区域植被指数分析与物候预测

作者：王子翔 叶岩宁

摘 要

经过数十年的进化，各地植物生长呈现规律。然而，每一年的季节变化不完全一样，会对植物生长造成影响。本文通过对某区域的植被指数进行分析与评价，对未来的物候情况进行提前预测，从而达到趋利避害的效果。

对问题一，我们首先确定使用的是 CNN 卷积神经网络对该区域的数据进行分析。我们将训练集中的图片手动划分为了一大一小两个部分，大的部分设置为训练集和小的则为测试集。通过训练集的数据对模型进行训练，实时输出 J(Θ)值反馈模型的误差量。同时，也可以通过将测试集的数据输入模型，以对当前模型进行准确度测试，反馈当前模型的识别准确率。

对于问题二，模型将图像信息与图像的序号标签同时作为模型训练的数据来源。将图像序号信息作为图像相对时间的时间戳，作为标签数据输入模型;将图像信息作为输出数据，反应对应时间标签下的图像信息。通过关联图像与时间戳，建立图像与时空的关系。

问题三为模型的输出结果与对结果的分析。通过对于建立的机器学习模型输入时间，模型就能够预测出此时间标签对应的图像的信息。我们再手动对于当前图像进行分析。因为模型经过损失值的计算和测试集的检验等条件的约束，得到的图像为可靠图像，能反映正确信息，此时的手动分析为可靠分析。

在问题四中，为了避免模型过拟合与欠拟合的发生，我们使用了 L2 正则化算法缓解过拟合的问题，同时提升模型的泛化性能。L2 正则化让参数向量的大部分元素都变得很小，压制了参数之间的差异性。L2 正则化的惩罚因子设置为 1e-3，制约能力较强，因此模型的泛型也较大。

对问题五而言，我们设计了损失计算函数分析 J(Θ)，测试集检验准确度来约束模型。同时我们针对距离数据集时间轴较远的时间的预测结果的参照 Amazon Mechanical Turk 实验，对于预测结果投票。为了避免歧义，我们将一段时间内的预测结果评分进行加权平均，最终得到预测结果的分数。

关键词： CNN 卷积神经网络模型　MSE 均方误差　ED 指数衰减学习　GDO 梯度下降优化器　EMA 滑动平均　L2 正则化　CKPT 断点续训　Lanczos 插值　AMT 土耳其机器人

一、问题重述

1.1 问题背景

经过数十亿年的进化，地球的生物圈适应了这颗行星的春夏秋冬。各地植物的生长节律与季节同步，春华秋实，这被称为植被的物候。然而，每一年的季节变换都不完全一样。有些年份的气温和雨水和其他年份都很不一样。这会给这颗行星的农林牧业带来一些波动，让植被的生长节奏变动：比如说干热带来的山火肆虐，比如说春天提前导致的生长季变长，比如说降水增大带来的洪涝，等等。如果我们能够提前预测各地未来的物候，那么就可以趋利避害。

1.2 问题要求

1. 设计合理的数学模型（包括但不限于机器学习模型），根据训练集中提供的某区域的物候时空数据训练模型参数，并对测试数据进行测试，分析模型在训练集和测试集中的效果；
2. 建立的数学模型需要同时合理的考虑同一时期不同区域上植被指数的空间关系，也需要考虑不同时期同一地区植被指数的演变情况，即需要建立能描述植被发展的时空模型；（重点说明如何对植被指数的时空关系进行建模）
3. 预测未来该区域的物候演变结果，并对预测结果进行分析；
4. 分析模型的泛化性能。
5. 建立合理的评价指标，如对预测结果的评价，对模型的评价。

二、问题假设

1. 所有数据集中的数据处于同一相对时间轴上。相对时间轴上的时间均为整数，时间轴上的范围为[0，500)的前开后闭区间。
2. 假定数据所对应的时间戳与图像的序号相同，对应相对时间轴上的时间节点。
3. 数据集的数据均真实可信。

三、符号说明

$z^{(j)}$ ：第 j 层神经元接收上层传入的刺激

a_i^j ：第 j 层第 i 个神经元获得的激活值

J(Θ)：损失函数

L ：神经网络总共包含的层数

S_l：第 l 层的神经元数目

K ：输出层的神经元数，亦即分类的数目

$||\omega||_2$：L2 正则化范数

$\Delta_{ij}^{(l)}$ ：权值梯度

$D^{(l)}$ ：各层权值的更新增量

α ：学习率

四、问题分析

4.1 问题一分析

该问题需要设计数学模型对某区域的物候时空数据训练模型。考虑到给出的数据为图片信息，因此我们设想通过建立机器学习模型来解决这一问题。通过使用 Python 的 Pillow 库打开 Tiff 图片文件并进行一些预操作。将图片缩放为更小尺寸，转换为数组，再通过数组尺寸变形拉伸成一维数组，转化数组的值的大小和类型。

经过对于图片的处理后，再使用 tensorflow 库建立 CNN 卷积神经网络模型，将图片的序号作为输入量，将处理后的图片作为输出量喂入神经网络训练模型。前向传播中设计神经网络模型的有关参数和基础方法。反向传播中定义损失函数 J(Θ)方法为 MSE 均方误差，定义 ED 指数衰减学习率，定义 GDO 梯度下降优化器，定义 EMA 滑动平均，定义 L2 正则化防止过拟合。最后定义包含 ckpt 断点续训的 STEPS 训练轮执行训练。

我们再将训练集中的图片手动划分为了一大一小两个部分，大的部分设置为训练集和小的则为测试集。通过训练集的数据对模型进行训练，实时输出 J(Θ)值反馈模型的误差量。同时，也可以通过将测试集的数据输入模型，以对当前模型进行准确度测试，反馈当前模型的识别准确率。

4.2 问题二分析

该问题利用到了 2.1 和 2.2 的假设内容。简单来说就是假设一个相对时间轴放置数据集中的数据。相对时间轴上的时间均为整数，时间轴上的范围为[0，500)的前开后闭区间。同时，数据所对应的时间戳与图像的序号相同，对应相对时间轴上的时间节点。

利用以上两点假设，我们建立了图像的相对时空关系。利用此关系，我们将时间戳作为标签数据输入模型，将图像信息作为输出数据输入模型。在模型中也建立图像与时空的关系。

4.3 问题三分析

该问题利用到了 2.3 的假设内容，即数据集的数据均真实可信。我们计算出模型的损失和准确度后，利用 2.3 即可推广得到在损失值/像素点的误差下，得到的结果为正确的。通过测试集得到模型的精度，表示在较短时间轴内精度

为此值，随着远离时间轴原点精度会逐渐下降。因此时间轴的长度不宜设置过大，因为远离原点的预测因为此时精度太小,将会失去意义。

通过对于建立的机器学习模型输入时间，模型就能够预测出此时间标签对应的图像的信息。我们再手动对于当前图像进行分析。因为模型经过损失值的计算和测试集的检验等条件的约束，此时得到的图像为可靠图像，能够反映正确的信息，因此我们认为此时的手动分析为可靠分析。

