Reconhecimento Facial com Redes Neurais Convolucionais

Higor Gabriel de Freitas
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo
São Paulo, Brasil
higor.freitas@usp.br

***Resumo*—Este trabalho relata a implementação e avaliação de uma Rede Neural Convolucional (CNN) para reconhecimento facial utilizando o conjunto de dados CelebA. Diferentemente da abordagem tradicional do trabalho anterior baseada em descritores manuais como HOG e LBP, a CNN já realiza automaticamente a extração de características diretamente das imagens. São analisados o impacto do pré-processamento das imagens, o desempenho do modelo em termos de acurácia e outras métricas, bem como uma comparação qualitativa com utilização de uma CGAN.**

I. Especificações da Máquina

O treino foi realizado em um computador significativamente melhor que no trabalho anterior, o dispositivo possui processador AMD Ryzen 5 3400G with Radeon Vega Graphics, operando a 3.70 GHz, com arquitetura x64 e suporte a sistema operacional de 64 bits, e conta com 16 GB de memória RAM instalada.

Para acelerar o treinamento, foi utilizada uma GPU dedicada AMD Radeon RX 7600, com 8 GB de memória de vídeo, que contribuiu significativamente para a redução do tempo de processamento durante o treinamento da rede neural.

Vale ressaltar que essa mudança de especificações afetou consideravelmente a análise, já que no teste anterior o número de épocas, tamanho da rede e amostragem dos descritores era limitada.

II. Base de Dados e Pré-processamento

A. Conjunto de Dados CelebA

O conjunto de dados CelebA é composto por mais de 200 mil imagens faciais de celebridades, cada uma associada a uma identidade. Apesar de sua grande dimensão, o conjunto apresenta desbalanceamento entre classes, com algumas identidades possuindo poucas imagens, o que pode prejudicar o treinamento de modelos supervisionados. Seguindo a mesma orientação adotada nos modelos tradicionais, foram selecionadas apenas as identidades pertencentes aos 20% com maior número de imagens, reduzindo o impacto do desbalanceamento extremo e tornando o treinamento viável em um computador doméstico. Essa seleção foi feita em um aqruivo separado `seletor.py` que exportava arquivos npzs, lido pelo `identificação.py` ou `cgan.py`.

B. Pré-processamento das Imagens

Diferentemente da abordagem baseada em descritores, na CNN as imagens são utilizadas diretamente como entrada do modelo. O pré-processamento consistiu em:

- Leitura das imagens coloridas no formato RGB;
- Redimensionamento para o tamanho fixo de 128×128 pixels;
- Normalização dos valores dos pixels para o intervalo $[0, 1]$.

Esse pré-processamento garante uniformidade nas entradas da rede e estabilidade numérica durante o treinamento. Sem normalização a convergência fica mais lenta e com maior instabilidade, afetando o desempenho do modelo.

III. Arquitetura da Rede Neural Convolucional

A implementação da arquitetura foi realizada utilizando a biblioteca TensorFlow, por meio do Keras, de alto nível, isso facilitou a construção do modelo e possibilitou a execução otimizada em GPU (Que foi um enorme diferencial em custo). Foi adotado o modelo Sequential, no qual as camadas são empilhadas de forma linear, seguindo exatamente a ordem e a configuração recomendadas nos slides do trabalho.

A CNN é descrita pelas seguintes camadas:

- Duas camadas convolucionais (Conv2D) com filtros 3×3, responsáveis pela extração automática de características locais;
- Camadas de *MaxPooling* para redução da dimensionalidade espacial e aumento de invariância a pequenas translações;
- Uma camada *Flatten* para converter os mapas de ativação em um vetor;
- Uma camada totalmente conectada com função de ativação ReLU;
- Uma camada de saída com função *softmax*, produzindo probabilidades sobre todas as identidades consideradas.

Essa arquitetura permitiu que a rede aprenda, até estruturas mais complexas relacionadas à identidade facial.

IV. Treinamento e Avaliação

A. Configuração Experimental

Os dados foram divididos em conjuntos de treino (70%), validação (15%) e teste (15%) conforme recomendado, utilizando *shuffle* aleatório para evitar vieses de ordenação.

O treinamento foi realizado em 20 épocas, utilizando o otimizador Adam e a função de perda de entropia cruzada também.

Durante o treinamento, foram salvas as curvas de perda e acurácia tanto no conjunto de treino quanto no de validação, para possibilitar a análise de sobreajuste.

