**基于LLM和卫星图像的社会公共卫生数据发掘与建模**

**项目背景、目的及意义**

**背景**

近年来，亚非拉地区的一些发展中国家因为长期的战乱、医疗资源匮乏以及基础设施不完善等多重因素，面临着严重的公共卫生挑战。同时，由于这些地区的社会公共卫生数据收集和监测体系相对薄弱，使得政府和国际组织在制定针对性的卫生资源分配和干预措施时面临巨大困难。目前，对于这些国家的公共卫生数据调研主要依赖于实地考察。例如，联合国等国际组织每年都会派遣研究团队，花费大量的经费和时间进入第三世界国家进行深入调研。然而，由于地域广阔、人力物力投入巨大等问题，这种方式的数据采样全面性和实时性往往难以得到保证。因此，我们亟需一种新的、高效的数据收集和分析方法，以解决这些问题。

**目的**

本研究的主要目标是通过运用大语言模型（LLM）强大的总结归纳能力，对已有的大量社会公共卫生数据进行深度挖掘和分析，同时结合实时的卫星图像数据，构建一种能从卫星图像中预测相关统计数据的深度学习模型。该模型将有助于我们更准确地预测各国或地区的公共卫生指标，如新生儿死亡率、传染病感染率等。通过该研究，我们希望能够提供针对亚非拉地区国家的公共卫生情况的准确预测，为政府和国际组织制定针对性的卫生干预措施提供科学依据。

**意义与实施必要性**

通过卫星图像数据，我们可以实时地获取第三世界国家的各项关键指标，如人口分布、城市化程度、公共卫生设施分布等。该方法不仅可以降低相关公共卫生研究和政策制订的成本，而且可以减少研究人员进行相关实地调研的频次，降低相关数据的调研成本。

此外，本研究有助于拓展卫星图像在公共卫生领域的应用。卫星图像的实时性、全覆盖性和客观性，使其成为了解和监测地球表面变化的重要工具。通过深度学习技术，我们可以从卫星图像中提取出更多有价值的信息，为解决亚非拉地区的公共卫生问题提供新的思路和方法。这种新的研究方法，不仅可以为政府和国际组织制定更有效的公共卫生政策提供科学依据，而且可以进一步推动国际社会对第三世界国家的关怀和援助。

**研究现状**

目前，一些研究团队已开始尝试利用卫星图像数据来预测公共卫生指标。例如，利用深度学习模型结合多源数据进行传染病预测、基于卫星遥感数据预测空气质量等研究已经展开。这些研究为利用先进技术解决公共卫生问题提供了范例，并为相关领域的研究提供了参考。

**发展动态**

随着大语言模型的不断升级，大语言模型在数据发掘和整理领域将会得到更深入的应用。同时，随着计算机视觉技术和卫星遥感技术的不断发展，利用深度学习结合卫星图像数据进行公共卫生指标预测的研究将更加深入和广泛。未来，可以进一步探索如何更好地整合多源数据、提高模型预测精度。此外，应关注研究成果的实际应用和政策推广，使其能够真正造福于亚非拉地区国家的公共卫生事业。

**内容简介**

我们希望验证基于卫星图像的神经网络对社会学统计数据预测的可行性。我们将选择1~3个典型社会学统计指标作为验证集，并选择合适的国家及地区作为模型验证的样本，利用LLM及多模态的LLM收集相关数据进行分析提取和整理，再对收集的卫星图像数据集进行处理和标注。最后在该数据集上完成神经网络模型的设计、训练和验证，评估该模型的预测准确度，最终达到对国家/地区进行连续的预测，生成Grid Map。

