UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Edivaldo Leitão de Albuquerque Neto

Análise de Comparação entre modelos de Compreensão Visual de Documentos usando LoRA

Recife

2026

Edivaldo Leitão de Albuquerque Neto

Análise de Comparação entre modelos de Compreensão Visual de Documentos usando LoRA

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciências da Computação

Orientador: Valmir Macario Filho

Recife

2026

Dedico este trabalho à minha família, pelo apoio, incentivo e compreensão ao longo de toda a minha trajetória acadêmica, e a todos que, de alguma forma, contribuíram para a realização deste sonho.

Agradecimentos

Agradeço à minha família pelo apoio, incentivo e compreensão ao longo de toda a minha trajetória acadêmica. Dedico um agradecimento especial ao meu falecido avô materno, cuja lembrança, ensinamentos e exemplo continuam sendo fonte de inspiração para mim. Este trabalho também representa o resultado do apoio e da confiança que sempre recebi de minha família. …

Resumo

O crescimento do volume de documentos digitais tem impulsionado o desenvolvimento de soluções baseadas em Inteligência Artificial para automatizar tarefas de extração e processamento de informações. Nesse contexto, este trabalho realiza uma análise comparativa dos modelos LayoutLMv3, DocFormer e Donut aplicados à tarefa de compreensão de documentos, investigando o impacto da técnica Low-Rank Adaptation (LoRA) no processo de ajuste fino dos modelos. A metodologia adotada consistiu na aplicação de diferentes configurações de LoRA, variando os módulos-alvo, o rank, o alpha, o dropout e a taxa de aprendizado, com o objetivo de reduzir o número de parâmetros treináveis sem comprometer o desempenho preditivo. Os experimentos foram conduzidos utilizando métricas de Precisão, Revocação e F1-Score para avaliar os resultados. Entre os modelos analisados, o LayoutLMv3 apresentou o melhor desempenho geral, alcançando Precisão de 92,02 %, Revocação de 90,01 % e F1-Score de 91,00% utilizando adaptação nos módulos Query e Value, com aproximadamente 1,19 milhão de parâmetros treináveis. O Donut também apresentou resultados competitivos, obtendo Precisão de 89,75%, Revocação de 92,13% e F1-Score de 90,92% com adaptação nos módulos queryj, value, key e output, empregando resolução de entrada de 960 × 720 pixels e aproximadamente 544 mil parâmetros treináveis. Os resultados demonstram que o uso de LoRA permite reduzir significativamente o número de parâmetros atualizados durante o treinamento, preservando elevados índices de desempenho. Além disso, observou-se que diferentes estratégias de adaptação influenciam diretamente o equilíbrio entre o custo computacional e a qualidade preditiva. Conclui-se que a combinação de modelos multimodais para a compreensão de documentos e de técnicas de ajuste fino eficientes representa uma alternativa viável para aplicações de Processamento Inteligente de Documentos, permitindo a obtenção de resultados competitivos com menor demanda computacional

Palavras-chave: Processamento Inteligente de Documentos, LayoutLMv3, DocFormer, Fine-tuning,Donut, LoRA ….

Abstract

The growth in the volume of digital documents has driven the development of Artificial Intelligence-based solutions to automate information extraction and processing tasks. In this context, this work compares LayoutLMv3, DocFormer, and Donut models for document comprehension, investigating the impact of the Low-Rank Adaptation (LoRA) technique on model fine-tuning. The methodology adopted consisted of applying different LoRA configurations, varying target modules, rank, alpha, dropout, and learning rate, to reduce the number of trainable parameters without compromising predictive performance. The experiments were conducted using Precision, Recall, and F1-Score metrics to evaluate the results. Among the models analyzed, LayoutLMv3 showed the best overall performance, achieving Precision of 92.02%, Recall of 90.01%, and an F1-score of 91.00% with adaptation in the Query and Value modules, using approximately 1.19 million trainable parameters. The Donut model also presented competitive results, achieving an accuracy of 89.75%, Recall of 92.13%, and an F1-score of 90.92% with adaptation in the query, value, key, and output modules, employing an input resolution of 960 × 720 pixels and approximately 544,000 trainable parameters. The results demonstrate that LoRA enables significant parameter reduction during training while preserving high performance. Furthermore, it was observed that different adaptation strategies directly influence the balance between computational cost and predictive quality. It is concluded that the combination of multimodal models for document comprehension and efficient fine-tuning techniques represents a viable alternative for Intelligent Document Processing applications, allowing for competitive results with lower computational demand.

Keywords: Intelligent Document Processing, LayoutLMv3,DocFormer, Donut, LoRA, Efficient fine-tuning ….

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

OCR	Optimized Character Recognition
Lora	Low-rank Adaption
RVL-CDIP	Ryerson Vision Lab Complex Document Information Processing
FUNSD	Form Understanding in Noisy Scanned Documents
CORD	Consolidated Receipt Dataset

Sumário

1 Introdução

O crescimento contínuo do volume de documentos digitais em organizações públicas e privadas têm impulsionado o desenvolvimento de soluções baseadas em Inteligência Artificial para automatizar tarefas de classificação, extração de informações e compreensão documental. Nesse contexto, a Visão Computacional e o Processamento Inteligente de Documentos têm desempenhado um papel fundamental ao possibilitar a interpretação automática de documentos que contêm texto, imagens, tabelas, formulários e outros elementos gráficos como em Mahadevkar et al. (2024). Entretanto, a diversidade de layouts, a qualidade variável das imagens, a presença de documentos manuscritos e a combinação de diferentes tipos de informação tornam essa tarefa significativamente mais complexa, exigindo modelos capazes de integrar simultaneamente características textuais, visuais e espaciais. As bases de dados utilizadas neste estudo ilustram essa heterogeneidade. O conjunto RVL-CDIP reúne documentos pertencentes a diferentes categorias, como cartas, formulários, relatórios, anúncios e memorandos, enquanto o FUNSD contém formulários escaneados com anotações semânticas e o CORD é composto por recibos comerciais. Exemplos desses documentos são apresentados na Figura 1. Nas últimas décadas, os avanços em

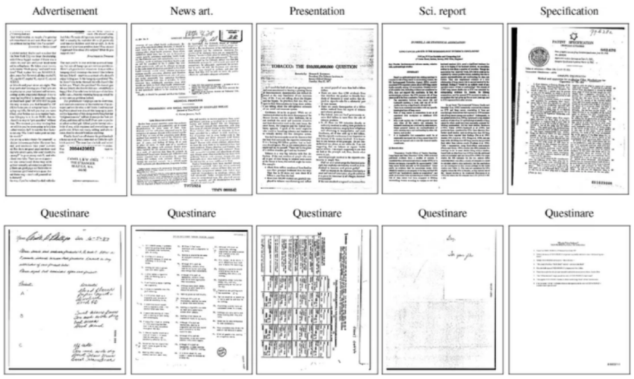


Figura 1 – Exemplos de documentos das bases de dados utilizadas no estudo (KV; MONDAL; JAWAHAR, 2022)

arquiteturas baseadas em *Transformers* impulsionaram o desenvolvimento de modelos multimodais especializados na compreensão de documentos. Entre os principais modelos propostos na literatura destacam-se o LayoutLMv3 proposto por Huang et al. (2022), que integra informações textuais, visuais e estruturais em um único processo de pré-treinamento; o DocFormer proposto por Appalaraju et al. (2021), que utiliza mecanismos de atenção multimodal para explorar as relações entre texto, imagem e layout; e o Donut (Document Understanding Transformer) proposto por Kim et al. (2022), que elimina a dependência explícita de sistemas de Reconhecimento Óptico

de Caracteres (OCR) ao processar diretamente a imagem do documento. Essas arquiteturas têm obtido resultados expressivos em tarefas como classificação documental, extração de informações e compreensão de formulários, consolidando-se como referências no estado da arte da área de Document AI. Apesar do elevado desempenho dessas arquiteturas, sua adaptação para tarefas específicas normalmente exige a realização do *fine-tuning* completo, processo no qual milhões de parâmetros do modelo são atualizados durante o treinamento. Essa estratégia demanda elevado poder computacional, grande quantidade de memória gráfica (VRAM) e longos tempos de processamento, tornando sua utilização inviável em muitos ambientes acadêmicos e profissionais com infraestrutura limitada. Como alternativa, a literatura recente tem proposto técnicas de *Parameter-Efficient Fine-Tuning* (PEFT), dentre as quais se destaca a *Low-Rank Adaptation* (LoRA) desenvolvido em Hu et al. (2022). Essa técnica introduz matrizes de baixa dimensionalidade em camadas específicas da arquitetura *Transformer*, permitindo que apenas uma pequena parcela dos parâmetros seja atualizada durante o treinamento. Dessa forma, reduz-se significativamente o consumo de memória e o custo computacional, preservando desempenho competitivo em relação ao *fine-tuning* tradicional. Diante desse contexto, o objetivo deste trabalho é realizar uma análise comparativa entre os modelos multimodais LayoutLMv3, DocFormer e Donut, investigando o impacto da aplicação da técnica *Low-Rank Adaptation* (LoRA) sobre o desempenho preditivo e a eficiência computacional dessas arquiteturas em tarefas de compreensão visual de documentos. Para isso, foram avaliadas diferentes configurações de adaptação, envolvendo módulos-alvo (*target modules*), valores de rank, alpha, dropout e taxa de aprendizado, utilizando os conjuntos de dados RVL-CDIP, FUNSD e CORD. O conjunto RVL-CDIP (*Ryerson Vision Lab Complex Document Information Processing*) é amplamente utilizado em tarefas de classificação documental e, neste trabalho, foi empregada sua versão RVL-CDIP-SMALL, disponibilizada na plataforma Kaggle, composta por aproximadamente 40 mil imagens distribuídas em 16 categorias documentais. O conjunto FUNSD (*Form Understanding in Noisy Scanned Documents*) contém formulários escaneados anotados com entidades semânticas e suas relações, sendo adequado para tarefas de compreensão de formulários e extração de informações. Já o conjunto CORD (*Consolidated Receipt Dataset*) reúne recibos comerciais anotados estruturalmente, permitindo avaliar modelos em tarefas de extração de informações de documentos semiestruturados. A avaliação experimental foi conduzida por meio das métricas de Precisão, Revocação e F1-Score, buscando analisar simultaneamente a qualidade preditiva dos modelos e a eficiência proporcionada pela redução do número de parâmetros treináveis decorrente da aplicação do LoRA. Os resultados demonstraram que a técnica *Low-Rank* Adaptation possibilita reduzir significativamente a quantidade de parâmetros atualizados durante o treinamento, preservando desempenho competitivo em relação ao *fine-tuning* completo. Observou-se ainda que o com-

