# 太阳黑子类型智能分类

水月苍穹

Asteroid2020 Group

版本：*0.01*

更新时间：*June 20, 2020*

# 1 赛题介绍

## 1.1 赛题背景

人类自诞生之日起，就饱受着极端天气的侵袭，如干旱、洪水、冰雹、台风等，威胁人类生命和财产的安全，阻碍社会经济的发展。随着人类科技文明的进步，大约在19世纪中期，人类发达地区开始遭受另一种灾害性空间天气——太阳风暴的危害。太阳风暴层造成当时无线电报网出现周期性大范围中断，太阳风暴引起地球磁场的剧烈变化，在地面诱生地磁感应电流，会损毁电网变压器，造成停电事故。而自人类进入太空以来，已多次经历卫星运行受太阳风暴影响的事例，卫星发生故障，材料退化，通信质量下降或中断，甚至导致卫星失效，提前退役。随着人类进行空间探索和空间活动不断深入，各种与人类生活息息相关的活动都依赖于导航、通信、广播、电视等空间技术系统，这些空间技术系统的安全已经不是单纯的科学问题，而是与国民经济、国家安全紧密相关的应用问题。

太阳黑子群又称为太阳活动区，是太阳风暴爆发的主要源区。在空间环境业务预报和研究中，活动区的形态特征和磁场特征通常被用作太阳风暴预报的主要参考因子，如威尔逊山磁分类、McIntosh分类、磁剪切、中性线长度等。在当前太阳观测数据量呈指数增长的情形下，传统的人工提取太阳活动特征的方法显然已经不能满足太阳风暴预报和空间环境预报的时效性需求，也不能满足广泛利用海量数据开展太阳风暴预报研究的要求。另一方面，太阳风暴预报模型主要通过统计关系来建立。基于人工经验提取的太阳风暴爆发的特征参量作为模型输入，观测数据中所包含的与太阳风暴相关的信息难以被充分利用，这限制了模式预报精度的进一步提高。

当前人工智能领域已发展出一系列处理大数据问题的方法。基于大数据和人工智能技术，很多行业的数据利用技术得到蓬勃发展，取得了许多前所未有的成果。面对当前太阳风暴预报受到的限制，有必要将海量太阳观测数据和人工智能技术相结合，发展新的太阳风暴识别和预报方法，提升我国空间环境预警预报能力。

## 1.2 赛程安排

### 1.2.1 线上赛（2020年6月28日——2020年7月26日）

- 每支参赛队伍可下载测试数据集（测试集标签不可见），本地调试算法，通过天池平台提交测试结果。
- 每支队伍每4日内有1次结果提交机会，若多次提交，以最后一次结果为准。排名更新日14时前提交结果，按照评测指标排序后，当天24时前更新得分及排行榜（排行榜将选择选手的历史最优成绩进行排名展示）。

### 1.2.2 线下颁奖及研讨会（拟2020年8月）

- 依据线上赛排行榜前三名，分别评为一、二、三等奖，举行线下颁奖仪式，并与国防科技创新特区相关专家组签订科研项目，参与下一阶段项目研究工作。
- 举办研讨会，邀请所有参赛队伍参加，作分享报告，展示模型分类、模型预警效果。主办方将从研讨会中选择具有创新思想和发展潜力的若干参赛队伍，进行项目支持。

## 1.3 竞赛题目

太阳活动区中黑子群的分类在太阳风暴预报中一直发挥着重要作用。然而，对该特征的识别还主要是通过人工判断的方式进行提取。针对当前太阳黑子群分类工作中面临的问题，本次比赛利用积累的海量太阳黑子磁场观测图像及其对应的磁类型标签，基于人工智能方法建立对黑子磁类型的自动分类模型。

