浙江大学

本科周报告

姓　　名：	赖梓林
学　　院：	计算机科学与技术
系：	计算机科学与技术
专　　业：	计算机科学与技术
学　　号：	3170104684

目录

第一章 摘要 3

第二章 正文 4

1.Topaz 补充内容 4

1.1 PU 学习和自动编码器 4

1.2 偏差(Bias) 8

1.3 分数阈值 9

1.4 Noise2Noise 10

1.5 文件格式 11

3.参考资料 12

摘要

报告为上次报告内容的补充，主要涉及在 Topaz 方法中的相关知识点，如 PU 学习、自动编码器等以及其他新内容。

二、正文：

1. Topaz 补充内容：

作者在 Topaz 方法总览中提到，Topaz 方法有以下几个优势：

1）取到更多颗粒，需要更少的手动标记。

2）取到的颗粒更具有代表性，综合上一条使得后序流程能取得更好结果。

3）偏差更小。

4）无需对显微图像进行完整的标记。

本次报告从这几个优点以及其他需要关注的点入手，简单讨论其中的相关内容

1.1. PU 学习和自动编码器：

首先要说明的是，作者上述提出的所有 Topaz Pipeline 方法的优势都是与传统流程的冷冻电镜单颗粒分析相比得出的。下面简单叙述传统流程取得颗粒偏少以及需要大量手动标记的原因。

我们知道，卷积神经网络通常需要尽可能多的样本来训练，这在冷冻电镜单颗粒分析中意味着我们需要大量的正(positive)标记样本，这导致研究人员需要花费时间和人力来标记足够的颗粒以及负(negative)区域。并且需要标记的负样本数量要大于正样本数量。（颗粒分布较疏）

在 Steve Lucky 的论文中提到，在大多冷冻电镜数据集中，负样本数量约为正样本数量的十倍。而我们又需要取得具有代表性的颗粒集合，这意味着大量的手动标记。在下图可以看到，有许多处于不同取向的颗粒很容易被忽视。

