APLICAÇÃO DE RECONHECIMENTO DE TEXTO MANUSCRITO PARA A DIGITALIZAÇÃO DE REDAÇÕES DO ENSINO MÉDIO EM PORTUGUÊS

APPLICATION OF HANDWRITTEN TEXT RECOGNITION (HTR) FOR THE DIGITIZATION OF HIGH SCHOOL ESSAYS IN PORTUGUESE

Lucas V. Dias
Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE), Recife – PE – Brasil
lucas.valentimdias@ufrpe.br

RESUMO

A digitalização de redações manuscritas é uma etapa crucial para sistemas de correção automática, mas representa um desafio significativo devido à variabilidade caligráfica e às particularidades linguísticas. Este artigo aborda essa lacuna ao realizar uma análise comparativa do desempenho de seis modelos de Reconhecimento de Texto Manuscrito de aprendizado profundo (HTR-Flor, StackMix, OrigamiNet, TrOCR, HTR-VT, DTrOCR) e dois serviços comerciais (AWS Textract, GCP Vision). Os modelos foram avaliados em um novo conjunto de dados composto por 22.927 linhas de texto extraídas de 1.071 redações manuscritas de alunos do ensino médio em português. Os resultados, medidos por Taxa de Erro de Caractere (CER) e Palavra (WER), indicam que os modelos HTR-VT (CER 9,20%) e Stackmix (CER 11,72%) demonstraram maior robustez e eficácia neste domínio específico. Notavelmente, modelos baseados em Transformer como TrOCR e DTrOCR, que são estado da arte em benchmarks padronizados, apresentaram desempenho significativamente inferior (CER > 48%), evidenciando uma fraca generalização para a caligrafia variada encontrada nas redações. O estudo conclui que a especialização do modelo ao domínio e a robustez a diferentes estilos de caligrafia são mais cruciais para o desempenho prático do que a performance em *datasets* genéricos, fornecendo um panorama sobre o estado da arte e os desafios do HTR para o cenário educacional brasileiro.

Palavras-chave: Reconhecimento de Texto Manuscrito, HTR, Redações Manuscritas, Aprendizado Profundo, CTC, Transformer.

ABSTRACT

The digitization of handwritten essays is a crucial step for automated grading systems, yet it poses a significant challenge due to calligraphic variability and linguistic particularities. This paper addresses this gap by conducting a comparative performance analysis of six deep learning-based Handwritten Text Recognition (HTR) models (HTR-Flor, StackMix, OrigamiNet, TrOCR, HTR-VT, DTrOCR) and two commercial services (AWS Textract, GCP Vision). The models were evaluated on a new *dataset* comprising 22,927 text lines extracted from 1,071 handwritten essays by Brazilian high school students in Portuguese. The results, measured by Character Error Rate (CER) and Word Error Rate (WER), indicate that the HTR-VT (9.20% CER) and Stackmix (11.72% CER) models demonstrated greater robustness and effectiveness in this specific domain. Notably, Transformer-based models like TrOCR and DTrOCR, which are state-of-the-art on standard benchmarks, showed significantly

inferior performance (CER > 48%), highlighting poor generalization to the varied handwriting found in the essays. The study concludes that model specialization to the domain and robustness to diverse handwriting styles are more critical for practical performance than performance on generic *datasets*, providing an overview of the state-of-the-art and challenges of HTR for the Brazilian educational scenario.

Keywords: Handwritten Text Recognition, HTR, Handwritten Essays, Deep Learning, CTC, Transformer.

Datas de submissão e aprovação do artigo: 08/08/2025

1 INTRODUÇÃO

A correção automática de redações representa um dos principais desafios contemporâneos na intersecção entre inteligência artificial e educação, especialmente considerando que os sistemas educacionais ainda dependem amplamente da produção manuscrita de texto (BANSAL et al., 2025; LIN; LI, 2025). O Reconhecimento de Texto Manuscrito (HTR - Handwritten Text Recognition) emerge como uma tecnologia fundamental para transpor essa barreira, convertendo imagens de texto manuscrito em texto digital processável computacionalmente (RASSUL et al., 2025; GARRIDO-MUNOZ et al., 2025). Esta necessidade torna-se particularmente crítica no contexto educacional brasileiro, onde a avaliação de redações manuscritas em larga escala apresenta desafios únicos relacionados à variabilidade caligráfica dos estudantes e às complexidades morfológicas da língua portuguesa (LEAL et al., 2021; DE ALENCAR; CUCONATO; RADEMAKER, 2018).

Recentes avanços em aprendizado profundo revolucionaram o campo do HTR, com modelos baseados em arquiteturas Transformer demonstrando capacidades notáveis na digitalização de documentos históricos e materiais manuscritos contemporâneos (LI et al., 2022). Nos últimos cinco anos, observou-se uma transição significativa dos métodos tradicionais baseados em heurísticas para sistemas neurais sofisticados que aproveitam mecanismos de atenção e pré-treinamento em larga escala (RASSUL et al., 2025; GARRIDO-MUNOZ et al., 2025). Modelos como TrOCR (Transformer-based Optical Character Recognition) e Vision Transformers adaptados para HTR alcançam desempenho superior em benchmarks padronizados, aproveitando o poder dos modelos de linguagem pré-treinados (LI et al., 2022; DOSOVITSKIY et al., 2020).

A digitalização precisa de redações manuscritas não apenas viabiliza a automação do processo de correção, mas também possibilita análises em larga escala do desempenho estudantil, identificação de padrões de erros e desenvolvimento de ferramentas pedagógicas mais eficazes (REDDY CHAVVA et al., 2024; SYED ABUTHAHIR et al., 2024). Sistemas de Avaliação Automática de Redações (AES - Automated Essay Scoring) têm demonstrado eficácia significativa quando integrados com modelos de linguagem avançados, alcançando correlações superiores a 0,8 com avaliações humanas (BANSAL et al., 2025; XIAO et al., 2025). Contudo, a precisão destes sistemas depende criticamente da qualidade da digitalização inicial do texto manuscrito, estabelecendo o HTR como componente essencial no *pipeline* de processamento (BANSAL et al., 2025).