portamento dos modelos varia conforme o tipo de documento analisado: o LayoutLMv3 apresentou melhor desempenho na tarefa de classificação documental (RVL-CDIP), o DocFormer destacou-se na compreensão de formulários (FUNSD) e o Donut obteve os melhores resultados no conjunto CORD, evidenciando que a escolha da arquitetura depende das características da tarefa e do domínio dos documentos. Além desta introdução, este trabalho está organizado da seguinte forma. A Seção 2 apresenta a fundamentação teórica, descrevendo as arquiteturas LayoutLMv3, DocFormer, Donut e a técnica LoRA. A Seção 3 detalha a metodologia, incluindo os conjuntos de dados, o pré-processamento, as configurações experimentais e as métricas de avaliação. A Seção 4 apresenta e discute os resultados obtidos. Por fim, a Seção 5 apresenta as considerações finais e as perspectivas para trabalhos futuros.

2 Base teórica e modelos avaliados

A seguir, são apresentados os principais modelos e técnicas que fundamentam o desenvolvimento deste estudo. Inicialmente, são abordadas arquiteturas voltadas à compreensão inteligente de documentos, com destaque para os modelos LayoutLMv3 desenvolvido por Huang et al. (2022), DocFormer Appalaraju et al. (2021) e Donut Kim et al. (2022), que se destacam na literatura recente por integrarem informações textuais, visuais e espaciais em tarefas de classificação documental e extração de informações. Embora compartilhem o objetivo de compreender documentos de forma multimodal, essas arquiteturas adotam estratégias distintas de representação e processamento de dados, tornando relevante a análise comparativa de seus desempenhos. Por fim, é apresentada a técnica Low-Rank Adaptation (LoRA), uma abordagem de ajuste fino eficiente que possibilita reduzir significativamente o número de parâmetros treináveis e os requisitos computacionais necessários para a adaptação desses modelos a tarefas específicas.

2.1 LayoutLMv3

O crescimento da área de *Document AI* impulsionou o desenvolvimento de modelos capazes de integrar informações textuais, visuais e espaciais para a compreensão de documentos. Nesse contexto, o LayoutLMv3 representa uma evolução da família LayoutLM ao adotar uma estratégia unificada de pré-treinamento multimodal, permitindo o aprendizado conjunto entre essas diferentes modalidades de informação. Diferentemente das versões anteriores, o LayoutLMv3 substitui a extração de características visuais baseada em redes neurais convolucionais por uma representação em *patches* inspirada no *Vision Transformer* (ViT) proposto por Dosovitskiy et al. (2021), simplificando a arquitetura e tornando o processamento mais eficiente. Nesse processo de unificação arquitetural, a imagem do documento é segmentada em pequenos blocos bidimensionais lineares e não sobrepostos (*patches*), os quais passam por uma projeção linear para gerar os *embeddings* visuais. Em paralelo, os termos textuais obtidos via OCR são mapeados pelo tokenizador de texto. A conexão estrutural entre essas duas frentes ocorre através da incorporação de *embeddings* espaciais 2D: tanto os tokens de texto quanto os patches de imagem recebem vetores de coordenadas que representam suas respectivas caixas delimitadoras (*bounding boxes*) normalizadas no plano x0, y0, x1, y1. Essa representação linear unificada de sequências visuais, textuais e espaciais é concatenada e injetada diretamente em um único codificador *Transformer* compartilhado, eliminando a necessidade de ramificações complexas para cada

modalidade. O modelo também combina três objetivos de pré-treinamento: *Masked Language Modeling* (MLM), *Masked Image Modeling* (MIM) e *Word-Patch Alignment* (WPA) para aprender representações multimodais mais consistentes. No processo de MLM, tokens de texto são ocultados aleatoriamente, desafiando o modelo a predizê-los utilizando não apenas o contexto textual restante, mas também o *layout* espacial e as pistas visuais vizinhas. No MIM, blocos contínuos de patches visuais são mascarados; a arquitetura deve então reconstruir essas regiões ocultas predizendo seus respectivos tokens visuais discretos através de um tokenizador de imagem. Para amarrar essas frentes, o WPA atua como um elo de alinhamento cruzado de grão fino: ele atribui um rótulo binário para indicar se uma palavra de texto está alinhada ou desalinhada em relação ao seu patch de imagem correspondente (omitindo a região visual quando a palavra é mascarada), forçando a rede a aprender explicitamente a correspondência mútua entre texto e representação gráfica. Conforme ilustrado na Figura 2, a arquitetura integra informações textuais obtidas por OCR, características visuais e informações de posicionamento (*bounding boxes*), permitindo capturar simultaneamente o conteúdo semântico e a organização espacial dos documentos. Essa abordagem proporcionou resultados de estado da arte em tarefas como classificação documental, compreensão de formulários como mostrado em Chanti et al. (2023).

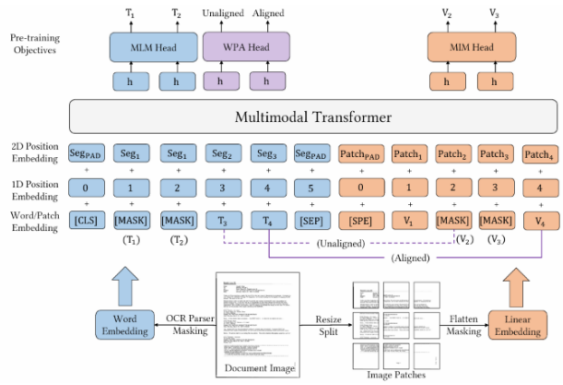


Figura 2 – Arquitetura do modelo LayoutLMv3 (HUANG et al., 2022)

2.2 DocFormer

O DocFormer foi proposto como uma arquitetura multimodal para compreensão de documentos, buscando integrar informações textuais, visuais e espaciais de forma mais eficiente que as abordagens anteriores a sua publicação em Appalaraju et al. (2021). Diferentemente de modelos que tratam essas modalidades separadamente, o DocFormer emprega mecanismos de atenção multimodal (*multi-modal self-attention*), permitindo que as diferentes representações interajam em todas as camadas do Transformer. Para viabilizar essa integração profunda, a arquitetura processa três fluxos

de entrada em paralelo: o conteúdo textual extraído via OCR, as características visuais locais e globais obtidas por meio de um *backbone* convolucional (geralmente uma rede ResNet), e as coordenadas espaciais das caixas delimitadoras (*bounding boxes*) de cada termo. O grande diferencial técnico da arquitetura está na mecânica de seu módulo de atenção multimodal desacoplada. Em vez de simplesmente concatenar os vetores de texto, visão e espaço na entrada da rede, o DocFormer calcula scores de atenção de forma independente para cada combinação de modalidade (interações texto-texto, visão-visão, texto-visão, texto-espaço e visão-espaço), fundindo essas informações de maneira geométrica e semântica ao longo de todas as camadas da rede. Conforme ilustrado na Figura 3, o modelo utiliza *embeddings* espaciais compartilhados entre texto e imagem, favorecendo o aprendizado conjunto do conteúdo textual e de sua localização no documento. Essa estratégia possibilita representar de maneira mais consistente a estrutura dos documentos, tornando o modelo adequado para tarefas como classificação documental, compreensão de formulários e reconhecimento de entidades. Além disso, o DocFormer adota uma estratégia de pré-treinamento autossupervisionado alicerçada em três tarefas fundamentais: o *Multi-modal Masked Language Modeling* (MM-MLM), voltado à predição de *tokens* e regiões visuais ocultas; o *Learn Reconstruction* (LR), focado em reconstruir os recursos visuais de áreas mascaradas; e o *Text-to-Image Alignment* (TIA), responsável por verificar a correspondência fina entre o texto e sua respectiva imagem. Essa abordagem permite o aprendizado de representações robustas sem depender de grandes volumes de dados anotados. Os resultados apresentados pelos autores demonstram desempenho competitivo em diferentes tarefas de compreensão documental, consolidando o modelo como uma alternativa relevante entre as arquiteturas multimodais para *Document AI*.

