



**计算机视觉**

**上机实验报告**

|  |  |
| --- | --- |
| 专 业： | 人工智能 |
| 班 级： | 2204 |
| 学 号： | 19220432 |
| 姓 名： | 陆昊宇 |

2025年04月01日

## 一、实验目的

1. 熟悉使用CLIP模型实现零样本图像分类。

2. 了解如何低代价地增强CLIP的零样本学习能力。

## 二、实验内容要求

#### 1. 零样本图像分类实验

（1）下载并部署 OpenAI 提供的预训练 CLIP 模型，自由构造测试集并使用 CLIP 进行零样本分类，评估 CLIP 在不同类别图像上的分类准确率；

（模型权重：https://github.com/openai/CLIP）

（2）参考论文《Visual Classification via Description from Large Language Models》，探索如何利用大语言模型（任选LLM）改进 CLIP 的分类能力，设计对比实验并分析 LLM 生成的描述文本对 CLIP 分类性能的影响；

（3）设计实验探究 CLIP 的局限性，包括但不限于细粒度分类能力、对数量相关描述的敏感度和复杂文本描述的理解能力等。

**提交入口：**

https://send2me.cn/XjLlYYFa/Q66T1LZzqcS2Jw

## 三、实验报告评分标准

1. 完成零样本图像分类实验1）（30’）

2. 完成零样本图像分类实验2）（25’）

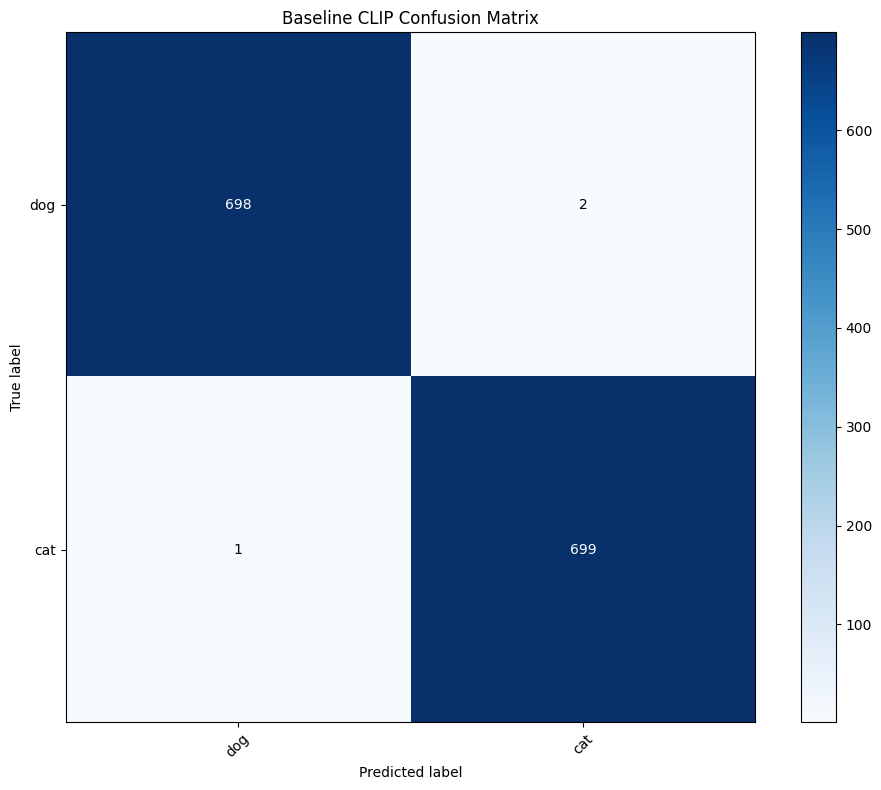
3. 完成零样本图像分类实验3）（25’）

4. 实验过程记录完整，表述逻辑清晰（10’）

