

# ChatBot para Verificação de Documentos

## 1 Introdução

Nos últimos anos, o aprendizado profundo emergiu como uma das áreas mais influentes da inteligência artificial, trazendo avanços significativos em visão computacional, processamento de linguagem natural e robótica. Este trabalho combina três componentes essenciais: aprendizado profundo, reconhecimento óptico de caracteres (OCR) com Tesseract e chatbots, para abordar desafios relacionados à automação de processos de verificação documental.

O aprendizado profundo, fundamentado em redes neurais artificiais, é particularmente eficaz na modelagem de padrões complexos em dados estruturados e não estruturados. De acordo com Goodfellow et al. (2016), modelos como redes neurais convolucionais (CNNs) e redes LSTM (Long Short-Term Memory) têm redefinido o estado da arte em reconhecimento de imagens e processamento de sequências textuais. Estas arquiteturas são amplamente utilizadas para tarefas que vão desde a classificação de imagens até a tradução automática de textos.

O Tesseract OCR, desde sua versão 4, integra redes LSTM como parte de seu pipeline de reconhecimento. Esse aprimoramento torna o Tesseract mais robusto em lidar com textos em fontes variadas e em contextos desafiadores, como documentos digitalizados de baixa qualidade (Smith, 2007). Além disso, o sistema permite a customização para melhorar o desempenho em cenários específicos, como a extração de informações de documentos oficiais.

Por outro lado, os chatbots têm se destacado como uma interface intuitiva entre usuários e sistemas computacionais. Segundo Brown et al. (2020), modelos baseados em aprendizado profundo, como transformers, possibilitam interações mais naturais e contextualizadas. Esses sistemas são particularmente úteis em aplicações como validação documental, onde o chatbot pode não apenas processar os dados, mas também interpretar e responder às consultas do usuário em tempo real.

Este trabalho explora a interação dessas tecnologias para desenvolver um sistema eficiente e automatizado de verificação de documentos. A abordagem proposta combina a robustez do OCR, a flexibilidade do aprendizado profundo e a interatividade dos chatbots, proporcionando uma solução integrada para os desafios de validação documental.

## Sumário

- |   |           |
|---|-----------|
| <b>1 Introdução</b>                           | <b>2</b>  |
| <b>2 Arquitetura do Sistema</b>               | <b>2</b>  |
| <b>3 Funcionamento do ChatBot</b>             | <b>3</b>  |
| <b>4 Implementação - ChatBot para RG</b>      | <b>4</b>  |
| <b>5 Implementação - ChatBot para Diploma</b> | <b>6</b>  |
| <b>6 Resultados - Verificação de RG</b>       | <b>9</b>  |
| <b>7 Resultados - Verificação de Diploma</b>  | <b>16</b> |
| <b>8 Desafios e Limitações</b>                | <b>19</b> |
| <b>9 Conclusão</b>                            | <b>19</b> |

## 2 Arquitetura do Sistema

O ChatBot é composto por diferentes módulos que trabalham de forma integrada para realizar a verificação dos documentos. A seguir, são descritos os principais componentes:

## Pré-processamento de Dados

O módulo de pré-processamento é responsável por:

- Carregar imagens de documentos (como RG) a partir de arquivos locais.

- Converter a imagem para escala de cinza, aplicar desfoco gaussiano para suavização e detectar bordas utilizando o algoritmo Canny.

- Localizar contornos na imagem para ajustar a perspectiva e melhorar a leitura dos caracteres.

### Reconhecimento Óptico de Caracteres (OCR)

Para extrair as informações textuais dos documentos processados, o sistema utiliza o Tesseract OCR, que integra redes LSTM (Long Short-Term Memory) como parte de seu pipeline de reconhecimento. Este é um componente fundamental do domínio de aprendizado profundo, permitindo o reconhecimento eficiente de caracteres em textos digitalizados.

### Interface do Usuário

A interface foi desenvolvida utilizando a biblioteca Gradio, integrando técnicas modernas de interação em linguagem natural. Esse módulo complementa o sistema, permitindo ao usuário interagir com o ChatBot de forma simples e intuitiva.