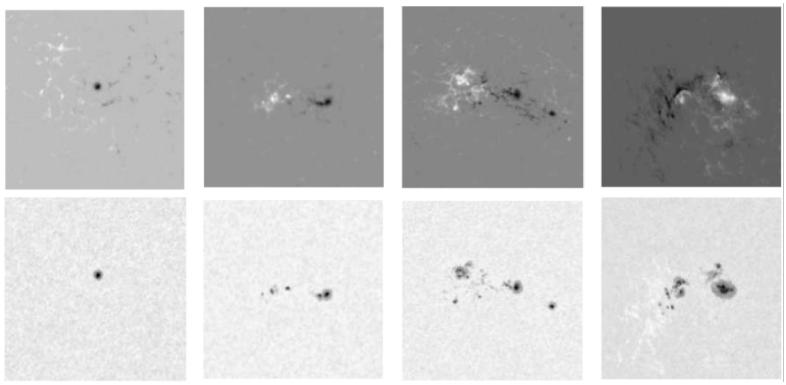


**图 1:** 如上分别为alpha（左1）、beta（左2）和更复杂的beta-x（左3、左4）磁类型的太阳黑子群磁场观测图像（上）和白光观测图像（下）。

## 1.4 评估指标

| | 识别的 A | 识别的非 A |
|---|---|---|
| 真实的 A | H | M |
| 真实的非 A | F | CN |

**图 2:** 评估指标

评分排名考虑的评估指数及优先级排序为：beta类的F1 score > betax类的F1 score > alpha类的F1 score。当多个队伍的优先指数相同时，比较次优先的评估参数进行排序。以下为各项评估参数的计算方法。计算单类样本的评估指数时，将三分类问题处理为二分类问题，以A类为例：

由此可计算召回率（Recall）、精确率（Precision）以及F1 score：

$$Recall = \frac{H}{H + M} \tag{1}$$

$$Precision = \frac{H}{H + F} \tag{2}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{3}$$

# 2 赛题分析

## 2.1 技术路线

- 每个太阳黑子群都有一张磁场观测图像和白光观测图像，是**双模态的图像分类问题**。每个太阳黑子对应的两张图像的大小是相同的，可以合并为一个双通道的输入。
- 每张输入图像的大小不一，是**任意尺寸图像的分类问题**。使用空间金字塔池化层(SPP)。由于框架不支持训练时输入不同大小的图像，可以使用TensorFlow提供的dynamic_pad将所有图像pad到相同的大小。
- 三个类别的图像数量不一样，是**类别不平衡问题**。备选方案有两种，一是过采样，二是更改损失函数。由于三个类别数据的数量差距不是很悬殊，估计这一块的影响不是很大，初期简单过采样即可。

## 2.2 参赛日记

### 2.2.1 2020.06.16

最初打算先尝试使用ResNet50进行finetune。为了满足ResNet50的输入条件，人工构造了三通道的输入数据，即con/75000，mag/5000+1以及两者相加con/75000+mag/5000+1。这样处理是因为continuum数据集中的min≈200、max≈73000，而mag数据集中min≈-5000、max≈5000。Finetune时，首先使用ImageNet上训练ResNet50作为BaseModel（去掉全连接），然后接上Global Max Pooling、Global Mean Pooling或SPP层，发现都不收敛。之后换了VGG-16、MobileNet等模型进行尝试，也均不收敛。此时排除了模型的问题，认为可能是数据导致这一问题。

### 2.2.2 2020.06.17

仔细观察数据集中的图像，发现magnetogram的图像存在很多噪声，严重影响了类别的辨识度，用人眼也很难分辨出各种类型的太阳黑子。而continuum的图像类别辨识度比较高，基本上alpha、beta和betax都有明显的区分度，于是修改TFRecord，仅使用continuum重新进行训练。此处设计了三个实验：

- 不对continuum数据做任何处理
- 对continuum数据进行归一化处理
- 对continuum数据进行截断、降噪、归一化

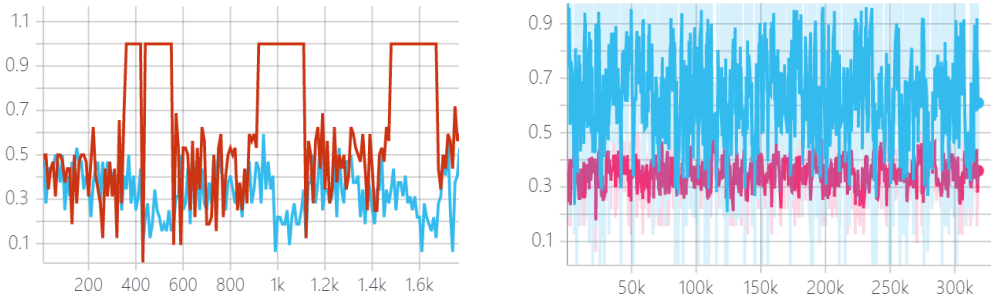
其中，实验二的公式为：

$$I' = \frac{I - min(I)}{max(I) - min(I)} \tag{4}$$

实验三的公式为：

$$I' = -\frac{clip(I, min(I), mean(I) - 2500)}{mean(I) - 2500 - min(I)} + 1 \tag{5}$$

