吉林大学计算机科学与技术学院

专业学位硕士研究生

《专业实践》报告

学号：2023534058

姓名：谭张顺

指导教师：王爱民教授

2024年 09月 08日

|  |
| --- |
| 1. 专业实践的目的和预期目标   专业实践的目的：  1、巩固基础理论和专业知识，增强对专业领域内最新技术动态和发展趋势的敏感性和适应能力。  2、通过实际操作和项目经验，提⾼运用所学知识，解决实际问题的能⼒。  3、培养解决复杂问题的能力，敢于创新，并能正确地将独创精神与科学态度相结合。  专业实践的预期目标：  1、加强和巩固理论知识，提升发现问题并运⽤所学知识分析问题和解决问题的能⼒。专业实践的预期目标还包括：  2、通过实践操作，提高专业技能和实际操作能力，为将来的职业生涯打下坚实基础。  3、培养独立思考和自主学习的能力，能够在没有直接指导的情况下探索和解决问题。  4、根据具体情况，选定合适的毕业设计题目。   1. 专业实践内容与成果   一:摘要  近年来，大模型发展如火如荼，多模态大模型是人工智能领域的一个重要发展方向，它通过融合多种类型的数据（如图像、文本、声音等），实现了更全面、准确的分析和决策。但多模态大模型通常需要更大规模的模型和更强的计算资源来进行训练和推理，需要更多的数据来进行预训练，这包括模型规模，计算资源和数据集构建的挑战。为了让大模型的多模态能力能够得到应用，我们细化了模型的领域，减少依赖的数据量以及模型大小，此外还使用langchain，量化，构建知识图谱等方式进行辅助，提高既定领域的模型能力。  二:研究背景  现实世界中的信息往往是以文本、图像、声音等多种模态形式存在的。传统的语言模型主要处理文本信息，而多模态大模型旨在整合这些不同类型的数据，提供更加丰富和自然的交互体验，但是目前仍然面临一系列的问题：   1. 数据限制：   高质量视频数据集稀缺：相较于图像数据集，高质量且标注丰富的视频数据集相对较少。这导致了视频生成模型在训练过程中可能面临数据不足的问题，从而影响模型的学习能力和泛化能力。  多样化内容缺失：现有数据集中往往缺少涵盖各种不同场景、人物、动作和情感的视频样本，这限制了模型生成内容的多样性和丰富度。   1. 算力资源的缺乏：   语音和图像视频相较于文本，   1. 控制能力：   内容控制：目前尚缺乏有效的方法来精准控制生成视频的内容，例如指定视频中的特定物体、人物或场景。  风格控制：难以在生成的过程中保持一致的艺术风格或视觉效果，尤其是在涉及复杂的运动和姿势时，保持风格的一致性变得更加困难。  到目前为止，各个互联网巨头也陆续推出自家的视频生成模型，比如：字节的magic animate、阿里的dreammoving等等对于生成漫画人物以及生成模仿视频效果一般，但是在各个领域仍然有着广泛的应用场景。  数字媒体行业：视频生成技术为广告、营销、社交媒体等领域提供了快速生产高质量视频内容的强大工具。这有助于创作者更快地将创意转化为可视化作品。  虚拟现实和增强现实：通过生成高度逼真的视频内容，可以显著提升虚拟现实和增强现实中的人机交互体验，使用户沉浸感更强。  电影制作：生成的视频可以用于辅助电影制作过程，比如快速原型制作、预览特效场景等，从而降低制作成本并加快创作流程。  游戏开发：在游戏开发中，可以利用视频生成技术来创建丰富的游戏环境和角色动画，提高玩家的游戏体验。  直播互动：在直播和实时互动娱乐中，生成的视频能够增加互动性，例如通过观众的选择来改变故事走向或角色行为。  综上所述，视频生成技术不仅为创意产业带来了革命性的变化，也为人机交互和娱乐体验提供了前所未有的可能性。随着技术的进步，这些技术的应用将会更加广泛，并持续推动相关领域的发展。  三:专业实践主要内容：   1. 多态大模型的使用   姿势提取是通过现成的姿势估计算法来完成的。这些算法可以将图像数据集中的人物姿态转化为对应的骨架点坐标，从而形成包含人物姿态信息的图像对。