4.4 问题四分析

针对该问题，为了避免模型过拟合与欠拟合的发生，我们降低的隐藏层参数的个数，让模型尽可能简单。模型复杂度下降的同时，过拟合现象也不容易产生，泛化性能得以提升。

另外，我们使用了规则项来约束模型的特征，即 L2 正则化算法。我们通过设置 L2 范数的规则项 $||\omega||_2$ 以约束每一个 ω 元素的大小。L2 正则化的惩罚因子设置为 1e-3，制约能力较强。通过权重衰减让参数向量的大部分元素都变得接近 0，压制了参数之间的差异性。从而缓解过拟合的问题，同时提升模型的泛化性能。

4.5 问题五分析

对该问题而言，我们对模型设计了先验性的损失计算函数，利用 MSE 均方误差算法，以分析模型当前 J(Θ)。不仅如此，我们划分出的测试集也能通过实时计算模型的准确度来显示模型的泛化性能。在计算出模型的损失和准确度后，即可推广得到在较短时间轴内精度为测试集得到模型的精度。随着远离时间轴原点精度会逐渐下降。因此预测时间距离数据集的偏移不宜设置过大，因为远离原点的预测因为此时精度将会低于测试得到的模型精度,预测将会失去意义。

我们针对距离数据集时间轴较远的时间的预测结果的参照 Amazon Mechanical Turk 实验，对于预测结果投票。为了避免歧义，我们将一段时间内的预测结果评分进行加权平均，最终得到预测结果的分数。

五、模型建立与求解

5.1 数据预分析

使用提供的图像数据输入 MATLAB 作图，得到一段时间内的植被指数图样。同一周期中的植被指数大约可分为四个阶段，我们分别定义为萌芽，茂盛，凋亡，次茂盛。(其中萌芽和凋亡阶段的图像相近，而次茂盛可以理解为茂盛的次级。)我们选取了某周期的四个阶段的图样，作为样例进行展示以说明每个周期的大致过程。

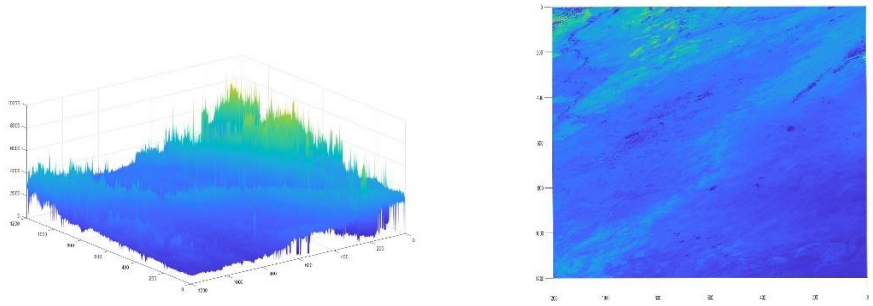


图 5. 1(1) – 同周期内第一阶段的植被指数图像(萌芽阶段)

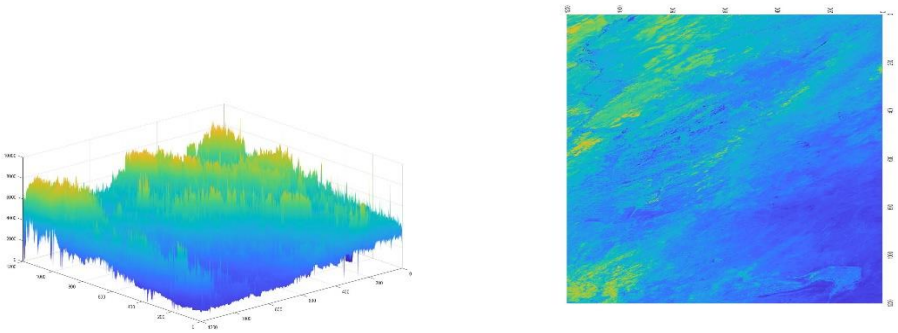


图 5. 1(2) – 同周期内第二阶段的植被指数图像(茂盛阶段)

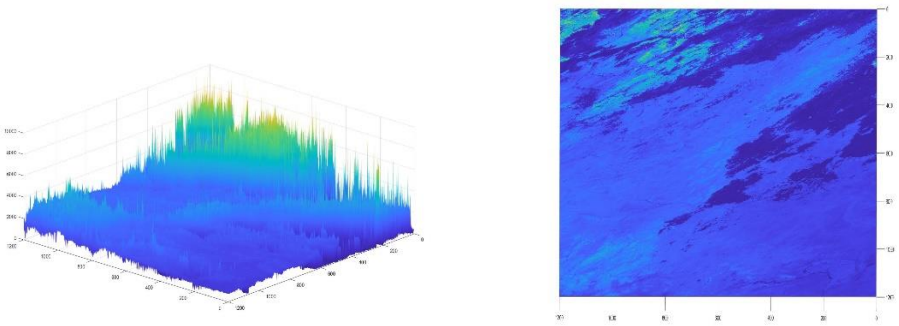


图 5. 1(3) – 同周期内第三阶段的植被指数图像(凋亡阶段)

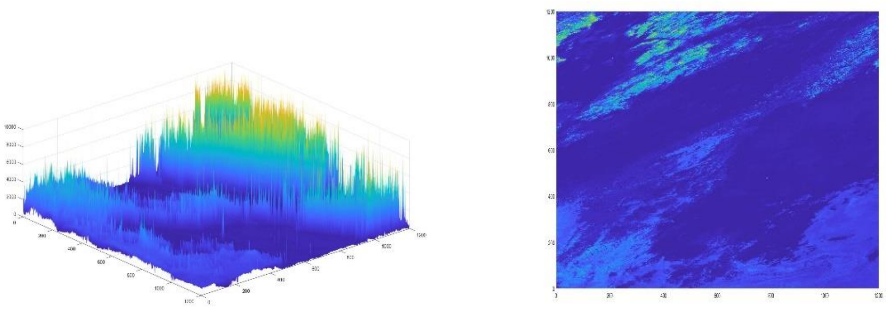


图 5. 1(4) – 同周期内第四阶段的植被指数图像(次茂盛阶段)