实验发现归一化和降噪处理不能解决该数据的收敛问题，训练过程中的准确率曲线呈现周期性大幅波动，如下图所示：



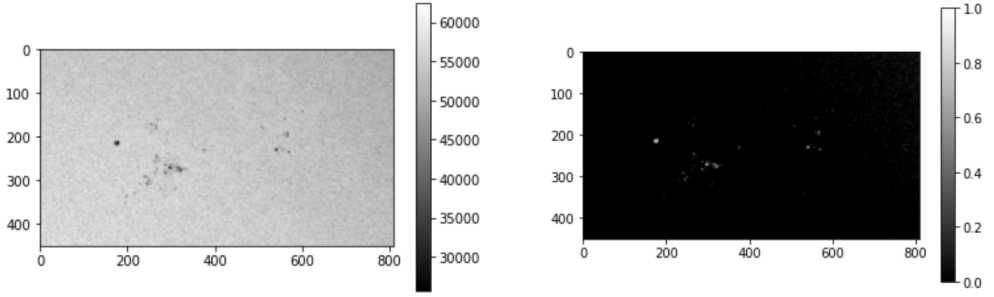

(a) 实验二的准确率曲线　　(b) 实验三的准确率曲线

**图 3:** 实验二和实验三的准确率曲线

### 2.2.3 2020.06.18

在昨天的实验三中，处理后的数据从视觉上看已经非常易于区分了，但模型训练仍然不收敛，怀疑是这个数据集在可变输入大小的设置下无法训练。查阅资料得知，无论是全局平均池化和SPP层都是可以训练收敛的，但这个数据集和别的数据集的区别是，别的数据集上输入的图像虽然每个批次大小都不同，但大体上整张图像都是待识别的目标；而在这个数据集中，待识别的太阳黑子仅占输入图像很小的一部分，输入的大部分都是噪声。可能就是这个原因导致了训练不收敛。

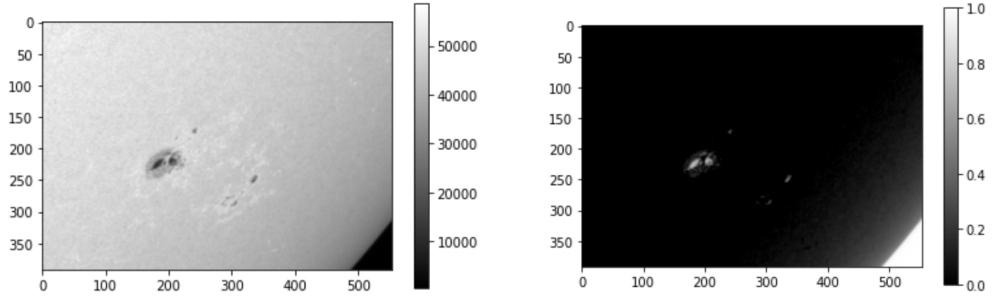
对于这一问题有两个方法可以尝试，一种是将图像切割成固定的大小，一种是人工从图像中提取特征。这里首先尝试的是第二种方法，如下图所示，左侧是原始输入图像，右侧是按公式5处理后的图像。



(a) 原始输入图像　　(b) 按公式5处理后的图像

**图 4:** 处理前后的太阳黑子图像 (continuum)

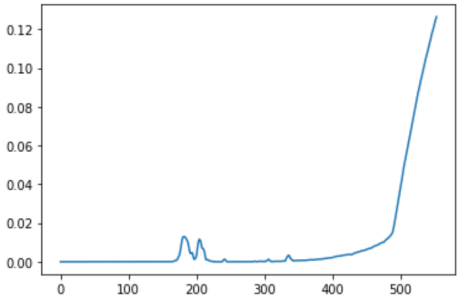
观察数据时还发现，有许多图像拍摄到了宇宙背景，宇宙背景和太阳黑子的观测数值类似，这会对图像的切割造成影响。如下图所示。



