基于 NDVI 与气象数据的时序预测模型报告

Yingqi WU, Shiyun GU

1. 数据处理（Data Processing）

1.1 NDVI 数据处理

NDVI 数据来源于多时相栅格（nc），一共有 25 个时间的数据，格式(t: 25, x: 76, y: 49)
在预处理阶段完成了以下操作：

- 去除损坏数据，经过检查其中有两个数据损坏
- 线性插值：由于 25 个数据用于训练过少，所以做了每三天一个 ndvi 的线性插值。通常来说，ndvi 的变化往往是近似线性的，也没有出现极端天气，所以这种情况下的线性插值用于训练是完全可以的。

1.2 气象数据处理（Weather Data）

气象数据包括多个环境协变量（如温度、降水），其处理流程为：

- **时间对齐**
  把时间步和 ndvi 对齐，把两个时间步之间的天气数据做平均或和，具体按照：
  土壤湿度：sm_30cm_mean
  降雨：RAIN_sum
  灌溉：irrig_mm_sum
  阳光：IRRAD_sum
  温度：TMIN_mean TMAX_mean
  大气：VAP_mean
  风：WIND_mean
- **空间化（Spatialization）**
  原始气象数据本身不具备空间维度在处理阶段：
  - 将每个气象变量扩展为 (x, y) 网格
  - 项目研究的农田被划分为四个区域，每个区域的灌溉方式不同，天气数据也各自独立，所以用四个像素点坐标框出一个区域后赋对应的天气数据

1.3 最终数据结构

整合后的完整数据张量为：
(time, x, y, channels)

- time = 37
- channels = NDVI + 气象变量 = 9

2. 模型输入与输出（Model Input & Output）

- 模型输入的是一段连续序列，输出则是将输入序列时间步往后平移一步的连续序列。我们以 test 集为例：test 的输入是 t0,t1...t5 六个时间步，那输出则是 t1,t2...t6。
- 输入格式(time, x, y, 9)， 输出格式(time, x, y, 1)，也就是输入的是 ndvi+天气共 9 个通道，输出则只有 ndvi 一个通道

3. 模型结构与训练策略

3.1 模型结构概述

模型采用 **ConvLSTM + Conv3D** 架构:

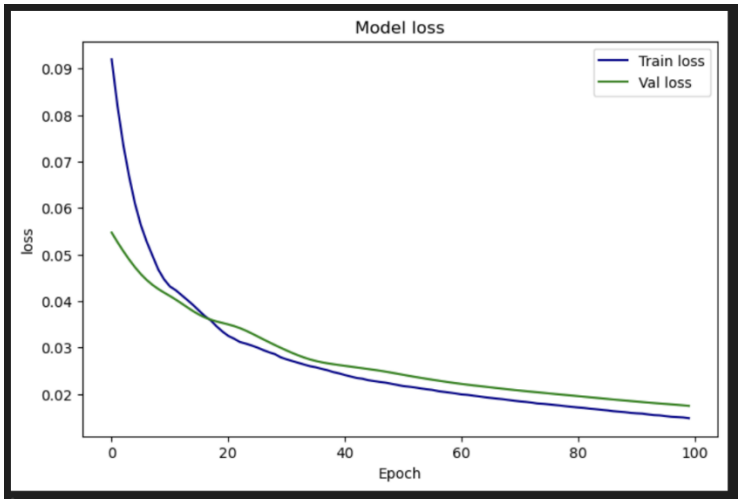
- **ConvLSTM 层**
  - 同时建模:
    - 空间结构（x, y）
    - 时间依赖（t）
  - 能捕捉:
    - 植被生长的空间连续性
    - NDVI 的时间演化规律
- **Conv3D 输出层**
  - 在时空特征上进行三维卷积
  - 直接输出 NDVI 预测序列

3.2 训练数据划分策略

实验中对比并尝试了多种方案:
最终版本 run_model_final 中， 按时间步顺序:

- 测试集：7 个连续时间步
- 训练集：22 个连续时间步（取中间段）
- 验证集：8 个连续时间步

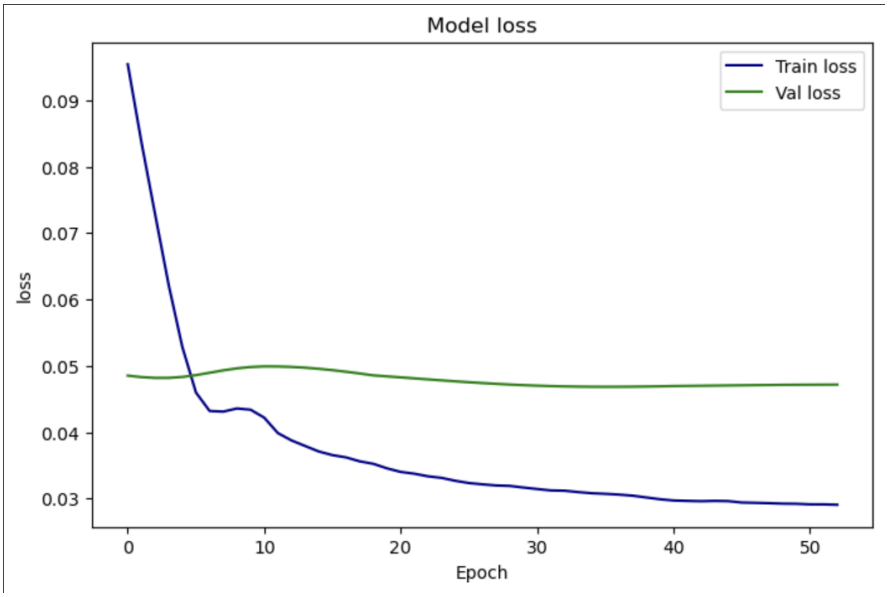
这样做的目的是同时保证模型能学到完整生长周期的中段特征



初始版本 run_model_origin 中， 按时间步顺序:

- 训练集：22 个连续时间步（取中间段）
- 验证集：7 个连续时间步
- 测试集：8 个连续时间步

训练集只覆盖了植物生长阶段，而验证集正好在衰退阶段。



另一个版本 run_model_rolling_train 中:

- 训练集：29 个连续时间步（取中间段）
- 验证集：8 个连续时间步

这是一个动态的训练集，第一折（fold）训练集：t0…t28，验证集 t29…t36。这一折训练 5 个 epoche 之后，进入第二折，训练集变为 t1…t29，验证集则是 t30…t0 继续训练。Fold2 训练 5 个 epoche 后，进入第三折，训练集变为 t2…t30，验证集则是 t31…t1。以此类推，总共训练指定的 fold 数量。

本意是考虑到数据集是一段可以前后衔接起来的植物生长+衰败的过程，想要让模型学习到所有阶段，但是结果不尽如人意。