O contexto da língua portuguesa apresenta desafios específicos para sistemas de HTR devido à sua rica morfologia, presença de caracteres acentuados e variações regionais na caligrafia (LEAL et al., 2021; GONÇALO OLIVEIRA, 2018). Pesquisas recentes em processamento de linguagem natural para português indicam que as particularidades morfológicas da língua requerem abordagens computacionais especializadas, especialmente em cenários de recursos limitados (DUARTE EHLERT et al., 2025; WIEMERSLAGE et al., 2022). A escassez de *datasets* específicos para português manuscrito, comparada à abundância de recursos para línguas como inglês, torna essencial o desenvolvimento de estudos comparativos que avaliem diferentes abordagens de HTR neste contexto específico.

Este trabalho contribui para o avanço do estado da arte ao apresentar uma análise comparativa sistemática de modelos de HTR de aprendizado profundo aplicados especificamente a redações manuscritas em português, incluindo a criação e avaliação de um novo *dataset* composto por 22.927 linhas de texto extraídas de 1.071 redações manuscritas de alunos do ensino médio brasileiro. A investigação abrange desde modelos consolidados baseados em redes convolucionais e recorrentes até arquiteturas Transformer de última geração, incluindo comparações com serviços comerciais estabelecidos. A relevância desta pesquisa estende-se além do domínio técnico, oferecendo *insights* fundamentais para o desenvolvimento de sistemas educacionais mais eficazes e acessíveis no contexto brasileiro.

2 DESENVOLVIMENTO

2.1 Trabalhos Relacionados

A diversidade de abordagens arquiteturais e metodológicas no campo do HTR tem resultado em uma ampla gama de modelos com características e desempenhos distintos (GARRIDO-MUNOZ et al., 2025). Enquanto modelos tradicionais baseados em CNNs (Convolutional Neural Networks) combinadas com RNNs (Recurrent Neural Networks) demonstram robustez em cenários específicos, as arquiteturas Transformer emergentes têm estabelecido novos paradigmas de desempenho em benchmarks padronizados, embora sua eficácia em domínios especializados ainda demande investigação mais aprofundada (LI et al., 2022; RASSUL et al., 2025).

Esta seção aborda os trabalhos mais relevantes que serviram de base para a análise comparativa apresentada neste artigo, destacando as contribuições específicas de cada modelo e sua relevância para o reconhecimento de texto manuscrito em português.

2.1.1 HTR-Flor

HTR-Flor (DE SOUSA NETO et al., 2020) é um sistema de HTR offline baseado em aprendizado profundo que propõe uma nova arquitetura utilizando Gated-CNNs (Convolutional Neural Networks com mecanismos de *gating*). Essa arquitetura é projetada para ser mais eficiente, com menos parâmetros e camadas, ao mesmo tempo em que supera o desempenho de sistemas HTR de última geração. O HTR-Flor integra ainda dois passos de modelo de linguagem, tanto em nível de caractere quanto de palavra, para refinar a transcrição. O sistema foi treinado e avaliado em

diversos *datasets* off-line de HTR, como IAM, Bentham, Rimes, Saint Gall e Washington, demonstrando sua robustez e capacidade de generalização.

2.1.2 StackMix e Blot Augmentations

O trabalho que introduz StackMix e Blot Augmentations (SHONENKOV et al., 2021) propõe um sistema de HTR que se destaca pelo uso de técnicas de aumento de dados inovadoras. Essas técnicas visam simular variações realistas encontradas em textos manuscritos, como rasuras (Blot Augmentations) e a geração de texto manuscrito a partir de texto impresso (StackMix). A aplicação dessas aumentações, em conjunto com uma rede Resnet-BiLSTM-CTC, demonstrou uma redução significativa nas taxas de erro de palavra (WER) e caractere (CER), melhorando a robustez do modelo a diferentes estilos de escrita e imperfeições do documento. Os experimentos extensivos em dez *datasets* de texto manuscrito validam a eficácia dessas abordagens na melhoria da qualidade do reconhecimento.

2.1.3 OrigamiNet

OrigamiNet (YOUSEF; BISHOP, 2020) representa um avanço significativo em direção ao reconhecimento de texto em página inteira, sem a necessidade de segmentação prévia em linhas ou palavras. A proposta central é um módulo de rede neural simples e inovador, denominado OrigamiNet, que pode ser adicionado a qualquer rede neural convolucional (CNN) existente treinada com CTC. Este módulo permite que o modelo implicitamente desdobre uma imagem de múltiplas linhas em uma única linha, transformando um arranjo 2D de caracteres em 1D. Isso permite que reconhecedores de texto de linha única sejam estendidos para lidar com páginas inteiras, aprendendo a 'desdobrar' o conteúdo. O modelo utiliza apenas camadas convolucionais 1-D, sem recorrência ou mecanismos de *gating* complexos, e alcança desempenho comparável a abordagens mais complexas.

2.1.4 TrOCR

TrOCR (Transformer-based Optical Character Recognition with Pre-trained Models) (LI et al., 2022) é uma abordagem de reconhecimento de texto de ponta a ponta que utiliza modelos Transformer pré-treinados para visão computacional e processamento de linguagem natural. Proposto pela Microsoft, o TrOCR é um modelo simples, mas eficaz, que pode ser pré-treinado com grandes volumes de dados sintéticos e ajustado com *datasets* rotulados por humanos. A arquitetura do TrOCR consiste em um *encoder* Transformer de imagem e um *decoder* Transformer de texto autorregressivo, permitindo o reconhecimento de caracteres ópticos de forma eficiente. Experimentos demonstram que o TrOCR alcança resultados impressionantes em diversas tarefas de reconhecimento de texto, incluindo texto manuscrito, devido à sua capacidade de alavancar o conhecimento adquirido em pré-treinamento em larga escala.

