文本

描述已自动生成

**开源大模型的下游应用开发**

（题目待定）

如果你对人工智能近期的发展感兴趣，你一定会知道GPT4（OpenAI）、Bard（Google）、NewBing（Microsoft）这类语言大模型，也知道SAM（MetaAI）这类图像大模型。对大模型的开发是现在人工智能大公司共同的发展趋势。在代码开发中，Caploit（Microsoft），Cursor（OpenAI）让人惊讶的发现，如果我们不会使用这些工具代替我们重复性的工作，我们将要花费几倍的时间来完成我们的代码。反之，如果我们只能生产低端的代码，我们很快将会被知识面更广、生产力更高的大模型所取代。在这个时代，如果我们不会使用AI，将会被会使用的人使用AI所取代。下面先介绍一下大模型是什么，大模型有什么问题，我们能做什么。

## 一、大模型是什么？

### 1.1 大模型的定义和特点

大模型（Large-scale Model）是指具有超大规模参数量和计算量的深度神经网络模型，通常需要海量的数据和算力来进行训练和推理。大模型具有以下几个显著的特点：

**泛化能力强**：大模型可以从海量数据中学习到丰富和通用的知识表示，从而提高在不同任务和领域上的泛化能力，减少对标注数据的依赖。

**可迁移性好**：大模型可以通过预训练和微调的方式，将预训练阶段学习到的知识迁移到下游任务上，从而降低开发成本和难度，提高开发效率和效果。

**可解释性差**：大模型由于参数量庞大，结构复杂，内部机制难以理解，因此很难对其进行有效的可解释性分析，导致其在一些敏感和关键的领域难以应用。

**资源消耗高**：大模型需要消耗大量的数据、算力、存储和能源等资源，不仅增加了训练和推理的成本和时间，也带来了环境和伦理等方面的问题。

### 1.2 大模型的分类和代表

根据大模型所涉及的数据类型和任务领域，可以将其分为以下几类：

**自然语言处理（NLP）领域的大模型**：这类大模型主要利用文本数据进行预训练，学习语言的语法、语义、逻辑等知识，然后在下游任务上进行微调或生成，如机器翻译、问答、文本摘要、对话系统等。代表性的大模型有BERT、GPT、T5等。

**计算机视觉（CV）领域的大模型**：这类大模型主要利用图像或视频数据进行预训练，学习视觉的特征、概念、关系等知识，然后在下游任务上进行微调或生成，如图像分类、目标检测、人脸识别、图像生成、风格迁移、视频理解等。代表性的大模型有ViT、CLIP、DALL-E等。

**多模态（Multimodal）领域的大模型**：这类大模型主要利用多种类型的数据进行预训练，学习不同模态之间的关联和融合，然后在下游任务上进行微调或生成，如图文检索、视觉问答、视频字幕等。代表性的大模型有UNITER、M2M-100、SimCLR等。

**强化学习（RL）领域的大模型**：这类大模型主要利用与环境的交互数据进行预训练，学习环境的动态和策略的优化，然后在下游任务上进行微调或生成，如游戏、机器人、自动驾驶等。代表性的大模型有AlphaGo、MuZero、D4PG等。

## 二、我们能做什么？

### 2.1 下游应用开发

大模型下游应用是指利用预训练好的大模型，通过微调或其他方法，来解决特定领域或任务的应用。大模型下游应用的优势是可以利用大模型学习到的丰富的通用知识和能力，提高应用的性能和效率。大模型下游应用的挑战是如何适应不同的数据分布和目标函数，保留和更新预训练知识。

实际生活中，我们要的不是一个能覆盖所有领域、解决所有问题的模型。大多数时候，人们只是想要一个生产力较高的，覆盖自身需要的领域的模型。而实际上，大多数大模型也无法完成复杂的工作，仅对某一领域有较强的泛化能力，例如GPT，目前仅在文字理解上表现出较大的潜力。然而，我们可以基于此，进行下游应用的开发，利用其在文字理解上的强大能力，帮助人们用自然语言和其他任务进行交互。

**自然语言处理**：自然语言处理是指使用计算机来理解和生成自然语言的技术。大模型在自然语言处理领域有很多应用，如机器翻译、自动摘要、情感分析、问答系统、对话系统等。例如，Google使用基于Transformer的大模型来实现其谷歌翻译服务1；微软使用基于GPT-3的大模型来实现其PowerPoint智能设计服务。

**计算机视觉**：计算机视觉是指使用计算机来理解和生成图像或视频的技术。大模型在计算机视觉领域也有很多应用，如目标检测、图像分割、图像生成、视频理解等。例如，Facebook使用基于ViT的大模型来实现其Instagram Reels视频推荐服务；OpenAI使用基于DALL-E的大模型来实现其图像生成服务。