B. Métricas

Após o treinamento do modelo, a avaliação foi realizada no conjunto de teste utilizando métricas padrão de classificação multiclasse. O modelo alcançou uma acurácia de 95,03%, indicando um desempenho elevadíssimo na identificação correta das identidades faciais.

Adicionalmente, foram obtidos valores elevados de precisão (95,37%), recall (94,86%) e F1-score (94,36%), demonstrando um bom equilíbrio entre a capacidade do modelo em minimizar falsos positivos e reconhecer corretamente quase as todas classes existentes. Isso indica que a Rede Neural Convolucional é bem eficiente em aprender representações discriminantes em rostos, mesmo em problemas multiclasse com grande número de identidades e distribuição desbalanceada de amostras. As métricas de desempenho foram calculadas com o auxílio da biblioteca *scikit-learn*, que fornece implementações rápidas para avaliação de classificadores.

C. Análise Gráfica

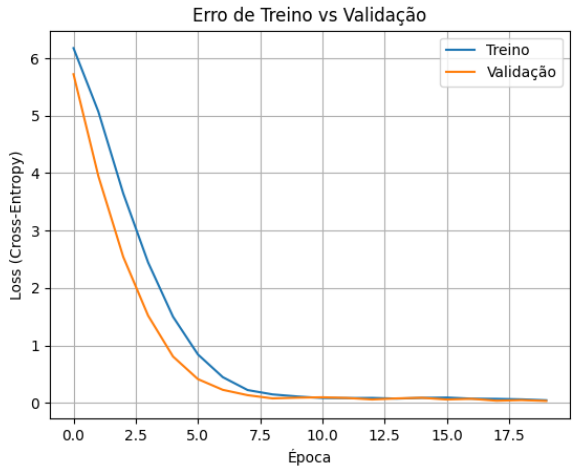

Fig. 1. Curva de erro médio ao longo das épocas

O gráfico apresentado mostra a evolução do erro (função cross-entropy) ao longo das épocas para os conjuntos de treinamento e validação. Observa-se uma convergência rápida e estável, com redução consistente do erro em ambas as curvas, indicando que o modelo aprendeu os padrões dos dados sem apresentar sinais relevantes de overfitting, já que as curvas permanecem próximas entre si. Essa convergência foi mais bem comportada e satisfatória quando comparada às abordagens baseadas em descritores manuais, como HOG e LBP, nas quais o erro apresentou maior instabilidade e menor capacidade de generalização. Esse comportamento mais estável da CNN refletiu-se diretamente em uma acurácia superior, o que indica uma vantagem do aprendizado automático em relação aos métodos baseados em extração manual.

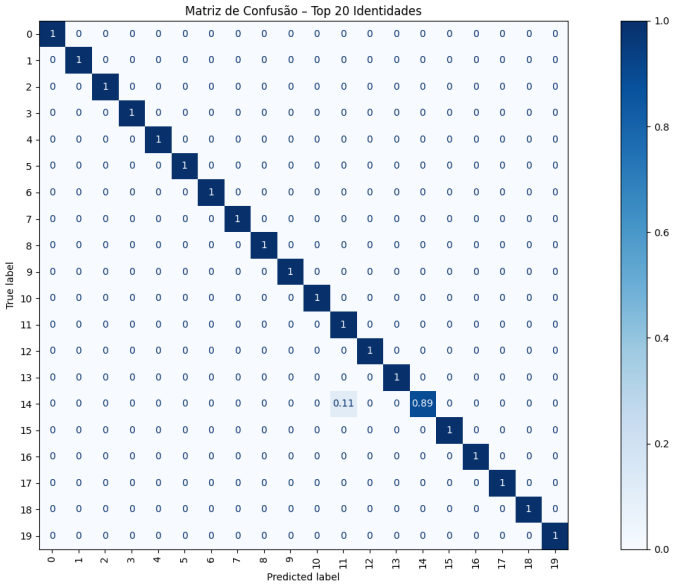

Fig. 2. Curva de erro médio ao longo das épocas

A matriz de confusão foi construída considerando apenas as 20 identidades com maior número de amostras no conjunto de dados, uma vez que a visualização completa das aproximadamente 2000 classes seria impossível. Mesmo com essa limitação intencional, a matriz mostra um desempenho extremamente consistente do modelo, com a maioria absoluta das predições concentrada na diagonal principal, indicando classificações corretas, observa-se apenas um erro pontual de confusão entre duas identidades específicas.