## 3 Funcionamento do ChatBot

O processo de verificação de documentos pelo ChatBot pode ser descrito nas seguintes etapas:

- 1. Entrada de Dados:** O usuário envia uma imagem de documento ou realiza perguntas sobre informações extraídas do documento.

- 2. Pré-processamento:** Técnicas de visão computacional, como suavização de imagens e detecção de bordas, são aplicadas para melhorar a qualidade da entrada.

- 3. Extração e Reconhecimento de Texto (OCR):** O texto é extraído usando o Tesseract OCR. Esse processo utiliza redes LSTM para realizar o reconhecimento de caracteres, uma abordagem baseada em aprendizado profundo.

- 4. Validação de Informações:** Os dados extraídos são analisados usando expressões regulares para identificar elementos como CPF e RG. Esses números são organizados e preparados para validação.

- 5. Resposta ao Usuário:** A interface do chatbot retorna respostas personalizadas com base nas informações validadas e nas perguntas realizadas.

## 4 Implementação - ChatBot para RG

Nesta seção, detalhamos o código para a implementação de um ChatBot para verificação de RG, explicando cada etapa do processo.

### Importação de Bibliotecas

O código inicia com a importação das bibliotecas necessárias:

```
1 import numpy as np
2 import cv2
3 import imutils
4 from google.colab.patches import cv2_imshow
5 from matplotlib import pyplot as plt
```

Essas bibliotecas permitem:

- Manipulação de arrays (NumPy);
- Processamento de imagens (OpenCV);
- Funções auxiliares para manipulação de imagens (imutils);
- Exibição de imagens no ambiente Google Colab (cv2\_imshow);
- Visualização de gráficos (Matplotlib).

### Montagem do Google Drive e Cópia de Arquivos

Os arquivos necessários são carregados do Google Drive:

```
1 from google.colab import drive
2 drive.mount('/content/gdrive')
3 !cp -R /content/gdrive/Mydrive/Cursos\ -\ recursos/0CR\ com\ Python/
  Imagens/Projeto2 imagens/
```

Esses comandos montam o Google Drive no ambiente Colab e copiam as imagens necessárias para o diretório de trabalho.

### Função para Exibição de Imagens

Uma função personalizada foi criada para exibir imagens:

```
1 def mostrar(img):
2     fig = plt.gcf()
3     fig.set_size_inches(20, 10)
4     plt.axis("off")
5     plt.imshow(cv2.cvtColor(img, cv2.COLOR_BGR2RGB))
6     plt.show()
```

Esta função utiliza Matplotlib para exibir imagens convertidas para o formato RGB.

## Pré-processamento da Imagem

A imagem do RG é carregada e processada:

```
1 img = cv2.imread('/content/images/RG.jpg')
2 original = img.copy()
3 mostrar(img)
4 (H, W) = img.shape[1:2]
5 print(H, W)
```

A imagem é convertida para tons de cinza, suavizada e as bordas são detectadas:

```
1 gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
2 blur = cv2.GaussianBlur(gray, (5, 5), 0)
3 edged = cv2.Canny(blur, 60, 160)
4 mostrar(edged)
```

Essas etapas facilitam a detecção de contornos no próximo passo.

## Detectão e Ordenação de Contornos

Os contornos são identificados e ordenados:

```
1 def encontrar_contornos(img):
2     contours = cv2.findContours(img, cv2.RETR_LIST, cv2.CHAIN_APPROX_SIMPLE
3     )
4     contours = imutils.grab_contours(contours)
5     contours = sorted(contours, key=cv2.contourArea, reverse=True)[:6]
6     return contours
```

O contorno mais significativo, com quatro vértices, é selecionado para correção de perspectiva.

## Correção de Perspectiva

A correção de perspectiva "achata" a imagem do RG:

```
1 matriz = cv2.getPerspectiveTransform(pts1, pts2)
2 transform = cv2.warpPerspective(original, matriz, (W, H))
3 mostrar(transform)
```

Isso ajusta a imagem para facilitar o reconhecimento de texto.

## Reconhecimento Óptico de Caracteres (OCR)

O texto é extraído da imagem corrigida:

```
1 import pytesseract
2 texto = pytesseract.image_to_string(transform, lang="por", config=
3 config_tesseract)
4 print(texto)
```

A biblioteca Tesseract é usada para reconhecimento óptico de caracteres (OCR).