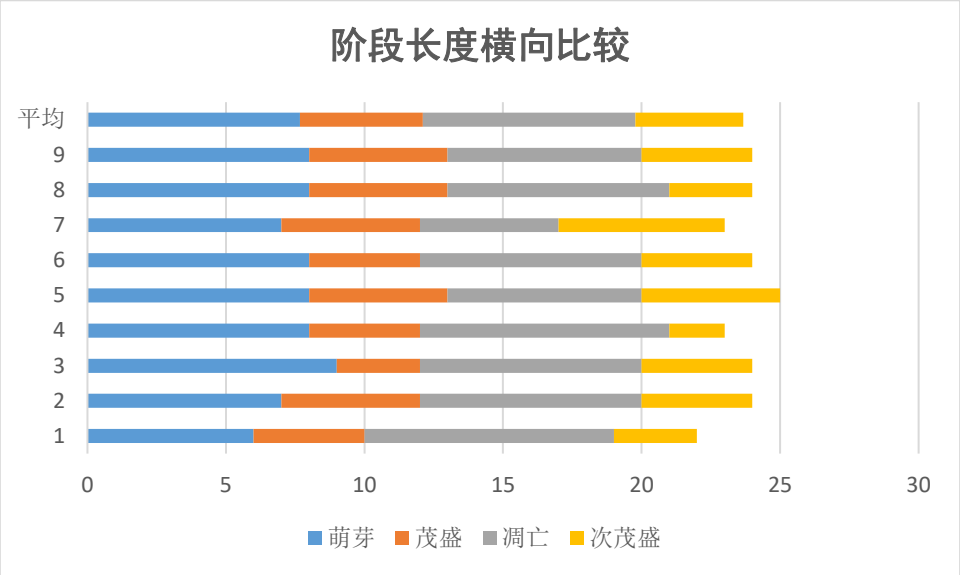


图 5. 1(5) – 阶段长度横向比较条形图

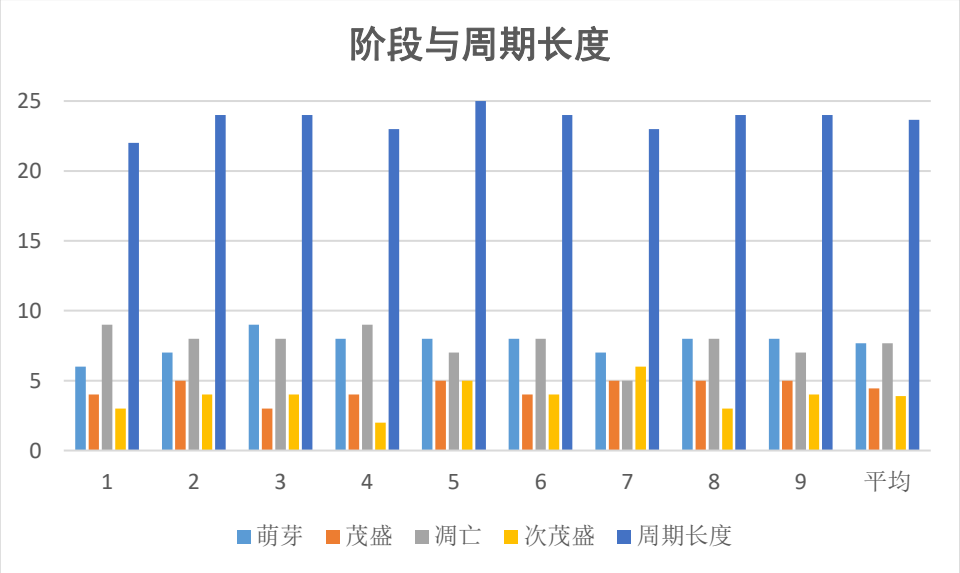


图 5. 1(6) – 阶段与周期长度簇状柱形图

可见，区域植被指数具有较为明显的周期性。针对植被指数进行物候预测具有一定的理论可行性。因此，我们使用处理过后的图像作为数据输入模型进行植被指数与物候预测。

5. 2 模型的建立和求解

5. 2. 1 问题一模型建立

CNN 卷积神经网络是一种前馈神经网络，它的人工神经元可以响应一部分覆盖范围内的周围单元，对于大型图像处理有出色表现。相比较其他深度、前馈神经网络，卷积神经网络需要考量的参数更少，使之成为一种颇具吸引力的深度学习结构。

利用 CNN 卷积神经网络，结合问题实际，植被指数分析模型建立过程如下：

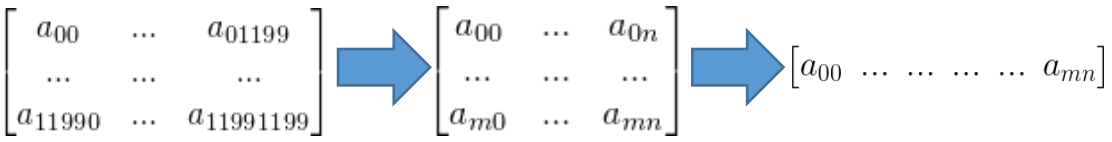
Step1. 图像信息预处理

在对神经网络输入数据之前，应该先将数据进行预处理，以获得相同规格的素材，方便神经网络的训练。区域植被指数分析过程中，只需考虑 2 个指标，即植被指数分布图与数据的时间坐标。

其中植被指数分布图数据的 Tiff 图片文件数据文件为 16 位（int16）。植被指数（NDVI）的有效值从 0 到 1，越高说明植被发育越好。通过使用 Python 的 Pillow 库打开 Tiff 图片文件并进行一些预操作。使用 ANTIALIAS 抗锯齿算法将图片缩放为更小尺寸，转换为数组，再通过数组尺寸变形拉伸成一维数组，转化数组的值的大小和类型。

再将一维数组类型转换为 float32 类型，同时将数据转化为 0-255 的数字。完成数据的归一化。

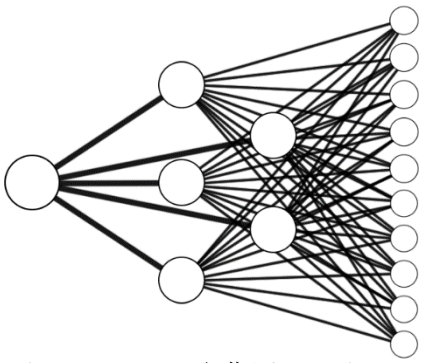
通过上述变换得到的数学矩阵为：



1200*1200 全尺寸图像　　缩放为 m*n 图像　　拉伸为一维数组

Step2. 前向神经网络的搭建

前向传播中设计神经网络模型的有关参数和基础方法。模型将时间轴坐标设为输入量，图像信息设置为输出量。相关参数设置遵循泛化原则，L 层数共设置 3 层，s_0输入层结点为 1，s_1 隐藏层结点为 5，s_2 输出层结点为 m*n。



输入层　　隐藏层　　输出层

神经网络每层都包含有若干神经元，层间的神经元通过权值矩阵Θ^l连接。一次信息传递过程可以如下描述：

1. 第 j 层神经元接收上层传入的刺激：　　$z^{(j)} = \Theta^{(j-1)} a^{(j-1)}$　　(1)

2. 刺激经激励函数 g 作用后，产生一个激活向量a^j。a_i^j表示的就是 j 层第 i 个神经元获得的激活值。

$$a^{(j)} = g(z^{(j)}) \qquad (2)$$

神经网络首先随机生成随机参数 w，根据神经网络的描述搭建前向传播规则，再使用隐藏层对输入进行偏移。同时神经网络引入正则化系数进行运算,控制偏

移量大小，进一步增加模型的泛化性能。

$$\begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} \rightarrow \begin{bmatrix} a_1^{(2)} \\ a_2^{(2)} \\ a_3^{(2)} \end{bmatrix} \begin{bmatrix} a_1^{(3)} \end{bmatrix} \rightarrow h_\theta(x) \qquad \begin{aligned} a^{(1)} &= x \\ z^{(2)} &= \Theta^{(1)} a^{(1)} \\ a^{(2)} &= g(z^{(2)}) \\ z^{(3)} &= \Theta^{(2)} a^{(2)} \\ a^{(3)} &= g(z^{(3)}) \\ h_\Theta(x) &= a^{(3)} \end{aligned} \tag{3}$$