V. Comparação com uso da cGAN

Conforme mencionado anteriormente, foi implementada uma *Conditional Generative Adversarial Network* (cGAN) com o objetivo de realizar *data augmentation*, permitindo que a rede neural classificadora tivesse acesso a um maior número de amostras durante o processo de aprendizado. As cGANs operam por meio do treinamento simultâneo de dois modelos distintos: um *gerador*, responsável por produzir imagens sintéticas condicionadas a rótulos, e um *discriminador*, treinado para distinguir entre imagens reais e artificiais. Ao longo do treinamento, o gerador busca enganar o discriminador, de modo que, idealmente, seja capaz de gerar imagens visualmente semelhantes às reais. Ambos os modelos seguiram o padrão descritao nos slides.

O gerador foi construído a partir da concatenação de um vetor de ruído com uma representação embutida (*embedding*) do rótulo correspondente, seguido por camadas densas, normalização em lote e convoluções transpostas, conforme apresentado a seguir:

- Uma camada de *Embedding* para os rótulos das classes, projetando as identidades em um espaço vetorial denso;

- Uma camada *Flatten* aplicada ao vetor de rótulos embutidos, permitindo sua concatenação com o vetor de ruído aleatório;
- Uma operação de concatenação entre o vetor de ruído latente e a representação do rótulo, condicionando explicitamente o processo de geração;
- Uma camada totalmente conectada responsável por expandir o vetor concatenado para um volume tridimensional inicial;
- Camadas de *Batch Normalization* e ativação ReLU, utilizadas para estabilizar o treinamento e acelerar a convergência;
- Uma camada de *Reshape*, transformando o vetor em um tensor com estrutura espacial;
- Três camadas de convolução transposta (*Conv2DTranspose*), responsáveis pela ampliação progressiva da resolução espacial da imagem;
- Camadas de normalização em lote e funções de ativação ReLU entre as convoluções transpostas;
- Uma camada de saída convolucional transposta com função de ativação *tanh*, produzindo imagens RGB normalizadas no intervalo $[-1, 1]$.

De forma análoga, o discriminador foi composto por camadas convolucionais condicionadas ao rótulo da imagem, utilizando um *embedding* expandido para o formato espacial da entrada, seguido de ativações *Leaky ReLU* e uma camada densa final com ativação sigmoide:

- Uma camada de *Embedding* para os rótulos das classes, mapeando cada identidade para uma representação espacial compatível com a dimensão da imagem;
- Uma camada de *Reshape*, ajustando o vetor de rótulos para um mapa bidimensional adicional;
- Uma operação de concatenação entre a imagem de entrada e o mapa de rótulos, permitindo a avaliação condicional;
- Três camadas convolucionais (*Conv2D*) com *stride* maior que um, responsáveis pela extração hierárquica de características visuais;
- Funções de ativação *Leaky ReLU*, utilizadas para evitar o problema de gradientes nulos durante o treinamento;
- Uma camada *Flatten*, convertendo os mapas de ativação em um vetor unidimensional;
- Uma camada totalmente conectada de saída com função de ativação sigmoide, responsável por estimar a probabilidade da imagem ser real ou gerada.

Os principais parâmetros utilizados no treinamento foram: dimensão do espaço latente igual a 100, 50 épocas, tamanho de lote 64 e imagens com resolução $64 \times 64 \times 3$. O treinamento completo teve duração aproximada de 11 horas. Ao final, observou-se que as imagens geradas ainda era basicamente ruído aleatório, indicando que o número de épocas não foi suficiente para a convergência do modelo. Considerando a elevada demanda computacional e o fato de que a rede neural classificadora já apresentava acurácia próxima de 95%, concluiu-se que não seria vantajoso estender significativamente o treinamento da cGAN. Ainda assim, foi realizada uma tentativa de utilização dessas imagens sintéticas no treinamento, não com o objetivo de ver alguma melhora, mas analisar qual seria o comportamento do modelo frente a dados ruidosos e verdadeiros.