## Extração de Dados

Os números relevantes (CPF, RG) são extraídos usando expressões regulares:

```
1 import re
2 digitos = re.findall(r'\d', texto)
3 CPF = digitos[:11]
4 RG = digitos[11:19]
```

Esses números são processados e armazenados para validação.

## Interface com ChatBot

Um chatbot é implementado para interação com o usuário:

```
1 def answer_question(question):
2     if question.lower() == "cpf":
3         return f"Ok, isto é um CPF valido: {texto}"
4     elif question.lower() == "rg":
5         return f"Ok, isto é um RG valido: {texto2}"
6
7 A interface é criada com a biblioteca Gradio, permitindo perguntas e respostas em tempo real:
```

```
1 with gr.Blocks() as app:
2     question_input = gr.Textbox(label="Digite sua pergunta")
3     answer_output = gr.Textbox(label="Resposta do chatbot", interactive=
4     False)
5     question_input.submit(answer_question, inputs=question_input,
6     outputs=answer_output)
7
8 app.launch()
```

## 5 Implementação - ChatBot para Diploma

Nesta seção, detalhamos a implementação de um ChatBot voltado para a validação de diplomas acadêmicos utilizando técnicas de OCR, visão computacional e processamento interativo. O sistema permite a extração de texto de imagens e PDFs de diplomas, bem como a detecção de termos específicos fornecidos pelo usuário.

## Instalação e Configuração

Para garantir a funcionalidade do ChatBot, são realizadas as seguintes instalações e configurações:

```
1 !sudo apt install tesseract-ocr
2 !pip install pytesseract
3 !pip install wordcloud
4 !pip install spacy
5 !pip install pdf2image
6
7 # Baixar o modelo de linguagem para português
8 !mkdir tessdata
9 !wget -O ./tessdata/por.traineddata https://github.com/tesseract-ocr/
       tessdata/blob/master/por.traineddata?raw=true
10 !apt-get install tesseract-ocr-por -y
11 !python -m spacy download pt_core_news_sm
```

Esses comandos garantem a instalação do Tesseract OCR, a biblioteca Pytesseract para integração em Python, o modelo de linguagem em português e o suporte para processamento de PDFs com a biblioteca pdf2image.

## Configuração do OCR

O diretório do modelo de linguagem é configurado para que o Tesseract possa reconhecer texto em português:

```
1 os.environ['TESSDATA_PREFIX'] = '/usr/share/tesseract-ocr/4.00/tessdata'
2
3 config_tesseract = "/usr/share/tesseract-ocr/4.00/tessdata/"
```

## Deteção e Marcação de Texto

O sistema permite a detecção de palavras-chave específicas no diploma e destaca visualmente as ocorrências:

```
1 def OCR-processa_imagem(img, termo_pesquisa, config_tesseract, min_conf,
                           fonte_dir):
2     resultado = pytesseract.image_to_data(img, config=config_tesseract,
                                           lang='por', output_type=Output.DICT)
3     num_ocorrencias = 0
4     for i in range(0, len(resultado['text'])):
5         confianca = int(resultado['conf'][i])
6         if confianca > min_conf:
7             texto = resultado['text'][i]
8             if termo_pesquisa in texto:
9                 x, y, img = caixa_texto(i, resultado, img, (0, 0, 255))
10                img = escreve_texto(texto, x, y, img, fonte_dir,
11                                     (50, 50, 225), 14)
12
13     return img, num_ocorrencias
```

- `image_to_data`: Extrai informações detalhadas como texto, coordenadas de localização e níveis de confiança.

- `min_conf`: Filtra texto com baixa confiança no reconhecimento.
- `caixa_texto` e `escreve_texto`: Desenham caixas e adicionam texto sobre as áreas detectadas.