2.1.5 HTR-VT

HTR-VT (Handwritten Text Recognition with Vision Transformer) (LI et al., 2025) explora a aplicação de Vision Transformers (ViT) para o reconhecimento de texto manuscrito. Dada a disponibilidade limitada de dados rotulados nesse domínio,

o HTR-VT propõe um modelo baseado em ViT que emprega apenas o componente *encoder* do Transformer padrão. Este trabalho destaca a eficácia de uma abordagem simples e eficiente em termos de dados, com modificações mínimas no ViT original. A pesquisa demonstra que o HTR-VT é capaz de alcançar precisão competitiva em *datasets* padrão de HTR, mesmo em cenários de escassez de dados, o que é particularmente relevante para domínios específicos como o reconhecimento de redações em português.

2.1.6 DTrOCR

DTrOCR (Decoder-only Transformer for Optical Character Recognition) (FUJITAKE, 2023) propõe um método mais simples e eficaz para o reconhecimento de texto, utilizando uma arquitetura Transformer apenas com o *decoder*. Este estudo demonstra que uma arquitetura Transformer focada exclusivamente no *decoder* pode superar modelos tradicionais de encoder-decoder para tarefas de OCR e STR (Scene Text Recognition). A simplicidade e a eficácia do DTrOCR o tornam uma alternativa promissora para o reconhecimento de texto, incluindo manuscritos, ao focar na geração de texto a partir das características visuais, sem a necessidade de um *encoder* complexo para extração de características visuais, que já podem ser aprendidas pelo *decoder*.

Apesar dos avanços significativos demonstrados por esses modelos em diversos *datasets* internacionais, poucos trabalhos focam especificamente na caligrafia de estudantes em língua portuguesa, uma lacuna que este estudo visa preencher através da comparação direta dos modelos HTR-Flor, StackMix, OrigamiNet, TrOCR, HTR-VT e DTrOCR no domínio específico de redações manuscritas de alunos do Ensino Médio brasileiro. A escolha desses modelos para análise comparativa baseia-se em suas diferentes abordagens arquiteturais e técnicas de treinamento, proporcionando uma visão abrangente do estado atual da tecnologia HTR aplicada a este contexto específico.

2.2 Fundamentação Teórica

Esta seção apresenta os conceitos fundamentais necessários para a compreensão dos modelos de HTR analisados neste trabalho, incluindo as métricas de avaliação utilizadas e as principais técnicas subjacentes aos sistemas modernos de reconhecimento de texto manuscrito.

2.2.1 Métricas de Avaliação: CER e WER

No contexto do Reconhecimento de Texto Manuscrito (HTR) e do Reconhecimento Óptico de Caracteres (OCR), a avaliação da precisão dos modelos é realizada por meio de métricas que quantificam a diferença entre o texto transcrito pelo sistema e o texto de referência (*ground truth*). As duas métricas mais comuns e relevantes são a Taxa de Erro de Caractere (CER - Character Error Rate) e a Taxa de Erro de Palavra (WER - Word Error Rate).

2.2.1.1 Métricas de Avaliação: CER e WER

A Taxa de Erro de Caractere (CER) mede a precisão do reconhecimento em nível de caractere. Ela é calculada com base na distância de Levenshtein (também conhecida como distância de edição) entre a sequência de caracteres predita pelo modelo e a sequência de caracteres de referência. A distância de Levenshtein contabiliza o número mínimo de operações de edição (inserções, deleções e substituições de caracteres) necessárias para transformar uma sequência na outra. A fórmula geral para o CER é:

$$CER = \frac{S + D + I}{N}$$

Onde:

- S é o número de substituições de caracteres.
- D é o número de deleções de caracteres.
- I é o número de inserções de caracteres.
- N é o número total de caracteres na sequência de referência.

Um CER de 0% indica uma transcrição perfeita em nível de caractere, enquanto valores mais altos indicam mais erros. Para texto manuscrito, estudos indicam que um bom CER geralmente varia entre 2% e 10%, dependendo da complexidade da caligrafia e do *dataset* (GARRIDO-MUNOZ et al., 2025). O CER é particularmente útil para avaliar a granularidade do reconhecimento e identificar problemas em caracteres individuais, o que é crucial para idiomas com morfologia rica como o português.

2.2.1.2 Word Error Rate (WER)

A Taxa de Erro de Palavra (WER) é uma métrica análoga ao CER, mas que opera em nível de palavra. Ela mede a proporção de palavras incorretamente reconhecidas em relação ao total de palavras na sequência de referência. Assim como o CER, o WER também utiliza a distância de Levenshtein, mas aplicada a palavras em vez de caracteres. A fórmula para o WER é:

$$WER = \frac{S + D + I}{N}$$

Onde:

- S é o número de substituições de palavras.
- D é o número de deleções de palavras.
- I é o número de inserções de palavras.
- N é o número total de palavras na sequência de referência.

Um WER de 0% significa que todas as palavras foram reconhecidas corretamente. O WER é uma métrica mais abrangente para avaliar a fluidez e a coerência do texto reconhecido, sendo amplamente utilizada em tarefas de reconhecimento de fala e HTR. Valores mais baixos de WER indicam melhor desempenho geral do sistema. Ambas as métricas são complementares e fornecem uma visão completa da qualidade da transcrição.

2.2.2 Técnicas Fundamentais em HTR

Os modelos de HTR modernos, especialmente aqueles baseados em aprendizado profundo, empregam diversas técnicas para lidar com a complexidade do reconhecimento de texto manuscrito. As principais incluem Connectionist Temporal Classification (CTC), mecanismos de Attention e a arquitetura Transformer.