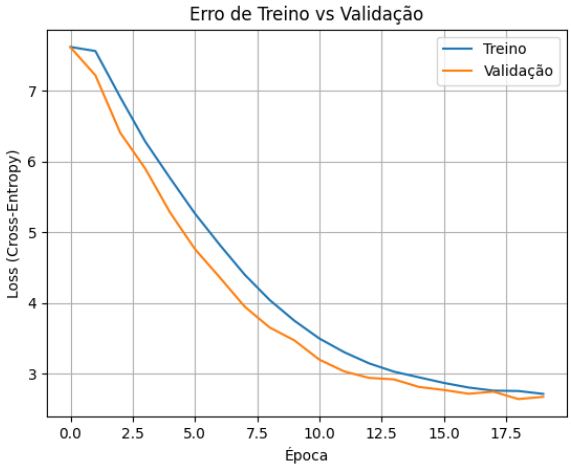

Fig. 3. Curva de erro médio ao longo das épocas com CGAN

Resultados obtidos nesses dados foram:

- **Acurácia (0.6278)**
- **Precisão (0.9386)**
- **Recall (0.6325)**
- **F1-score (0.7371)**

No geral, mesmo com a presença de ruídos nos dados gerados sinteticamente, a rede neural apresentou um desempenho não tão ruim, conseguindo identificar corretamente uma parcela significativa do conjunto de dados e mantendo elevada confiabilidade nas predições realizadas.

VI. Comparação com Abordagens Tradicionais

Comparando-se a CNN com os modelos tradicionais baseados em HOG e LBP do outro trabalho, existe uma diferença conceitual fundamental, enquanto os descritores manuais impõem uma representação fixa das imagens, a CNN aprende automaticamente as características mais relevantes para a tarefa.

Nos testes anteriores, os modelos tradicionais apresentaram desempenho limitado, especialmente na tarefa de identificação multiclasse, com acurácias relativamente baixas (menores que 2% em 3 dos 4 testes). A CNN, apesar de mais custosa computacionalmente, mostrou claramente uma maior capacidade de ajuste aos dados, atingindo acurácias muito altas e exibindo padrões mais claros de separação entre classes, mesmo quando inserido ruído. Além disso, as CNNs apresentam são bem mais generalista quando comparadas a métodos manuais, que são projetados especificamente para dados de imagem e dependem fortemente de heurísticas definidas a priori. Enquanto esses descritores extraem padrões locais fixos, limitados a bordas, texturas ou gradientes, as CNNs aprendem padrões mais

abstratos, que se aplicam não apenas a domínios de imagens, mas também a outras modalidades de dados estruturados em grades, tornando-o mais flexível.

VII. Análise de Desempenho e Limitações

Pelo que foi observado nos experimentos, além do elevado custo computacional percebido durante o treinamento, a CNN, exatamente por possuir uma arquitetura mais complexa, mostrou-se mais sensível à escolha de hiperparâmetros e à quantidade de dados disponíveis. Em conjuntos de dados menores, seu desempenho foi significativamente inferior, com redução perceptível da acurácia, indicando uma maior dependência de bases extensas e bem balanceadas para que o modelo consiga generalizar adequadamente. Contudo, de forma geral, o desempenho da CNN foi excelente. Mesmo com apenas 20 épocas de treinamento, o modelo alcançou uma acurácia de aproximadamente 95%, conforme observado nos resultados, o que indica que a rede foi capaz de aprender representações relevantes e discriminativas de maneira eficiente.

VIII. Conclusões

Neste trabalho foi implementada uma Rede Neural Convolucional para reconhecimento facial utilizando o conjunto de dados CelebA. Observou-se que a CNN elimina a necessidade de descritores manuais, aprendendo automaticamente representações diretamente das imagens. Em comparação com abordagens tradicionais, o modelo convolucional apresentou desempenho superior, ao custo de maior complexidade computacional e calibração fina de parâmetros.

Embora tenha sido discutido o uso de Redes Gerativas Adversariais (GANs) como estratégia de *data augmentation*, essa abordagem não foi tão eficiente. O objetivo era aumentar a quantidade de amostras disponíveis e avaliar o impacto de dados sintéticos no desempenho da CNN. Apesar do ruído puro observado nas imagens geradas, a utilização delas permitiu investigar o comportamento do modelo frente a dados ruins, contribuindo para uma análise mais completa da robustez da rede.

Mas, no geral, os resultados obtidos reforçam a eficácia das Redes Neurais Convolucionais para tarefas de reconhecimento facial, evidenciando a importância de um pré-processamento adequado, de uma arquitetura bem definida e da análise criteriosa das curvas de treinamento. O modelo apresentou elevada capacidade de aprendizado e generalização, consolidando a CNN como uma abordagem robusta e eficiente quando comparada a métodos tradicionais baseados em descritores manuais.

Referências

[1] OpenAI. ChatGPT: Large Language Model. Formatação do texto em LaTeX, correção de código e dúvidas gerais no desenvolvimento Disponível em: https://chat.openai.com. Acesso em: 2026.