强化学习：强化学习是指使用计算机来学习复杂的决策策略的技术。大模型在强化学习领域也有很多应用，如游戏、机器人控制、自动驾驶等。例如，DeepMind使用基于MuZero的大模型来实现其AlphaGo Zero和AlphaZero等人工智能棋手；百度使用基于PaddlePaddle的大模型来实现其Apollo自动驾驶平台。

**增强现实**：增强现实是指将虚拟信息与现实世界结合的技术。大模型在增强现实领域也有很多应用，如物体识别、场景理解、虚拟试衣等。例如，Snapchat使用基于SnapML的大模型来实现其Snap Camera滤镜服务；阿里巴巴使用基于PAI的大模型来实现其天猫魔镜虚拟试妆服务。

医疗影像处理：医疗影像处理是指使用计算机来分析和诊断医疗影像的技术。大模型在医疗影像处理领域也有很多应用，如肿瘤检测、病理诊断、骨龄评估等。例如，腾讯使用基于Tencent AI Lab MedicalNet的大模型来实现其腾讯觅影医疗影像平台；阿里巴巴使用基于ET Medical Brain的大模型来实现其阿里健康医疗影像平台。

台。

**语音识别**：语音识别是指使用计算机来识别和转换语音的技术。大模型在语音识别领域也有很多应用，如语音搜索、语音输入、语音合成等。例如，百度使用基于Deep Speech 2的大模型来实现其百度语音服务；苹果使用基于Siri的大模型来实现其Siri智能助理服务。

推荐系统：推荐系统是指使用计算机来根据用户的兴趣和行为，推荐相关的内容或产品的技术。大模型在推荐系统领域也有很多应用，如电商、社交网络、视频平台等。例如，阿里巴巴使用基于ET Industrial Brain的大模型来实现其淘宝、天猫等电商平台的商品推荐服务；腾讯使用基于Tencent AI Lab DeepCTR的大模型来实现其微信、QQ等社交网络平台的内容推荐服务。

音频处理：音频处理是指使用计算机来分析和生成音频的技术。大模型在音频处理领域也有很多应用，如音乐分类、音乐生成、语音转换等。例如，Spotify使用基于Spotify Music Transformer的大模型来实现其Spotify音乐服务；OpenAI使用基于Jukebox的大模型来实现其音乐生成服务。

智能客服：智能客服是指使用计算机来提供客户服务的技术。大模型在智能客服领域也有很多应用，如自动回复、智能导航、智能预约等。例如，京东使用基于京东AI Lab JIMI的大模型来实现其京东智能客服平台；美团使用基于美团AI Lab Meituan Chatbot的大模型来实现其美团智能客服平台。

**金融风控**：金融风控是指使用计算机来预测和控制金融市场、借贷风险等的技术。大模型在金融风控领域也有很多应用，如股票预测、信用评估、欺诈检测等。例如，平安集团使用基于平安科技PAI的大模型来实现其平安金融风控平台；蚂蚁集团使用基于蚂蚁科技Ant Financial Services Group的大模型来实现其芝麻信用评估平台。

**城市智能交通：**城市智能交通是指使用计算机来优化和管理城市交通的技术。大模型在城市智能交通领域也有很多应用，如交通信号控制、交通拥堵预测、交通事故分析等。例如，百度使用基于百度AI Cloud的大模型来实现其百度城市智能交通平台；高德地图使用基于高德AI Lab GaoDe Traffic Prediction Model的大模型来实现其高德地图交通拥堵预测服务。

社交网络分析：社交网络分析是指使用计算机来分析和利用社交网络中的用户行为、关系等信息的技术。大模型在社交网络分析领域也有很多应用，如用户画像、社交推荐、舆情分析等。例如，Facebook使用基于Facebook AI Research Graph Neural Network的大模型来实现其Facebook社交网络分析平台；微博使用基于微博AI Lab Weibo Sentiment Analysis Model的大模型来实现其微博舆情分析服务。

**自动化设计**：自动化设计是指使用计算机来自动生成设计图、建筑模型等的技术。大模型在自动化设计领域也有很多应用，如平面设计、室内设计、建筑设计等。例如，Adobe使用基于Adobe Sensei的大模型来实现其Adobe Photoshop、Adobe Illustrator等平面设计服务；AutoDesk使用基于AutoDesk AI Lab Dreamcatcher的大模型来实现其AutoCAD、Revit等建筑设计服务。

**跨领域知识迁移**：跨领域知识迁移是指使用计算机将在一个领域学到的知识和能力迁移到另一个领域中的技术。大模型在跨领域知识迁移领域也有很多应用，如跨语言文本理解、跨媒体内容生成等。例如，OpenAI使用基于GPT-3的大模型来实现其OpenAI Codex代码生成服务；百度使用基于ERNIE的大模型来实现其百度翻译跨语言文本理解服务。