## ChatBot para Validação

O ChatBot é projetado para interagir com o usuário e processar diplomas em imagens ou PDFs. Ele permite:

- Upload do diploma para validação.
- Pesquisa por palavras-chave fornecidas pelo usuário.
- Retorno de resultados em tempo real, com destaque visual das ocorrências encontradas.

```
1 def chatbot_validador():
2     print("Bem-vindo ao validador de diplomas!")
3
4     # Upload do arquivo
5     uploaded = files.upload()
6
7
8
```

A função `OCR_processa` utiliza o Tesseract para extrair texto da imagem, enquanto `mostrar` exibe o resultado processado.

Imagen Original do Documento



Figura 1: Imagem original do RG.

A Figura 1 apresenta a imagem original do RG, utilizada como entrada no sistema. Esse documento contém informações relevantes, como número do RG, CPF, data de emissão e assinatura. A qualidade dessa imagem é essencial, pois influencia diretamente na eficiência do processamento e na precisão da extração de dados.

## 6 Resultados - Verificação de RG

Nesta seção, apresentamos os resultados detalhados obtidos ao longo do processo de verificação de um documento de identidade (RG) utilizando o sistema implementado. Cada etapa do processo é ilustrada e explicada em detalhes, demonstrando as funcionalidades e potenciais do sistema, desde a entrada do documento até a validação final por meio do chatbot.

## Conversão para Escala de Cinza

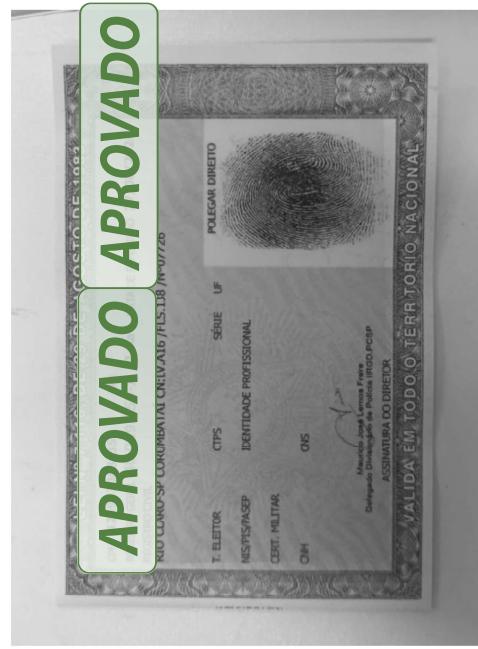


Figura 2: Imagem convertida para tons de cinza.

## Aplicação de Desfoque Gaussiano

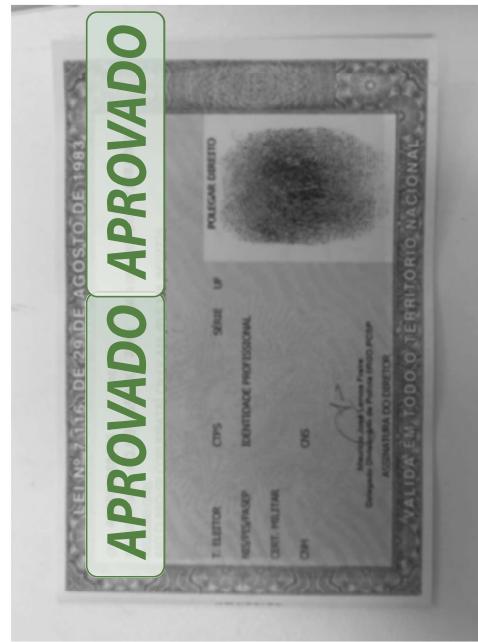


Figura 3: Imagem suavizada com desfoco gaussiano.

A Figura 2 mostra a imagem original após a conversão para escala de cinza. Essa etapa elimina informações de cor que não são necessárias para o reconhecimento óptico de caracteres (OCR), simplificando a complexidade visual. A maior distinção entre os elementos textuais e o fundo melhora a detecção de padrões textuais.

A Figura 3 apresenta o resultado da aplicação do desfoco gaussiano. Essa técnica reduz ruídos e detalhes irrelevantes, suavizando a imagem e destacando as bordas dos caracteres. Isso é crucial para minimizar interferências na próxima etapa, que envolve a detecção de bordas.