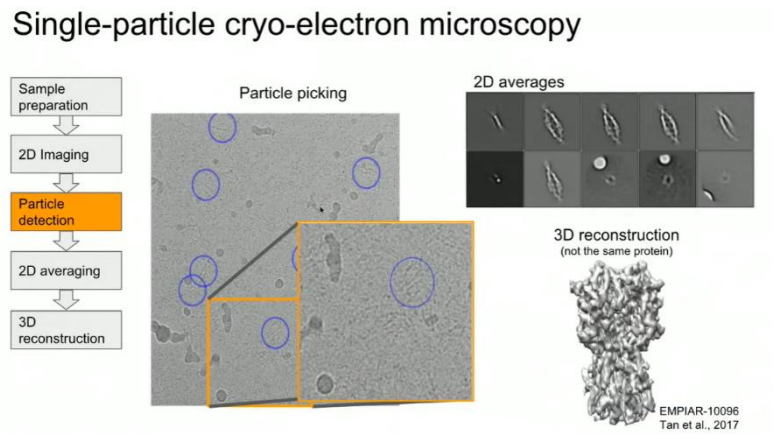

图 单颗粒分析

1.1.1 PU 学习：

Topaz 通过 PU 学习（Positive-Unlabeled Learning）解决了传统流程的困难。PU 学习的定义和简介是上周报告内容，这里不再提及。

在 Topaz 的 PU 学习中，我们从少量正标记数据和其他可能是正或负的未标记样本中，学习分类器的参数。

如下图所示，橙色为正标记样本，灰色为未标记样本，我们通过 PU 学习得到的结果是一条区分正负样本的决策边界(Decision Boundary)。

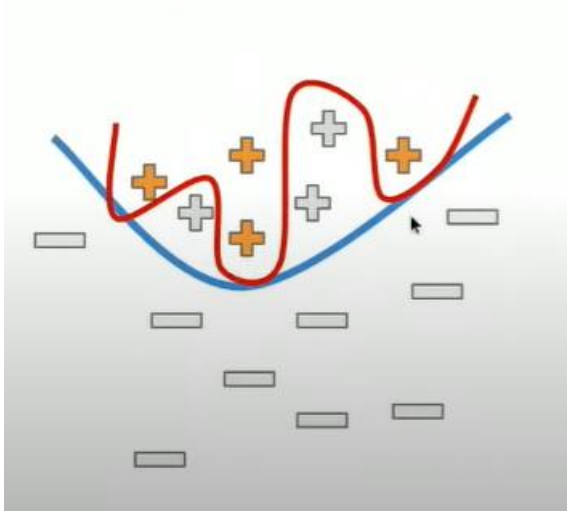

图 PU 学习

而实现 PU 学习的具体方法也会导致结果，也就是这条决策边界的表现的不同。如下图，作者对比了几种不同的实现方法。可以看到非负风险评估器(Non-negative Risk Estimator)代表的蓝线的区分效果要好于前两种方法，前两种方法似乎过度看重已知的正样本，导致最后的决策边界只是简单地将已标记样本和未标记样本分开，这样的结果类似于 PN 学习(Positive-Negative Learning)取得的结果。而在 PN 学习中，由于假定未标记样本皆为负样本，在学习过程中会有无法检测出某些颗粒的情况，所以表现也较差。这三种方法的想法都是估计样本的正负经验损失。

作者提出了另一种方法，在这种方法中，假设我们知道正样本的比例，并且正确地分类了标签数据。就可以使用这个比例为未标记样本添加约束，使网络最小化损失加上 KL 散度项。

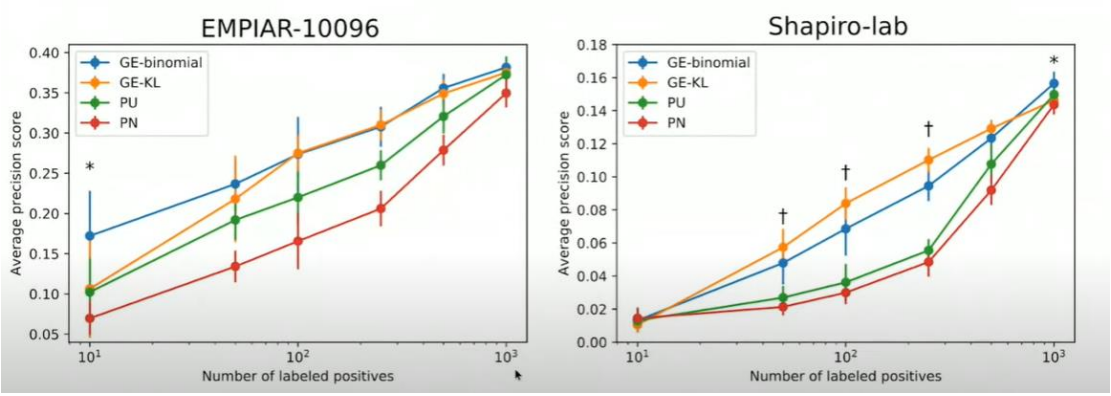

图 表现对比

了解了作者的想法之后，我们可以看一下表现对比图，其中横坐标为标记颗粒数目，纵坐标为准确度评分。

我们首先分析 EMPIAR-10096 数据集中的表现，整体来看，GE-binomial 和 GE-KL 的表现要好于 PU 以及 PN，并且排名约为 GE-binomial>GE-KL>PU>PN。

在已标记颗粒数目极少，如图中的十个时，GE-binomial 的表现远远超过其他三种方法，而在颗粒标记数目较多，如图中的一千个时，GE-binomial 的表现也是最好的。

再看 Shapiro-lab 数据集中的表现，在这个数据集中，整体表现是 GE-binomial 和 GE-KL 最优，其中 GE-KL 在颗粒数目较少的情况表现更好，GE-binomial 在颗粒数目较多的情况表现更好。

值得注意的一点是，在 Shapiro-lab 数据集中，仅有十个标记颗粒的情况下，四种方法的表现较为接近，且准确评分较低，而在 EMPIAR-10096 数据集中，四种方法的表现相差甚远。形成这两个数据集中表现的差距的原因并不明确，作者提到有可能是因为 Shapiro-lab 数据集中目标颗粒的三维结构较为复杂，所以在颗粒数目偏低时，会漏掉一些重要的视角的颗粒图形，导致最终表现较差。

从结果可以看到，Topaz 中的 PU 学习优于前沿的非负风险评估器，并且远远优于 PN 学习。

1.1.2 自动编码器：

作者补充时提到，引入混合自动编码器模型会取得更好的整体表现。

在定义中，自动编码器是一种在无监督下学习高效的数据编码的神经网络，目的是通过训练网络忽略信号噪声来学习数据的表示（编码），通常用于降维。自动编码器尝试从简化编码中生成尽可能接近原始输入的表现形式。