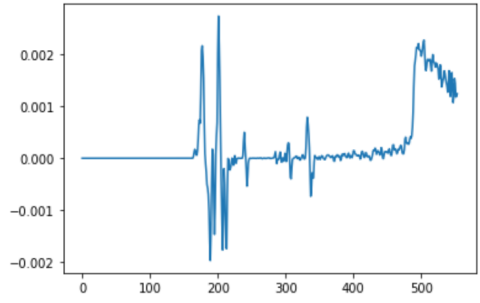
(a) 原始输入图像　　(b) 按公式5处理后的图像

**图 5:** 处理前后的太阳黑子图像-宇宙背景 (continuum)

图5(b)沿x轴方向的方差以及方差的梯度如下所示：




(a) 图5(b)的x轴方差

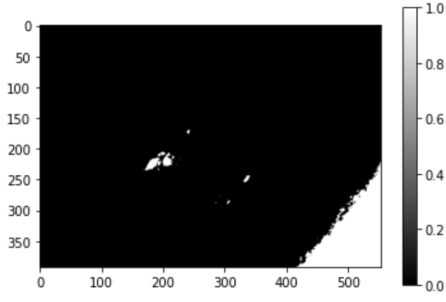



(b) 图5(b)的x轴方差的梯度

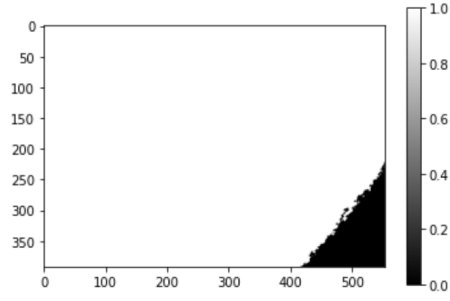
**图 6:** 图5(b)沿x轴方向的方差以及方差的梯度

为消除宇宙背景对方差和梯度造成的影响，首先对图5(b)进行二值化（太阳=0，太阳黑子和宇宙背景=1），再对该图像的四个角取10*10的小区域计算均值，若均值>0.9，则认为这个角是宇宙背景。

于是从这个角出发，使用洪泛法进行填充（太阳=0，太阳黑子=1，宇宙背景=2）。再次二值化，得到一个蒙版（太阳与太阳黑子=1，宇宙背景=0）。将蒙版与图5(b)相乘，得到了最终的图像（图8）。




(a) 图5(a)二值化




(b) 图7(a)洪泛法得到的蒙版

**图 7:** 二值化与洪泛法处理

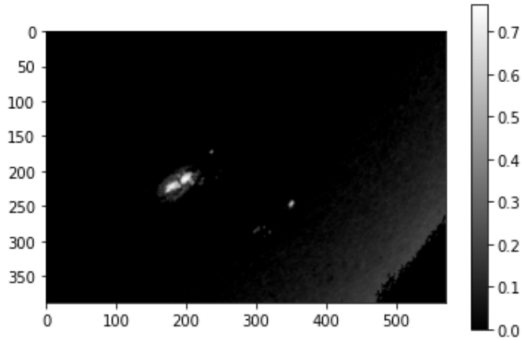
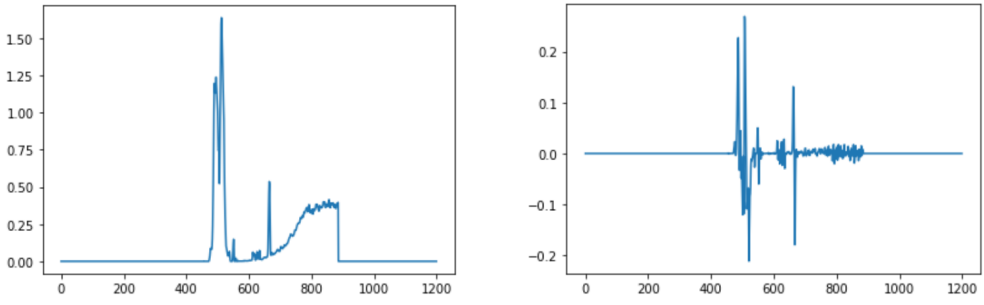



**图 8:** 最终得到的图像

对图8重新计算沿x轴方向的方差以及方差的梯度，结果如下图所示。经过这样处理，每张图像都转换为了长度为1200的两个向量，可以使用一个简单的全连接神经网络进行分类。




(a) 图8的x轴方差　　(b) 图8的x轴方差的梯度

**图 9:** 图8沿x轴方向的方差以及方差的梯度

### 2.2.4 2020.06.19

今天对数据集的具体信息进行了统计：

- alpha、beta、betax三类样本的文件数目分别4709、7353、2407
-
-