**一、项目名称：**

基于图像描述模型的视觉场景语音解说

**二、项目简介：**

图像描述模型以及文本转语音技术日趋成熟，视觉功能障碍者可以借此使得生活更加便利。然而，目前移动端的视觉辅助应用的技术尚未及时更新。因此，我们试图将图像描述模型技术与文本转语音技术结合，为这部分人群提供更加方便的无障碍辅助服务。同时我们探索人类在机械辅助下更多、更广泛的视角所带来的新的可能性。

**三、成果预期：**

一款可以即时对所拍摄场景进行语音解说的性能优秀的移动应用客户端。

**四、功能预期：**

|  |  |
| --- | --- |
|  |  |
| 通过拍照或从相册中选择图片，生成描述 | 描述图片中的内容 |
| 通过tts接口生成语音解说并播放 | 可以包括对图片中物体、场景、人物的描述，以及对图片的整体理解。 |
| 视频流描述 | 描述视频中的内容 |
| 多语言支持 | 至少支持中文和英文文字输出和解说 |
| 离线使用 | 在未联网环境下也能使用 |
|  |  |

**五、实现步骤：**

|  |
| --- |
|  |
| Image Caption模型部署 |
| TTS语音转文字接口 |
| 客户端开发 |
| 视频流场景描述 |
| 准确率测试 |
| 性能测试 |
|  |

**六、性能预期：**

|  |  |
| --- | --- |
|  |  |
| 响应速度 | 每秒1到10帧,生成文字描述不超过0.05s，文字转语音不超过0.025s，整个过程不超过0.2s。 |
| 准确率 | 生成的描述完全且准确，对高频场景的描述准确率达95%。 |
| 稳定性 | 对同一场景的解说具有一致性 |
| 兼容性 | 支持市场上大部分手机机型 |

**七、时间计划：**

|  |  |  |
| --- | --- | --- |
| **日期** | **计划** | **总结** |
| 7月22日 | 尝试部署开源图像描述模型BLIP，为之后与付学长的对接工作做准备。 | 测试模型部署完成，但该模型在本地运行效果较差，响应时间长达5s，只能生成英文描述，且对场景的描述有限。 |
| 7月23日 | 寻找tts模型，并学习安卓应用开发。 | TTS拟采用安卓自带api，同时也寻找到更多在线免费api以待选择。了解了安卓应用开发的基本流程。 |
| 7月24日 | 尝试跟着官方文档完成安卓调用PyTorch模型，并使用之前找到的性能较差的模型，完成对**静态图片**生成**语音描述**的应用**V1**。 | 已有最基本的交互界面，但测试模型未成功运行，寻找原因中。 |
| 7月25日 | 模型转换为Torchscript时出现问题，疑似模型太复杂转换不了，拟转换ONNX Runtime方法 |
| 7月26日 |  |
| 7月27日 | **V2：**对**视频流**进行描述。 |  |
| 7月28日 |  |
| 7月29日 |  |
| 7月30日 | 性能测试与优化。  记录数据，撰写报告。 |  |
| 7月31日 |  |
| 8月1日 |  |
| 8月2日 |  |
| 8月3日 | 丰富功能。  优化报告。 |  |
| 8月4日 |  |
| 8月5日 |  |
| 8月6日 | 制作PPT演示。 |  |
| 8月7日 |  |
| 8月8日 |  |  |
| 8月9日 |  |  |