Deteção de Bordas com Canny

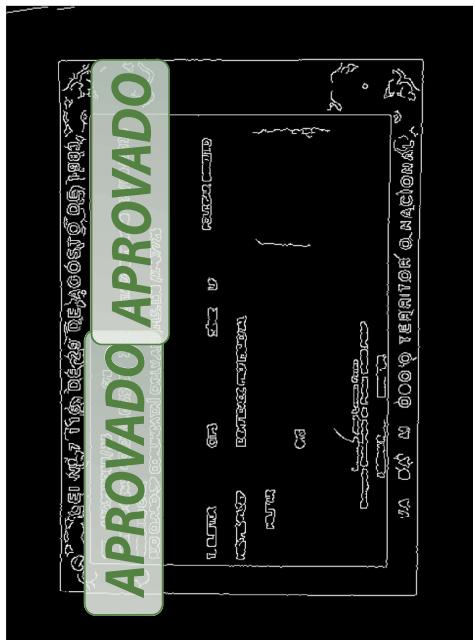


Figura 4: Detecção de bordas utilizando o algoritmo Canny.

A Figura 4 exibe o resultado da detecção de bordas, realizada utilizando o algoritmo de Canny. Essa etapa identifica os contornos mais significativos da imagem, permitindo a delimitação precisa das áreas de interesse, como campos de texto ou outras seções relevantes do documento.

Identificação e Correção de Perspectiva

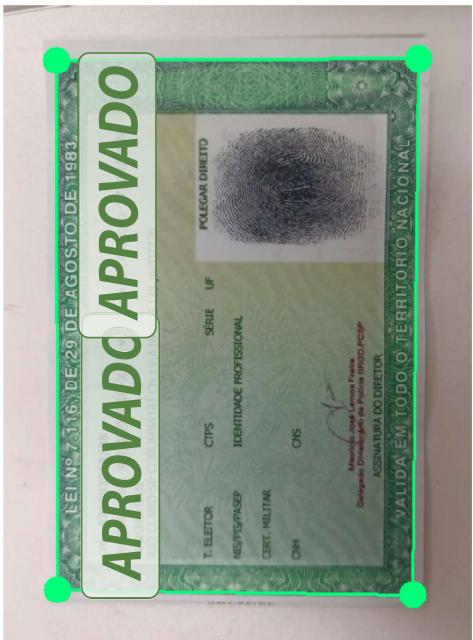


Figura 5: Imagem corrigida para ajuste de perspectiva.

A Figura 5 ilustra a imagem após a correção de perspectiva. Essa transformação é crucial para alinhar o documento, especialmente quando ele foi fotografado em ângulos inclinados. O alinhamento uniforme dos campos de texto facilita a extração de dados pelo OCR, garantindo maior precisão.

#### Extração e Validação com Gradio



Figura 6: Interface Gradio para validação de RG.

A última etapa envolve a extração do texto com o Tesseract OCR e sua validação por meio da interface Gradio. Na Figura 6, observa-se a interface interativa onde o

usuário insere consultas (como "RG"), e o sistema valida os dados extraídos, retornando informações relevantes sobre o documento.

## Possíveis Aplicações na UNESP

A integração desse sistema no ambiente administrativo da UNESP pode trazer inúmeras vantagens, agilizando processos e reduzindo a necessidade de intervenções manuais. Destacam-se as seguintes aplicações:

### 1. Validação de Documentos Acadêmicos

O sistema pode ser utilizado para verificar automaticamente documentos de identidade em processos administrativos, como matrícula de novos alunos, inscrição em programas de bolsas e submissão de documentos em plataformas digitais. Isso economiza tempo e reduz erros humanos.

### 2. Atendimento Digital

Integrado aos sistemas de atendimento ao aluno, como o SUCEM, o chatbot pode realizar a validação de documentos enviados online, permitindo respostas rápidas e automatizadas para consultas sobre validade e autenticidade de informações.

### 3. Processos de Pesquisa

Em atividades de pesquisa que exijam identificação e validação de participantes, o sistema pode agilizar a verificação de documentos, garantindo maior eficiência e confiabilidade nos dados coletados.

### 4. Organização de Eventos Acadêmicos

Em congressos e simpósios organizados pela UNESP, o sistema pode ser empregado para validar documentos de identidade de participantes, automatizando a conferência e acelerando o credenciamento.