Figura 3 – Arquitetura do modelo Docformer
(APPALARAJU et al., 2021)

2.3 Donut

Donut (*Document Understanding Transformer*) foi proposto como uma abordagem para compreensão de documentos que elimina a dependência de sistemas de Reconhecimento Óptico de Caracteres (OCR), adotando uma arquitetura denominada OCR-free proposta por Kim et al. (2022). Diferentemente das abordagens tradicionais, o modelo realiza o processamento diretamente sobre a imagem do documento, reduzindo a propagação de erros provenientes da etapa de reconhecimento de texto. Essa estratégia busca superar limitações comuns dos sistemas OCR, como dificuldades no processamento de documentos com baixa qualidade de imagem, de diferentes idiomas ou com templates complexos, fatores que podem comprometer o desempenho das etapas subsequentes de compreensão documental como mostrado em Babu et al. (2024). Conforme ilustrado na Figura 4, o Donut utiliza uma arquitetura do tipo *encoder-decoder*, na qual um *Swin Transformer* é responsável pela extração das características visuais da imagem, enquanto um Transformer autorregressivo gera sequencialmente a saída correspondente à tarefa desejada.Segundo Kim et al. (2022) essa abordagem permite tratar tarefas como classificação de documentos, extração de informações e compreensão de formulários como problemas de geração de texto . Os resultados apresentados pelos autores demonstram que o Donut alcança desempenho competitivo em diferentes tarefas de compreensão documental, simplificando o pipeline de processamento ao eliminar a necessidade de integração com mecanismos de OCR.

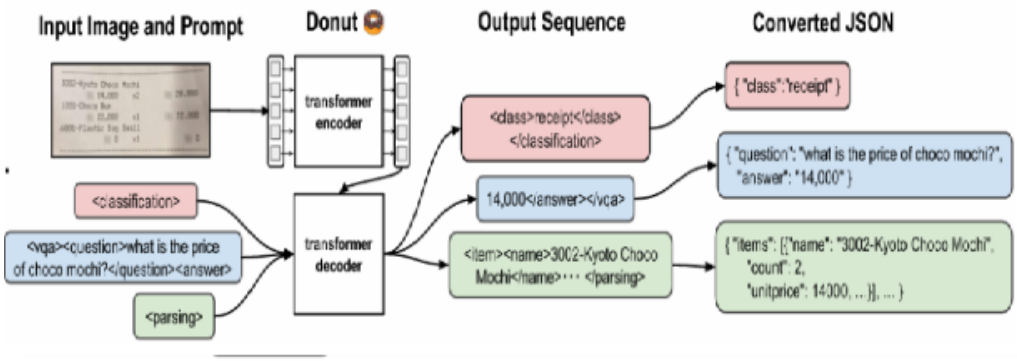


Figura 4 – Arquitetura do modelo Donut
(KIM et al., 2022)

2.4 Low-Rank Adaptation (LoRA)

O crescimento do número de parâmetros dos modelos Transformer trouxe desafios relacionados ao custo computacional e a necessidade de memória durante o processo de ajuste fino. Em muitos casos, o treinamento completo de modelos com centenas de milhões ou bilhões de parâmetros torna-se inviável para pesquisadores e instituições com recursos computacionais limitados. A técnica Low-Rank Adaptation

(LoRA) é uma abordagem de *Parameter-Efficient Fine-Tuning* (PEFT) que busca reduzir o número de parâmetros treináveis durante o ajuste fino de modelos de grande porte como realizado em Hu et al. (2022). A ideia central do LoRA consiste em manter congelados os pesos originais do modelo pré-treinado e em introduzir matrizes de baixa dimensionalidade nas projeções lineares das camadas Transformer. Durante o treinamento, apenas essas matrizes adicionais são atualizadas, reduzindo drasticamente o número de parâmetros treináveis. Os autores em Hu et al. (2022) demonstraram que essa estratégia pode reduzir o número de parâmetros ajustados em até 10.000 vezes e diminuir significativamente o consumo de memória, mantendo desempenho equivalente ou superior ao do ajuste fino tradicional. Do ponto de vista matemático, o LoRA assume que a atualização necessária para adaptar um modelo a uma nova tarefa pode ser representada por uma decomposição de baixa ordem (*low-rank decomposition*). Assim, em vez de atualizar diretamente uma matriz de pesos (W), o método aprende com duas matrizes menores (A) e (B), cuja multiplicação aproxima a atualização desejada. Essa estratégia reduz o número de parâmetros treináveis e permite maior eficiência computacional. A aplicação do LoRA em modelos multimodais para compreensão documental tornou-se especialmente relevante devido ao elevado custo de treinamento dessas arquiteturas. Estudos recentes demonstram que o método permite adaptar modelos de grande porte com uma pequena fração dos parâmetros originais, preservando elevados níveis de desempenho e tornando viável sua utilização em ambientes com recursos computacionais limitados. No escopo desta pesquisa, a introdução do LoRA direciona-se especificamente à otimização das arquiteturas dos modelos em foco. Embora a eficácia dessa técnica de adaptação eficiente seja documentada na literatura para modelos de linguagem puros e redes dedicadas à visão computacional de propósito geral , observa-se uma oportunidade no mapeamento comparativo e sistemático de seu impacto quando aplicada simultaneamente a diferentes paradigmas de *Document AI*. A investigação do comportamento do LoRA frente a modelos baseados em OCR tradicional em contraste com abordagens OCR-free constitui o núcleo da originalidade metodológica deste estudo. As vantagens técnicas dessa abordagem estão relacionadas principalmente à maior eficiência do processo de ajuste fino e à redução dos requisitos computacionais durante o treinamento. Ao manter congelados os pesos do modelo pré-treinado e atualizar apenas as matrizes introduzidas pelo LoRA, reduz-se significativamente a demanda por memória de vídeo (VRAM) durante a retropropagação do gradiente, possibilitando a realização dos experimentos em plataformas com recursos computacionais limitados. Diante desse cenário, este trabalho parte da hipótese de que a aplicação da técnica Low-Rank Adaptation (LoRA) às projeções lineares de atenção das arquiteturas avaliadas permite reduzir significativamente o número de parâmetros treináveis, mantendo-o inferior a aproximadamente 1% do total de parâmetros do modelo, sem comprometer de forma relevante sua capacidade de generaliza-

ção. Espera-se, assim, preservar o desempenho preditivo em tarefas de classificação documental e extração de informações, avaliado por meio das métricas de Precisão, Revocação e F1-Score, ao mesmo tempo em que se reduz o custo computacional necessário para o ajuste fino.

3 Metodologia

Esta seção descreve os procedimentos metodológicos adotados para a realização desta pesquisa. Inicialmente, são apresentadas as bases de dados utilizadas nos experimentos, destacando suas características e adequação às tarefas de compreensão visual de documentos. Em seguida, são descritas as etapas de pré-processamento aplicadas aos documentos, as configurações de fine-tuning com a técnica *Low-Rank Adaptation* (LoRA) para os modelos LayoutLMv3, DocFormer e Donut, bem como as métricas empregadas na avaliação do desempenho dos modelos.

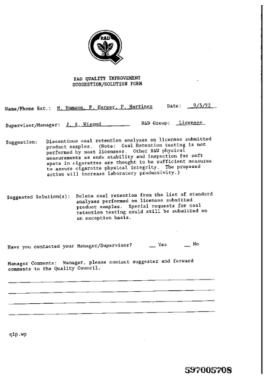
3.1 Bases de Dados

Para fundamentar a validação experimental desta pesquisa e viabilizar uma análise comparativa rigorosa do impacto do mecanismo LoRA (Low-Rank Adaptation), o desenho metodológico incorpora três conjuntos de dados que já foram usados para validação destes mesmos modelos: o FUNSD, o RVL-CDIP (em sua versão amostral reducional RVL-CDIP-SMALL) e o CORD. A seleção coordenada dessas bases de dados foi projetada estrategicamente para submeter as arquiteturas LayoutLMv3, DocFormer e Donut a múltiplos cenários operacionais, testando seus limites de generalização sob diferentes níveis de complexidade estrutural, densidade de ruído e formatos de anotação. Além de cobrir diferentes taxonomias e desafios de leitura visual, a escolha e a moderação volumétrica dessas bases de dados a exemplo da adoção do subconjunto de 40.000 imagens para a tarefa de classificação alinham-se diretamente às restrições práticas de hardware disponíveis para execução deste trabalho, permitindo manter o equilíbrio entre a economia de recurso computacional e a preservação do desempenho.

3.2 FUNSD

O FUNSD (*Form Understanding in Noisy Scanned Documents*) é um conjunto de dados desenvolvido para tarefas de compreensão de formulários digitalizados. A base de dados é composta por 199 documentos anotados manualmente, contendo aproximadamente 9.707 entidades semânticas e 31.485 palavras, organizados nas categorias question, answer, header e other, além das relações existentes entre essas entidades. As entidades semânticas referem-se a blocos de texto ou palavras agrupadas que compartilham uma unidade de significado coerente dentro do documento (como um nome, uma data ou uma pergunta específica), funcionando como as peças

fundamentais para a extração de informações. Por sua vez, a presença de ruídos — que vão desde distorções físicas no papel, manchas, rasgos e dobras, até falhas de digitalização, baixa resolução e imperfeições geradas por ferramentas de OCR que impõem um desafio realista aos modelos computacionais. De acordo com Chanti et al. (2023) a principal característica do FUNSD é preservar tanto o conteúdo textual quanto a organização espacial dos elementos presentes nos formulários, permitindo avaliar a capacidade dos modelos de compreender a estrutura lógica do documento e identificar relações entre seus componentes, mesmo diante das degradações visuais e textuais provocadas por esses ruídos.



R&D QUALITY IMPROVEMENT
SUGGESTION/SOLUTION FORM

Name/Phone Ext.: M. Hamann, P. Farmer, P. Martinez Date: 9/5/93

Supervisor/Manager: J. D. Wigand R&D Group: Licensure

Suggestion: Discontinue seal retention analyses on licenses submitted product samples. (Note: Seal Retention testing is not performed on licenses submitted for R&D purposes as seal stability and integrity for soft spots in cigarettes are thought to be sufficient measure to assure cigarette physical integrity. The proposed action will increase laboratory productivity.)