2.2.2.1 Connectionist Temporal Classification (CTC)

CTC (GRAVES et al., 2006) é uma função de custo (*loss function*) utilizada em redes neurais recorrentes (RNNs) para treinar modelos que mapeiam sequências de entrada para sequências de saída, onde o alinhamento entre as sequências de entrada e saída é desconhecido. No contexto do HTR, o CTC permite que a rede aprenda a transcrever uma imagem de texto diretamente para uma sequência de caracteres, sem a necessidade de segmentar explicitamente cada caractere na imagem. Isso é particularmente vantajoso para texto manuscrito, onde a segmentação de caracteres pode ser ambígua devido à caligrafia cursiva e variações de espaçamento. O CTC introduz um caractere 'blank' que permite que a rede ignore regiões não informativas da imagem e lide com repetições de caracteres, simplificando o processo de treinamento e tornando-o mais robusto a variações na entrada.

2.2.2.2 Mecanismos de Attention

Os mecanismos de Attention (BAHDANAU; CHO; BENGIO, 2016) revolucionaram as arquiteturas de sequência a sequência (sequence-to-sequence), permitindo que o modelo foque em partes relevantes da sequência de entrada ao gerar cada elemento da sequência de saída. Em HTR, isso significa que, ao transcrever um caractere ou palavra, o modelo pode dar mais peso às regiões da imagem de entrada que são mais relevantes para aquele caractere ou palavra específica. Isso melhora significativamente a capacidade do modelo de lidar com sequências longas e complexas, além de proporcionar uma maior interpretabilidade, pois é possível visualizar quais partes da imagem o modelo está 'prestando atenção' em cada etapa da transcrição. Modelos baseados em Attention são particularmente eficazes para lidar com a variabilidade e a natureza não linear do texto manuscrito.

2.2.2.3 Arquitetura Transformer

A arquitetura Transformer (VASWANI et al., 2017), introduzida em 2017, eliminou a necessidade de redes recorrentes (RNNs) e convolucionais (CNNs) em muitas tarefas de processamento de linguagem natural e visão computacional, incluindo HTR. A principal inovação do Transformer é o uso extensivo de mecanismos de auto-atenção (*self-attention*), que permitem que o modelo processe todas as partes da sequência de entrada em paralelo, capturando dependências de longo alcance de forma mais eficiente. No HTR, os Transformers podem ser aplicados tanto no *encoder* (para extrair características da imagem) quanto no *decoder* (para gerar a sequência de texto). Modelos como TrOCR e HTR-VT são exemplos da aplicação bem-sucedida de Transformers em HTR, demonstrando sua capacidade de aprender representações robustas e alcançar resultados de ponta, especialmente quando combinados com pré-treinamento em larga escala. A capacidade de processamento paralelo dos Transformers também os torna ideais para aplicações que exigem alta eficiência computacional.

2.3 Metodologia Experimental

Esta seção detalha a metodologia empregada no estudo, abrangendo a descrição do conjunto de dados utilizado para treinamento, validação e teste dos modelos de HTR, o ambiente computacional onde os experimentos foram conduzidos e os parâmetros de treinamento específicos para cada modelo.

2.3.1 Conjunto de Dados

O conjunto de dados utilizado para este estudo é composto por redações manuscritas em português, produzidas por alunos do Ensino Médio. A escolha desse tipo de dado é crucial para o desenvolvimento de um serviço de correção automática de redações, pois reflete a realidade do material a ser processado.

O *dataset* foi cuidadosamente dividido para garantir a robustez e a generalização dos modelos, seguindo uma proporção padrão para tarefas de aprendizado de máquina:

- **Divisão Inicial**: As redações foram inicialmente divididas em 70% para treinamento, 15% para validação e 15% para teste.
- **Imagens de Redações**: O conjunto total compreende 1.071 imagens de redações completas.
- **Imagens de Linhas Extraídas**: A partir das redações, foram extraídas 22.927 imagens de linhas individuais, por meio de um processo de anotação manual. Essa extração em nível de linha é fundamental para muitos modelos de HTR que operam nessa granularidade.

A distribuição específica das redações e linhas extraídas por conjunto (treinamento, validação e teste) é a seguinte:

Redações:

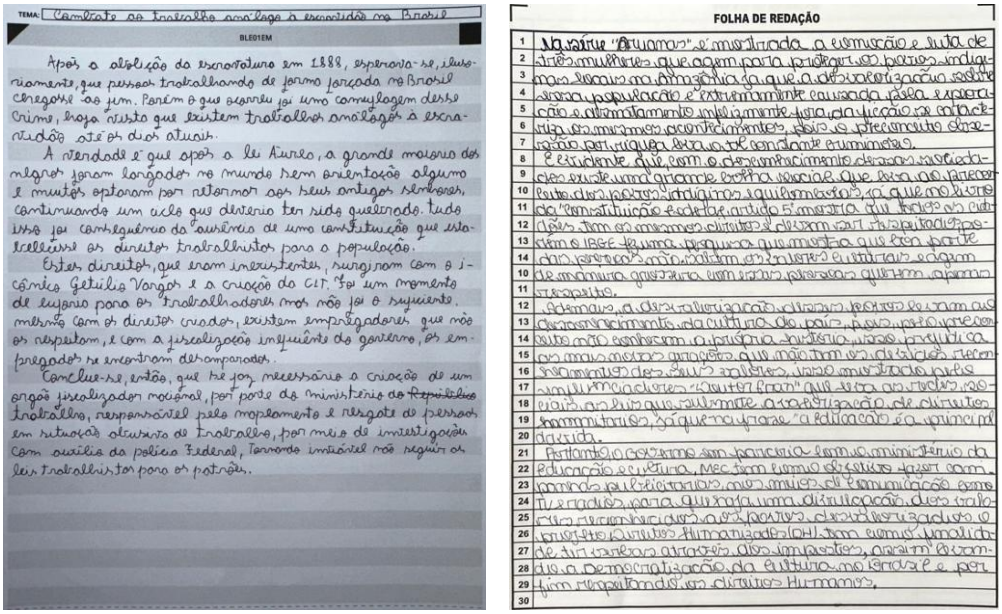
- Treinamento: 749 redações
- Validação: 161 redações
- Teste: 161 redações

Linhas Extraídas:

- Treinamento: 16.064 linhas
- Validação: 3.471 linhas
- Teste: 3.392 linhas

Essa divisão assegura que os modelos sejam treinados em um volume significativo de dados, validados em um conjunto independente para ajuste de hiper parâmetros e avaliados em um conjunto de teste totalmente inédito, garantindo uma estimativa imparcial do desempenho.

