácia**, destacando-se pela sua simplicidade e eficiência. Barua et al. (2023) compararam várias arquiteturas (ANN, CNN, RNN, LSTM e GRU) e **confirmaram que as CNNs apresentam o melhor equilíbrio entre precisão e custo computacional**.

As CNNs 2D são eficazes porque extraem padrões espaciais locais — como texturas e harmónicos — nas representações tempo-frequência do áudio, tornando-se o padrão dominante para classificação acústica.

5. Modelos Baseados em RNN (Recurrent Neural Networks)

Modelação Temporal — LSTM

Apesar do sucesso das CNNs, estas redes tratam o som como uma imagem estática, perdendo parte da informação sobre a evolução temporal. Para contornar essa limitação, Tyagi et al. (2023) propuseram o uso de **redes LSTM (Long Short-Term Memory)** aplicadas a **sequências temporais de MFCCs**.

O modelo conseguiu **86% de acurácia** e **F1-score de 0.87** no UrbanSound8K, revelando que as LSTMs são capazes de capturar padrões rítmicos e dependências de longo prazo entre frames, essenciais para distinguir sons dinâmicos como sirenes, passos ou motores. Os autores concluíram que a combinação entre CNN (para extração espacial) e LSTM (para dependências temporais) — conhecida como **CRNN (Convolutional Recurrent Neural Network)** — constitui a abordagem mais completa.

6. Modelos End-to-End — CNNs 1D e Aprendizagem Direta do Áudio

Enquanto as abordagens anteriores dependem de MFCCs, Abdoli, Cardinal e Koerich (2019) propuseram um modelo **1D-CNN end-to-end** que aprende diretamente do **sinal de áudio bruto (raw waveform)**. Essa abordagem elimina completamente o pré-processamento manual, deixando a rede aprender as representações relevantes a partir do som original.

O modelo atingiu **89% de acurácia** no UrbanSound8K e mostrou filtros internos semelhantes aos padrões auditivos humanos, confirmando a capacidade das CNNs 1D de aprender características perceptualmente significativas.

Em 2024, Zhao, Ye, Shen e Liu (Zhao et al., 2024) aplicaram uma arquitetura 1D-CNN semelhante à deteção de manipulação de áudio com ENF signals, reforçando o potencial destas redes na análise direta de sinais acústicos.

As CNNs 1D oferecem vantagens como **simplicidade, baixo número de parâmetros e eliminação da etapa de feature engineering**, embora possam ser mais sensíveis ao ruído e necessitem de normalização cuidadosa.

7. Discussão e Trabalho Futuro

A literatura revela uma clara evolução metodológica:

- `2014–2015`: Criação do dataset e introdução das primeiras CNNs aplicadas a espectrogramas (Salamon, Piczak).
- `2019–2021`: Consolidação de CNNs com MFCCs e surgimento de abordagens end-to-end (Abdoli, Massoudi).
- `2023–2024`: Ênfase na modelação temporal (Tyagi) e no processamento direto do áudio (Zhao).

As tendências atuais indicam que a **combinação de CNNs e RNNs (CRNN)** representa o estado da arte na classificação de som urbano, pois combina:

- **CNNs** → Captação de padrões espaciais e harmónicos.
- **RNNs/LSTMs** → Modelação temporal e rítmica.
- **Abordagens end-to-end** → Eliminação da dependência de features manuais.

Para trabalhos futuros, sugerem-se **arquiteturas híbridas**, técnicas de **aumento de dados (data augmentation)** e o uso de **atenção temporal-frequencial** para

melhorar a robustez e interpretabilidade dos modelos.

8. Conclusão

Com base nas obras analisadas, observa-se que os maiores avanços em classificação de som urbano derivam da transição entre **modelos tradicionais baseados em características fixas e modelos de aprendizagem profunda** capazes de extrair representações diretamente do sinal. A adoção de **CNNs e LSTMs** revelou-se decisiva, sendo que o uso combinado destas arquiteturas permite explorar simultaneamente as dimensões **frequencial e temporal** do som.

Assim, a escolha de desenvolver neste projeto dois modelos — um **CNN** e um **RNN** (LSTM) aplicados ao **UrbanSound8K** — está totalmente alinhada com o estado atual e representa uma abordagem sólida, fundamentada e relevante no contexto da inteligência artificial aplicada ao som.

(este é o conteúdo que coloquei - para verem no VSCode é só entrar no ficheiro Summary.md e fazer Ctr+Shift+V)