Suggested Solution(s): Delete seal retention from the list of standard analyses performed on licenses submitted product samples. Special requests for seal retention testing could still be submitted on an exception basis.

Have you contacted your Manager/Supervisor? ___ Yes ___ No

Manager Comments: Manager, please contact suggester and forward comments to the Quality Council.

_____

_____

_____

_____

qip.wp

597005708

Figura 5 – Exemplo de formulário no funsd

Fonte: Elaborado pelo autor (2026).

3.3 RVL-CDIP

O RVL-CDIP (*Ryerson Vision Lab Complex Document Information Processing*) é uma das bases de dados mais utilizadas para avaliação de algoritmos de classificação documental. O conjunto original contém 400.000 imagens distribuídas em 16 categorias, incluindo cartas, formulários, memorandos, relatórios científicos, faturas, artigos, anúncios e documentos manuscritos. Devido às limitações de memória e processamento das plataformas utilizadas nos experimentos, como Google Colab e Kaggle, foi empregada a versão RVL-CDIP-SMALL, disponibilizada na plataforma Kaggle, composta por aproximadamente 40.000 imagens. Neste trabalho, essa base de dados foi utilizada para avaliar a capacidade dos modelos de realizar classificação automática de documentos, analisando o impacto da técnica *Low-Rank Adaptation* (LoRA) sobre o desempenho dessa tarefa.

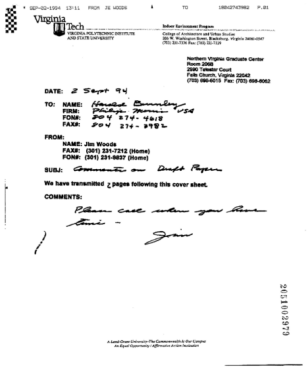


Figura 6 – Exemplo de documento no rvl-cdip

Fonte: Elaborado pelo autor (2026).

3.4 CORD

O CORD (*Consolidated Receipt Dataset*) é um conjunto de dados voltado para tarefas de extração de informações de recibos comerciais. A base de dados reúne aproximadamente 1.000 imagens contendo anotações estruturadas de informações como nome do estabelecimento, produtos adquiridos, quantidades, preços, impostos e valor total da compra.Sua principal finalidade é avaliar modelos capazes de compreender documentos semiestruturados, nos quais o significado das informações depende não apenas do texto, mas também de sua disposição espacial no documento. Neste trabalho, o CORD foi utilizado para analisar a capacidade dos modelos de realizar tarefas de extração estruturada de informações, permitindo avaliar a eficácia da adaptação por LoRA em documentos comerciais.

Figura 7 – Imagem de exemplos dos recibos no CORD

Fonte: Elaborado pelo autor (2026).

3.5 Pré-processamento

O pré-processamento dos dados foi realizado com o objetivo de adequar os documentos aos requisitos de entrada das arquiteturas avaliadas, preservando as informações visuais, textuais e estruturais necessárias para as tarefas de classificação documental e extração de informações. Essa etapa também buscou padronizar a representação dos dados e reduzir o consumo de recursos computacionais durante o treinamento dos modelos. Para os modelos LayoutLMv3 e DocFormer, foram utilizadas as informações textuais e espaciais já disponibilizadas nas anotações dos conjuntos de dados, como as palavras reconhecidas e suas respectivas coordenadas de localização na página. Essas informações não foram extraídas manualmente durante o pré-processamento; nesta etapa, elas foram apenas organizadas e convertidas para o formato de entrada exigido por cada arquitetura, como sequências de tokens e coordenadas normalizadas. Essa representação permite associar o conteúdo textual à sua posição no documento, favorecendo a exploração conjunta de características visuais, textuais e de layout, aspecto fundamental para a interpretação semântica de documentos estruturados como é demonstrado em Tay et al. (2022). No caso do Donut, que adota uma abordagem OCR-free, o processamento foi realizado diretamente sobre a imagem do documento, sem a necessidade de extração prévia de texto por OCR. Inicialmente, as imagens foram convertidas para o formato RGB e redimensionadas para as resoluções definidas experimentalmente, conforme apresentado na Tabela 1. Foram avaliadas diferentes configurações de resolução, permitindo analisar o impacto da quantidade de informação visual disponível no desempenho do modelo. As resoluções adotadas variaram de acordo com as características de cada arquitetura e dos conjuntos de dados utilizados. Para o LayoutLMv3, as imagens foram padronizadas para 224 × 224 pixels nos conjuntos FUNSD, CORD e RVL-CDIP. Para o DocFormer, foi utilizada a resolução de 500 × 384 pixels em todos os conjuntos de dados. Já para o Donut, foram empregadas as resoluções de 960 × 720 *pixels* nos conjuntos FUNSD e CORD, e 640 × 480 pixels no conjunto RVL-CDIP.

Tabela 1 – Redimensionamento dos datasets por modelo

Modelo	FUNSD	CORD	RVL-CDIP
LayoutLMv3	224 × 224	224 × 224	224 × 224
DocFormer	500 × 384	500 × 384	500 × 384
Donut	960 × 720	960 × 720	640 × 480

Fonte: Elaborado pelo autor (2026).

O mesmo procedimento de pré-processamento foi aplicado aos conjuntos de treinamento, validação e teste, respeitando a separação entre eles para evitar vaza-

mento de informações, seguindo a proporção de 80%,10% e 10%. Dessa forma, cada amostra foi processada apenas dentro do conjunto ao qual pertencia, garantindo a integridade da avaliação experimental, e que no teste final de o modelo não tivesse visto a mesma imagem anteriormente.

3.6 Configurações de Fine-Tuning

O fine-tuning consiste na etapa de adaptação de um modelo previamente treinado para uma tarefa ou domínio específico. Os modelos LayoutLMv3, DocFormer e Donut utilizados nesta pesquisa foram originalmente pré-treinados em grandes coleções de documentos, aprendendo representações multimodais capazes de integrar informações textuais, visuais e espaciais. Entretanto, essas representações possuem caráter geral e, por si só, não são suficientes para maximizar o desempenho em conjuntos de dados específicos, como RVL-CDIP, FUNSD e CORD. Os trabalhos originais dessas arquiteturas empregam o fine-tuning como etapa final para especializar os modelos em diferentes tarefas de compreensão documental, tais como classificação de documentos, extração de informações e reconhecimento de entidades semânticas. Dessa forma, este trabalho segue a mesma estratégia metodológica, diferenciando-se pela adição da utilização da técnica *Low-Rank Adaptation* (LoRA) como mecanismo de adaptação eficiente dos parâmetros. Ao contrário do *fine-tuning* tradicional, no qual todos os parâmetros da arquitetura são atualizados durante o treinamento, o LoRA mantém congelados os pesos do modelo pré-treinado e adiciona matrizes de baixa dimensão às projeções lineares das camadas Transformer. Durante o ajuste fino, apenas essas matrizes adicionais são treinadas, reduzindo significativamente o número de parâmetros atualizados durante o treino, o consumo de memória gráfica (VRAM) e o custo computacional, sem modificar a arquitetura original do modelo Hu et al. (2022). As configurações empregadas neste trabalho foram definidas a partir de experimentos preliminares, nos quais foram avaliadas diferentes combinações de hiperparâmetros e módulos adaptados, buscando estabelecer um equilíbrio entre desempenho preditivo e eficiência computacional. Inicialmente, foram avaliadas diferentes resoluções de entrada das imagens e diferentes limites máximos de tokens textuais. Essa etapa mostrou-se particularmente importante para o modelo Donut, cuja arquitetura OCR-free realiza o processamento diretamente sobre a imagem do documento. Resoluções mais elevadas preservam uma maior quantidade de detalhes visuais, porém aumentam significativamente o consumo de memória durante o treinamento. Dessa forma, buscou-se definir configurações que mantivessem informações relevantes do documento sem exceder a capacidade das GPUs utilizadas. Em seguida, foram avaliados diferentes valores para o hiperparâmetro Rank (r), responsável por definir a dimensão das matrizes de baixa ordem introduzidas pelo LoRA. Foram realizados experimentos

com valores iguais a 8 e 16, permitindo analisar a influência da quantidade de parâmetros treináveis na capacidade de adaptação dos modelos. De forma complementar, foi investigado o hiperparâmetro *Alpha* (α), que controla a contribuição das matrizes do LoRA sobre os pesos congelados do modelo durante o treinamento. Foram avaliados os valores 16 e 32, buscando estabilizar o processo de otimização e favorecer a convergência do treinamento. Também foram conduzidos experimentos variando a taxa de aprendizado (learning rate), o dropout e o batch size, de modo a obter uma adaptação estável sem comprometer a capacidade de generalização dos modelos. O batch size foi mantido reduzido em razão das limitações de memória das GPUs empregadas nos experimentos. Outro aspecto investigado foi a seleção dos módulos-alvo (target modules) para aplicação do LoRA. Como a técnica insere matrizes treináveis apenas nas camadas selecionadas, a escolha desses módulos influencia diretamente a quantidade de parâmetros atualizados e o desempenho obtido. No LayoutLMv3, foram avaliadas adaptações nas projeções *Query* e *Value*, responsáveis pelo cálculo da atenção entre os tokens textuais e visuais. No DocFormer, adotou-se a estratégia *all-linear*, aplicando o LoRA em todas as projeções lineares da arquitetura. Já no Donut, foram investigadas as projeções *query, key, value* e *output*, pertencentes aos mecanismos de atenção presentes no codificador visual baseado em *Swin Transformer* e no decodificador textual. As configurações finais utilizadas em cada arquitetura e conjunto de dados são apresentadas na Tabela 2, contendo os valores adotados para *Rank, Alpha, Dropout, Learning Rate, Batch Size e Target Modules*.

