

# 毕设讲稿

## 讲稿部分

【PPT 1】

[开头语]

【PPT 2】

我的汇报主要分成 4 个部分：**研究背景**；**研究过程**；**实验结果** 和 **总结分析**。

【PPT 3】

首先介绍 “研究背景”。

【PPT 4】

台风作为一种自然灾害经常带来广泛的经济损失、生命损失和社会活动的中断。因此高准确度的台风强度和路径的预测的研究工作就显得尤为重要。传统的台风预报方法的问题之一在于单一的数据源，这可能是预测结果与实际相比出现偏差的来源之一。

本文研究的基于多源数据融合的台风强度和路径预报可以在一定程度上提高台风预报的准确性和可靠性，帮助相关部门更及时地做出灾害应对和减灾措施。

【PPT 5】

现有的对于台风路径和强度的预测的方法可分为 3 种：动力学方法，统计方法和统计-动力学方法。

【PPT 6】

统计模型的发展历程包括 2016 年提出并使用的 CNN 模型和 RNN 模型；以及 2019 年提出的 LSTM，即长短期记忆和同年提出的卷积长短期记忆。

【PPT 7】

本研究的主要贡献在于：

1. 将单一的数据源扩展为多源数据；

2. 将单一的模型扩展为多种模型集成的架构；
3. 最后本研究在 MSE 和 MAE 的评估指标的基础上引入了台风相似度定量分析方法对预测台风和实际台风进行对比。

#### 【PPT 8】

接下来介绍 “研究过程”

#### 【PPT 9】

研究过程主要分为 4 步：

1. 论文研读
2. 数据预处理
3. 模型实现（以及）
4. 模型评估

#### 【PPT 10】

本文使用来自两个来源的数据：**SHIPS 开发数据库** 和 **在线发布的用于强度回归的基准TC图像数据集（TCIR）**。将这两个来源的数据分成图像数据和季后分析数据作为模型训练的数据集。

#### 【PPT 11】

4 种类型的 TC 图像包括红外图像、水汽图像、可见光谱图像和微波图像。

#### 【PPT 12】

季后分析数据如PPT上的图片所示，除了基本的台风信息以外还有一些 SHIPS 环境参数，例如，SHRD：垂直风切变（Shear）等。

#### 【PPT 13】

在训练模型之前对训练数据的预处理主要分为以下几个步骤：

1. 将上述搜集到的数据储存在一个格式为 HDF5 的文件中；
2. 将数据集划分为 “训练集” “验证集” 和 “测试集” 3 个部分；
3. 接着，格式化单个台风的记录序列，并对路径数据用 `MinMaxScaler` 函数进行归一化；
4. 切割时间序列数据集，使得输入满足 LSTM 模型的期望；

然后预处理完成。预处理之后的数据将作为模型输入。

## 【PPT 14】

这是最终使用的预测模型的架构图。我们的模型主要分为三个部分：编码和CCA模块、RNN模块、以及输出回归器和路径预测器。

### 首先看 编码和CCA模块

从左侧输入的多时序卫星图像进入模型。每张图像尺寸为 $128 \times 128 \times 2$ ，表示在不同时间点获取的图像序列。经过两个卷积神经网络（CNN）层进行特征提取，每层提取到的特征维度为 $15 \times 15 \times 32$ 。然后，这些特征进入通道注意力机制（CCA）模块，进一步增强关键特征的表达。

### 接下来是 RNN 模块

特征提取和增强后，数据输入到RNN模块。RNN模块由多个卷积长短期记忆网络（convLSTM）层组成，这些层能够捕捉时序特征，进一步处理图像序列的时间依赖关系。每个时间步的输出特征维度为 $10 \times 12 \times 64$ 。

### 最后是 输出回归器

在RNN模块处理后的特征与 SHIPS 环境参数 通过concat操作合并，进入输出回归器部分。输出回归器包含一个CNN层和多个全连接（Dense）层，用于预测24小时后的台风最大风速（Vmax）。预测结果通过一个全连接层输出，维度为[1]。