具体来说，研究人员利用了如MMpose这样的开源姿势估计工具，从LAION这样的大型图像数据集中提取出包含人体关键点的人像图像。这些关键点包括但不限于头部、肩膀、手肘、手腕、臀部、膝盖和脚踝等部位的位置。通过这种方式，可以从原始图像中提取出姿态特征，并与图像配对，形成用于后续训练的图像-姿态对数据集。  由于原始数据集中可能包含一些不符合要求的图像，例如缺少关键点或关键点检测错误的图像，因此需要进行数据集清洗。数据清洗的过程包括过滤掉那些无法正确检测到足够多关键点的图像，以及去除那些姿态信息不清晰或不完整的图像。这样做可以确保最终使用的训练数据集的质量，从而提高模型的性能。   1. 领域能力的加强：   第一阶段：使用带有姿态信息的图像对来训练模型，以便模型学会如何根据姿态信息生成相应的图像。这一阶段的重点在于让模型掌握姿态控制的能力，因此只更新与姿态编码相关的参数。  第二阶段：利用无姿态信息的视频数据集来进一步训练模型，使其学习如何生成具有时间一致性的视频。在这个阶段，主要更新与时间一致性相关的参数，例如时间自注意力层和跨帧注意力层。   1. 其他的优化方式   为了确保生成的视频具有良好的时间一致性，即视频中的运动平滑且连贯，主要使用了时间自注意力技术以及连续采样的方法。  首先，在模型中加入了额外的时间自注意力层，以捕捉视频帧之间的依赖关系，从而确保视频中的动作连贯。然后，通过引入跨帧注意力机制，使相邻帧之间的内容保持一致，减少闪烁现象。最后，通过重复使用去噪过程中的噪声，确保长视频的生成过程中每一帧之间有良好的过渡。  为了保持空间一致性，即确保视频中同一物体在不同帧之间的位置和形状一致，主要使用了空间注意力机制以及特征融合的方式。  首先，通过空间注意力机制，模型可以专注于图像中的特定区域，从而在生成过程中保持物体的空间位置不变。然后，在视频生成过程中，通过将来自参考图像的特征与当前帧的特征进行融合，可以确保生成的视频在视觉上与参考图像保持一致。   1. 模型的最终效果   通过我们自训练的多模态动漫领域大模型可以正确识别语音指令，以及识别动漫中的人物特点。   1. 专业实践总结   专业实践让我们切身体验从事工程任务的艰辛，不仅补充了课堂上难以触及的实用知识，更让我们的视野跨越了书本的界限，见识了更加多元的世界。这段旅程初期虽以浩瀚文献的阅读为起点，但每一次不解的求解，都化作了我前行的动力，无论是向校内外的导师请教，还是与经验丰富的工程师交流，都让我深刻体会到知识的无界与合作的力量。  实践，作为我们学习的第二课堂，它不仅仅是知识的检验场，更是能力的孵化器。在这里，我学会了如何将理论知识转化为实际行动，如何在解决问题的过程中不断反思与成长。我深知，未来的道路上，唯有持续将所学与实践相结合，勇于探索，敢于创新，才能不断提升自我，完善综合素质，为未来的事业奠定坚实的基础。  最后，我衷心感谢那些在实习过程中给予我无私帮助与关怀的每一位老师和同学，是你们的鼓励与支持，让我更加坚定了前行的步伐，也让我明白了团队合作与个人努力同样重要。这段宝贵的经历，将成为我人生旅途中一道亮丽的风景线。  研究生签名：谭张顺  2024 年 9 月 8 日 |

注：宋体、小四号字，行距（固定值22磅），不少于3000字。

|  |  |
| --- | --- |
| 指导教师鉴定意见 | 这份实践计划展示了良好的方法论和技术应用，通过充分的数据预处理和多种保证时间一致性以及空间一致性等算法的比较选择，成功实现了模仿视频的生成。模型验证过程的细致和全面性也为结果的可靠性提供了坚实的保障。总体而言，这是一个很有建设性和实用性的实践计划，对数据处理和深度学习领域的学习与发展具有积极的推动作用。 |
| 专业实践  成  绩 | 专业实践成绩：（成绩按百分制评定。评分依据：实习报告、实习内容、出勤情况、校外导师意见等。）  指导教师签名：王老师图片2  2024 年 9 月 8 日 |