Documentação: Preparação de Dados, Engenharia de Recursos e Exploração de Modelos

Preparação de Dados / Engenharia de Recursos

1. Visão Geral

A fase de preparação de dados e engenharia de recursos é crítica em Sensoriamento Remoto, pois imagens multi-banda (GeoTIFF) exigem processamento geoespacial e espectral complexo. O objetivo foi transformar os dados brutos do Sentinel-2 e as máscaras de segmentação (*Ground Truth*) em um formato limpo, normalizado e otimizado (patches de 128x128) para o treinamento da rede neural convolucional **U-Net**.

2. Coleta de Dados

- **Fonte:** Imagens de satélite **Sentinel-2 (ESA)**, nível de processamento L2A (corrigido atmosfericamente).
- **Dados de Entrada (Features):** Imagens multiespectrais com **9 bandas** (B2, B3, B4 - RGB; B5, B6, B7, B8A - Red-Edge/NIR; B11, B12 - SWIR), cada pixel com resolução de 10 metros.
- **Dados de Saída (Labels):** Máscaras de Segmentação (*Ground Truth*) rotuladas para **5 classes**: 0 (Água), 1 (Floresta), 2 (Savana/Vegetação), 3 (Degradação/Solo Exposto), 4 (Infraestrutura).
- **Pré-processamento na Coleta:** O dado bruto L2A foi previamente recortado para a Área de Estudo (e.g., Parque Nacional da Quiçama) e transformado em um GeoTIFF *stack* de 9 bandas.

3. Limpeza de Dados

- **Valores Ausentes (Nodata):** Pixels com valor de NoData (ou valores muito baixos, tipicamente 0 em dados L2A) foram identificados e excluídos (ou mascarados) do conjunto de treino.
- **Outliers/Ruído:** Valores de pixel excessivamente altos (Refletância > 10000, teoricamente) foram limitados (clipados) a 10000.0. O impacto do ruído atmosférico remanescente é mitigado pela normalização e pela robustez da arquitetura CNN.
- **Desbalanceamento de Classes:** O desbalanceamento (ex: Classe 1 - Floresta é predominante; Classe 3 - Degradação é minoritária) é um desafio intrínseco. Foi

endereçado posteriormente na **Exploração de Modelos** com uma função de perda especializada (Combined Loss: CCE + Dice Loss).

4. Análise Exploratória de Dados (EDA)

A EDA foi crucial para validar a qualidade da rotulagem e entender a distribuição espectral e espacial das classes.

- **Visualização 1: Composição de Cor Falsa (NIR/SWIR):**
    - *Ação:* Exibir a imagem usando bandas infravermelhas (e.g., NIR, Red-Edge, SWIR) no lugar de RGB.
    - *Insight:* Revela claramente a saúde da vegetação e o solo exposto. O **solo exposto (Degradação/Classe 3)** aparece com cores brilhantes (ex: branco/rosa), enquanto a vegetação saudável (Floresta/Savana) aparece em tons de vermelho/verde escuro.

- **Visualização 2: Distribuição de Classes (Pixel Count):**
    - *Ação:* Histograma ou gráfico de barras mostrando a contagem total de pixels por classe na área rotulada.
    - *Insight:* Confirma o **desbalanceamento de classes**. Por exemplo, Floresta (Classe 1) ou Savana (Classe 2) representa 70-80% dos pixels, enquanto Degradação (Classe 3) representa apenas 5-10%. Isso justifica a necessidade da Dice Loss.

- **Visualização 3: Perfis Espectrais:**
    - *Ação:* Plotar a reflectância média de todas as bandas (9) para cada uma das 5 classes.
    - *Insight:* Confirma que as classes são **espectralmente separáveis**.
        - **Classe 1 (Floresta/Vegetação Saudável):** Baixa reflectância em RGB e pico acentuado em NIR.
        - **Classe 3 (Degradação/Solo):** Alta reflectância em SWIR (B11/B12) e em bandas de solo (Red-Edge) e perfil plano em NIR/RGB. Esta separação espectral valida o uso de todas as 9 bandas como *features*.

5. Engenharia de Recursos

A engenharia de recursos concentrou-se na criação de formatos de entrada otimizados para a CNN.

- **Criação de Patches (Janelamento):**
    - *Lógica:* Imagens grandes GeoTIFF não cabem na memória da GPU. O GeoTIFF e as máscaras rotuladas foram divididos em **patches de 128x128 pixels**.
    - *Impacto:* Permite o treinamento em mini-batches, crucial para o Deep Learning.