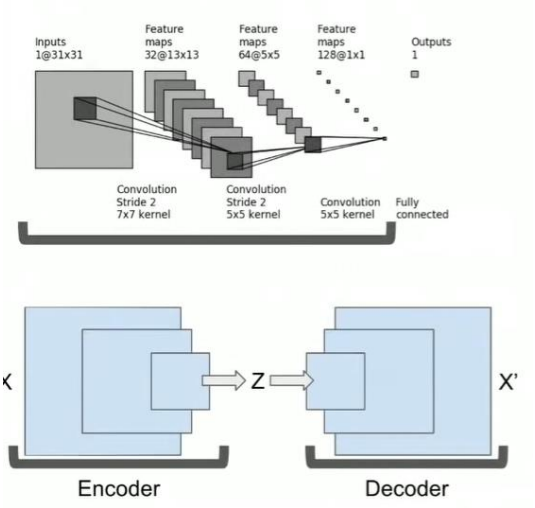

图 自动编码器

在 Topaz 中，给定一个显微图像窗口，自动编码器将其编码，如图中的 z，再解码为显微图像窗口，并在简化编码中生成尽可能接近原始输入的表现形式。这样做的好处是可以通过极少量的标记样本来提高分类器的表现。

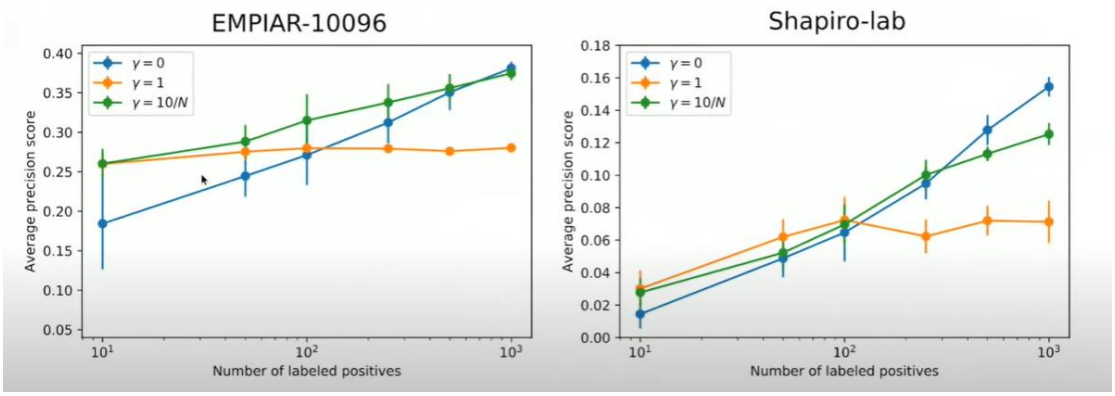

图 表现对比图

从表现对比图中可以看到，在有较少正标记样本的情况下，引入带权的自动编码器会大幅提升整体的表现，但在正标记样本上升到一定数目后，权重为 1 的自动编码器反而有负面影响。

从整体来看，自动编码器适用于正标记样本较少的情况，而在样本偏多，如图中的一千个或以上时，使用自动编码器只会带来反效果。

1.2. 偏差(Bias)：

如下图所示，在传统流程的单颗粒分析中，很多步骤都可能为最终结果带来偏差。例如在 2D 或 3D 的类平均时，我们需要用人眼来判断，决定丢弃哪些可能是无用的类，这时如果丢弃了一个实际为目标物质的某个视角的类，会对整体结果带来很大的偏差，并且我们知道，单颗粒分析包含了一个不断迭代更新的过程，偏差也会在这个迭代过程中被一步步放大。

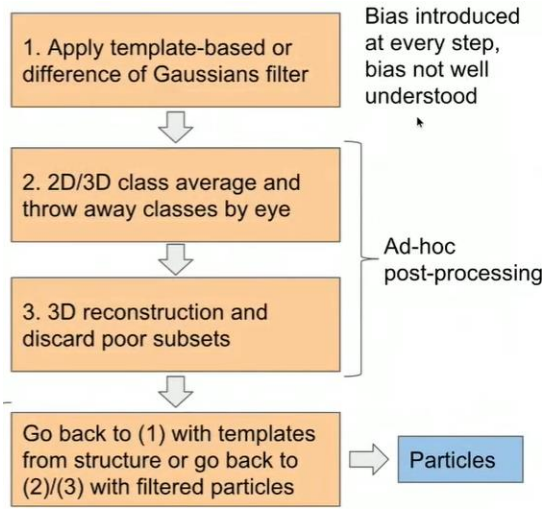

图 单颗粒分析

而在 Topaz 中，只有第一步，也是唯一的一步可能引入偏差。在这一步中，我们标记少量的具有代表性的颗粒，如果我们错误标记，或是标记的颗粒不具有代表性，就可能产生偏差，但与单颗粒分析流程相比可能性较小，更易于理解且更可控。

