Código: PPGEC-DOUTORADO_2026_2_BLDB1

Universidade de Pernambuco
Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Tese de Doutorado

Área: Inteligência Computacional

Título: Algoritmos Multimodais e Raciocínio para Document Visual Question Answering

Orientador – Byron Leite Dantas Bezerra (byron.leite@upe.br)

Descrição

O problema de **Document Visual Question Answering (DocVQA)** consiste em responder perguntas em linguagem natural sobre imagens de documentos, coleções de documentos ou infográficos, combinando leitura de texto (OCR), compreensão de layout e interpretação visual/numérica. Competições promovidas no ICDAR [1] propõem geralmente três tarefas principais no contexto de DocVQA: *Single Document VQA*, *Document Collection VQA* e *Infographics VQA*. Através destas competições os pesquisadores tem acesso a datasets que evidenciam tanto o progresso quanto as lacunas atuais na área. Esses recursos e resultados mostram que, embora abordagens extrativas baseadas em OCR + modelos de linguagem atinjam desempenho razoável em respostas que aparecem textualmente no documento, há desafios a serem superados em questões que exigem interpretação visual, operações aritméticas, multi-span answers e evidência em coleções.

Os desafios centrais que justificam investigação aprofundada incluem: **(i)** integração robusta entre OCR (com ruído e variação de domínio) e representações multimodais; **(ii)** mecanismos de raciocínio simbólico e numérico (contagem, soma, comparação) acoplados a modelos neurais; **(iii)** tratamento de respostas não-extrativas (non-span, multi-span) e recuperação de evidências em coleções; e **(iv)** explicabilidade e avaliação fina além de métricas de similaridade de string. A evolução observada nas últimas competições aponta para a eficácia de arquiteturas multimodais pré-treinadas (por exemplo, modelos que combinam texto, layout e visão) em superar pipelines puramente textuais, mas também revela quedas de desempenho em categorias que exigem raciocínio e operações sobre elementos visuais e tabulares [2].

Este projeto de doutorado consiste em investigar **novos algoritmos multimodais** e estratégias de pré-treinamento e fine-tuning para DocVQA, com três linhas possíveis a explorar: **(A)** desenvolvimento de módulos híbridos que combinam transformadores multimodais com componentes simbólicos para operações aritméticas e lógica discreta; **(B)** técnicas de fusão robusta entre OCR incerto e sinais visuais (features de layout, ícones, gráficos) incluindo aprendizado contrafactual e adaptação de domínio; **(C)** métodos de recuperação e justificação de evidências em coleções, integrando recuperação densa e re-rankers contextuais. Ainda como desdobramento, a pesquisa poderá avaliar o impacto de pré-treinamentos específicos para infográficos, estratégias de data augmentation sintético para gráficos/tabelas [3], anotações manuscritas [4], e métricas que capturem capacidade de raciocínio, propondo benchmarks e protocolos experimentais para medir generalização e interpretabilidade.

A proposta envolve uma equipe multidisciplinar e faz parte do projeto de pesquisa e inovação "***Algoritmos e Modelos de Inteligência Artificial e Visão Computacional para Processamento Inteligente de Documentos***" fomentado pelo CNPQ, e em parceria com a empresa Di2Win (www.di2win.com). Para conhecer mais sobre o orientador e seus temas de pesquisa, convido a assistir a entrevista aqui.

Referências Bibliográficas

1. R. Tito et al., *ICDAR 2021 Competition on Document Visual Question Answering*, arXiv:2111.05547.

2. Powalski, R., Borchmann, L., Jurkiewicz, D., Dwojak, T., Pietruszka, M., Palka, G.: Going full-tilt boogie on document understanding with text-image-layout transformer. arXiv preprint arXiv:2102.09550 (2021).
3. Alhassan Mumuni, Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. Array, 16 (2022). https://doi.org/10.1016/j.array.2022.100258.
4. de Sousa Neto, A.F., Bezerra, B.L.D., de Moura, G.C.D. *et al.* Data Augmentation for Offline Handwritten Text Recognition: A Systematic Literature Review. *SN COMPUT. SCI.* **5**, 258 (2024). https://doi.org/10.1007/s42979-023-02583-6 .