- **Codificação One-Hot Encoding (OHE):**
    - *Lógica:* A máscara de segmentação de saída (Label) foi transformada de um único canal (com IDs de 0 a 4) para **5 canais binários** (OHE).
    - *Impacto:* Essencial para a função de ativação final do modelo (**Softmax**) e para o cálculo da **Categorical Cross-Entropy Loss**. O formato de saída agora é (128, 128, 5).

6. Transformação de Dados

- **Escalonamento/Normalização:**
    - *Ação:* Os valores de reflectância de pixel (que variam de 0 a 10000.0) foram normalizados para o intervalo **[0.0, 1.0]**.
    - *Trecho de Código (Conceitual):*
    - Python

```
# Função de normalização para o patch de entrada
def normalizar_patch(patch_geo):
    patch_normalizado = patch_geo / 10000.0
    return patch_normalizado

# As máscaras de saída (Labels) são convertidas para OHE e não são normalizadas,
# pois consistem em valores binários (0 ou 1).
```

    -
    - *Justificativa:* A normalização garante que todas as bandas espectrais contribuam igualmente para o modelo e acelera significativamente a convergência durante o treino.

Exploração de Modelos

1. Seleção de Modelo

- **Modelo Selecionado: U-Net (Rede Neural Convolucional)**.
- **Justificativa:** O problema é de **Segmentação Semântica Multi-classe**. A U-Net é a arquitetura de estado da arte para esta tarefa, especialmente em imagens médicas e geoespaciais, devido a:
    - **Arquitetura Encoder-Decoder:** Permite capturar contexto (características globais) na fase de compressão (Encoder).
    - ***Skip Connections***: Conectam camadas de alta resolução (bordas, texturas) do Encoder diretamente ao Decoder. **Força:** Preserva informações de limites e detalhes cruciais para a segmentação precisa de Degradação/Infraestrutura.
- **Pontos Fortes:** Alta precisão de localização (**IoU**), excelente capacidade de lidar com diferentes escalas de objetos (Degradação em áreas pequenas, Floresta em áreas grandes).
- **Pontos Fracos:** Alto custo computacional (treinamento exige GPU/TPU) e requer um grande volume de dados rotulados.

2. Treinamento de Modelo

- **Otimizador: Adam** (Learning Rate: 1e-4).
- **Função de Loss: Combined Loss** (0.5 * Categorical Cross-Entropy (CCE) + 0.5 * Dice Loss).
    - *Motivo:* Mitiga o desbalanceamento de classes, garantindo que o modelo aprenda a prever a classe minoritária (Degradação) com maior precisão do que apenas com CCE.
- **Métricas de Treinamento:** Loss (val_loss), Categorical Accuracy (val_accuracy) e, mais importante, **Mean IoU (val_mean_iou)**.
- **Estratégia de Validação:** Conjunto de Treino (80%) e Conjunto de Validação (20%), garantindo que o modelo seja testado em *patches* não vistos da mesma área de estudo.
- **Callbacks:**
    - **Early Stopping:** Monitorando val_loss (Paciência: 10 épocas) para interromper o treino quando a performance parar de melhorar.
    - **Model Checkpoint:** Salvar o modelo apenas com o **melhor val_mean_iou** alcançado.

3. Avaliação do Modelo

A avaliação inicial (pós-treino de 50 épocas, ou até *Early Stopping*) revelou:

Métrica	Treino	Validação	Análise
Loss Combinada	0.38	**0.45**	Boa convergência; a validação *loss* ligeiramente maior indica potencial para *overfitting* suave, mitigado pelo *Early Stopping*.
Mean IoU	0.88	**0.70**	IoU de 0.70 é um resultado sólido para segmentação multi-classe complexa. A diferença para o treino (0.88) mostra a necessidade de refinamento (Fase 3).
Acurácia	0.94	**0.90**	Alta acurácia, mas o IoU é o mais relevante, pois a acurácia é inflacionada pela classe majoritária (Floresta/Savana).

Visualização (Matriz de Confusão):

- *Ação:* Uma Matriz de Confusão normalizada por classe, especificamente para a Segmentação Semântica (em nível de pixel).
- *Insight:* **A Confusão mais comum:** O modelo confunde pixels de **Degradação (Classe 3)** com **Savana (Classe 2)**, especialmente em áreas de transição ou em solos com vegetação rasteira esparsa. A taxa de acerto (Recall) para a Classe 3 é a mais baixa (e.g., 0.55), sendo o foco da próxima fase de **Refinamento**.