工作汇报

冯浩哲

2018 3.4

展示大纲

- 近期工作内容
- 工作中所遇到的困难
- 工作计划

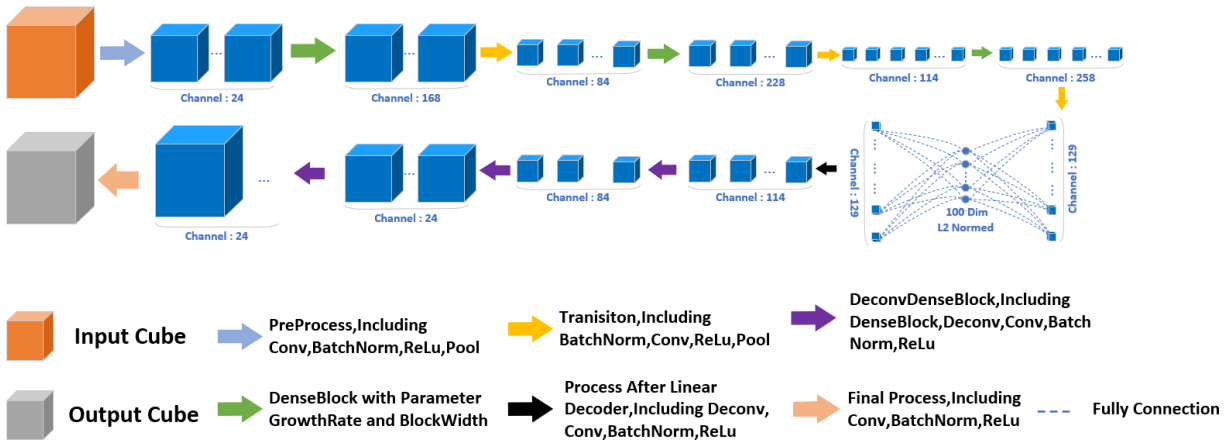
近期工作内容

- 研究基于聚类算法的 Suggestive Annotation 策略
  我们基于特征聚类的策略提出了一个完全无监督的 Suggestive Annotation 流程，这个流程全过程并不需要医生参与就会给出推荐标注的数据集结果。我们在 LIDC-IDRI 数据集上检测了这个策略，结果表明，这个策略能够显著优越于随机选取。我们将在下文中详细阐述这个流程。

近期工作内容

基于聚类算法的 Suggest Annotation 流程展示:

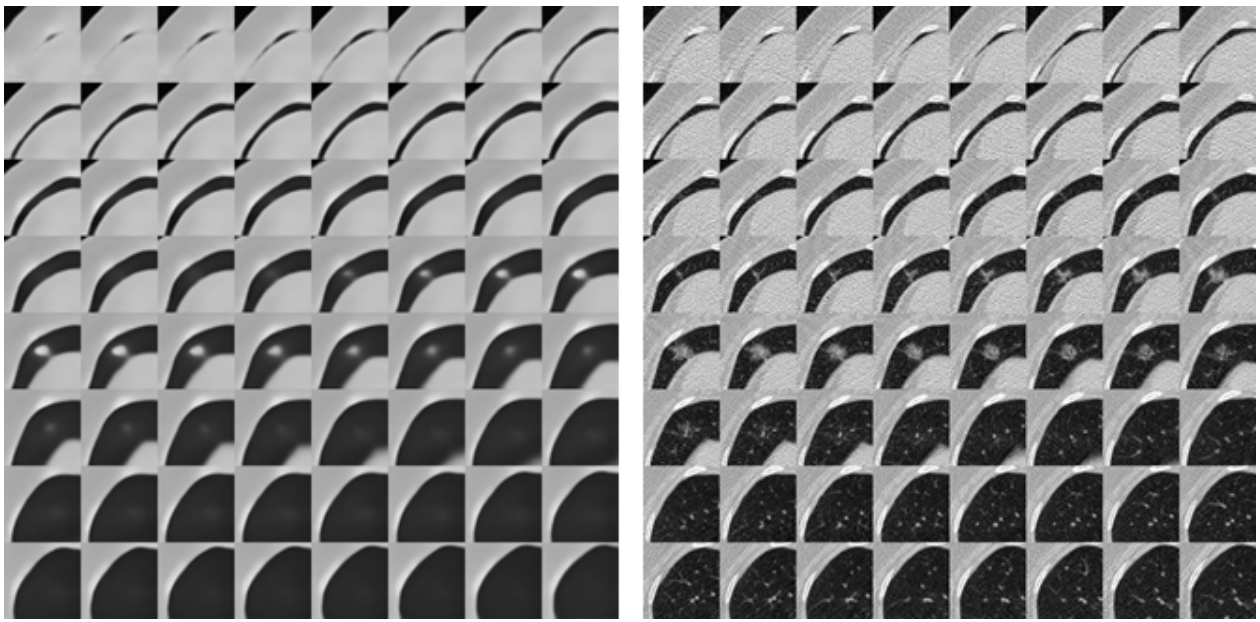
1 构建并训练一个深度自动编码器
首先我们训练一个无监督的深度自动编码器，这个编码器将输入的图像编码为一个 100 维的特征。基于 LIDC-IDRI 数据集的肺结节分割问题实例，我们设计的基于 Dense Block 的自动编码器如下图所示:



近期工作内容

基于聚类算法的 Suggest Annotation 流程展示:

1. 构建并训练一个深度自动编码器
   深度自动编码器采用 $loss = ||X - Encoder(X)||_p$ 这样的损失函数来进行训练，整个过程是无监督的。这个网络结构在我们的肺结节分割问题实例上的结果如图所示:



我们用这个深度自编码器将图像映射到一个 100 维的向量并用它来聚类。

近期工作内容

基于聚类算法的 Suggest Annotation 流程展示:

2 采用密度聚类方法分离孤立点
将图像映射到一个 100 维的特征向量之后，我采用密度聚类法对整个数据集进行聚类，将数据集分为孤立点类与其它几类。我们的策略是所有属于孤立点的数据集都采纳作为推荐标注的数据集，而同时对密度聚类得到的非孤立点类 $(A_1, ..., A_n)$ 选取具有代表性的 $(k_1, ..., k_n)$ 个数据集点。

近期工作内容

基于聚类算法的 Suggest Annotation 流程展示:

3 采用谱聚类方法选取具有代表性的数据集点
采用递归谱聚类方法对密度聚类得到的非孤立点集进行划分，直到划分的每一个子集所含有的数据个数小于某个阈值。在 LIDC 数据集中我们要求每一个子集所含有的数据集点都小于 40。如是我们就得到了若干个含有的数据个数小于 40 的子集集合。此时我们将整个非孤立点数据集划分为 $B_1, ..., B_m$ 个子集，同时我们需要按比例从这些子集中选取 $l_1, ..., l_m$ 个数据集点。我们采用 Danny Chen 教授在 Suggest Annotation 文章中所提到的最大 k 覆盖算法来从每一个子集集合中选取具有代表性的数据集点。