### Benefícios Gerais

A implementação do sistema na UNESP proporcionaria:

- Maior eficiência em processos administrativos.
- Redução de erros na validação de documentos.
- Atendimento ágil e acessível para a comunidade acadêmica.

- Redução da carga de trabalho dos funcionários, permitindo maior foco em atividades estratégicas.

Os resultados apresentados demonstram a robustez do sistema em processar e validar documentos de identidade. A aplicabilidade do sistema em instituições como a UNESP reforça sua viabilidade para modernizar processos administrativos e acadêmicos, trazendo eficiência e segurança.

### Análise Final

Os resultados demonstram a eficácia do sistema em processar imagens de documentos de identidade. A combinação de técnicas de pré-processamento, detecção de bordas e OCR permitiu a extração precisa do texto. A integração com a interface Gradio facilitou a interação com o usuário, proporcionando uma experiência intuitiva e eficiente. Os principais desafios identificados foram:

- Qualidade da imagem original: Imagens de baixa qualidade podem afetar o desempenho do OCR.
- Complexidade visual: Elementos como fundos texturizados e marcas d'água podem introduzir ruído.

Mesmo com essas limitações, o sistema demonstrou robustez em condições variadas, validando sua aplicação prática em cenários reais.

## 7 Resultados - Verificação de Diploma

Nesta seção, apresentamos os resultados obtidos ao aplicar o sistema de validação no documento de diploma fornecido. O objetivo foi verificar a capacidade do sistema em identificar palavras específicas no documento utilizando técnicas de OCR (Reconhecimento Óptico de Caracteres) e visão computacional.

### Documento Original

A imagem original do diploma foi carregada no sistema como entrada para o processo de validação. O diploma contém informações textuais importantes, como o nome da instituição, o nome do aluno, a data de emissão e o número de registro, além de elementos decorativos que adicionam complexidade ao documento.

- Qualidade da Imagem: Pequenas distorções causadas por iluminação desigual ou reflexos impactaram a precisão do OCR, principalmente em textos menores.
- Estilo da Fonte: Palavras com fontes estilizadas ou decorativas foram mais difíceis de identificar, resultando em falhas no reconhecimento.
- Parâmetros do OCR: As configurações padrão do Tesseract, como o idioma e a confiança mínima, podem não estar totalmente otimizadas para este tipo de documento.



Figura 7: Imagem original do diploma.

A Figura 7 apresenta o diploma fornecido, que foi utilizado como base para o processo de validação.

### Resultados da Validação

Durante o processamento da imagem, o sistema utilizou as etapas de pré-processamento (conversão para escala de cinza, suavização, detecção de bordas e correção de perspectiva) e a aplicação do OCR para extração de texto. O principal resultado foi:

- A única palavra corretamente delimitada e identificada foi "diploma".
- A palavra foi localizada no corpo principal do texto, indicando que o sistema conseguiu processar parcialmente o conteúdo textual do diploma.

No entanto, informações adicionais, como nomes ou datas, não foram reconhecidas com precisão devido às limitações descritas a seguir.

### Limitações Observadas

O teste revelou algumas limitações do sistema ao processar o diploma fornecido:

- Complexidade Visual: O diploma apresenta elementos decorativos, selos e bordas que introduzem ruídos e dificultam a identificação precisa de todas as palavras.

O sistema demonstrou ser funcional ao identificar palavras específicas, como "diploma", mas apresentou limitações na identificação de informações mais detalhadas. Isso evidencia a necessidade de melhorias no pré-processamento de imagens e na configuração do OCR.

### Possíveis Melhorias

Com base nos resultados, sugerimos algumas melhorias para aprimorar o sistema:

- Treinamento de Modelos Específicos: Redes neurais treinadas para diplomas nacionais podem melhorar a identificação de padrões textuais e reduzir erros.
- Aprimoramento do Pré-processamento: Técnicas adicionais de segmentação e remoção de ruído podem isolar melhor as áreas textuais.
- Ajustes no OCR: Otimizar os parâmetros do Tesseract, como o idioma e o nível de confiança mínima, para se adequar melhor às características dos diplomas.
- Uso de Redes Convolucionais: Incorporar CNNs para segmentação e reconhecimento de texto pode complementar o OCR tradicional, aumentando a precisão.