Figura 1 – Folha de redação 1 e 2 disponíveis no conjunto de dados.



Fonte: Elaborado pelo autor (2025).

A Figura 1 evidencia os diferentes níveis de legibilidade, estruturas das folhas de redação e variações caligráficas presentes no *dataset*. Essa diversidade é fundamental para avaliar a robustez dos modelos de HTR e a sua capacidade de generalização frente aos desafios do contexto educacional brasileiro.

2.3.2 Ambiente de Experimentos

Todos os experimentos e o treinamento dos modelos foram realizados em um ambiente de computação de alto desempenho para garantir a eficiência e a viabilidade do processo. O ambiente utilizado foi o Google Colab, uma plataforma baseada em nuvem que oferece recursos computacionais robustos, incluindo acesso a GPUs (Graphics Processing Units).

As especificações do ambiente de experimentos são as seguintes:

- **Plataforma**: Google Colab Pro
- **Memória RAM**: 83,5 GB
- **GPU**: NVIDIA A100 com 40GB de memória VRAM

A utilização de uma GPU A100 com 40GB de VRAM é particularmente importante para o treinamento de modelos de aprendizado profundo, especialmente aqueles baseados em arquiteturas Transformer, que são computacionalmente intensivos. A grande quantidade de memória VRAM permite o processamento de batches maiores de dados, acelerando o treinamento e possibilitando a experimentação com modelos mais complexos.

2.3.3 Configuração de Treinamento dos Modelos

Para garantir a comparabilidade dos resultados e a reprodutibilidade dos experimentos, foram adotados parâmetros de treinamento consistentes, baseados nas configurações propostas nos artigos de referência ou nos repositórios oficiais de cada modelo. De forma geral, todos os modelos utilizaram um *batch size* de 16 para treinamento e validação, foram treinados por 50 *epochs* (exceto StackMix e HTR-VT,

que usaram 50.000 iterações), além de uma estratégia de *Early Stopping* com *patience* de 5 *epochs* (ou 5.000 iterações) para evitar *overfitting*. Os hiperparâmetros específicos para cada modelo estão detalhados na Tabela 1.

Tabela 1 – Parâmetros de treinamento dos modelos de HTR

Modelo	Taxa de Aprend.	Otimizador	Função de Perda	Scheduler
HTR-Flor	0,001	RMSprop	CTCLoss	Reduce LR on Plateau
StackMix and Blot Augmentations	0,001 (máx)	AdamW	CTCLoss	OneCycleLR c/ Warmup
OrigamiNet	0,01 → 0.001	Adam	CTCLoss	Exponential Decay
TrOCR	0,00002	Adam	Cross-entropy	Inv. Sqrt. c/ Warm-up
HTR-VT	0,001 (máx)	AdamW + SAM	CTCLoss	Cosine c/ Warm-up
DTrOCR	0,0001	AdamW	Cross-entropy	–

Fonte: Elaborado pelo autor (2025).

2.3.4 Configuração dos Serviços Comerciais

Para a avaliação dos serviços comerciais de reconhecimento de texto, foram utilizadas as APIs oficiais dos provedores, garantindo acesso às versões mais atualizadas dos algoritmos.

2.3.4.1 AWS Textract

O serviço AWS Textract (AMAZON WEB SERVICES, 2025) foi acessado através da biblioteca boto3 para Python. Cada imagem de linha do conjunto de teste foi processada individualmente através da API, utilizando a funcionalidade de detecção de texto padrão. O serviço foi configurado para processar imagens em formato JPEG com resolução original, sem aplicação de filtros adicionais.

2.3.4.2 GCP Vision

O GCP Vision (GOOGLE CLOUD PLATAFORM, 2025) foi acessado através da biblioteca google-cloud-vision para Python. Similar ao AWS Textract, cada imagem foi processada individualmente, utilizando a API de detecção de texto com configurações padrão. O serviço foi configurado para retornar o texto detectado em formato de *string* simples, facilitando a comparação com os outros modelos.

2.3.4.3 Considerações de Custo

É importante destacar que ambos os serviços comerciais apresentam custos associados ao uso. Com base nas informações de preços consultadas em julho de 2025:

- **GCP Vision:** US$ 1,50 por 1.000 chamadas de API até 5 milhões de chamadas por mês, reduzindo para US$ 0,60 por 1.000 chamadas adicionais.

- **AWS Textract:** US$ 1,50 por 1.000 chamadas de API até 1 milhão de chamadas por mês, reduzindo para US$ 0,60 por 1.000 chamadas adicionais.

Esses custos foram considerados na análise comparativa, especialmente ao avaliar a viabilidade econômica de cada solução para aplicações em larga escala no contexto educacional.

2.4 Resultados e Discussão

Esta seção apresenta e discute os resultados obtidos pelos diferentes modelos de HTR e serviços comerciais no conjunto de dados de redações em português. A análise abrange tanto aspectos quantitativos, com a apresentação das métricas de desempenho, quanto aspectos qualitativos, com a discussão dos tipos de erros observados e exemplos visuais.

2.4.1 Desempenho Quantitativo dos Modelos

Para contextualizar os resultados obtidos neste estudo, é importante primeiro apresentar o desempenho dos modelos avaliados no conjunto de dados IAM Handwriting Database (IAM) (MARTI; BUNKE, 2002), um benchmark amplamente utilizado na literatura de HTR. A Tabela 2 apresenta os resultados dos modelos no IAM Dataset, servindo como referência para comparação com os resultados obtidos no *dataset* de redações em português.

