MAPEAMENTO DE BOAS PRÁTICAS E FERRAMENTAS LIVRES PARA A CONSTRUÇÃO DE DATASETS DE IMAGENS COM FOCO EM APRENDIZADO SUPERVISIONADO

Victor Gabriel Leite da Costa (PIC/UEM), Marcelo de Gomensoro Malheiros (Orientador). E-mail: ra133070@uem.br

Universidade Estadual de Maringá, Departamento de Informática, Maringá, PR.

Área e subárea do conhecimento: Ciência da Computação / Metodologia e Técnicas da Computação.

Palavras-chave: Qualidade de dados; Ciência Aberta; Reprodutibilidade.

RESUMO

A construção de datasets de qualidade é uma etapa fundamental para o desenvolvimento de aplicações de Aprendizado de Máquina e para a reprodução de resultados científicos. Este trabalho tem como objetivo mapear boas práticas, padrões de anotação e ferramentas livres para a construção e organização de datasets públicos de imagens voltados ao aprendizado supervisionado. Foi realizada uma revisão da literatura utilizando termos relacionados à qualidade de datasets, dados abertos, boas práticas e anotação de imagens. Dos 107 trabalhos inicialmente recuperados, 14 foram selecionados para análise aprofundada. A partir do levantamento, as práticas identificadas foram organizadas em etapas relacionadas à preparação e organização dos dados, anotação, verificação de qualidade, curadoria e documentação. Como estudo de caso, algumas dessas práticas foram aplicadas ao Brazilian Coin Detection Dataset. A análise de sua composição identificou desbalanceamento entre as classes monetárias, com aproximadamente 8,7 vezes mais instâncias da classe mais frequente em relação à menos frequente. A sistematização proposta fornece uma visão geral das principais práticas que podem auxiliar pesquisadores na construção e manutenção de datasets de imagens.

INTRODUÇÃO

A disponibilidade de dados de qualidade é fundamental para o desenvolvimento e a avaliação de métodos de Aprendizado de Máquina. Em problemas supervisionados, a construção de um dataset envolve não apenas a obtenção das amostras, mas também sua organização, anotação, validação e documentação. Essas características influenciam tanto a utilização dos dados quanto sua posterior reutilização por outros pesquisadores.

No contexto de imagens, diferentes tarefas demandam formas específicas de anotação e ferramentas adequadas para produzi-las. Além disso, problemas relacionados à qualidade, como inconsistências, distribuição inadequada das classes ou documentação insuficiente, podem comprometer o uso do conjunto.

Dessa forma, este trabalho tem como objetivo mapear boas práticas, padrões abertos de anotação e ferramentas de software livre para a construção e organização de datasets públicos de imagens voltados ao aprendizado supervisionado, buscando sistematizar essas práticas em uma visão geral do processo.

MATERIAIS E MÉTODOS

O trabalho possui caráter teórico-experimental e foi dividido em duas etapas principais: o levantamento e a sistematização de boas práticas para construção de datasets de imagens e a aplicação de parte dessas práticas em um conjunto de dados público como estudo de caso.

Inicialmente, foi realizada uma revisão da literatura nas bases ACM Digital Library, IEEE Xplore e Google Scholar. As buscas utilizaram os termos “best practices”, “dataset quality”, “open datasets”, “open software” e “image annotation”, escolhidos de forma a abranger aspectos relacionados à qualidade, organização, compartilhamento e anotação de conjuntos de dados. Foram recuperados 107 trabalhos, sendo 9 da ACM Digital Library, 47 da IEEE Xplore e 51 do Google Scholar. A seleção ocorreu inicialmente pela análise dos títulos e, posteriormente, pela leitura dos resumos, descartando-se trabalhos sem relação direta com boas práticas, curadoria ou ferramentas de anotação aplicáveis ao contexto de aprendizado supervisionado. Em uma última etapa, foi considerada a relevância dos trabalhos e a reputação dos respectivos locais de publicação, resultando em 14 artigos selecionados para análise aprofundada.

Durante a análise, os trabalhos foram inicialmente classificados quanto ao seu caráter conceitual, apresentação de boas práticas, aplicações, frameworks ou ferramentas. A partir do conteúdo identificado, as recomendações foram posteriormente reorganizadas em etapas de mais alto nível, buscando representar uma visão geral do processo de construção e manutenção de um dataset de imagens. Essa organização compreende: preparação e organização dos dados, relacionada à estrutura dos arquivos e associação entre dados e metadados; anotação, envolvendo definição das classes, escolha da representação dos rótulos, ferramentas e procedimentos de revisão; verificação da qualidade, incluindo integridade, consistência e composição do conjunto; curadoria, voltada à identificação e revisão de possíveis problemas; e documentação e compartilhamento, relacionada ao registro das características, limitações e procedimentos empregados no dataset.