Tabela 2 – Parâmetros utilizados por cada modelo nas bases de dados

Modelo	Target Modules	Rank; Alpha	Dropout	Taxa de aprendizado	Dataset
LayoutLMv3	query; value	8; 16	0.1	1e-4	RVL-CDIP
LayoutLMv3	query; value; key; dense	16; 32	0.1	1e-4	FUNSD
LayoutLMv3	query; value; key; dense	16; 32	0.1	1e-4	CORD
Donut	query; value	8; 16	0.1	3e-4	RVL-CDIP
Donut	query; value; key; output	16; 32	0.1	3e-4	CORD
Donut	query; value; key; output	16; 32	0.1	3e-4	FUNSD
DocFormer	All-linear	8; 16	0.1	3e-4	RVL-CDIP
DocFormer	All-linear	16; 32	0.1	2e-4	FUNSD
DocFormer	All-linear	16; 32	0.1	2e-4	CORD

Fonte: Elaborado pelo autor (2026).

3.7 Métricas

A avaliação do desempenho dos modelos foi realizada por meio das métricas Precisão, Revocação e F1-Score, amplamente utilizadas em tarefas de extração de informações e classificação de documentos (HUANG et al., 2022). A Precisão indica a proporção de predições corretas entre todas as predições realizadas pelo modelo,

sendo calculada pela razão entre o número de verdadeiros positivos e a soma de verdadeiros positivos e falsos positivos, de acordo com a Equação 3.1:

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{3.1}$$

em que VP representa os verdadeiros positivos e FP os falsos positivos.

A Revocação, também denominada Recall, mede a capacidade do modelo de identificar corretamente os elementos relevantes presentes no conjunto de referência. Essa métrica é expressa pela razão entre os verdadeiros positivos e a soma de verdadeiros positivos e falsos negativos, conforme apresentado na Equação 3.2:

$$\text{Revocação} = \frac{VP}{VP + FN} \tag{3.2}$$

em que FN representa os falsos negativos.

O F1-Score corresponde à média harmônica entre Precisão e Revocação, sendo uma métrica apropriada para avaliar o equilíbrio entre ambas, especialmente em cenários com distribuição desigual de classes (ADEGUN; VIRIRI; TAPAMO, 2023). Sua formulação é apresentada na Equação 3.3:

$$F1 = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{3.3}$$

A análise conjunta dessas medidas possibilita uma avaliação mais completa do desempenho dos modelos, considerando tanto a qualidade preditiva quanto a capacidade de recuperar corretamente as informações relevantes.

Por fim, a quantidade de parâmetros treináveis foi observada como indicador de eficiência computacional, especialmente no contexto do uso da técnica Low-Rank Adaptation (LoRA), conforme discutido por Zhang et al. (2023). Esse aspecto é relevante para avaliar a viabilidade prática da adaptação dos modelos a ambientes com recursos computacionais limitados, sem comprometer significativamente o desempenho obtido.

3.8 Fine-tuning do LayoutLMv3 com LoRA

O processo de ajuste fino do LayoutLMv3 foi realizado utilizando a biblioteca *Transformers*, em conjunto com a biblioteca PEFT (*Parameter-Efficient Fine-Tuning*), responsável pela implementação da técnica *Low-Rank Adaptation* (LoRA). Inicialmente, foi carregado o modelo pré-treinado LayoutLMv3ForSequenceClassification, disponível na biblioteca *Hugging Face*, juntamente com o processador correspondente, res-

ponsável pela normalização das imagens, tokenização do texto e geração das informações de *layout* necessárias ao modelo. Para adequar o modelo às tarefas de classificação documental, foi substituída a camada de classificação original por uma nova camada totalmente conectada (Linear), cuja dimensão de saída foi ajustada de acordo com o número de classes de cada conjunto de dados utilizado. Em seguida, todos os parâmetros do modelo foram congelados, permitindo que apenas os parâmetros inseridos pelo LoRA fossem atualizados durante o treinamento. A adaptação por LoRA foi aplicada exclusivamente às projeções de atenção do mecanismo Multi-*Head Self-Attention* do *Transformer*. Foram avaliadas diferentes configurações de módulos-alvo, contemplando adaptações apenas nas matrizes *Query* e *Value*, bem como uma configuração ampliada envolvendo *Query, Key, Value* e a projeção de saída (*Dense*). Para cada experimento foram definidas as combinações de hiperparâmetros apresentados na TABELA 2, incluindo valores para o posto (rank), fator de escala (alpha), taxa de dropout e taxa de aprendizado. Após a inserção das camadas LoRA, apenas os parâmetros pertencentes aos adaptadores permaneceram treináveis, enquanto os pesos originais do modelo permaneceram congelados. Essa estratégia reduz significativamente a quantidade de parâmetros atualizados durante o treinamento, diminuindo o consumo de memória e o custo computacional do processo de ajuste fino. O treinamento foi conduzido utilizando o algoritmo AdamW, com taxa de aprendizado definida conforme cada experimento, batch size igual a um documento por dispositivo e acumulação de gradientes para simular lotes maiores sem exceder a memória disponível da GPU. Foram executadas doze épocas de treinamento, realizando-se avaliação ao término de cada época e armazenando automaticamente o modelo com melhor desempenho segundo a métrica F1-score. Ao final de cada época, o modelo foi avaliado utilizando as métricas de Precisão *(Precision*) e Revocação (*Recall)*, F1-score, possibilitando comparar o impacto das diferentes configurações de LoRA sobre o desempenho do LayoutLMv3 nos conjuntos RVL-CDIP, FUNSD e CORD.

3.9 Fine-tuning do DocFormer

O DocFormer foi utilizado como um dos modelos base para os experimentos devido à sua arquitetura multimodal, capaz de integrar simultaneamente informações textuais, visuais e espaciais presentes em documentos digitalizados. Diferentemente de arquiteturas que processam apenas texto ou imagem, o DocFormer combina embeddings textuais provenientes do tokenizer, características visuais extraídas da imagem do documento e informações de posicionamento (*bounding boxes*), permitindo representar de forma conjunta o conteúdo e o layout da página. O modelo original foi desenvolvido para tarefas de compreensão visual de documentos e avaliado em diferentes bases de dados, incluindo RVL-CDIP, FUNSD e CORD, realizando ajuste fino

(*fine-tuning*) após a etapa de pré-treinamento para adaptação às características específicas de cada tarefa (APPALARAJU et al., 2021). Dessa forma, neste trabalho, o *fine-tuning* também foi aplicado, pois apesar do modelo possuir representações multimodais previamente aprendidas, as bases avaliadas apresentam diferentes domínios e objetivos. O RVL-CDIP possui foco em classificação documental, exigindo que o modelo adapte suas representações visuais e estruturais para distinguir diferentes categorias de documentos. Já as bases FUNSD e CORD envolvem tarefas de compreensão documental, nas quais é necessário capturar relações entre texto, posição espacial e elementos visuais, tornando necessária uma adaptação mais específica ao formato dos documentos. Para reduzir o custo computacional desse processo, foi empregada a técnica *Low-Rank Adaptation* (LoRA), permitindo ajustar o modelo sem atualizar todos os seus parâmetros. No DocFormer, a adaptação foi aplicada às projeções lineares do modelo (*all-linear*), mantendo os pesos originais congelados e atualizando apenas matrizes de baixa dimensão. Essa estratégia possibilita uma adaptação eficiente às características das diferentes bases de dados, reduzindo a quantidade de parâmetros treináveis e o consumo de memória em comparação ao *fine-tuning* completo da arquitetura.

3.10 Fine-Tuning do Donut utilizando LoRA

Foi utilizado como modelo base o Donut-base, disponibilizado pela NAVER Clova AI no *Hugging Face*, com aproximadamente 201 milhões de parâmetros. Sua arquitetura combina um encoder visual baseado em *Swin Transformer* e um decoder autoregressivo, permitindo a compreensão de documentos sem a necessidade de OCR. O modelo e o processador DonutProcessor foram carregados com resolução de entrada de 960 x 720 pixels, definida após experimentos preliminares por apresentar melhor preservação dos detalhes dos documentos da base CORD. Em seguida, foram configurados os tokens especiais da tarefa, incluindo o token de início. Para o ajuste fino, foi aplicada a técnica LoRA nas matrizes de projeção Query e Value das camadas de atenção do *decoder Transformer*, adicionando aproximadamente 524 mil parâmetros treináveis. Essa abordagem foi adotada devido ao elevado número de parâmetros do Donut, permitindo adaptar o modelo ao domínio dos documentos avaliados sem a necessidade de atualizar todos os seus pesos, reduzindo o custo computacional e o risco de sobreajuste. Os experimentos utilizaram as bases CORD-v2, FUNSD e RVL-CDIP, sendo esta última utilizada em sua versão reduzida (rvl-cdip-small). O pré-processamento envolveu a conversão das imagens para RGB, redimensionamento para a resolução definida e transformação das anotações estruturadas em sequências textuais no formato JSON esperado pelo *decoder*. As sequências foram tokenizadas com tamanho máximo de 256 tokens, aplicando truncamento quando necessário e

substituindo tokens de preenchimento por -100 para evitar sua influência no cálculo da função de perda. Devido à arquitetura *encoder-decoder*, foi implementado um *Data-Loader* personalizado para agrupar as imagens e os rótulos utilizados no treinamento supervisionado. Após o *fine-tuning*, o modelo foi avaliado nos conjuntos de teste, utilizando o método generate() para produzir a saída textual correspondente aos documentos analisados.