**项目研究条件与创新之处**

1研究积累

LLM大语言模型具备多模态数据挖掘、分析的能力

深度学习模型的评估和优化方法较为完善

已经有一些团队在基于卫星遥感图片的数据研究上取得成果

2条件

与斯坦福大学沟通，达成项目合作并初步验证了项目的可实现性

获取了印度地区精确到村庄级的公共卫生调研散点数据

3优势

本项目多学科交叉，结合公共卫生、地理信息、统计、计算机等多个领域

当前在这一领域的交叉研究较为有限

本项目与斯坦福大学的团队进行合作

对于亚非拉地区的公共卫生情况评估和决策具有广阔的应用前景和意义

4风险

公共卫生等社会学数据可能涉及数据安全问题

建立的模型可能存在泛化能力不足过拟合的问题

模型的可解释性不足，难以解释模型的预测结果和决策过程

数据发掘得到的数据量不足、精确度不够

5创新

构建卫星图像和社会公共卫生数据结合的训练模型，拓展了传统方法和技术的交叉应用

利用大语言模型等先进技术，挖掘处理数据

成果可应用于公共卫生决策、社会发展等多个领域

多学科相交叉，涉及公共卫生、地理信息、统计、计算机等多领域

**项目预期成果**

1. 论文成果：总结项目中涉及的卫星图像处理技术、大语言模型应用技术和统计数据挖掘方法，为相关领域的研究和应用提供参考和借鉴。

2. 系统开发：建立基于实时卫星图像的社会公共卫生数据预测模型，开发亚非拉等地区公共卫生指标的预测系统。

3. 验证评估：验证该模型的准确度和泛化能力，展示项目成果在社会学数据统计领域的应用潜力和价值。

4. 推广成果：将模式推广到社会学研究领域，为该领域研究开辟新的道路。

5. 社会效益：实现对亚非拉等地区社会学数据的预测，为政府和国际组织制订针对性的政策提供科学依据，推动对第三世界国家的人文关怀。

项目方案

**拟解决的问题**

当下社会公共卫生数据的统计成本较高，尤其是在一些经济条件相对落后的亚非拉地区发展中国家。我们的目标是利用现代的大语言模型（LLM）和深度学习技术，通过LLM从网络批量收集数据进行提取和整理，并结合卫星图像利用深度学习来进行相关数据的预测与估计，从而大大降低数据统计的成本。我们希望这种模式能够得到推广，进而推动国际社会对第三世界国家的人文关怀与援助。

**主要内容**

1. **前期准备**

在导师的指导下，讨论选出1~3个典型社会人口统计指标作为验证集，并选择合适的国家及地区作为模型验证的样本。

1. **数据收集与整理**
   1. 利用LLM收集与该统计学数据相关报告与文献。
   2. 利用多模态的LLM对收集到的内容进行分析，提取其中的具体数据并进行整理。
2. **数据标注**

从GoogleEarth等卫星地图平台上下载对应地区的卫星图片，对卫星图片的尺度进行统一调整，并根据划分的区域进行分组，完成图片的数据标注。

1. **模型设计与训练**
   1. 阅读与学习神经网络相关论文，选择合适的神经网络的模型，完成该项目的模型设计
   2. 将目前已有的高精度散点数据划分为训练集和测试集，在训练集上进行模型预训练，并根据训练结果及时调整模型及参数，达到更好的训练结果。
   3. 利用预训练模型，估计从网络上收集的数据的准确度。
   4. 加入从网络收集的大量数据扩充数据集，对模型进行训练和优化。

**5. 模型验证**

在测试集上评估该模型的预测准确度。

**6. 模型应用**

通过该模型对国家/地区进行连续的预测，生成Grid Map。

**计划目标**

利用LLM搭建一个数据发掘系统，收集足量的有效数据集，进行最终模型的设计和训练，完成从实时卫星图像到公共卫生指标的预测，并检验其准确度和泛化能力，验证该模式可以推广到实际应用。

**思路方法**

我们的方法主要基于两个相对成熟的技术：LLM和神经网络。

* 我们将借助当前LLM的强大资源检索能力，收集各地区的各种模态的数据；并利用多模态大模型的总结归纳能力，系统性地发掘已有的社会公共卫生统计数据。
* 我们将使用神经网络，以卫星图像作为输入，发掘卫星图像和相关统计数据的联系，并进行预测和估计。
* 最后我们将使用精细采样的数据进行验证，评估模型的准确度和泛化能力。

**组织实施**

* 每周小组内进行讨论，交流本周的成果，并制定下一周的工作目标。
* 每月向导师汇报项目进度，听取导师的指导与规划。

**进度安排**

* 2024年8月前，编写脚本，完成足量的数据收集与整理。同时进行模型的初步设计。
* 2024年10月前，编写脚本，完成卫星图片的下载、整理与标注，完成数据集的制作。
* 2024年12月前，完成神经网络模型在离散数据点上的预训练。
* 2025年3月前，完成神经网络模型的最终训练，并完成最终的数据验证。使用模型生成最终成果Grid Map。
* 2025年5月前，完成论文和结题报告撰写