EDGE VISION EDA

ANÁLISE EXPLORATÓRIA DE DADOS PARA VISÃO COMPUTACIONAL

Relatório Técnico

Londrina, Dezembro de 2025

1. Introdução

O avanço das técnicas de visão computacional e aprendizado profundo tem ampliado significativamente a capacidade de sistemas computacionais interpretarem informações visuais de forma automática. Em aplicações de detecção de objetos, a qualidade, consistência e distribuição dos dados utilizados para treinamento exercem papel fundamental no desempenho final dos modelos.
Nesse contexto, a Análise Exploratória de Dados (Exploratory Data Analysis – EDA) constitui uma etapa crítica no ciclo de desenvolvimento de sistemas baseados em visão computacional, permitindo identificar inconsistências estruturais, compreender padrões estatísticos e antecipar potenciais problemas antes da etapa de modelagem.
Este relatório apresenta a documentação técnica do Edge Vision EDA, uma Prova de Conceito (PoC) dedicada à análise exploratória de datasets de detecção de objetos. O projeto tem como foco a avaliação estrutural de um dataset externo, a geração de métricas agregadas sobre bounding boxes e a produção de evidências visuais que subsidiam decisões técnicas anteriores ao treinamento de modelos.
O sistema desenvolvido é exclusivamente analítico, não realizando qualquer etapa de treinamento, inferência ou modificação do dataset original.

2. Objetivos

2.1 Objetivo Geral

Desenvolver um pipeline de Análise Exploratória de Dados (EDA) para datasets de detecção de objetos, capaz de validar estruturalmente os dados, gerar métricas estatísticas agregadas e produzir visualizações analíticas que apoiem a tomada de decisão antes da fase de modelagem.

2.2 Objetivos Específicos

- Validar a consistência estrutural entre imagens e arquivos de label.
- Identificar labels inválidos, imagens sem anotação e inconsistências de formatação.
- Calcular métricas estatísticas agregadas das bounding boxes.
- Analisar a distribuição de tamanhos dos objetos anotados.
- Gerar gráficos analíticos a partir das métricas calculadas.
- Garantir organização, rastreabilidade e reprodutibilidade do processo de EDA.

3. Visão Geral do Sistema

O Edge Vision EDA foi projetado como um pipeline analítico reexecutável, operando sobre um dataset externo sem realizar qualquer alteração nos dados originais. A solução segue uma abordagem modular, na qual cada componente possui responsabilidade bem definida, favorecendo clareza, manutenção e evolução futura.
O arquivo main.py atua exclusivamente como orquestrador do fluxo de execução, coordenando as etapas de validação, cálculo de métricas e geração de visualizações, sem incorporar lógica analítica específica.

4. Arquitetura do Sistema

A arquitetura do sistema foi organizada de forma modular, visando baixo acoplamento e alta coesão entre os componentes. Os principais módulos são:

- **Settings**: centraliza configurações e paths do projeto.
- **Validator**: realiza a validação estrutural do dataset.
- **Metrics**: calcula métricas estatísticas agregadas das bounding boxes.
- **Plots**: gera visualizações analíticas a partir das métricas calculadas.
- **Logging Global**: padroniza o registro de eventos e mensagens do sistema.
- **Main**: orquestra a execução sequencial do pipeline de EDA.

Essa separação garante clareza arquitetural e alinhamento com boas práticas de engenharia de software.

5. Metodologia de Desenvolvimento

5.1 Validação Estrutural do Dataset

A validação estrutural é realizada por meio da verificação da correspondência entre imagens e arquivos de label, identificando três categorias principais de inconsistência:

- Labels sem imagem correspondente.
- Imagens sem arquivo de label (imagens negativas).
- Labels inválidos ou mal formatados.

Essa etapa não realiza interpretação semântica das bounding boxes, limitando-se à verificação estrutural e sintática dos dados.

5.2 Cálculo de Métricas Estatísticas

Após a validação, o sistema calcula métricas estatísticas agregadas relacionadas às bounding boxes presentes no dataset. As métricas incluem:

- Largura, altura e área médias das bounding boxes.
- Valores mínimos e máximos dessas dimensões.
- Proporção média entre largura e altura.
- Quantidade total de bounding boxes.

Essas métricas permitem compreender o comportamento global do dataset e identificar padrões ou anomalias estruturais.

5.3 Análise de Distribuição de Tamanhos

As bounding boxes são classificadas em categorias de tamanho (small, medium e large), possibilitando avaliar o balanceamento do dataset em relação à escala dos objetos anotados. Essa análise é fundamental para antecipar desafios relacionados ao treinamento de modelos de detecção.

5.4 Geração de Visualizações

A partir das métricas agregadas, o sistema gera gráficos analíticos que representam:

- Estatísticas geométricas das bounding boxes.
- Distribuição agregada de tamanhos dos objetos.

As visualizações produzidas têm caráter exploratório e suportam a interpretação dos resultados estatísticos.

6. Registro e Monitoramento

O sistema utiliza o módulo de logging padrão do Python, com configuração centralizada. São registrados eventos de inicialização, validação, cálculo de métricas, geração de plots e encerramento do pipeline, garantindo rastreabilidade completa da execução.

7. Resultados Obtidos

Durante a execução do pipeline de EDA, o sistema foi capaz de:

- Identificar inconsistências estruturais no dataset.
- Gerar métricas estatísticas agregadas representativas das bounding boxes.
- Produzir gráficos analíticos que evidenciam padrões de dimensão e distribuição dos objetos.
- Persistir métricas em formato CSV para análise posterior.

Os resultados obtidos fornecem subsídios técnicos relevantes para decisões relacionadas à limpeza de dados, balanceamento do dataset e escolha de estratégias de modelagem.

8. Limitações

Por se tratar de uma Prova de Conceito focada em EDA, o sistema apresenta algumas limitações:

- As análises são baseadas majoritariamente em métricas agregadas, não em dados amostrais individuais.
- Não são realizadas análises estatísticas avançadas ou inferência probabilística.
- O sistema não modifica, corrige ou normaliza o dataset original.
- A análise semântica das bounding boxes não faz parte do escopo atual.

Essas limitações são coerentes com o objetivo exploratório do projeto.

9. Conclusão

O Edge Vision EDA demonstrou a viabilidade de um pipeline estruturado e reprodutível para análise exploratória de datasets de detecção de objetos. A arquitetura modular adotada facilitou a organização do código e a separação clara de responsabilidades, enquanto a geração de métricas e visualizações forneceu uma base sólida para decisões técnicas anteriores ao treinamento de modelos.
A solução apresentada atende aos objetivos propostos e estabelece uma fundação consistente para projetos subsequentes de modelagem e inferência em visão computacional.