Tabela 2 – Desempenho dos modelos de HTR no IAM Dataset

Modelo	CER (%)	WER (%)
HTR-Flor	3,72	11,18
StackMix and Blot Augmentations	3,77	–
OrigamiNet	4,76	–
TrOCR (Base)	3,42	–
HTR-VT	4,7	14,9
DTrOCR	2,38	–

Fonte: Elaborado pelo autor (2025).

Os resultados no IAM Dataset demonstram que o DTrOCR alcançou o melhor desempenho em termos de CER (2,38%), seguido pelo TrOCR (3,42%) e HTR-Flor (3,72%). Esses valores servem como baseline para avaliar como os modelos se comportam em um *dataset* padronizado e amplamente reconhecido na comunidade científica.

A Tabela 3 apresenta os resultados dos modelos e serviços comerciais avaliados especificamente no conjunto de dados de redações manuscritas em português desenvolvido neste estudo.

Tabela 3 – Desempenho dos modelos de HTR e serviços comerciais no conjunto de dados de redações em português

Modelo	CER (%)	WER (%)
HTR-VT	9,20	25,75
StackMix and Blot Augmentations	11,72	31,29

HTR-Flor	14,10	39,90
OrigamiNet	14,18	35,66
GCP Vision	23,83	59,94
AWS Textract	36,19	81,38
DTrOCR	48,01	70,21
TrOCR	50,64	72,05

Fonte: Elaborado pelo autor (2025).

A comparação entre as Tabelas 2 e 3 revela diferenças significativas no desempenho dos modelos quando aplicados ao domínio específico de redações manuscritas em português. Enquanto o DTrOCR demonstrou excelente performance no IAM Dataset, sua eficácia foi consideravelmente reduzida no contexto de redações, evidenciando a importância da especificidade do domínio e da diversidade da caligrafia para o desempenho dos modelos de HTR.

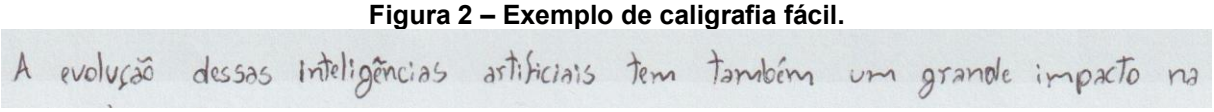
2.4.2 Análise Qualitativa dos Erros e Exemplos Visuais

A discussão dos resultados ganha profundidade ao analisar qualitativamente os tipos de erros cometidos pelos modelos. A variabilidade da caligrafia em redações de ensino médio, aliada às particularidades da língua portuguesa, como a presença de caracteres acentuados e cedilha, representa um desafio significativo para os sistemas de HTR. Para ilustrar esses desafios, apresentamos exemplos visuais de transcrições de diferentes modelos em casos de caligrafia fácil e difícil.

2.4.2.1 Exemplo de Caligrafia Fácil

A Figura 2 apresenta um exemplo de caligrafia considerada fácil, com boa legibilidade e espaçamento regular. Abaixo da imagem, são mostradas as transcrições de referência (*ground truth*) e as obtidas por diferentes modelos e serviços comerciais.

Figura 2 – Exemplo de caligrafia fácil.



Fonte: Elaborado pelo autor (2025).

Texto de Referência (*Ground Truth*): 'A evolução dessas inteligências artificiais tem também um grande impacto na'

Transcrições:

- **AWS Textract**: 'A evolução dessas inteligências artificiais tem também um grande impacto na'
- **GCP Vision**: 'A evolução dessas inteligências artificiais tem também\nUm\ngrande impacto\nna'
- **HTR-Flor**: 'A evolução desas inteligências artificiais tem tambénm um grande impacto no'
- **HTR-VT**: 'A evolução dessas inteligências artificiais tem também um grande impacto no'
- **OrigamiNet**: 'A volução dessas inteligências artrcas tem também um grande impecto no'

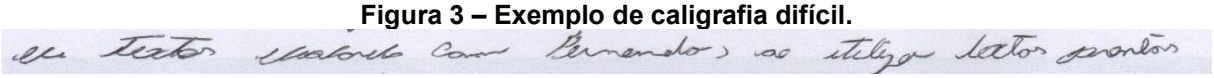
- **Stackmix**: 'A evlução dessas inteligências artificiais tem também um grande impacto na'
- **TrOCR**: 'A dessas inteligência inteligência tem também um in in in impacto na'
- **DTrOCR**: 'A evo inteli art t u gran impac n'

Neste exemplo, o AWS Textract e o HTR-VT demonstraram um desempenho superior, com transcrições muito próximas ou idênticas ao *ground truth*. O GCP Vision, embora tenha reconhecido a maioria das palavras, inseriu quebras de linha indevidas. Modelos como TrOCR e DTrOCR apresentaram dificuldades significativas, com erros de reconhecimento de caracteres e palavras, resultando em transcrições incoerentes. Isso sugere que, mesmo em caligrafias consideradas fáceis, a robustez dos modelos varia consideravelmente, e a capacidade de generalização para diferentes estilos de escrita é um fator crítico.

2.4.2.2 Exemplo de Caligrafia Difícil

A Figura 3 ilustra um caso de caligrafia difícil, caracterizada por traços menos definidos, espaçamento irregular e possíveis junções de caracteres. A complexidade desse tipo de escrita representa um desafio ainda maior para os sistemas de HTR.

Figura 3 – Exemplo de caligrafia difícil.



Fonte: Elaborado pelo autor (2025).