4 Resultados e dicussão

Os experimentos realizados tiveram como objetivo comparar o desempenho dos modelos LayoutLMv3, Donut e DocFormer em tarefas de compreensão visual de documentos, avaliando tanto o comportamento dos modelos originais quanto suas versões adaptadas com a técnica Low-Rank Adaptation (LoRA). As avaliações foram conduzidas utilizando as métricas Precisão, Revocação e F1-score, permitindo analisar simultaneamente a capacidade de identificação correta das informações e a robustez geral dos modelos em diferentes tipos de documentos.

4.1 Comparação dos Modelos Base

A Tabela 3 apresenta os resultados de F1-score obtidos pelos modelos em suas configurações originais para os conjuntos RVL-CDIP, FUNSD e CORD. A delimitação analítica desta seção inicial exclusivamente à métrica de F1-score justifica-se pela necessidade de padronização reportados nos artigos originais de proposição das arquiteturas avaliadas, que utilizam essa métrica como principal *benchmark* .

Tabela 3 – F1-Score dos modelos padrão nas bases de dados avaliadas

Modelo	Parâmetros	RVL-CDIP	FUNSD	CORD
LayoutLMv3	127 M	91,00%	88,87%	79,47%
Donut	201 M	89,97%	85,27%	90,92%
DocFormer	183 M	91,89%	90,70%	80,39%

Fonte: Elaborado pelo autor (2026).

Observa-se que os três modelos apresentaram desempenho semelhante nos conjuntos RVL-CDIP e FUNSD, com valores de F1 superiores a 88%. O melhor resultado foi obtido pelo Docformer no conjunto RVL-CDIP, alcançando a F1 de 91,89%. No conjunto CORD, que apresenta maior variabilidade estrutural nos documentos, todos os modelos sofreram redução de desempenho. Dessa forma o Donut se mostrou acima ao DocFormer e LayoutLMv3 que mantiveram resultados próximos, com F1 de aproximadamente 80% ou menos. Os resultados indicam que os três modelos são capazes de lidar adequadamente com documentos de diferentes formatos, embora arquiteturas multimodais mais recentes apresentam maior robustez em cenários de extração de informações complexas.

4.2 Avaliação das Métricas de Precisão, Recall e F1

As TABELAS 4,5 e 6 apresentam uma comparação detalhada das métricas de precisão, revocação e F1 para os modelos em suas configurações originais, para cada uma das bases de dados usadas.

Tabela 4 – Precisão, Revocação e F1-Score dos modelos no dataset RVL-CDIP

Modelo	Precisão	Revocação	F1-Score
LayoutLMv3	92,02%	90,01%	91,00%
Donut	90,03%	89,91%	89,97%
DocFormer	90,73%	93,09%	91,89%

Fonte: Elaborado pelo autor (2026).

Tabela 5 – Precisão, Revocação e F1-Score dos modelos no dataset FUNSD

Modelo	Precisão	Revocação	F1-Score
LayoutLMv3	89,45%	88,34%	88,87%
Donut	84,46%	86,11%	85,27%
DocFormer	91,57%	89,84%	90,70%

Fonte: Elaborado pelo autor (2026).

Tabela 6 – Precisão, Revocação e F1-Score dos modelos no dataset CORD

Modelo	Precisão	Revocação	F1-Score
LayoutLMv3	82,14%	79,47%	80,77%
Donut	89,75%	92,13%	90,92%
DocFormer	79,62%	81,18%	80,39%

Fonte: Elaborado pelo autor (2026).

De forma geral, os resultados demonstram equilíbrio entre precisão e revocação para os três modelos. O DocFormer apresentou a maior revocação no conjunto RVL-CDIP (93,09%), indicando maior capacidade de recuperação das informações relevantes. Em contrapartida, o LayoutLMv3 apresentou comportamento mais equilibrado entre as três métricas, resultando em maior estabilidade entre os diferentes conjuntos avaliados. O Donut destacou-se por manter simultaneamente elevados valores de Precisão e *Recall* em CORD, evidenciando sua eficiência em tarefas de entendimento documental sem dependência explícita de OCR, onde os ruídos das imagens de exemplo no CORD, dificultaram para os outros modelos.

4.3 Resultados do Fine-tuning completo

A Tabela 7 apresenta os resultados obtidos pelos modelos LayoutLMv3, DocFormer e Donut após o processo de fine-tuning completo, considerando as métricas de Precisão, Revocação, F1-Score e o tempo aproximado de treinamento para cada conjunto de dados. A análise desta seção tem como objetivo estabelecer uma linha de base para posterior comparação com a adaptação utilizando a técnica *Low-Rank Adaptation* (LoRA). Observa-se que os três modelos apresentaram desempenho elevado em todos os conjuntos de dados, alcançando valores de F1-Score superiores a 80%. Esses resultados confirmam a capacidade das arquiteturas multimodais em aprender simultaneamente características textuais, visuais e espaciais durante o processo de ajuste fino. No conjunto RVL-CDIP, destinado à classificação documental, o LayoutLMv3 obteve o melhor desempenho, registrando 91,41% de F1-Score, além do menor tempo de treinamento entre os modelos avaliados (10 horas e 35 minutos). O DocFormer apresentou desempenho próximo, com 90,34%, demandando aproximadamente 11 horas e 14 minutos de treinamento. Já o Donut alcançou 89,77% de F1-Score, porém exigiu o maior tempo computacional (11 horas e 56 minutos), refletindo a maior complexidade de sua arquitetura baseada em *encoder-decoder*. No conjunto FUNSD, voltado para tarefas de compreensão de formulários e extração de entidades, o DocFormer apresentou o melhor resultado, alcançando 91,29% de F1-Score, seguido pelo LayoutLMv3, com 88,76%, e pelo Donut, com 88,44%. Apesar das diferenças relativamente pequenas entre os modelos, observa-se que o DocFormer explorou de forma mais eficiente as relações espaciais presentes nesse tipo de documento, característica esperada devido ao seu mecanismo de atenção multimodal desacoplada. No conjunto CORD, composto por recibos comerciais com elevada variabilidade visual, o Donut destacou-se ao atingir 89,68% de F1-Score, superando o DocFormer (82,58%) e o LayoutLMv3 (81,63%). Esse comportamento pode ser atribuído à arquitetura OCR-free, que processa diretamente a imagem do documento e reduz a propagação de erros provenientes do reconhecimento óptico de caracteres, tornando-se mais robusta diante de documentos com ruídos, baixa qualidade ou layouts pouco padronizados. Em relação ao custo computacional, observa-se uma tendência consistente entre as arquiteturas avaliadas. O LayoutLMv3 apresentou os menores tempos de treinamento, variando entre aproximadamente 10 horas e 35 minutos e 10 horas e 48 minutos, seguido pelo DocFormer, com tempos entre 11 horas e 9 minutos e 11 horas e 27 minutos. O Donut foi a arquitetura que demandou maior tempo de treinamento, permanecendo cerca de 12 horas em todos os experimentos. Esse comportamento é coerente com sua arquitetura *encoder-decoder* baseada em *Swin Transformer* e com a utilização de resoluções de entrada superiores às empregadas pelos demais modelos, fatores que aumentam a quantidade de operações realizadas durante o treinamento. De forma geral, os resultados demonstram que o LayoutLMv3 apresentou o melhor equilíbrio

entre desempenho e eficiência computacional, alcançando as maiores métricas em tarefas de classificação documental com o menor tempo de treinamento. O DocFormer destacou-se na compreensão de formulários do conjunto FUNSD, enquanto o Donut obteve os melhores resultados no conjunto CORD, evidenciando que sua abordagem OCR-free oferece maior robustez para documentos comerciais com maior variabilidade visual. Esses resultados constituem a linha de base para a comparação apresentada na seção seguinte, na qual é avaliado o impacto da aplicação da técnica *Low-Rank Adaptation* (LoRA) sobre o desempenho e a eficiência computacional dessas arquiteturas.

Tabela 7 – Resultados dos modelos submetidos ao fine-tuning completo

Modelo	Precisão	Revocação	F1-Score	Dataset	Tempo de treino
LayoutLMv3	92,02%	90,01%	91,41%	RVL-CDIP	10h 35min
LayoutLMv3	86,41%	91,24%	88,76%	FUNSD	10h 48min
LayoutLMv3	80,56%	82,73%	81,63%	CORD	10h 41min
Donut	91,74%	87,89%	89,77%	RVL-CDIP	11h 56min
Donut	88,04%	87,80%	87,92%	CORD	11h 42min
Donut	84,53%	92,74%	88,44%	FUNSD	11h 49min
DocFormer	91,18%	89,52%	90,34%	RVL-CDIP	11h 14min
DocFormer	91,89%	90,70%	91,29%	FUNSD	11h 27min
DocFormer	82,11%	83,05%	82,58%	CORD	11h 09min

Fonte: Elaborado pelo autor (2026).

4.4 Impacto da Aplicação de LoRA

A aplicação do LoRA permitiu reduzir significativamente a quantidade de parâmetros treináveis sem necessidade de atualização completa dos modelos, como mostra a TABELA 8.

Tabela 8 – Parâmetros treináveis dos modelos com LoRA nos datasets

Modelo	RVL-CDIP	FUNSD	CORD
LayoutLMv3	1,192 M	2,66 M	2,66 M
Donut	0,262 M	0,544 M	0,544 M
DocFormer	1,45 M	2,90 M	2,90 M

Fonte: Elaborado pelo autor (2026).