Os trabalhos selecionados abrangem diferentes aspectos relacionados à construção e manutenção de datasets, incluindo qualidade dos dados (Gong et al., 2023), estratégias e ferramentas de anotação de imagens (Liao, Kar e Fidler, 2021; Pande et al., 2022), curadoria e revisão de rótulos (Bernhardt et al., 2022), construção responsável de datasets (Orr e Crawford, 2024) e documentação e compartilhamento de dados abertos (Towse, Ellis e Towse, 2021). Essa variedade de abordagens permite reunir práticas aplicáveis a diferentes etapas do processo de preparação de conjuntos de dados para aprendizado supervisionado.

Como estudo de caso, foram selecionadas algumas das práticas passíveis de aplicação sobre o Brazilian Coin Detection Dataset[1], composto por imagens de moedas brasileiras e suas respectivas anotações. A utilização do conjunto tem caráter ilustrativo, permitindo demonstrar como as práticas mapeadas podem orientar verificações reproduzíveis sobre um dataset existente. Entre as análises realizadas estão verificações estruturais dos arquivos e anotações e a caracterização da composição do conjunto. Para o presente resumo, a distribuição das classes monetárias foi escolhida como exemplo simples da aplicação prática do processo sistematizado.

RESULTADOS E DISCUSSÃO

O levantamento mostrou que as boas práticas relacionadas a datasets estão distribuídas ao longo de diferentes etapas, abrangendo desde sua organização e anotação até atividades posteriores de avaliação, curadoria e documentação. A sistematização dessas recomendações em etapas de mais alto nível permite estabelecer uma visão geral do processo e relacionar diferentes tipos de verificações a aspectos específicos da qualidade dos dados.

Como exemplo de aplicação, foi analisada a distribuição das classes monetárias do Brazilian Coin Detection Dataset. Foram identificadas 8.233 instâncias da classe de 5 centavos, 7.844 de 10 centavos, 3.007 de 25 centavos, 1.771 de 50 centavos e 944 de 1 real. Dessa forma, a classe mais frequente possui aproximadamente 8,7 vezes mais instâncias que a menos frequente.

Esse resultado não caracteriza, isoladamente, o dataset como inadequado, uma vez que a distribuição esperada depende de sua finalidade e do processo de coleta. Entretanto, em tarefas de aprendizado supervisionado, o desbalanceamento pode favorecer o aprendizado das classes mais frequentes, fazendo com que o modelo apresente maior dificuldade na identificação das classes com menor número de exemplos. Além disso, a acurácia global pode não refletir adequadamente o desempenho obtido nas classes minoritárias. Assim, a análise da distribuição das classes exemplifica como uma das práticas mapeadas pode revelar características relevantes antes da utilização do conjunto, permitindo que possíveis limitações sejam conhecidas e documentadas.

CONCLUSÕES

O levantamento realizado permitiu identificar e organizar práticas relacionadas à construção e manutenção de datasets públicos de imagens, abrangendo etapas de preparação, anotação, avaliação da qualidade, curadoria, documentação e compartilhamento. Essa organização fornece uma visão mais estruturada do processo e ajuda a compreender que a qualidade de um dataset depende de diferentes aspectos, e não de uma única característica isolada.

[1] https://ieee-dataport.org/open-access/brazilian-coin-detection-dataset

A aplicação inicial das práticas em um dataset existente mostrou que essas verificações podem ser utilizadas para caracterizar aspectos relevantes antes do uso dos dados em tarefas de aprendizado supervisionado. No estudo de caso, a análise da distribuição das classes evidenciou um desbalanceamento considerável entre as denominações, ilustrando como uma etapa simples de avaliação pode revelar limitações ou características que devem ser consideradas durante a utilização do conjunto.

Os resultados obtidos até o momento indicam que a sistematização dessas práticas podem contribuir para uma avaliação mais estruturada de datasets de imagens, auxiliando na identificação e documentação de características relevantes antes de sua utilização em tarefas de aprendizado supervisionado. A consolidação dessas práticas deverá resultar em um material de apoio voltado à construção, organização, avaliação e compartilhamento de datasets públicos de imagens.

REFERÊNCIAS

BERNHARDT, M. et al. **Active label cleaning for improved dataset quality under resource constraints.** Nature Communications, v. 13, art. 1161, 2022. DOI: 10.1038/s41467-022-28818-3.

GONG, Y. et al. **A survey on dataset quality in machine learning.** Information and Software Technology, v. 162, p. 107268, 2023.

LIAO, Y.-H.; KAR, A.; FIDLER, S. **Towards Good Practices for Efficiently Annotating Large-Scale Image Classification Datasets.** In: 2021 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2021. Anais [...]. Nashville: IEEE, 2021. p. 4348–4357.

ORR, W.; CRAWFORD, K. **Building Better Datasets:** Seven Recommendations for Responsible Design from Dataset Creators. arXiv, 2024.

PANDE, B. et al. **A Review of Image Annotation Tools for Object Detection.** In: 2022 INTERNATIONAL CONFERENCE ON APPLIED ARTIFICIAL INTELLIGENCE AND COMPUTING (ICAAIC), 2022. Anais [...]. 2022. p. 976–982.

TOWSE, A. S.; ELLIS, D. A.; TOWSE, J. N. **Making data meaningful:** guidelines for good quality open data. The Journal of Social Psychology, v. 161, n. 4, p. 395–402, 2021.