前向传播偏移过程　　前向传播过程

Step3. 反向神经网络的搭建

反向传播中需要定义损失函数 J(Θ)和反向传播方法。反向传播定义批数量，学习率，学习衰减率，正则化系数，训练轮数，滑动平均衰减率等一系列参数，以及使用 MSE 均方误差计算损失函数 J(Θ)方法，定义 ED 指数衰减学习率，定义 GDO 梯度下降优化器，定义 EMA 滑动平均，定义 L2 正则化防止过拟合。最后定义包含 ckpt 断点续训的 STEPS 训练轮执行训练。

由于神经网络允许多个隐含层，即各层的神经元都会产出预测，因此，就不能直接利用传统回归问题的梯度下降法来最小化 J(Θ)，而需要逐层考虑预测误差，并且逐层优化。

$$\delta^{(l)} = \begin{cases} a^{(l)} - y & l = L \\ (\Theta^{(l)} \delta^{(l+1)})^T .* g'(z^{(l)}) & l = 2, 3, \ldots, L-1 \end{cases} \tag{4}$$

其中：

$$g'(z^{(l)}) = a^{(l)} .* (1 - a^{(l)}) \tag{5}$$

使用 MSE 均方误差计算损失函数 J(Θ)值，其中 MSE 均方误差公式为：

$$MSE = \frac{\sum_{i=1}^{r} (n_i - 1) s_i^2}{N - r} \tag{6}$$

根据优化过程在不同阶段的特点，一个大体的思路就是前期使用较大的学习率加速收敛，后期用较小的学习率保证稳定，这就是学习率衰减背后的思想。使用 ED 指数衰减学习率，可以在迭代初期得到较高的下降速度，可以在较小的训练轮数下获得更好的收敛度。以下是 ED 指数衰减学习率公式：

$$lr = lr_{base} \cdot \gamma^{\left\lfloor \frac{step}{stepsize} \right\rfloor} \tag{7}$$

GDO 梯度下降法的核心，是最小化目标函数 J(θ)，其中 θ 是模型的参数，θ ∈Rd。它的方法是，在每次迭代中，对每个变量，按照目标函数在该变量梯度的相反方向，更新对应的参数值。其中，学习率 η 决定了函数到达（局部）最小

值的迭代次数。以下是 GDO 梯度下降优化器公式：

$$E[g^2]_t = 0.9E[g^2]_{t-1} + 0.1g_t^2 \tag{8}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} \odot g_t \tag{9}$$

EMA 滑动平均用来估计变量的局部均值，使得变量的更新与一段时间内的历史取值有关。变量 v 在 t 时刻记为v_t，θ_t为变量 v 在 t 时刻的取值，即在不使用滑动平均模型时$v_t = \theta_t$，在使用滑动平均模型后，v_t的更新公式如下：

$$v^{(t)} = \alpha v^{(t-1)} + (1-\alpha)\theta^{(t)} \tag{10}$$

求各层权值的更新增量 $D^{(l)}$，连接偏置的权值不进行正规化

$$D_{i,j}^{(l)} = \begin{cases} \frac{1}{m}(\Delta_{i,j}^{(l)} + \lambda\Theta_{i,j}^{(l)}), \text{if } j \neq 0 \\ \frac{1}{m}\Delta_{ij}^{(l)}, \text{if } j = 0 \end{cases} \tag{11}$$

更新各层的权值矩阵 Θ(l)，其中 α 为学习率：

$$\Theta^{(l)} = \Theta^{(l)} + \alpha D^{(l)} \tag{12}$$

5. 2. 2 问题二模型建立

该问题利用到了 2. 1 和 2. 2 的假设内容。简单来说就是假设一个相对时间轴放置数据集中的数据。相对时间轴上的时间均为整数，时间轴上的范围为[0，500)的前开后闭区间。

利用以上两点假设，我们建立了图像的相对时空关系。同一周期中的植被指数大约可分为四个阶段，我们分别定义为萌芽，茂盛，凋亡，次茂盛。(其中萌芽和凋亡阶段的图像相近，而次茂盛可以理解为茂盛的次级。) 根据以上划分，我们统计了 9 个周期各个阶段的长度和周期长度。

表 5. 2. 2 阶段划分与周期长度

阶段	平均周期长度
萌芽	7.67
茂盛	4.44
凋亡	7.67
次茂盛	3.89
周期长度	23.67

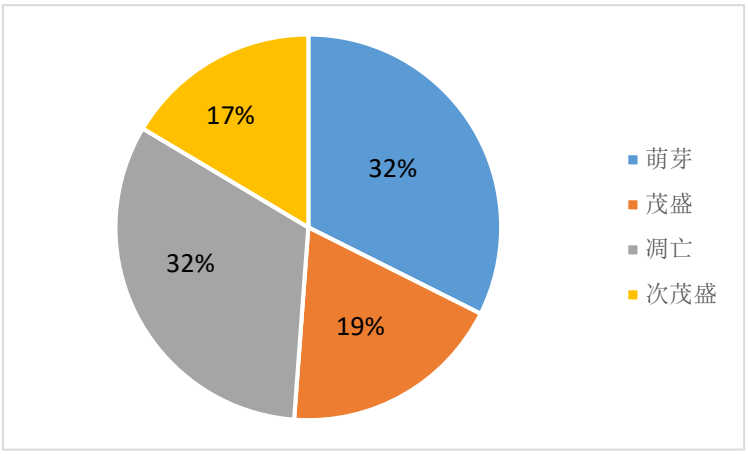


图 5. 2. 2(1) – 阶段平均长度饼状图

利用此关系，我们将时间戳作为标签数据输入模型，将图像信息作为输出数据输入模型。在模型中也建立图像与时空的关系。

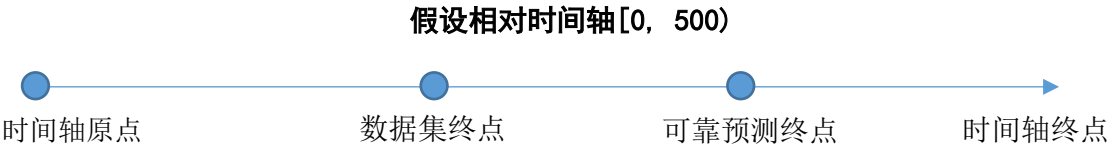


图 5. 2. 2(2) – 假设相对时间轴

我们在此基础上，认为数据所对应的时间戳与图像的序号相同。由于图像数据已知按照某个特定地区的数据时间排序。因此我们可以认为图像的(相对时间轴标签)时间戳，即为图像的序号，并与相对时间轴上的时间节点一一对应。