Observa-se que o Donut apresentou a menor quantidade de parâmetros ajustáveis, com apenas 262 mil parâmetros treináveis em sua configuração mais leve, enquanto o LayoutLMv3 e o DocFormer exigiram aproximadamente 1,19 milhão e 1,45

milhão de parâmetros, respectivamente. Essa redução representa uma diminuição expressiva em relação ao número total de parâmetros dos modelos, que variam entre 127 milhões e 201 milhões de acordo como reportado nos trabalhos originais e testes realizados durante este trabalho, evidenciando uma das principais vantagens do LoRA para cenários com recursos computacionais limitados. A TABELA 9 apresenta os resultados detalhados dos modelos após aplicação do LoRA em cada conjunto de dados.

Tabela 9 – Resultados dos modelos utilizando LoRA

Modelo	Precisão	Revocação	F1-Score	Dataset	Tempo de treino
LayoutLMv3	91,37%	90,03%	90,69%	RVL-CDIP	31 min
LayoutLMv3	84,21%	89,16%	86,62%	FUNSD	24 min
LayoutLMv3	78,07%	80,84%	79,43%	CORD	22 min
Donut	90,22%	85,97%	88,00%	RVL-CDIP	49 min
Donut	88,04%	87,80%	87,92%	CORD	45 min
Donut	82,02%	91,30%	86,40%	FUNSD	47 min
DocFormer	89,72%	87,72%	88,71%	RVL-CDIP	38 min
DocFormer	91,13%	89,74%	90,48%	FUNSD	34 min
DocFormer	79,95%	81,13%	80,54%	CORD	36 min

Fonte: Elaborado pelo autor (2026).

Os melhores resultados com LoRA foram obtidos pelo LayoutLMv3 no conjunto RVL-CDIP, com F1 de 90,69%. No conjunto FUNSD, o DocFormer apresentou desempenho superior, com F1 de 90,48%, enquanto no conjunto CORD os modelos baseados em OCR apresentaram resultados semelhantes, variando entre 79% e 81%, destacando-se nesse cenário o Donut. Esses resultados demonstram que o LoRA apesar de diminuir a capacidade de generalização dos modelos,consegue diminuir o tempo de treino em até pelo menos 10 vezes.

4.5 Comparação entre os desempenhos em relação ao LoRA

A Figura 8 apresenta a diferença de desempenho entre os modelos originais e suas respectivas versões ajustadas com LoRA.

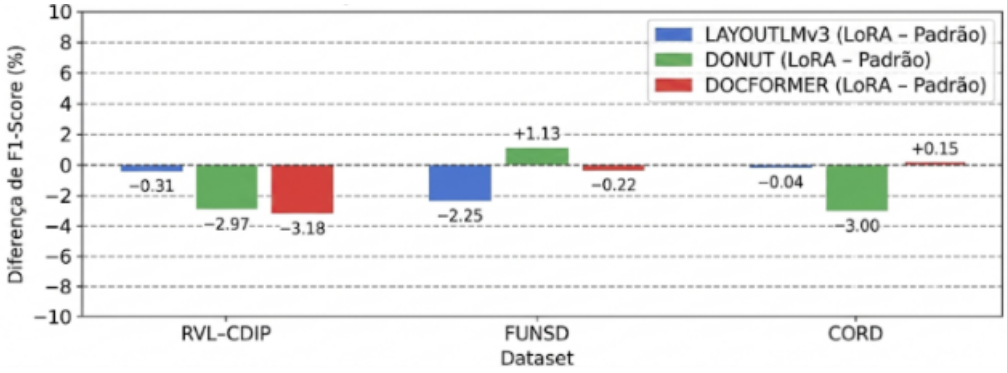


Figura 8 – Precisão, Recall e F1-Score (Modelo Padrão)

Fonte: Elaborado pelo autor (2026).

Os resultados mostram que a utilização do LoRA produziu reduções relativamente pequenas no desempenho quando comparada ao ajuste completo dos modelos. Em diversos cenários, a diferença observada foi inferior a 3 pontos percentuais em F1. O comportamento mais consistente foi observado no LayoutLMv3, cuja perda média permaneceu inferior a 3,5 pontos percentuais. Resultados semelhantes foram identificados para Donut e DocFormer, demonstrando que a adaptação por baixa decomposição de posto preserva grande parte do desempenho original. Esses resultados corroboram para apontar o LoRA como uma alternativa eficiente para ajuste fino de grandes modelos multimodais, especialmente em ambientes com limitações de hardware, a seguir na Tabela 2 os melhores resultados foram compilados com os junto com os parâmetros escolhidos. A Tabela 9 apresenta o tempo aproximado de treinamento dos modelos considerando ambas as estratégias. Observa-se que a utilização do LoRA reduziu expressivamente o tempo necessário para a adaptação das arquiteturas em todos os conjuntos de dados avaliados. Enquanto o fine-tuning completo demandou aproximadamente 10 horas e 35 minutos até 11 horas e 56 minutos, a adaptação utilizando LoRA reduziu esse tempo para um intervalo entre 22 e 49 minutos, dependendo da arquitetura e do conjunto de dados utilizado. O LayoutLMv3 apresentou os menores tempos de treinamento, variando entre 22 e 31 minutos, seguido pelo DocFormer, com tempos entre 34 e 38 minutos. O Donut permaneceu como a arquitetura de maior custo computacional, necessitando entre 45 e 49 minutos, comportamento esperado em função de sua arquitetura encoder-decoder baseada em *Swin Transformer* e da utilização de imagens em resoluções superiores às empregadas pelos demais modelos. Essa redução representa uma economia superior a 90% no tempo de treinamento, tornando o processo de adaptação significativamente mais viável para ambientes com recursos computacionais limitados, como os utilizados nesta pesquisa.

Tabela 10 – Comparação entre os tempos de treinamento dos modelos

Modelo	Fine-tuning completo	LoRA	Redução aproximada
LayoutLMv3	10h35–10h48	22–31 min	$\approx$ 95%
Donut	11h09–11h27	34–38 min	$\approx$ 94%
DocFormer	11h42–11h56	45–49 min	$\approx$ 93%

Fonte: Elaborado pelo autor (2026).

Embora o ganho em tempo seja expressivo, ele decorre principalmente da redução da quantidade de parâmetros que precisam ser atualizados durante o treinamento. A Tabela 8 apresenta a quantidade de parâmetros treináveis utilizada em cada arquitetura após a aplicação do LoRA. Enquanto nas Tabela 10 e 11 são apresentadas as diferença entre os f*ine-tuning* usados na pesquisa na métrica de *benchmark* f1-score e quantiadade de parâmetros de treinamento. Observa-se que o LayoutLMv3, originalmente composto por aproximadamente 127 milhões de parâmetros, passou a atualizar entre 1,19 milhão e 2,66 milhões de parâmetros, o que representa aproximadamente 0,94% a 2,09% do modelo original. O DocFormer, com cerca de 183 milhões de parâmetros, utilizou entre 1,45 milhão e 2,90 milhões de parâmetros treináveis, correspondendo a aproximadamente 0,79% a 1,58% do total da arquitetura. A maior redução foi observada no Donut, cuja arquitetura possui aproximadamente 201 milhões de parâmetros. Após a aplicação do LoRA, apenas 262 mil parâmetros foram atualizados no conjunto RVL-CDIP, e 544 mil parâmetros nos conjuntos FUNSD e CORD, correspondendo a apenas 0,13% e 0,27% do número total de parâmetros do modelo. Esse resultado evidencia a elevada eficiência da técnica, uma vez que praticamente toda a rede permaneceu congelada durante o treinamento. É importante destacar que essa redução não significa que o modelo se tornou menor ou que houve compressão da arquitetura. O LoRA mantém todos os pesos do modelo pré-treinado inalterados e introduz apenas matrizes adicionais de baixa dimensionalidade nas projeções lineares selecionadas. Dessa forma, apenas uma pequena fração dos parâmetros passa a ser otimizada durante o ajuste fino, reduzindo significativamente o consumo de memória durante a retropropagação dos gradientes e acelerando o treinamento. Mesmo com essa redução nas métricas de eficiência a diminuição expressiva dos parâmetros treináveis e do tempo de treinamento, os resultados apresentados na seção anterior demonstram que as perdas de desempenho permaneceram inferiores a aproximadamente 3% em relação ao fine-tuning completo. Esse comportamento confirma a hipótese de que a adaptação eficiente promovida pelo LoRA é capaz de produzir generalização das arquiteturas multimodais, tornando sua utilização particularmente atrativa em cenários com restrições de infraestrutura computacional. De forma geral, os experimentos evidenciam que o LoRA constitui uma alternativa ao fine-tuning tradicional, reduzindo drasticamente o custo computacional do treinamento em troca de redução

acentuada do capacidade computacional durante o treino do modelo.

Tabela 11 – Comparação da redução da quantidade de parâmetros treináveis

Modelo	Fine-tuning completo	LoRA (menor)	LoRA (maior)	Redução aproximada
LayoutLMv3	127 M	1,192 M	2,66 M	99,06% a 97,91%
Donut	183 M	1,45 M	2,90 M	99,21% a 98,42%
DocFormer	201 M	0,262 M	0,544 M	99,87% a 99,73%

Fonte: Elaborado pelo autor (2026).

Tabela 12 – Comparação dos ganhos de F1-Score entre fine-tuning completo e LoRA

Modelo	$\triangle$ Base $\rightarrow$ FT (F1-Score)	$\triangle$ FT $\rightarrow$ LoRA (F1-Score)	Dataset
LayoutLMv3	+0,41%	-0,72%	RVL-CDIP
LayoutLMv3	-0,11%	-2,14%	FUNSD
LayoutLMv3	+0,86%	-2,20%	CORD
Donut	-0,20%	-1,77%	RVL-CDIP
Donut	+3,17%	-2,04%	CORD
Donut	-3,00%	+0,04%	FUNSD
DocFormer	-1,55%	-1,63%	RVL-CDIP
DocFormer	+0,59%	-0,81%	FUNSD
DocFormer	+2,19%	-2,04%	CORD

Fonte: Elaborado pelo autor (2026).