5. 排版工整自洽，图表说明完整（10’）

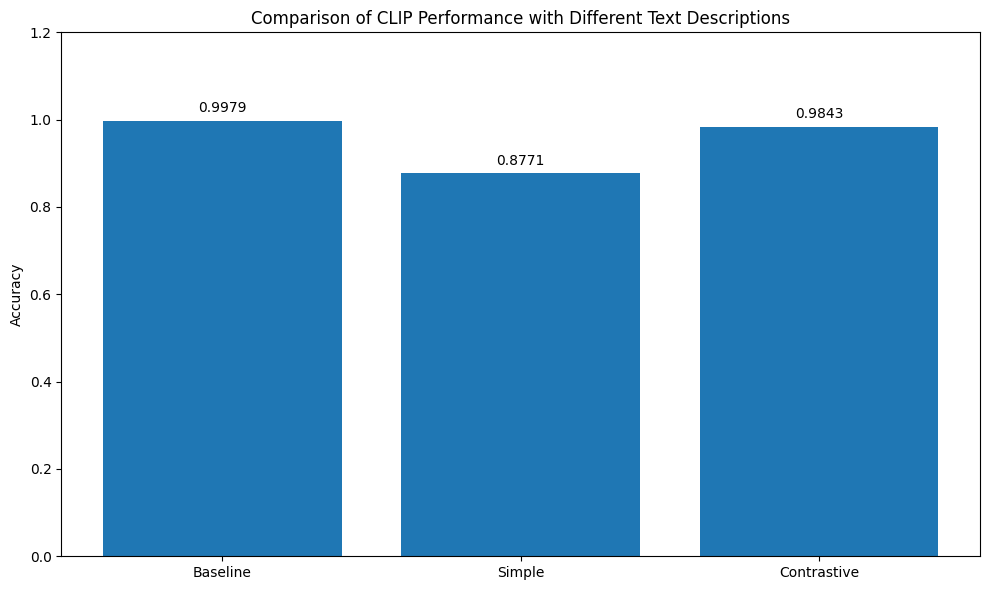
## 四、实验过程

（1） 利用clip在猫狗测试集上进行零样本分类的结果如下：



结果表明，clip在测试的数据集上有着不错的表现。

（2）Baseline是没有提示词输入的模型；Simple是利用LLM进行简单的文字描述；Contrastive是利用LLM进行对比式的文字描述。最终效果如下：

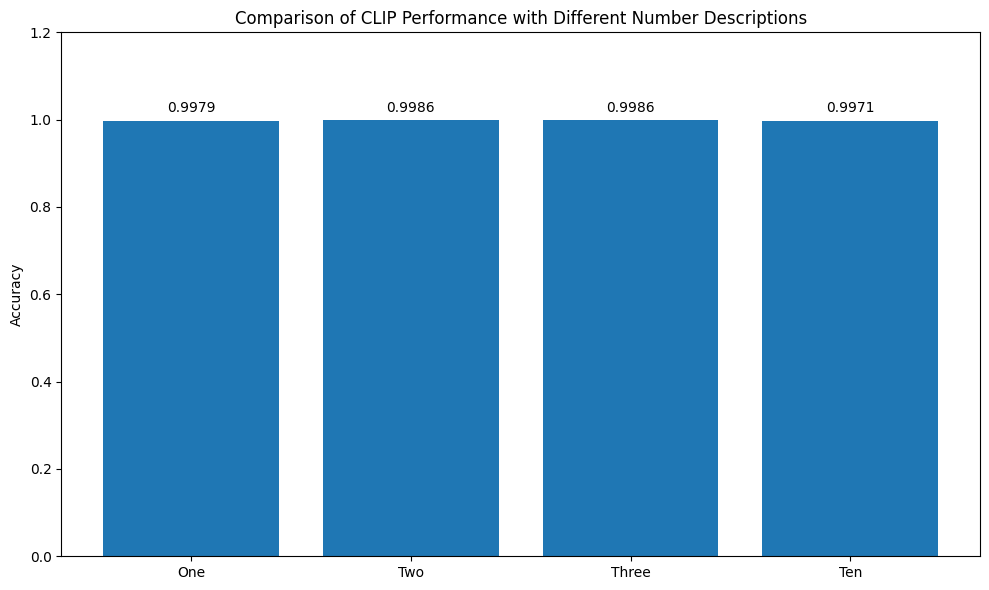
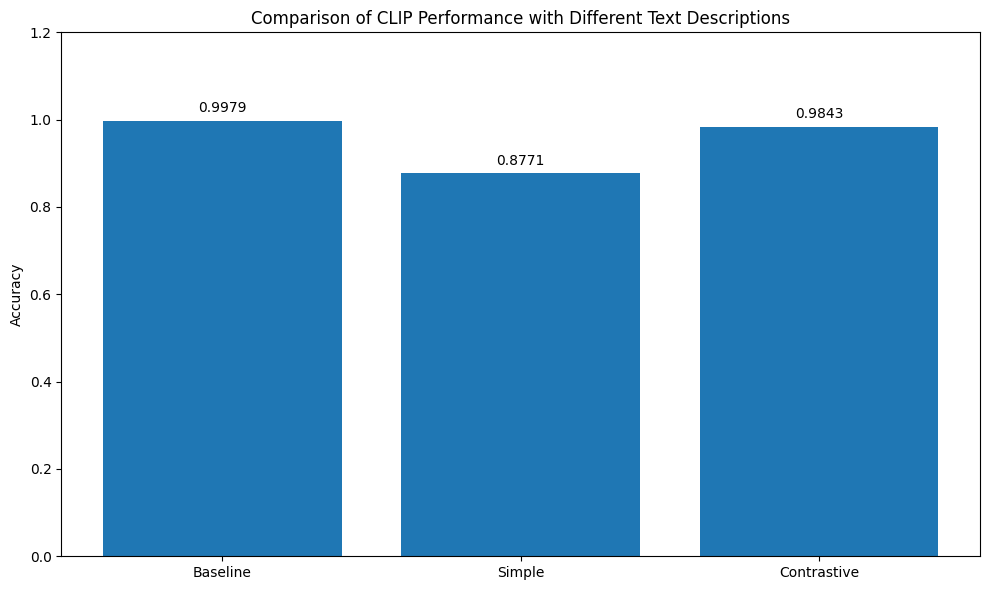


这里发现，利用llm的模型似乎不如baseline，可能的原因是clip只能接受77 tokens的输入，因此llm的描述并不准确。在同样不准确的情况下，contrastive所包含的信息比simple多也是自然的。

同时测试时也注意到，clip对于输入的描述文本非常敏感，不同的描述文本acc差异可能达到30%。  
注：这里的LLM模型采用的是deepseek-chat；temperature设置为0.7.

1. 细粒度测试：在CIFAR-100数据集上进行零样本分类，准确率为0.6065.  
   

数量描述敏感度：在猫狗数据集上，给予不同的数量输入，准确率差异不大，说明CLIP模型对于数量不太敏感。

  
复杂文本测试：参考第二项任务，限制在77 token的CLIP模型，并不能很好的理解复杂文本信息。  


## 五、实验结论

CLIP在测试用的猫狗数据集上表现良好，对于描述的数量不敏感。

CLIP在CIFAR-100数据集表现不好，细粒度分类能力差。

CLIP对输入的提示信息处理能力弱，不能很好的理解复杂文本。