Texto de Referência (*Ground Truth*): e textos vobardo com termendo, ao itiliza lxtos prontãs

Transcrições:

- **AWS Textract**: 'ee Textor Cam Pernando, as itiliza latos prenton'
- **GCP Vision**: 'textos\nMalore Comm\nPernandos\n500\nitely lector prantors'
- **HTR-Flor**: 'e textos pocece com lunandos ao itilga foxtas proneás'
- **HTR-VT**: 'ede totos vobardo com termendo, ao itiliza lxtos prontãs'
- **OrigamiNet**: 'eu textos vobardo com cumendos ao itiliza lextas pranhos'
- **Stackmix**: 'eou textos opade com lonandos o etelga loltas pants'
- **TrOCR**: 'textosss o comssvidas ao aog,g textoss'
- **DTrOCR**: 'textt o pe a iti lats pants'

Neste cenário de caligrafia difícil, a performance de todos os modelos e serviços comerciais foi visivelmente comprometida. Embora alguns modelos, como o HTR-VT, tenham conseguido capturar algumas palavras-chave, a maioria das transcrições apresenta um alto índice de erros, com palavras incorretas, caracteres trocados e frases sem sentido. Os modelos baseados em Transformer (TrOCR e DTrOCR), que demonstraram bom desempenho em *datasets* padronizados, tiveram uma performance particularmente fraca neste tipo de caligrafia, sugerindo uma menor robustez a variações extremas na escrita. Isso reforça a necessidade de *datasets* mais diversos e técnicas de aumento de dados específicas para lidar com a complexidade da caligrafia em português.

2.4.3 Discussão dos Tipos de Erro

A análise dos exemplos visuais revela padrões de erros comuns:

- **Erros de Segmentação**: Modelos como o GCP Vision frequentemente inserem quebras de linha indevidas, indicando dificuldades na segmentação de linhas ou na interpretação do layout do texto.
- **Erros de Caracteres Similares**: A confusão entre caracteres visualmente semelhantes (ex: 'o' e 'a', 'm' e 'n', 'i' e 'l') é um erro recorrente, especialmente em caligrafias menos legíveis.
- **Erros de Palavras Completas**: Em caligrafias difíceis, a transcrição de palavras inteiras pode ser comprometida, resultando em palavras inexistentes ou com significado completamente diferente do original.
- **Sensibilidade à Caligrafia Cursiva**: A caligrafia cursiva, comum em redações, representa um desafio particular. Modelos que não foram extensivamente treinados em dados cursivos podem ter dificuldade em conectar caracteres e reconhecer palavras completas.
- **Particularidades da Língua Portuguesa**: A presença de caracteres acentuados ('á', 'é', 'í', 'ó', 'ú') e da cedilha ('ç') pode introduzir erros adicionais, caso os modelos não tenham sido adequadamente expostos a esses caracteres durante o treinamento.

Em resumo, a análise qualitativa complementa a avaliação quantitativa, fornecendo insights valiosos sobre as limitações e os pontos fortes de cada modelo. Fica evidente que, para o domínio específico de redações manuscritas em português, a robustez a diferentes estilos de caligrafia e a capacidade de lidar com as particularidades da língua são cruciais para o sucesso do HTR.

3 CONSIDERAÇÕES FINAIS

Este trabalho explorou a aplicação do Reconhecimento de Texto Manuscrito (HTR) no contexto desafiador de redações de ensino médio em português, realizando uma análise comparativa de modelos de aprendizado profundo e serviços comerciais. Os resultados demonstraram que, embora os modelos baseados em Transformer (TrOCR e DTrOCR) apresentem excelente desempenho em *datasets* padronizados, sua performance pode ser significativamente comprometida em caligrafias mais complexas e variadas, como as encontradas em redações. Em contraste, modelos como HTR-VT e Stackmix mostraram maior robustez em diferentes estilos de escrita, destacando a importância da generalização para aplicações práticas.

A análise qualitativa dos erros revelou padrões como dificuldades na segmentação, confusão entre caracteres similares e sensibilidade à caligrafia cursiva e às particularidades da língua portuguesa. Esses achados reforçam a necessidade de abordagens que considerem a diversidade da escrita manual e a riqueza morfológica do português para o desenvolvimento de sistemas de HTR mais eficazes no contexto educacional.

3.1 Limitações do Estudo

É importante reconhecer as limitações inerentes a este estudo. Embora o *dataset* de redações manuscritas em português seja significativo e representativo do domínio, sua diversidade pode não abranger todas as variações de caligrafia e estilos de escrita presentes na população estudantil brasileira. Além disso, a análise

comparativa foi restrita a um conjunto específico de modelos de HTR e serviços comerciais, o que pode não cobrir todas as abordagens existentes no campo. A anotação manual das linhas de texto, embora garanta alta precisão, é um processo demorado e não escalável, o que limita o tamanho do *dataset* que pode ser processado.

3.2 Trabalhos Futuros

Com base nos resultados e nas limitações identificadas, diversas direções para trabalhos futuros podem ser exploradas:

- **Expansão do Conjunto de Dados**: A criação de um *dataset* maior e mais diversificado de redações manuscritas em português, com anotações semi-automáticas ou colaborativas, poderia aprimorar ainda mais o treinamento e a avaliação dos modelos.
- **Fine-tuning de Modelos Transformer**: Investigar o *fine-tuning* de modelos como o TrOCR e DTrOCR com um *dataset* mais específico de português manuscrito, utilizando técnicas de aumento de dados avançadas, pode melhorar significativamente seu desempenho em caligrafias desafiadoras.
- **Técnicas de Pré-processamento Avançadas**: Explorar o impacto de técnicas de pré-processamento de imagens mais sofisticadas, como binarização adaptativa, correção de inclinação e normalização de traços, na performance dos modelos de HTR.
- **Integração com PLN**: Combinar o HTR com módulos de Processamento de Linguagem Natural (PLN) para criar um *pipeline* completo de correção automática de redações, que não apenas transcreva o texto, mas também analise o conteúdo, a gramática e o estilo.
- **Análise de Custos-Benefícios**: Realizar uma análise mais aprofundada dos custos-benefícios da utilização de serviços comerciais de HTR em comparação com o desenvolvimento e manutenção de modelos próprios, considerando o volume de dados e a frequência de uso.