**人机交互**：人机交互是指使用计算机与人类进行交流和协作的技术。大模型在人机交互领域也有很多应用，如语音交互、图像交互、手势交互等。例如，Amazon使用基于Alexa的大模型来实现其Amazon Echo智能音箱服务；微软使用基于Kinect的大模型来实现其Xbox游戏手势交互服务。

**元学习**：元学习是指使用计算机来学习如何学习的技术。大模型在元学习领域也有很多应用，如快速适应、少样本学习、零样本学习等。例如，DeepMind使用基于MetaMind的大模型来实现其Meta-Learning Shared Hierarchies元学习框架；OpenAI使用基于Reptile的大模型来实现其One-Shot Imitation Learning少样本学习框架。

**可解释性机器学习**：可解释性机器学习是指使用计算机来解释和理解机器学习模型的决策过程和输出结果的技术。大模型在可解释性机器学习领域也有很多应用，如注意力机制、可视化工具、因果推理等。例如，Google使用基于Google Brain Attention and Augmented Memory的大模型来实现其Google Neural Machine Translation注意力机制；IBM使用基于IBM Research AI Explainability 360 Toolkit的大模型来实现其IBM Watson OpenScale可视化工具。

### 2.2最新技术

**百度文心千帆**：百度文心千帆是百度基于其大模型文心的一系列下游应用，涵盖了电力、燃气、金融、航天、传媒、城市、影视、制造、社科等11个行业，提供了智能写作、智能阅读、智能翻译等多种功能。

**360大模型**：360大模型是360基于其大模型360AI的一系列下游应用，涵盖了安全、教育、医疗、娱乐等多个领域，提供了智能对话、智能创作、智能诊断等多种功能。

**Securepocilot**：Securepocilot是一款基于大模型的安全事件分析和响应平台，可以帮助安全人员快速定位和解决安全问题，提高安全效率。

**BloombergGPT**：BloombergGPT是一款基于GPT模型的金融文本生成和分析平台，可以帮助金融人员生成和理解金融相关的文本，提高金融效率。

**Microsoft365 Copilot：**Microsoft365 Copilot是一款基于OpenAI Codex的生成插件，可以一句话生成一个word、一个PPT。效果非常惊人。

### 2.3 我们能做什么

**选择合适的大模型：**根据我们的应用领域、任务类型、数据规模和质量等因素，选择一个合适的大模型作为基础模型，如BERT、GPT、XLNet等。我们可以参考一些大模型的评测和比较，如GLUE、SQuAD、SuperGLUE等，来了解不同大模型的优势和局限。

**利用有效的微调方法：**根据我们的应用需求、数据特点和计算资源等因素，选择一个有效的微调方法对基础模型进行再训练，如全局微调、局部微调、逐层微调或混合微调等。我们可以参考一些微调方法的研究和实践，如HuggingFace、PyTorch Lightning等，来了解不同微调方法的原理和效果。

LangChain是一项最新的技术，效果上等同于向语言模型打上思想限制，又在这个限制中提升到原模型没有的高度，使它成为人人的本地智库。这是我认为可以做的方向之一。

**采用优化技术提升性能：**根据我们的应用场景、性能指标和成本预算等因素，采用一些优化技术对微调后的模型进行改进，如参数压缩、结构简化、计算并行或数据并行等。我们可以参考一些优化技术的研究和实践，如TensorRT、ONNX Runtime、DeepSpeed等，来了解不同优化技术的方法和效果。

加速：如stable diffusion，ChatGLM，LLAM这类模型，都是比原型快出上千倍的模型，而能力却不输那些原生大模型。这是大模型应用的一个新的方向。想象一下，在我们的电脑、我们的手机上，部署一个离线的模型，一个属于自己的“贾维斯”，而它占用内存不超过5GB！

提示词库：提示词（prompt）优化是一项重要的技术，也是我们能做的方向。随着PLM体量的不断增大，对其进行fine-tune的硬件要求、数据需求和实际代价也在不断上涨。除此之外，丰富多样的下游任务也使得预训练和微调阶段的设计变得繁琐复杂，因此研究者们希望探索出更小巧轻量、更普适高效的方法，Prompt就是一个沿着此方向的尝试。融入了Prompt的新模式大致可以归纳成”pre-train, prompt, and predict“，在该模式中，下游任务被重新调整成类似预训练任务的形式。例如，通常的预训练任务有Masked Language Model， 在文本情感分类任务中，对于 "I love this movie." 这句输入，可以在后面加上prompt "The movie is \_\_\_" 这样的形式，然后让PLM用表示情感的答案填空如 "great"、"fantastic" 等等，最后再将该答案转化成情感分类的标签，这样以来，通过选取合适的prompt，我们可以控制模型预测输出，从而一个完全无监督训练的PLM可以被用来解决各种各样的下游任务。因此，合适的prompt对于模型的效果至关重要。大量研究表明，prompt的微小差别，可能会造成效果的巨大差异。研究者们就如何设计prompt做出了各种各样的努力——自然语言背景知识的融合、自动生成prompt的搜索、不再拘泥于语言形式的prompt探索等等。我们能做的就是设计这样的一个表格，里面包含着应用场景，与对应的prompt。