5. 2. 3 问题一二模型求解

有了模型后，使用训练集来训练参数。说准确点，一般是用来降参数进行梯度下降运算的。测试集没有参与梯度下降的过程，也就是说是没有经过训练的。既不用测试集梯度下降，也不用它来控制超参数，只是在模型最终训练完成后，用来测试一下最后准确率。

我们将训练集中的图片手动划分为了一大一小两个部分，大的部分设置为训练集和小的则为测试集。通过训练集的数据对模型进行训练，实时输出 J(Θ)值反馈模型的误差量。同时，也可以通过将测试集的数据输入模型，以对当前模型进行准确度测试，反馈当前模型的识别准确率。

其中，训练集和测试集没有交集且测试集由训练集中合理挑选，能够代表整个数据集的准确度，这样的数据才具有统计学意义。通常来说，训练集和测试集比例约为 8：1～9：1，且测试集符合随机均匀分布。但是考虑到本模型注重预测

既有时间轴范围外的数据而非预测范围内的数据，因此测试集从训练集的尾部挑选而非从训练集全局随机挑选。这样测试出的精确度反馈结果虽然可能会低于随即均匀选点所计算出的精确度，且不符合常规设计原理，但是这样的设计更能代表预测结果的实际情况，能更好的反馈模型的预测能力。

通过 J(Θ)值的反馈，描述模型在训练集中的效果。通过准确度的测试，描述模型在测试集中的识别准确率。

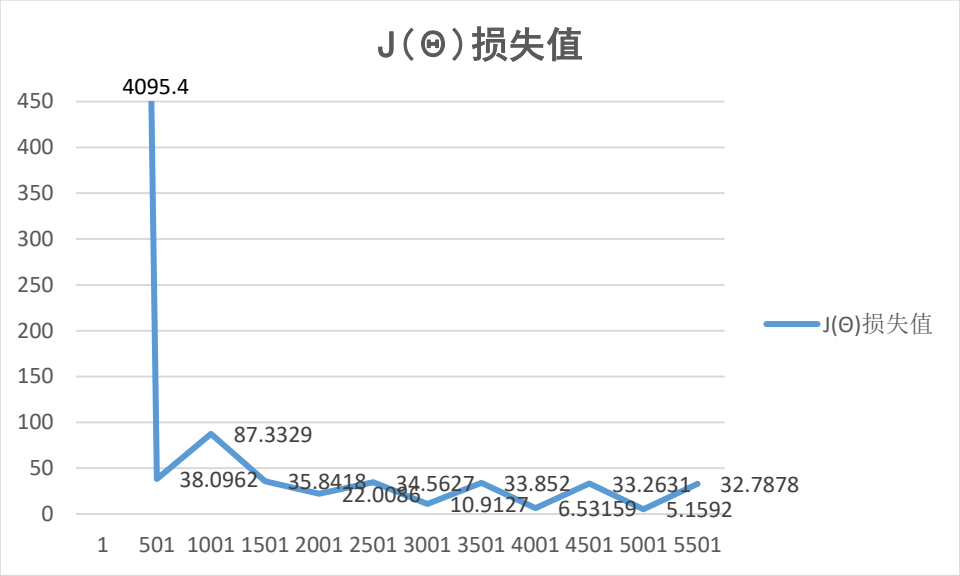
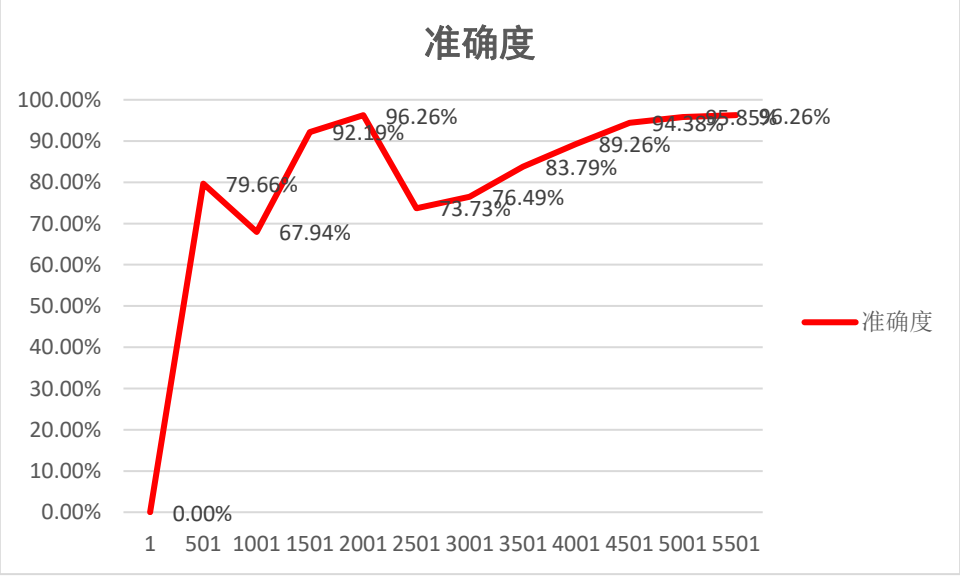


图 5. 2. 3(1) – J(Θ)损失值与轮数关系折线图



5. 2. 3(2) –精确度与轮数关系折线图

5. 2. 4 问题三模型求解

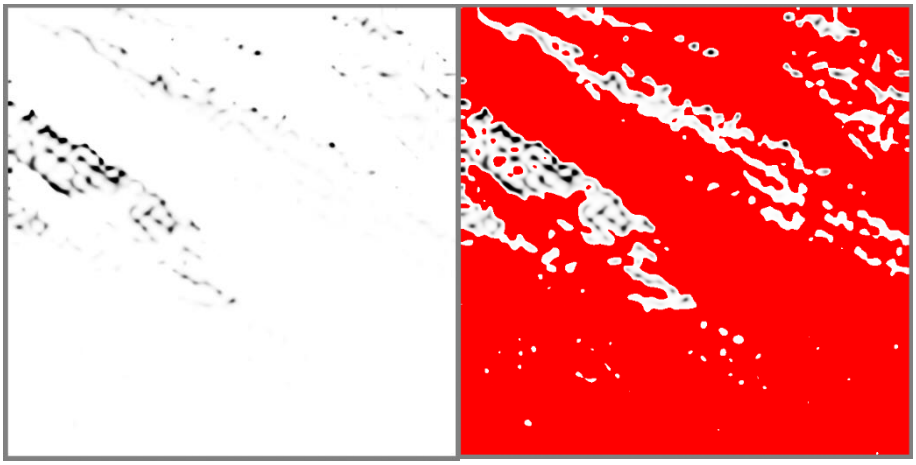
经过上述步骤，我们已经获得了一个对于区域植被指数分析与物候预测有一

定能力的模型。根据 2.3 的假设内容，即数据集的数据均真实可信。我们计算出模型的损失和准确度后，利用 2. 3 即可推广得到在损失值/像素点的误差下，得到的结果为正确的。

通过测试集得到模型的精度，表示在较短时间轴内精度为此值，随着远离时间轴原点精度会逐渐下降。因此时间轴的长度不宜设置过大，因为远离原点的预测因为此时精度太小,将会失去意义。