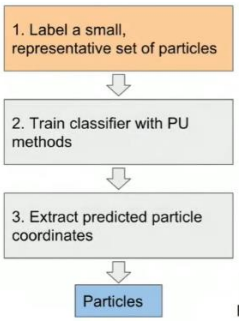

图 TOPAZ

1.3. 分数阈值：

正如之前的报告中提到的，Topaz 对显微图像区域进行评分，尽可能多地找到可能的正标记区域，并通过极大值抑制方法来减少候选区域。在重新观看作者的讲解时发现，Topaz 可以通过分数阈值来控制取得的颗粒数目。

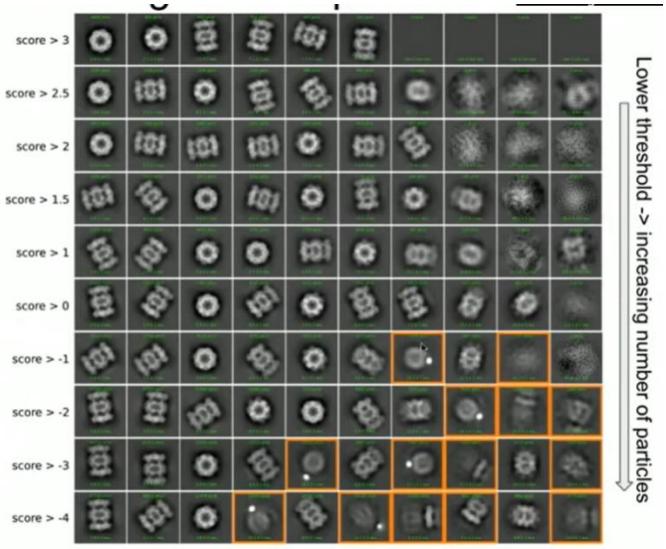

图 不同阈值下取得的颗粒

从上面的图我们可以看到，随着分数的阈值逐渐下降，取得的颗粒逐渐增多。但这也带来了一个问题：在阈值较低时，Topaz 会开始取回假的正样本（如图中橙色框的颗粒），并且随着阈值的降低，假的正样本的数量也越来越多。我们可以根据实际问题和使用来权衡取回数量以及实际正样本和假正样本的比例。

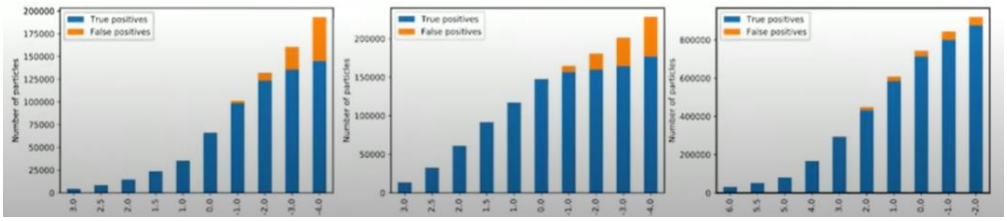

图 不同阈值表现对比图

从表现对比图中可以看到，降低阈值基本上会为我们取得更多颗粒，假正样本仅在阈值降低到一定程度时才会出现。并且对于不同的数据集，实际情况也不同，在第三张图的数据集中，阈值的下降显著地提升了取回的颗粒数目，并且假正样本的数量也一直保持在较低的水平。

1.4. Noise2Noise：

Topaz 通过 Noise2Noise 框架模型来去噪，并且模型已在一些冷冻电镜数据集上训练过。常规的去噪模型通常通过原始图像和添加了噪声的图像学习，而 Noise2Noise 是通过一系列成对的添加了噪声的图像学习。

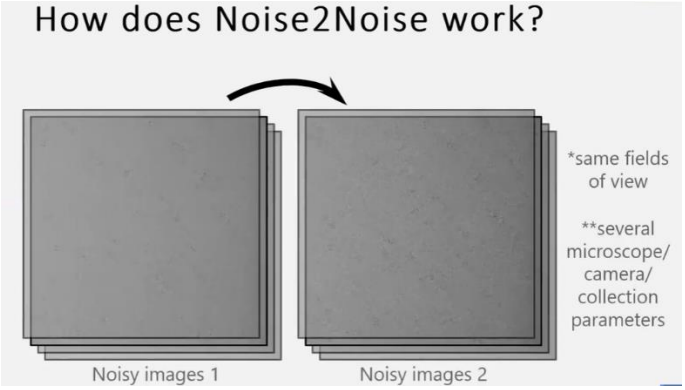

图 Noise2Noise 示意图

Noise2Noise 在两张添加了噪声的图形中学习，由于图像中的信号(Signal)是固定的，只有噪声(Noise)不同，在学习之后，Noise2Noise 可以将添加了一种形式噪声的图像转换为添加另一种形式噪声的图像。最后将某种特定形式噪声的图像转换为符合我们要求的图像(Clean Target)

