1.总结一下开发集、测试集的划分原则

首先定义训练集、开发集与测试集

- 训练集：运行学习算法
- 开发集：用于调整参数，选择特征，以及对学习算法做出其他决定
- 测试集：用于评估算法性能，但不会据此改变学习算法或参数

以下为开发集与测试集划分原则总结：

- 开发集和测试集应该服从同一分布，并应该与计划解决的问题的数据的分布一致。
- 开发集的规模应该尽可能大，而测试集的大小应该大到能够对系统性进行高度可信的评估。但是大数据时代，开发集和测试集规模也并不是越大越好，实际上，开发集和测试集的比例远低于30%。
- 使用单值评估指标进行优化，可以更快速做出决定。
- 考虑多项目标优化时候，可以选择将其整合到一个表达式，或设置满意度指标，在这一标准下优化另外的目标。
- 当开发集和评估指标不能提供正确导向的时候，尽快修改。

2.你在ImageNet数据集上训练、测试结果很好，但是对你自己拍摄的图片效果很差。你觉得该怎么办？

1.尝试理解数据属性在训练集和开发集分布之间的差异

2.尝试找到更多的训练数据，以更好匹配算法遇到的开发集样本。

3.对比讨论“端到端”和“流水线组件”两种思路的优缺点、适用场景

端到端：

优点：训练集大的时候，不会受到人工因素影响，更容易达到较优的效果；可以输出比标签更丰富的内容，例如句子、图像、音频等。

缺点：训练集小的时候，表现可能比人工设计的流水线差；适用于端到端的数据部分场合更不容易获取。

适用场景：有大量适用数据；需要输出比直接标签更为复杂的内容，例如句子、图像等。

流水线组件：

优点：数据可用性更强；独立的组件可能会使得每个组件的任务比较简单，总而每一个组件需要的训练数据都更少；使用一些人类经验，能帮助学习算法更快速理解数据中的某些特征。

缺点：过多的人工参与可能影响算法的性能上限；

适用场景：数据量不足；将任务划分为多个小任务的时候大大降低每个小任务难度；