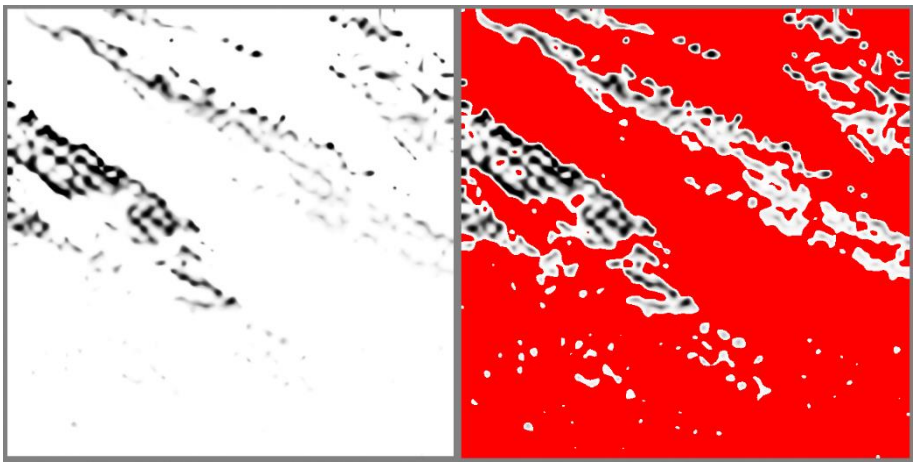
通过对于建立的机器学习模型输入时间，模型就能够预测出此时间标签对应的图像的信息。我们再手动对于当前图像进行分析。因为模型经过损失值的计算和测试集的检验等条件的约束，此时得到的图像为可靠图像，能够反映正确的信息，因此我们认为此时的手动分析为可靠分析。

以下为部分预测结果图像：



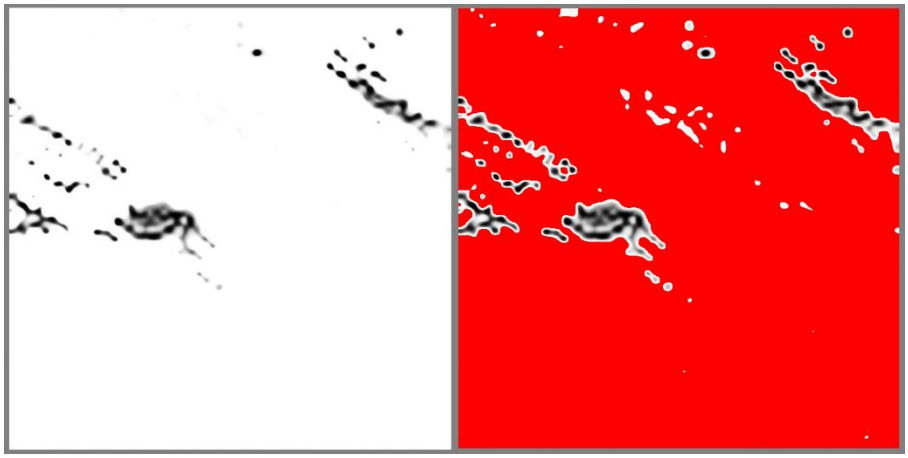
5. 2. 4(1) –第 264 个时间戳植被指数预测结果

左图为 Lanczos 插值算法等运算放大结果。右图为添加高光截切预览的结果。



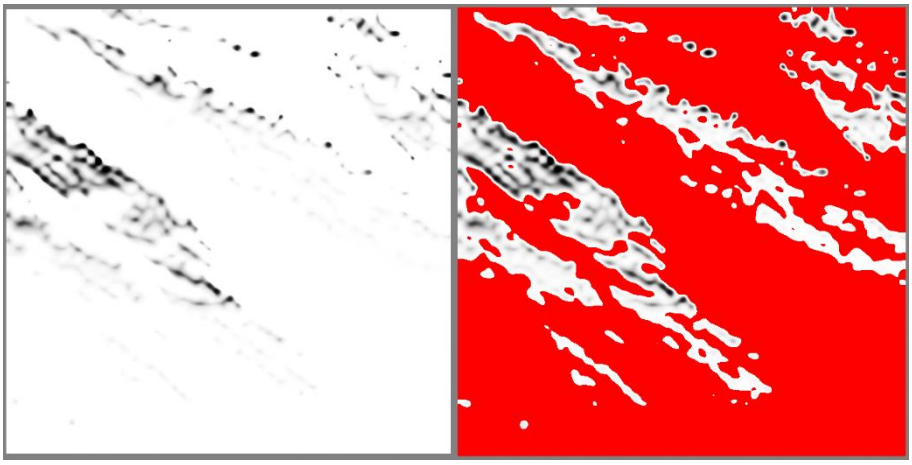
5. 2. 4(2) –第 270 个时间戳植被指数预测结果

左图为 Lanczos 插值算法等运算放大结果。右图为添加高光截切预览的结果



5. 2. 4(3) –第 276 个时间戳植被指数预测结果

左图为 Lanczos 插值算法运算放大结果。右图为添加高光截切预览的结果



5. 2. 4(4) –第 280 个时间戳植被指数预测结果

左图为 Lanczos 插值算法运算放大结果。右图为添加高光截切预览的结果

经过与 5. 1 数据预分析的对比，我们认为该预测结果接近第 11 周期的特征。植被指数图像符合 5. 1 的预分析结果。模型预测结果较为准确，对于区域植被指数的反应有现象级结论。虽然模型运算过程因为算力的问题，不可避免的存在有一定的精度损失，但是对于大致趋势有较为准确的反馈，即分为萌芽，茂盛，凋亡，次茂盛四个阶段，且每一阶段的时间符合预期结果。

5. 2. 5 问题四模型建立

拟合过程中通常都倾向于让权值尽可能小，最后构造一个所有参数都比较小的模型。因为一般认为参数值小的模型比较简单，能适应不同的数据集，也在一定程度上避免了过拟合现象。可以设想一下对于一个线性回归方程，若参数很大，那么只要数据偏移一点点，就会对结果造成很大的影响；但如果参数足够小，数据偏移得多一点也不会对结果造成什么影响，即『抗扰动能力强』。

因此，我们使用了规则项来约束模型的特征，即 L2 正则化算法。我们通过设置 L2 范数的规则项 $||\omega||_2$ 以约束每一个 ω 元素的大小。L2 正则化的惩罚因子设置为 1e-3，制约能力较强。通过权重衰减让参数向量的大部分元素都变得接近 0，压制了参数之间的差异性。从而缓解过拟合的问题，同时提升模型的泛化性能。

$$\omega^* = arg\min_{\omega} \sum_i L(y_i, f(x_i, \omega)) + \lambda||\omega||_2 \tag{13}$$