4.6 Análise visual das saídas dos modelos

Esta seção apresenta uma comparação das saídas geradas pelos modelos LayoutLMv3, DocFormer e Donut a partir de uma mesma imagem de recibo comercial pertencente ao conjunto de dados CORD. A utilização do mesmo documento como entrada para os três modelos permite uma análise visual direta do comportamento de cada arquitetura diante de uma tarefa de extração de informações estruturadas. As Figuras 9, 10 e 11 apresentam as saídas produzidas pelos modelos para o mesmo recibo. Observa-se que o LayoutLMv3 e o DocFormer apresentaram maior dificuldade na identificação e organização das informações relevantes do documento, resultando em saídas menos completas ou com menor aderência à estrutura esperada. Essas dificuldades podem estar relacionadas à complexidade do layout do recibo e à necessidade de integração precisa entre informações textuais, visuais e espaciais.

Figura 9 – Resultado do modelo Donut

Fonte: Elaborado pelo autor (2026).

Figura 10 – Resultado do modelo LayoutLMv3

Fonte: Elaborado pelo autor (2026).

Figura 11 – Resultado do modelo Docformer

Fonte: Elaborado pelo autor (2026).

Em contraste, o Donut demonstrou melhor desempenho nesse exemplo, gerando uma saída mais próxima da estrutura de informações esperada para o documento. Por utilizar uma abordagem OCR-free, o modelo processa diretamente a imagem do recibo e produz uma representação textual estruturada, o que favoreceu a

extração de campos relevantes presentes no documento analisado, mesmo sem o elemento da coordenada espacial do texto. Essa comparação reforça os resultados quantitativos apresentados anteriormente, evidenciando que o Donut apresentou maior capacidade de lidar com a estrutura visual do recibo selecionado, enquanto LayoutLMv3 e DocFormer demonstraram limitações na extração das informações contidas na mesma imagem do conjunto CORD.

5 Considerações finais

Este trabalho apresentou uma análise comparativa entre os modelos de compreensão visual de documentos LayoutLMv3, DocFormer e Donut, investigando o impacto da aplicação da técnica Low-Rank Adaptation (LoRA) no processo de ajuste fino. Os experimentos foram realizados utilizando os conjuntos de dados RVL-CDIP, FUNSD e CORD, contemplando tarefas de classificação documental e extração de informações, com avaliação baseada nas métricas de Precisão, Revocação e F1-Score. Os resultados demonstraram que os três modelos apresentaram desempenho competitivo nas tarefas avaliadas, alcançando valores superiores a 80% de F1-Score em todos os conjuntos de dados. Entre eles, o LayoutLMv3 obteve os melhores resultados gerais no conjunto RVL-CDIP, enquanto o DocFormer apresentou melhor desempenho no FUNSD. Já o Donut destacou-se pela robustez em documentos comerciais presentes no CORD, evidenciando as vantagens da abordagem OCR-free em cenários com maior variabilidade visual. A principal contribuição deste trabalho foi demonstrar que a aplicação do LoRA reduz substancialmente o custo computacional do ajuste fino. Nos experimentos realizados, o tempo de treinamento foi reduzido de aproximadamente 11 horas para um intervalo entre 22 e 49 minutos, além de diminuir a quantidade de parâmetros treináveis de centenas de milhões para poucas centenas de milhares ou poucos milhões de parâmetros, dependendo da arquitetura e do conjunto de dados utilizado. Entretanto, os resultados também evidenciaram uma limitação importante da técnica. Embora o LoRA tenha proporcionado ganhos expressivos em eficiência computacional, não foram observados ganhos de desempenho em relação ao fine-tuning completo. Em todos os experimentos, os modelos adaptados apresentaram desempenho igual ou ligeiramente inferior ao treinamento convencional, com perdas de F1-Score variando entre 0,72 e 2,20 pontos percentuais, permanecendo dentro de uma margem relativamente pequena quando comparadas ao ganho obtido em tempo de treinamento e redução dos parâmetros atualizados. Outro aspecto observado foi que o aumento da quantidade de parâmetros treináveis ou a utilização de configurações mais complexas de LoRA não resultaram, necessariamente, em melhor desempenho. Em diversos experimentos, configurações mais simples produziram resultados semelhantes aos obtidos com adaptações mais extensas, indicando que a escolha adequada dos módulos-alvo e dos hiperparâmetros exerce maior influência na qualidade do ajuste fino do que simplesmente aumentar a capacidade de adaptação do modelo. Dessa forma, conclui-se que o LoRA não substitui o fine-tuning completo quando o objetivo é obter o maior desempenho possível, mas constitui uma alternativa eficiente quando existem restrições de infraestrutura computacional. Nesses cenários, a técnica permite preservar a

maior parte da capacidade preditiva dos modelos enquanto reduz significativamente o tempo de treinamento, sendo alcançada redução de até 99,87Por fim, este trabalho contribui para o entendimento do comportamento da técnica LoRA em diferentes arquiteturas multimodais de compreensão visual de documentos, evidenciando que sua adoção deve considerar o equilíbrio entre desempenho e custo computacional. Em aplicações nas quais pequenas reduções de desempenho são aceitáveis frente aos ganhos de eficiência, o LoRA apresenta-se como uma solução prática e viável para a adaptação de modelos de grande porte.

6 Conclusão

A realização dos experimentos permitiu observar que o desempenho das arquiteturas multimodais varia de acordo com o tipo de documento analisado. Embora os três modelos tenham apresentado resultados superiores a 80% de F1-Score, cada arquitetura demonstrou maior adequação para cenários específicos. O LayoutLMv3 destacou-se em tarefas de classificação documental, o DocFormer apresentou melhor desempenho em formulários estruturados e o Donut mostrou maior robustez em documentos comerciais contendo ruídos visuais, em função de sua abordagem OCR-free. Os experimentos também evidenciaram que o desempenho dos modelos depende não apenas da arquitetura empregada, mas também das configurações adotadas durante o ajuste fino. A escolha dos módulos adaptados pelo LoRA, bem como dos hiperparâmetros de rank e alpha, influenciou diretamente os resultados obtidos. Em diversos cenários, configurações mais simples produziram desempenho semelhante ao de configurações mais complexas, indicando que o aumento do número de parâmetros treináveis não implica, necessariamente, melhor capacidade preditiva. Outro aspecto observado refere-se ao comportamento da técnica LoRA. Os resultados mostraram que seu principal benefício está relacionado à eficiência computacional. A redução do número de parâmetros treináveis permitiu diminuir significativamente o tempo de treinamento e o consumo de memória, tornando viável a adaptação de modelos multimodais em ambientes com recursos computacionais limitados. Entretanto, essa economia não se refletiu em ganhos de desempenho. Comparado ao fine-tuning completo, o LoRA preservou grande parte da capacidade preditiva, mas apresentou pequenas perdas nas métricas avaliadas, evidenciando um compromisso entre desempenho e custo computacional. Esses resultados reforçam que a escolha entre fine-tuning completo e LoRA deve considerar o objetivo da aplicação. Quando a prioridade é obter o maior desempenho possível, o treinamento convencional continua sendo a alternativa mais adequada. Por outro lado, quando existem restrições de infraestrutura ou necessidade de reduzir tempo de treinamento, o LoRA apresenta-se como uma solução eficiente, mantendo desempenho competitivo com custo computacional significativamente inferior.

7 Trabalhos futuros

Como trabalhos futuros, sugere-se a ampliação dos experimentos para outras bases de dados de documentos com maior diversidade estrutural e linguística, bem como a investigação de técnicas mais recentes de Parameter-Efficient Fine-Tuning (PEFT), incluindo QLoRA e AdaLoRA. Também podem ser exploradas estratégias de compressão de modelos, quantização e destilação de conhecimento, visando possibilitar a implantação dessas soluções em dispositivos com restrições ainda mais severas de processamento e memória.

Referências

ADEGUN, A. A.; VIRIRI, S.; TAPAMO, J. R. Review of deep learning methods for remote sensing satellite images classification: experimental survey and comparative analysis. *Journal of Big Data*, Springer, v. 10, p. 93, 2023.

APPALARAJU, S. et al. *DocFormer: End-to-End Transformer for Document Understanding*. [S.l.], 2021.

BABU, P. V. et al. Extraction of text from images using document understanding transformer (donut). In: *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*. [S.l.]: IEEE, 2024.

CHANTI, S. et al. Documentqa-leveraging layoutlmv3 for next-level question answering. In: *2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*. Erode, India: IEEE, 2023. p. 1–8.

DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR)*. [S.l.: s.n.], 2021.

HU, E. J. et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2022.

HUANG, Y. et al. Layoutlmv3: Pre-training for document ai with unified text and image masking. In: *Proceedings of the 30th ACM International Conference on Multimedia*. Lisboa, Portugal: ACM, 2022.

KIM, G. et al. Ocr-free document understanding transformer. In: *European Conference on Computer Vision (ECCV)*. [S.l.: s.n.], 2022.

KV, J.; MONDAL, A.; JAWAHAR, C. V. Document image analysis using deep multi-modular features. *SN Computer Science*, Springer, v. 4, 2022.

MAHADEVKAR, S. V. et al. Exploring ai-driven approaches for unstructured document analysis and future horizons. *Journal of Big Data*, Springer, v. 11, p. 92, 2024.

TAY, Y. et al. Efficient transformers: A survey. *ACM Computing Surveys*, 2022.

ZHANG, L. et al. Lora-fa: Efficient and effective low rank representation fine-tuning. *arXiv preprint*, 2023.