### Conclusão da Verificação do Diploma

Apesar das limitações enfrentadas, o sistema foi capaz de identificar a palavra-chave "diploma" no documento fornecido. Isso demonstra a viabilidade da aplicação do sistema em cenários reais, com a possibilidade de melhorias futuras para torná-lo mais robusto e eficiente em diferentes contextos, como validação acadêmica e administrativa.

## 8 Desafios e Limitações

Durante o desenvolvimento e implementação do sistema de verificação de RG, foram encontrados diversos desafios que impactaram o desempenho do sistema. Entre os principais, destacam-se:

- **Qualidade da Imagem Original:** Documentos digitalizados ou fotografados em condições de baixa iluminação, com sombras ou baixa resolução dificultaram o reconhecimento dos caracteres pelo OCR. Isso impactou a precisão do sistema, especialmente em documentos com fontes muito pequenas ou desgastadas.
- **Complexidade Visual do Documento:** Elementos decorativos, marcas d'água e fundos texturizados introduziram ruídos que afetaram a detecção de bordas e o reconhecimento de texto. Esses elementos aumentaram a dificuldade para o Tesseract identificar os caracteres corretamente.
- **Variações de Perspectiva:** Fotografias tiradas em ângulos inclinados ou com distorção exigiram uma etapa de correção de perspectiva. Apesar de eficiente, essa correção depende da qualidade da detecção de contornos, que pode ser comprometida em documentos muito danificados ou com bordas mal definidas.
- **Configuração do OCR:** O Tesseract requer parâmetros específicos, como idioma e confiança mínima, para otimizar o reconhecimento. Ajustes inadequados desses parâmetros podem comprometer a precisão do sistema.
- **Limitações de Dados:** O sistema foi testado com um conjunto limitado de exemplos. Documentos com formatos muito diferentes ou fontes não treinadas pelo Tesseract podem apresentar desempenho inferior.

- **Dependência de Intereração Humana:** Apesar da interface Gradio facilitar a interação, a dependência do usuário para fornecer termos de pesquisa e interpretar os resultados ainda representa uma limitação para automação completa.
- Embora esses desafios não tenham inviabilizado o sistema, eles destacam áreas de aprimoramento, especialmente na robustez do OCR e no pré-processamento de imagens.

## 9 Conclusão

O desenvolvimento do sistema de verificação de RG e Diploma utilizando aprendizado profundo, técnicas de visão computacional e uma interface interativa demonstrou ser uma abordagem promissora para a automação de processos de validação documental. O sistema combina:

- Técnicas de pré-processamento de imagens, como conversão para escala de cinza, suavização e detecção de bordas.
- Reconhecimento Óptico de Caracteres (OCR) com o Tesseract, auxiliado por redes LSTM, para extração confiável de texto.
- Interface Gradio, que proporciona uma interação simples e eficaz com o usuário, validando documentos em tempo real.

Os resultados obtidos mostraram uma precisão satisfatória, especialmente em imagens de qualidade adequada e documentos padronizados. A integração do sistema com aplicações administrativas, como matrículas e validações em eventos acadêmicos na UNESP, destaca seu potencial de aplicabilidade prática.

No entanto, os desafios enfrentados, como ruídos em imagens e complexidade visual dos documentos, apontam para possíveis melhorias futuras:

- Treinamento de modelos OCR específicos para padrões de documentos nacionais.
- Implementação de técnicas de aprendizado profundo mais avançadas, como redes neurais convolucionais (CNNs) para segmentação de texto e identificação de padrões.
- Automação completa do processo, reduzindo a dependência de interações手工.
- Portanto, o sistema proposto apresenta uma base sólida para futuras expansões e pode ser um importante aliado em processos administrativos e acadêmicos que envolvam validação de documentos na UNESP e em outras instituições.

## Referências

- GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. *Deep Learning*. MIT Press, 2016.
- SMITH, Ray. An Overview of the Tesseract OCR Engine. *Proceedings of the Ninth International Conference on Document Analysis and Recognition*, IEEE, 2007.
- BROWN, Tom et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020.