### 同时另一部分

路径预测器 接收两个输入：一个是时间序列路径数据（维度为 $[9 \times 2]$ ），这里的 9 是时间序列的长度，2 代表每个序列的维度，维度为 2 对应 [经度, 纬度] 两个变量，另一个是辅助特征数据（维度为 x）。路径数据通过 LSTM 层 处理，并通过全连接层输出，预测24小时后的台风经纬度位置（lon和lat），输出维度为[2]。

### 最终输出

最后，我们将输出回归器和路径预测器的结果合并，生成最终的输出：24小时后的台风最大风速（Vmax）和台风的经纬度位置（lon和lat），其维度为[3]。

## 【PPT 15】

在每个训练周期结束之后，我们以 MAE（Mean Absolute Error）和 MSE（Mean Square Error），即绝对均方误差和平方均方误差作为模型的评估指标。

#### 【PPT 16】

此外我们还使用了台风相似度作为评估的指标之一。下面将介绍相似度的计算方法。

1. 首先假定有两个台风向量，标识为 B, C. 每个向量的第  $i$  个分量表示为某一时刻台风各种属性集成的子向量.
2. 然后我们定义每个子向量（包含台风在某一时刻下的不同属性）之间的距离为各个属性分量的加权平方和（ $f_i$  是权重函数）
3. 接着我们构造一个关于台风 B, C 的距离矩阵. 矩阵某一个位置  $(i, j)$  上的值对应 B 的第  $i$  个分量和 C 的第  $j$  个分量的距离，也就是我们上一步定义的距离函数
4. 在得到距离矩阵 Dis 之后，我们在此基础上按照给出的规则进行更新，得到一个新的积累矩阵 Acc.
5. 最后对积累矩阵的最后一个元素开根号，就得到我们想要的 B, C 两个台风的相似距离

#### 【PPT 17】

上一页 PPT 介绍的计算方法中最后得到的是两个台风之间的相似距离。在得到相似距离之后，还要使用归一化函数将相似距离归一化。归一化函数选择双曲正切函数。

最后将归一化后的相似距离转化为相似度。

PPT 左边的图展示了相似度计算的 Python 代码。

#### 【PPT 18】

这一页 PPT 展示了实验设置的参数以及实验使用的计算资源的情况。

#### 【PPT 19】

实验一共进行了 30 个周期，这是实验的日志信息。

#### 【PPT 20】

最后展示一下实验结果。

#### 【PPT 21】

首先来看路径预测的性能表现。图片展示了预测路径的 MSE 随时间的变化曲线。颜色较深的曲线为验证集上的 MSE，颜色较浅的曲线为测试集上的 MSE。

可以看出，随着训练周期的增加，模型路径预测的路径的损失值逐渐下降，说明模型的性能在不断提高。

同时也可以看出，模型在验证集上的损失值与在测试集上相差不大，说明模型没有出现过拟合或欠拟合的现象。

#### 【PPT 22】

这是编号为“200402W”的台风的预测路径和实际路径的对比图。红色路径是台风的真实数据，蓝色部分是模型的预测路径。虽然模型在 MAE 和 MSE 指标上均有收敛，但是由于台风位置（经度、纬度）对于模型的精度要求较高（相差为1的经度在地图上已经是很远的距离了），所以尽管预测路径已经十分靠近实际路径，但是拟合度仍有待提高。

#### 【PPT 23】

接着来看台风强度预测的性能表现。这里选取了两个不同的台风，201303C 和 201402C 台风。可以看到模型预测的性能表现在 201303C 中表现较好，但是在 201402C，尤其是从第 7 个点开始，TC 的强度急剧上升，但是模型缺乏对这种极端天气的预测能力。