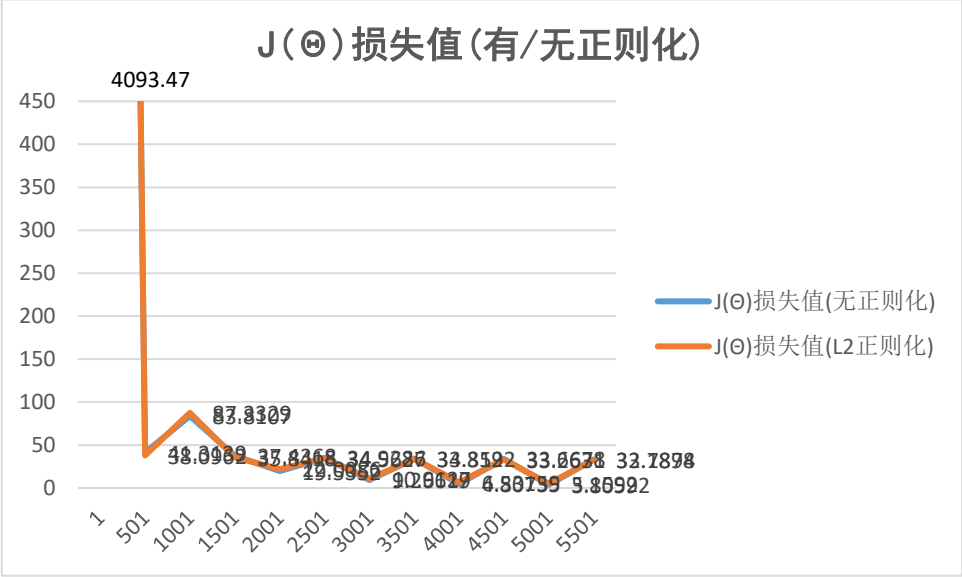
其中 L2 范数：

$$||\omega||_2 = \sqrt{\sum_{i=1}^{n} |\omega_i|^2} \tag{14}$$

5. 2. 6 问题四模型求解

如果没有使用正则化算法，通常会产生过拟合的现象，即过分拟合数据集结果。这样一来会造成模型损失值极小，但是实际上模型的预测能力却非常差。因为模型过分强调拟合，却忽略了可能存在的噪点和其他可能的情况。因此，不带正则化的模型因其预测能力的缺点，通常是不被使用的。

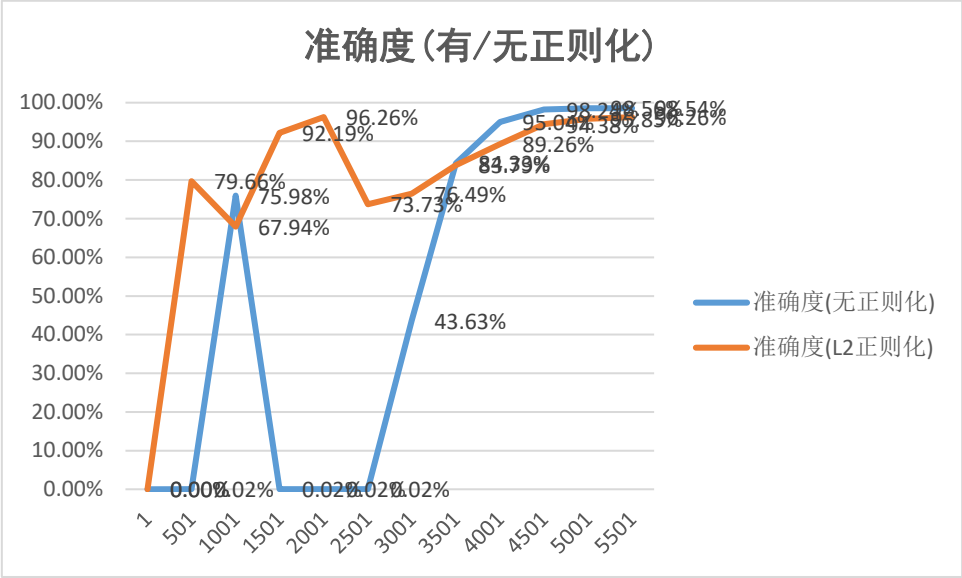
为了探究正则化对于模型的影响，我们分别对加入 L2 正则化算法和没有加入 L2 正则化算法进行建模。建模结果如下：



5. 2. 6(1) –损失值与轮数关系折线图

由上图可见，不使用正则化后时，损失值相较于使用 L2 正则化时稍微降低。每个取样点的 **J(Θ)**损失值大约小了 1. 5~2. 0。虽然看起来并不是非常大，但是在 **J(Θ)**值较小的点 **J(Θ)**值降幅甚至超过了 30%。由此可见，在不使用正则化的情况下，运算出的模型相对于真实值还是有非常明显的拟合上的进展。

但是，这是否也能体现在准确度上呢?



5. 2. 6(2) –精确度与轮数关系折线图

由上图可见，虽然不使用正则化可以在一定程度上降低 **J(Θ)**损失值，但是却会造成整体准确度不足，即模型训练效果很好，但是预测结果一塌糊涂的情况。因此我们在建立模型时，应该使用正则化算法。通过权重衰减让参数向量的大部分元素都变得接近 0，压制了参数之间的差异性。从而缓解过拟合的问题，提升模型的泛化性能。

5. 2. 7 问题五模型建立

假设数据集的数据均真实可信。我们计算出模型的损失和准确度后，利用 2.3 即可推广得到在损失值/像素点的误差下，得到的结果为正确的。通过测试集得到模型的精度，表示在较短时间轴内精度为此值，随着远离时间轴原点精度会逐渐下降。

当距离时间轴原点的距离在可靠范围内，因为模型经过损失值的计算和测试集的检验等条件的约束，此时得到的图像为可靠图像，能够反映正确的信息，因此我们认为此时的手动分析为可靠分析。

针对距离数据集时间轴较远的极端时间，预测结果将参照 Amazon Mechanical Turk 实验，使用人工对于预测结果投票。为了避免歧义，我们将一段时间内的预测结果评分进行加权平均，最终得到预测结果的分数。

因此模型结果大体可分两个部分，分别是可靠范围内的预测和远离可靠范围的极端预测。针对可靠范围的预测结果，我们采用直接的人工分析结果的方法，结合临近的时间结点的预测进行整体分析预测，最终获得分析结果。针对远离可靠范围的极端预测结果，将会加入土耳其机器人实验。先人工对于预测结果进行评分，判断预测结果是否符合实际情况，并对于预测结果进行投票评分。评分结果会与一段时间内其他预测结果的评分进行加权平均，最后确认改预测得到的分