错误的提示词与好的提示词效果相差很远很远。

**关注可解释性和伦理道德：**在使用大模型进行应用开发时，我们也要关注大模型的可解释性和伦理道德问题，如大模型的工作机制、决策依据、不确定性量化、数据清洗、偏差消除、公平性评估、隐私保护、版权保护、道德审查、社会影响评估等。我们可以参考一些可解释性和伦理道德的研究和实践，如LIME、SHAP、AI Ethics Guidelines等，来了解不同可解释性和伦理道德问题的重要性和解决方案。

## 三、大模型优化与微调技术

### 3.1 大模型微调的概念和方法

大模型微调（Fine-tuning）是指在已经预训练好的大模型的基础上，针对特定的下游任务，对部分或全部参数进行再训练，以适应新任务的数据分布和目标函数。大模型微调可以有效地利用预训练阶段学习到的通用知识，减少训练时间和数据量，提高下游任务的性能。

大模型微调的方法主要有以下几种：

**全局微调**：全局微调是指对预训练好的大模型的所有参数进行再训练，通常需要较小的学习率和较多的训练步数，以避免破坏预训练阶段学习到的知识。全局微调可以充分地适应新任务的数据分布和目标函数，但也容易导致过拟合和灾难性遗忘。

**局部微调**：局部微调是指对预训练好的大模型的部分参数进行再训练，通常需要较大的学习率和较少的训练步数，以快速地适应新任务的数据分布和目标函数。局部微调可以减少过拟合和灾难性遗忘的风险，但也可能损失预训练阶段学习到的知识。局部微调可以根据参数在网络中的位置或重要性进行选择，如只微调最后一层、最后几层或最重要的层等。

逐层微调：逐层微调是指对预训练好的大模型的参数进行分层再训练，通常需要动态地调整每一层的学习率和训练步数，以平衡新任务的适应性和预训练知识的保留。逐层微调可以根据参数在网络中的位置或重要性进行分层，如从底层到顶层或从顶层到底层逐渐增加或减少学习率和训练步数。

**混合微调**：混合微调是指对预训练好的大模型的参数进行不同方式的再训练，根据不同层或不同任务的特点，选择合适的微调方法，如全局微调、局部微调或逐层微调等。混合微调可以灵活地适应新任务的数据分布和目标函数，同时平衡预训练知识的保留和更新。

### 3.2 大模型优化的概念和方法

大模型优化（Optimization）是指在预训练或微调阶段，对大模型的参数、结构或计算过程进行改进，以提高大模型的训练效率、推理速度或泛化能力。大模型优化可以有效地解决大模型面临的计算资源、存储空间或内存限制等问题。

大模型优化的方法主要有以下几种：

**参数压缩**：参数压缩是指对大模型的参数进行降维、量化、稀疏化或剪枝等操作，以减少大模型的参数量和存储空间。参数压缩可以降低大模型的计算复杂度和内存占用，提高大模型的推理速度和能耗效率。

**结构简化**：结构简化是指对大模型的结构进行裁剪、分解、蒸馏或融合等操作，以减少大模型的层数、通道数或连接数等。结构简化可以降低大模型的计算复杂度和内存占用，提高大模型的推理速度和能耗效率。

**计算并行：**计算并行是指对大模型的计算过程进行切分、分配或同步等操作，以利用多个处理器或设备同时进行计算。计算并行可以提高大模型的训练效率和扩展性，克服单个处理器或设备的计算资源或存储空间限制。

**数据并行**：数据并行是指对大模型的训练数据进行切分、分配或同步等操作，以利用多个处理器或设备同时进行训练。数据并行可以提高大模型的训练效率和扩展性，克服单个处理器或设备的计算资源或存储空间限制。

**自适应学习**：自适应学习是指对大模型的学习过程进行动态地调整、选择或组合等操作，以根据不同层、不同任务或不同数据的特点，采用合适的学习方法。自适应学习可以提高大模型的泛化能力和鲁棒性，避免过拟合或欠拟合等问题。

## 四、我们现在有什么？

基于ChatGLM与LangChain的本地智库

基于StableDiffusion、Dreambooth、lora大模型的本地高质量图像生成系统

GPT调用API权限

NewBing请求转发（可以当API用）

文心一言调用API权限

AIStudio、BML算力支持

本地深度学习开发环境

Bard使用申请中

上述各种技术的使用方法与原理，大致成本与开发周期