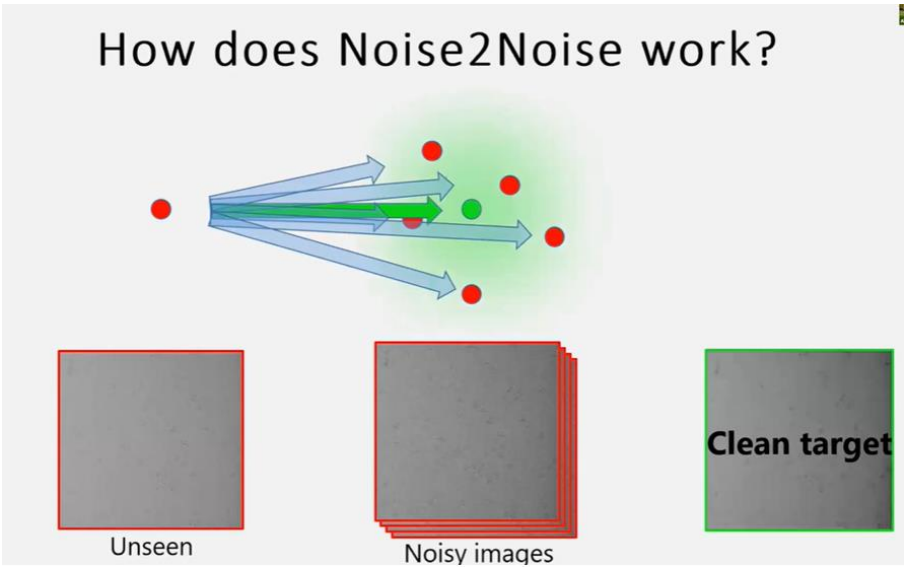

图 Noise2Noise 原理示意图

1.5. 文件格式：

在 Topaz 中大体涉及到 MRC, TIFF 以及 PNG 格式的文件。其中 MRC 文件格式为冷冻电镜以及电子断层扫描的行业标准，由英国医学研究理事会分子生物学实验室开发，在 2014 年被标准化。

MRC 格式的文件中包含一个三维网格，其中每个体素(voxel)包含对应电子密度或电势的值。tiff 是存储光栅图形（位图、点阵图）的计算机文件格式，最小单位由像素构成，只包含点的信息。

MRC 文件描述了一个三部分组成的二进制文件：

第一部分是主标题，包含有关图像/体积的元数据的固定的格式值。主标头上限为 1024 字节，包含为未来扩展预留的未分配空间。

第二部分是可变长度的扩展标头，这部分设计以适配晶体学应用。

第三部分包含实际的图像/体积数据其中网格的值表示为一系列可能的数据类型之一，具体取决于地图(Map)的“模式”(Mode)。

原始的格式规范列出了五种模式，模式 0 代表一个字节长的整型，有符号或无符号不固定，但使用者的共识是当作有符号整型。模式 1 和 2 分别表示两字节长的整型和四字节长的实数。模式 3 和 4 分别代表包含一对两个字节长的整型的复数以及四字节长的实数。模式 5 从未被使用过。在这个原始规范之后也有其他的模式被提出，例如模式 6：无符号两字节整型。

3. 参考资料：

1. CNN Training Loop Refactoring - Simultaneous Hyperparameter Testing

2.Pytorch Explained

3.Pytorch install - Quick and Easy

4.Tensors for Deep Learning - Broadcasting and Element-wise

5.CNN Image Preparation Code Project - Learn to Extract,Transform…

6.Build Pytorch CNN - Object Oriented Neural Networks

https://www.youtube.com/watch?v=ozpv_peZ894&list=PLZbbT5o_s2xrfNyHZsM6ufI0iZENK9xgG&index=33

（网址同系列）

7.Topaz - tbepler 项目

8. Topaz - Bepler, T., Morin, A., Brasch, J., Shapiro, L., Noble, A.J., Berger, B. (2019). Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs. Nature Methods. https://doi.org/10.1038/s41592-019-0575-8

9. Topaz-Denoise Bepler, T., Kelley, K., Noble, A.J., Berger, B. (2020). Topaz-Denoise: general deep denoising models for cryoEM and cryoET. bioRxiv. https://doi.org/10.1101/838920

10.RELION - SBGrid Consortium

https://www.youtube.com/watch?v=j0ml_GJ2368

11.Topaz - SBGrid Consortium

以及学姐提供的相关资料和论文。