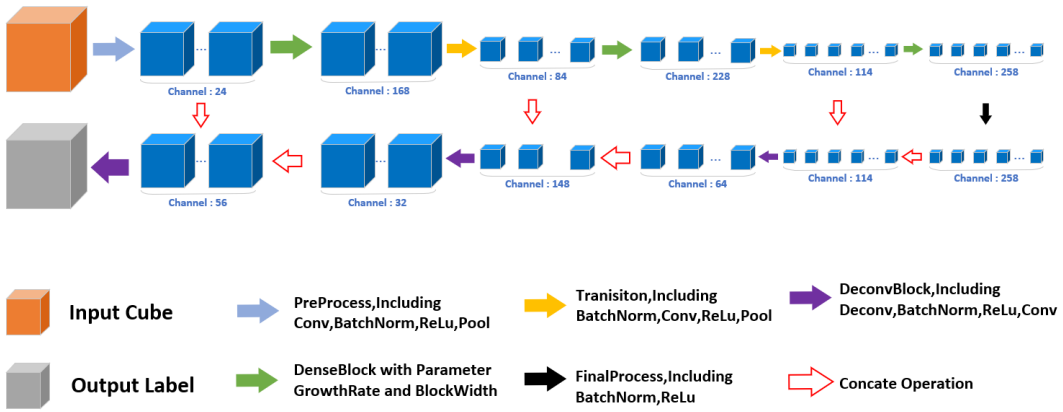
近期工作内容

以上过程就是我们所提出的基于聚类方法的完全无监督 Suggest Annotation 策略。我们用这个策略在 1148 个数据集中选取 400 个数据集作为训练集，并用剩下的数据集做验证，训练结果在验证集上达到了 Dice=0.748，而随机选取策略则只能达到 Dice=0.736。

工作中遇到的困难

我们在工作中主要遇到了 3 个困难：

1. 当前用于 3D 语义分割的 U-Net 耗费内存太大，训练收敛速度太慢。针对这个困难，我们现在暂时采用的是自己设计的基于 Dense Block 的 3D 语义分割网络，如图所示：



这个网络只需要花原来一半的内存以及 5 个小时的训练时间就能收敛。3D U-Net 的分割结果能够达到平均 Dice:0.763,VoE:0.37，而该网络能够达到平均 Dice:0.757，VoE:0.373，两个网络结果基本等同。

工作中遇到的困难

我们在工作中主要遇到了 3 个困难：

2 LIDC 数据集没有提供测试集与验证集
因为 LIDC 数据集并没有提供训练集，验证集，测试集的划分，如果我们需要得到一个实验结果，我们需要用 5 折交叉检验法，这个方法非常消耗时间。我们现在采用的方法是先用我们的完全无监督 Suggest Annotation 策略选取一些数据，然后把剩下的数据集作为测试集。

工作中遇到的困难

我们在工作中主要遇到了 3 个困难：

3 对于自动编码器所得到的特征与分割结果好坏仍然无法建立显著性联系
我们策略的基本假设是基于自动编码器所将原图 (64*64*64) 所映射到 1 个 100 维的向量能够描述原图的特征，即如果两个图像的编码向量彼此非常相似，那么两个图像的原图则具有非常相似的特征，从而我们能断言这两个图像只用标注一个就够了。我们经过实验发现，在测试集中分割得好的图像所对应的编码向量离训练集编码向量的平均距离一般会小于分割得不好的图像的平均距离，但是也有一些图像离训练集编码向量的平均距离很小，但是它们被分割得同样很糟糕。也就是两个图像的编码向量彼此相似并不代表着两个图像的分割结果也相似。

Danny Chen 教授给出的建议

同时，Danny Chen 教授也给我们的策略提出了 4 条问题与建议：

1. 我的聚类方法是否允许不同聚类之间的 Overlapping
2. 对于那些只有几个数据的小类，比如孤立点问题，我的聚类方法如何实现类别均衡的。
3. 对于从聚类结果中的数据选取，比如从 600 个数据集中选取 60 个数据集，这个选取的策略可以用最小支配集的方案来选取。
4. 尝试在选取原数据集 60%,50%,40%,...10% 的条件下对每一个百分比的最后结果画图，看看在选取不同百分比数据的时候我们的策略的表现。

工作计划

- 短期工作目标
  完成毕业论文，预计 5.15 号之前完成
- 中期工作目标
  对 Danny Chen 教授所提出的建议进行尝试与实施。现在我们已经基于第一，三条建议对算法进行了改进，但是对第二个和第四个建议，即实现类别均衡问题以及在不同的百分比下绘图我们还会进行尝试。
  我们希望能对原算法进行改进，我们希望达到我们的策略比随机选的结果 Dice 要高出 0.02 以上，且测试结果 Dice 高于 0.755。
- 长期工作目标
  将聚类算法用于 Danny 提出的其它重要课题上，重新审视对抗策略。
- 投稿目标
  暂时目标定为 11 月的 CVPR 会议，但是我不太确定自己这些基于医学影像的结果能不能投 CVPR。