数。如果分数高于可信阈值 60，则预测结果可信，否则预测结果不可信。我们拟定了加权分数计算公式：

$$s_{(i)} = 0.05 * s_{(i\pm2)} + 0.2 * s_{(i\pm1)} + 0.5 * s_{(i)} \qquad (15)$$

5. 2. 8 问题五模型求解

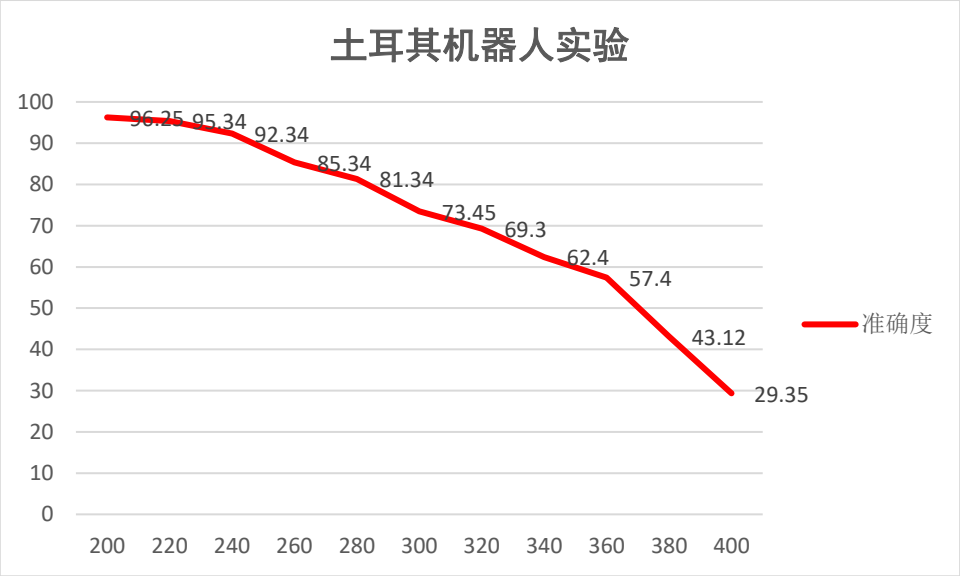
我们设计了损失计算函数分析 J(Θ)，测试集检验准确度来约束模型。同时我们针对距离数据集时间轴较远的时间的预测结果的参照 Amazon Mechanical Turk 实验，对于预测结果评分进行加权平均，最终得到预测结果的分数。

针对靠近数据集的数据，我们通过测试集运算获得近似结果。在靠近数据集的部分，我们可以认为当前测试得到的准确度能够反映正确的信息。因此，我们参考计算得到的精确度信息，即图 5. 2. 3(2)的终点数据 96.25%。

针对远离可靠范围的极端预测结果，使用土耳其机器人实验。人工评分加权平均，最后确认改预测得到的分数。如果分数高于可信阈值 60，则预测结果可信，否则预测结果不可信。

表 5. 2. 8 土耳其机器人实验结果

时间轴	得分
200	96.25
240	92.34
260	85.34
300	73.45
340	62.40
380	43.12



5. 2. 8 –土耳其机器人实验结果

由上表及上图可见，在远离时间轴原点的地方，所得预测结果的准确率会逐渐下降，并随着远离距离的增加，下降速度进一步增大。在靠近训练集的时间结点的预测结果得分较高，相对时间位置为 240 时，依然具有 90%以上的正确率;当时间轴进一步延伸后，当相对时间位置大于 300 时，具有低于 70%的正确率;此时，再将时间轴进一步拉伸，当相对时间位置达到 350 左右时，达到我们设置的可靠阈值。我们认为 350 附近的某点，存在阈值分界点。靠近时间原点精确度为预测值，当时间点延伸时精确度逐渐下降。当靠近阈值分界点时，精确度接近阈值 60。越过阈值分界点的预测值被认为不可靠预测，我们认为此时的预测结果不可信。

5.3 模型评价和改进

优点：模型使用 CNN 卷积神经网络模型对区域植被指数分析与物候进行预测，提出效用值代替效用函数，并结合统计调查的方法进行了验证。可以看出改进的 CNN 卷积神经网络模型是比较简便和有效的方法。

缺点：使用 MSE 均方误差计算 J(Θ)损失值，无法区分出此时的结果与实际情况的关系。该算法比较适用于线性的输出(如回归问题)，因为其对与真实值差别越大，则惩罚力度越大，这并不适用于分类问题。

改进：使用 CE 交叉熵算法计算代替。CE 交叉熵具有 MSE 均方误差不具有的优点，即通过避免 σ′ (·)的出现，避免学习速率降低的情况。CE 交叉熵算法在分类问题中有良好的应用。

缺点：数据集为 1200*1200 的图片，如果全尺寸进行运算，则会消耗大量的运算资源。但是如果使用 ANTIALIAS 抗锯齿算法将图片缩放为更小尺寸进行运算，则存在一定的精度损失。

改进：使用 ANTIALIAS 抗锯齿算法将图片缩放为更小尺寸进行运算后，使用 Lanczos 插值算法等一系列算法将图片进行一系列运算。计算出图片丢失的像素，可以一定程度上弥补缩放图片进行运算带来的精度损失，同时拥有很高的运算效率。

缺点：时间轴长度远大于假定的训练集时间跨度，因此只能通过土耳其机器人实验的思想划分偏离可靠范围的预测结果是否可靠，但是无法简单且准确的量化此时的准确度。如果大量使用土耳其机器人实验的思想，虽然可以获得较高的准确度，但是无法控制运算成本。

六、物候预测的影响

经济影响

物候预测可有效的对气候，水文及土壤等受环境影响而出现的自然现象进行

有效的预测，准确的预测可以有效减少物候现象带来的经济损失，在农作物种植方面，农作物的区划是推广栽培作物，合理配置作物的先决条件，而区划又受物候现象的影响，因此物候预测对解决农作物的种植与划分具有极大的参考价值，可以有效地根据物候来合理进行农作物的划分。同样的，物候预测对于养蜂，捕鱼，狩猎等与生物界有关的各种经济建设都有着实际影响，除此之外，在对山区土地的资源利用上物候预测可以很好的对很多没有进行调查的地方进行物候预测，在减少实地观测成本的同时还能合理的利用山区土地。对国民经济起着极大促进的作用。

社会影响

物候与人类社会息息相关，首先是人类种植业，农民在什么时段播种，在什么时段收获，都要根据物候来制定，在两千多年前，中国古代人民就把一年四季函数的变换分为所谓二十四节气，而物候预测对物候进行有效的预测可以对人类社会的耕种周期进行妥善的预测以便农民及其他与气候息息相关职业的安排工作进度与情况。对社会有着深远而积极的影响。

生态影响

物候预测得到的物候结果可以有效地改善生态系统，因为物候现象对生态系统功能与结构有着显著的影响，在气候变化背景下，植物的物候发生明显变化，克氏针茅草原近 20 年温度升高，降水量减少，土壤水分含量减少，羊草和克氏针茅返青期在该条件下明显滞后，相关分析显示返青前期土壤水分是导致反青滞后的重要原因。因此，通过物候预测提前对物候进行准确的预测，就能够提前对即将出现的物候及可能造成的生态影响采取有效的应对措施，使得物候对生态系统的负面影响降低。

七、参考文献

[1] 吴恩达. Coursera. 机器学习公开课.

[2] Google Tensorflow API r1.14(stable)。

[3] C. Rashtchian，P. Young，M. Hodosh，and J. Hockenmaier. Collecting image annotations using amazon’s mechanicalturk. In NAACL HLT Workshop on Creating Speech andLanguage Data with Amazon’s Mechanical Turk, pages 139–147，2010.