Essas direções futuras visam não apenas aprimorar a precisão do HTR em redações de ensino médio, mas também contribuir para o avanço da pesquisa em reconhecimento de texto manuscrito em português e para o desenvolvimento de ferramentas educacionais inovadoras.

REFERÊNCIAS

AMAZON WEB SERVICES. **Amazon textract**, 2025. Disponível em: https://aws.amazon.com/pt/textract/. Acessado em: 25 jul. 2025.

BAHDANAU, D.; CHO, K.; BENGIO, Y. **Neural machine translation by jointly learning to align and translate**. *arXiv preprint arXiv:1409.0473*, 2016.

BANSAL, B. et al. **Automated essay scoring: A comparative study of machine learning and deep learning approaches**. In: INTERNATIONAL CONFERENCE ON ADVANCES IN ELECTRICAL, COMPUTING, COMMUNICATION AND SUSTAINABLE TECHNOLOGIES (ICAECT), 5., 2025. Anais... 2025. p. 1–7.

DE ALENCAR, L. F.; CUCONATO, B.; RADEMAKER, A. **MorphoBr: An open source large-coverage full-form lexicon for morphological analysis of portuguese**. Texto Livre: Linguagem e Tecnologia, v. 11, n. 3, p. 1–25, 2018.

DE SOUSA NETO, A. F. et al. Htr-flor: **A deep learning system for offline handwritten text recognition**. In: SIBGRAPI CONFERENCE ON GRAPHICS, PATTERNS AND IMAGES (SIBGRAPI), 33., 2020. Anais... 2020. p. 54–61.

DOSOVITSKIY, A. et al. **An image is worth 16x16 words: Transformers for image recognition at scale**. *arXiv preprint arXiv:2010.11929*, 2020.

DUARTE EHLERT, A.; DA ROCHA JUNQUEIRA, J.; ASTROGILDO DE FREITAS, L.; BRISOLARA CORRÊA, U. **A review of recent advances on automatic question generation for the portuguese language**. The International FLAIRS Conference Proceedings, v. 38, n. 1, 2025.

FUJITAKE, M. **Dtrocr: Decoder-only transformer for optical character recognition**. *arXiv preprint arXiv:2308.15839*, 2023.

GARRIDO-MUNOZ, C.; RIOS-VILA, A.; CALVO-ZARAGOZA, J. **Handwritten text recognition: A survey**. *arXiv preprint arXiv:2303.01804*, 2025.

GONÇALO OLIVEIRA, H. **Distributional and knowledge-based approaches for computing portuguese word similarity**. Information, v. 9, n. 2, 2018.

GOOGLE CLOUD PLATAFORM. **Cloud vision api – handwriting detection**, 2025. Disponível em: https://cloud.google.com/vision/docs/handwriting. Acessado em: 25 jul. 2025.

GRAVES, A. et al. **Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks**. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML), 23., 2006, New York. Proceedings... New York: Association for Computing Machinery, 2006. p. 369–376.

LEAL, S. E. et al. **Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese**. *arXiv preprint arXiv:2109.11181*, 2021.

LI, M. et al. Trocr: **Transformer-based optical character recognition with pre-trained models**. *arXiv preprint arXiv:2109.10282*, 2022.

LI, Y.; CHEN, D.; TANG, T.; SHEN, X. **Htr-vt: Handwritten text recognition with vision transformer**. Pattern Recognition, v. 158, p. 110967, 2025.

LIN, J.; LI, J. **Harnessing large language models for college-level english language assessment: Opportunities, challenges, and future directions**. In: INTERNATIONAL CONFERENCE ON INFORMATICS EDUCATION AND COMPUTER TECHNOLOGY APPLICATIONS (IECA), 2., 2025. Anais... 2025. p. 58–63.

MARTI, U.-V.; BUNKE, H. **The iam-database: An english sentence database for offline handwriting recognition**. International Journal on Document Analysis and Recognition, v. 5, n. 1, p. 39–46, 2002.

RASSUL, Y. H. et al. **Advancing offline handwritten text recognition: A systematic review of data augmentation and generation techniques**. *arXiv preprint arXiv:2401.07153*, 2025.

REDDY CHAVVA, R. K. et al. **A transformer-based approach for enhancing automated essay scoring**. In: INTERNATIONAL CONFERENCE ON ADVANCED COMPUTING AND EMERGING TECHNOLOGIES (ACET), 1., 2024. Anais**...** 2024. p. 1–6.

SHONENKOV, A. et al. **Stackmix and blot augmentations for handwritten text recognition**. *arXiv preprint arXiv:2109.01422*, 2021.

SYED ABUTHAHIR, S. et al. **Building a deep neural network for automated essay scoring in english language teaching**. In: INTERNATIONAL CONFERENCE ON COMPUTING COMMUNICATION AND NETWORKING TECHNOLOGIES (ICCCNT), 15., 2024. Anais**...** 2024. p. 1–6.

VASWANI, A. et al. **Attention is all you need. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS**, 30., 2017. Anais**...** 2017.

WIEMERSLAGE, A. et al. **Morphological processing of low-resource languages: Where we are and what's next**. *arXiv preprint arXiv:2210.15575*, 2022.

XIAO, C. et al. **Human-ai collaborative essay scoring: A dual-process framework with llms**. In: INTERNATIONAL LEARNING ANALYTICS AND KNOWLEDGE CONFERENCE (LAK), 15., 2025. Proceedings**...** ACM, 2025. p. 293–305.

YOUSEF, M.; BISHOP, T. E. **Origaminet: Weakly-supervised, segmentation-free, one-step, full page text recognition by learning to unfold**. *arXiv preprint arXiv:2005.02107*, 2020.