#### 【PPT 24】

最后来看预测台风和实际台风相似度的变化曲线。可以观察到台风相似度从最开始的低点开始上升，最后稳定在 87% 以上。说明模型总体来说性能良好。

#### 【PPT 25】

最后总结一下本研究的工作以及自我的评价。

#### 【PPT 26】

本课题基于多源数据，包括图像数据和季后分析的 TC 相关信息作为训练数据，集成了不同的模型，提出了一种新颖的机器学习的模型架构来对台风强度和路径进行预测。

创新性地使用了对台风相似度评估的量化方法来对预测台风和历史台风进行相似度方面的评估。

总体而言，基本上达到了课题的预期。

#### 【PPT 27】

尽管总体而言基本上达到了实验的预期成果，但是本次工作还有非常多的不足。其中包括：

##### 1. 对比实验的欠缺

由于计算资源的限制，本实验相对缺乏与其他预测模型的对比实验。

## 2. 损失函数有待进一步优化

本实验中，模型的损失函数为强度和位置数据的均方误差（MSE）的简单求和。然而，这两个方面的比率可能会相互干扰，导致某一方面的预测性能表现不够理想。未来的工作可以考虑采用更复杂的损失函数，如加权和或逐步训练的方式，以更好地平衡强度和位置的预测性能。

## 3. 地理信息的处理方面的优化

台风路径的预测对模型精度要求非常高，而地球表面的经度和纬度之间的差异可能导致预测路径与实际路径之间的较大偏差。为了解决这个问题，未来的研究可以考虑引入更精细的地理信息处理方法。

【PPT 28】

我的汇报到此结束，谢谢大家

# 提问部分

### 问题 1：模型中使用了哪些数据源？这些数据源是如何融合的？

**回答：**模型主要使用了两类数据源：多时序卫星图像和路径数据。卫星图像通过CNN进行特征提取，并经过通道注意力机制（CCA）模块进行特征增强；路径数据则通过LSTM层进行处理。特征融合主要在RNN模块中实现，通过卷积长短期记忆网络（convLSTM）处理时间序列数据，并在输出阶段通过concat操作将特征融合用于预测。

### 问题 2：为什么选择convLSTM来处理时序特征？

**回答：**convLSTM结合了卷积神经网络（CNN）和长短期记忆网络（LSTM）的优势，既能捕捉图像的空间特征，又能处理时间序列的依赖关系。因此，convLSTM特别适合用于处理时序卫星图像数据，能够更好地提取时空特征，提高预测的准确性。

### 问题 3：CCA模块的作用是什么？

**回答：**通道注意力机制（CCA）模块用于增强关键特征的表达。它通过对不同通道的特征进行加权，突出重要的特征，抑制不重要的特征，从而提升特征提取的效果。这对于后续的时序处理和预测非常重要。

### 问题 4：模型如何处理异常数据或噪声？

**回答：**在数据预处理阶段，我们对输入的数据进行了标准化处理，并使用了异常值检测和清洗方法，尽量减少噪声和异常数据的影响。此外，模型的多层结构和注意力机制也有助于增强对噪声的鲁棒性，提高预测的稳定性和准确性。

### 问题 5：模型预测的准确性如何评估？

**回答：** 我们使用了多个评估指标来衡量模型的性能，包括平均绝对误差（MAE）、均方误差（MSE）和R平方（ $R^2$ ）等。通过对历史台风数据的预测结果与真实值进行比较，计算这些评估指标，从而评估模型的预测准确性。

## **问题 6：在训练模型时，如何处理过拟合问题？**

**回答：** 我们采用了多种方法来防止过拟合，包括：

- **数据增强：** 通过对图像进行旋转、翻转等操作增加训练数据的多样性。
- **正则化：** 在模型的卷积层和全连接层中添加L2正则化项。
- **Dropout：** 在LSTM层和全连接层中加入Dropout层，随机丢弃一部分神经元，防止模型过度依赖某些特征。

## **问题 7：模型的训练和推理时间如何？**

**回答：** 模型的训练时间依赖于数据量和计算资源。一般来说，训练需要数小时到一天不等，具体时间取决于硬件配置（如GPU）。推理时间相对较短，一般在秒级以内，能够满足实时预测的需求。